Machine Learning

Supervised

Unsupervised

Active

Collaborative filtering

Machine Translation

out of many approaches in ML this course will focus on probabilistic (bayesian) methods

Textbook - Bishop 2006

Terminology

model

Sample

likelihood, max likelihood

prior

posterior

MAP

predictive distribution

Model a coin flip

$P(\text{Head}) = p$

$P(\text{Tail}) = 1-p$

assumption : no. of flips ~~...~~

$\Rightarrow$ consecutive flips are independent

Model explains how data is generated.

Sample ($\subseteq$ Data).

H T H H T H T T

x_1 x_2 x_3 ...

i^{th} coin flip x_i

$$x_i = \begin{cases} 1 & \text{Head} \\ 0 & \text{Tail} \end{cases}$$

bernoulli variable

$P(\text{Head}) = P(x_i = 1) = p$

scenario 1	200 H	300 T
scenario 2	2 H	3 T
scenario 3	5 H	0 T

<u>likelihood</u> = P(data | model)

HTHHT $\quad p(1-p)\,p\,p(1-p)$

$p^3(1-p)^2$

if data $x_1\ x_2\ x_3 \ldots x_n$

$$\boxed{p(x_i) = p^{x_i}(1-p)^{1-x_i}}$$

$$= \begin{cases} p & x_i = 1 \\ 1-p & x_i = 0 \end{cases}$$

likelihood : $L = \prod_{i=1}^{N} p(x_i)$

$$= \prod_{i=1}^{N} p^{x_i}(1-p)^{1-x_i}$$

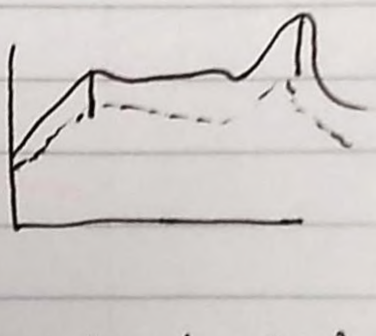
$$= p^{\Sigma x_i} (1-p)^{1-x_i}$$

$$p^{\# H} \quad p^{\# T}$$

max likelihood : pick the model that maximizes the value of likelihood.

$$\boxed{L = p^{\# H} \, p^{\# T}}$$

Find p s.t. L is max^m.



To maximize f

maximize $\log f$ instead (monotonic transformation)

$$L = p^{\#H}(1-p)^{\#T}$$

Take log on both sides

$$\log L = \#H \cdot \log p + \#T \log(1-p)$$

Differentiate wrt p.

$$\frac{d \log L}{dp} = \frac{\#H}{p} + \frac{\#T}{1-p}(-1)$$

set derivative to 0.

$$\frac{\#H}{p} = \frac{\#T}{1-p}$$

$$\cancel{\#H = \frac{p}{1-p}\#T}$$

$$\#H(1-p) = p \cdot \#T$$

$$\#H - \#Hp = p \cdot \#T$$

$$\#H = p(\#H + \#T)$$

$$\boxed{p = \frac{\#H}{\#H + \#T} = \frac{\#H}{N}}$$

max. likelihood estimate of p

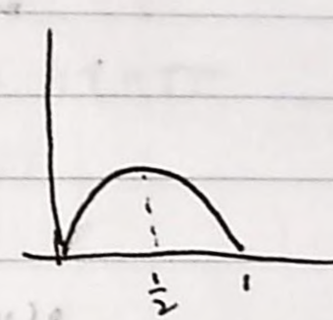
$$\hat{p} = \frac{\#H}{N}$$

prior = distribution over possible models.

distribution in game over p.

Choice of prior should reflect our belief / knowledge about the problem.

choose $P_r(p) = C p^2 (1-p)^2$
arbitrarily



$$\int C p^2 (1-p)^2 \, dp = 1$$

$$\cancel{C\left[\frac{p^3}{3} \frac{(1-p)^3}{3}\right]_0^1 = 1}$$

$$C\int_0^1 p^2 (1 + p^2 - 2p) = 1$$

$$C\int_0^1 (p^2 + p^4 - 2p^3) = 1$$

$$C = 30$$

Posterior = distribution over model that reflects our belief on probability of models, after having seen data.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$ Bayes theorem

A: model, B: data

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

$$p(\text{model}\,|\,\text{data}) = \frac{\overbrace{\boxed{p(\text{model})}}^{\text{prior}}\ \overbrace{\boxed{p(\text{data}\,|\,\text{model})}}^{\text{Likelihood}}}{p(\text{data})}$$

$$\propto p(\text{model})\ P(\text{data}/\text{model})$$

$$= \text{prior} \times \text{likelihood}$$

here,

$$p(\text{model}\,|\,\text{data}) \propto 30p^2(1-p)^2 \quad p^H(1-p)^T$$

$$\propto p^{H+2}(1-p)^{T+2}$$

$$P(\text{Data}) = \int_p 30\, p^2 (1-p)^2\, p^{200} (1-p)^{300}\, dp,$$

All possible ways to generate → for 200H 300T

2nd possible prior (discrete)

$$p = \begin{cases} \frac{1}{2} & 0.9 \\ \frac{1}{4} & 0.05 \\ \frac{3}{4} & 0.05 \end{cases}$$

Head

Data HTHTT

$$P(\text{Data}) = \underbrace{0.9 \left(\tfrac{1}{2}\right)^5}_{p=\frac{1}{2}} + \underbrace{0.05 \left(\tfrac{1}{4}\right)\left(\tfrac{1}{4}\right) \tfrac{3}{4}\, \tfrac{3}{4}\, \tfrac{3}{4}}_{p=\frac{1}{4}}$$

$$+ \underbrace{0.05 \left(\tfrac{3}{4}\right)\left(\tfrac{3}{4}\right)\left(\tfrac{1}{4}\right)\left(\tfrac{1}{4}\right)\left(\tfrac{1}{4}\right)}_{p=\frac{3}{4}}$$

next flip

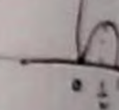
Data HTHTH ...

Data HHHHH ...

next flip

Prior : distribution over model which expresses our belief (before seeing data) about what are likely models.

choosing arbitrary prior $Pr(p) = 30p^2(1-p)^2$



How to predict without data?

predicting with data

$$p(\text{model} \mid \text{Data}) = \frac{p(\text{Data} \mid \text{model})\, p(\text{Model})}{p(\text{Data})}$$

$$\propto p(\text{Data} \mid \text{model})\, p(\text{model})$$

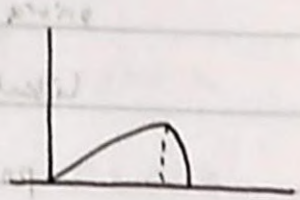
Prior $\propto p^2(1-p)^2$

Situation – HTHTH Likelihood $= p^3(1-p)^2$

prior $\times$ likelihood $\propto p^5(1-p)^4$

NOTE: If we know the functional form of a distribution then we can calculate its normalizing constant.

$$\text{Posterior} = C_2\, p^5 (1-p)^4$$



Maximum A Posteriori (MAP) : Pick the model (p) which maximizes the Posterior.

$$\text{Posterior} = C_2\, p^5 (1-p)^4$$

$$\log \text{posterior} = \log C_2 + \log p^5 + \log (1-p)^4$$

$$= \log C_2 + 5 \log p + 4 \log 1-p$$

Differentiate wrt p on both sides

$$\frac{d}{dp} \log \text{posterior} = 0 + \frac{5}{p} + \frac{4}{1-p}$$

set derivative to zero.

$$\frac{5}{p} + \frac{4}{1-p} = 0$$

$$5(1-p) = 4p$$

$$\frac{5}{9} = p$$

v/s $\frac{3}{5}$ from max likelihood

Situation HHHHH

prior $\propto p^2 (1-p)^2$

likelihood $\propto p^5$

posterior $= C_3 \, p^7 (1-p)^2$

log & d/dp

$$\frac{d}{dp} \log \text{posterior} = \frac{d}{dp} \log C_3 + 7 \frac{d}{dp} \log p + 2 \frac{d}{dp} \log 1-p$$

$$= 0 + \frac{7}{p} + \frac{2}{1-p}(-1)$$

set derivative to zero.

$$\frac{7}{p} = \frac{2}{1-p}$$

$$7(1-p) = 2p$$

$$\frac{7}{9} = \hat{p}$$

v/s $\hat{p} = 1$ from maximum likelihood

Given my current belief : $Pr(p)$
what is the prob. that next coin in H.

Ex. current belief $Pr(p) = \begin{cases} 0.9 & p = 0.2 \\ 0.1 & p = 0.3 \end{cases}$

$$0.9 \times 0.2 + 0.1 \times 0.3 = 0.21$$

Predictive Distribution :
$P(\text{next head} \mid \text{Data}) =$

$$\int_p \underbrace{Pr(p \mid \text{Data})}_{\text{posterior}} \underbrace{P(\text{Next head} \mid p)}_{\text{likelihood}} \, dp$$

Coin Problem :
H Heads $\quad$ Prior $\propto p^2 (1-p)^2$
T Tails.

Posterior $\propto p^{H+2} (1-p)^{T+2}$

$$\hat{p}_{\text{predictive}} = \int C p^{H+2} (1-p)^{T+2} \, p \, dp$$

Probability that $Pr(\text{next head}) = p$ given data.

prob. that next flip is head given prob head is p.

conjugate prior :

prior and posterior have the same dtype of dependance on the parameter(s).

(belong to same family of distributions)

Model - say how data was generated ; prior = dist. over model

Sample - subset of data

method of prediction - max. likelihood

max. a posteriori

predictive dist

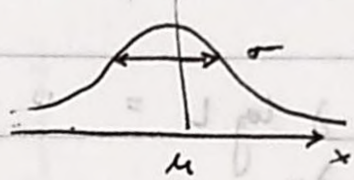
prior :

do not rule out options.

as the amount of data increases, impact of prior goes down

prior ≠ bias .

Normal Distribution.



$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$= \sqrt{\frac{\beta}{2\pi}}\; e^{-\frac{\beta}{2}(x-\mu)^2}$$

precision $\beta = \frac{1}{\sigma^2}$

Assume that each x_i is drawn from $N(\mu, \frac{1}{\beta})$ and that different x_i's are independant

	x_1	x_2	x_3	x_4	x_5
Data :	10.3	5.1	12.6	11.1	-2.5

Estimation : pick mean & precision.

L = Likelihood $= P(\text{Data} \mid \text{model})$

$$= p(x_1)\cdot p(x_2)\cdot p(x_3)\cdot p(x_4)\cdot p(x_5)$$

$$= \prod_i \sqrt{\frac{\beta}{2\pi}}\; e^{-\frac{\beta}{2}(x_i-\mu)^2}$$

$$\log L = \cancel{\sum (2\pi)^{-\frac{1}{2}} \beta^{\frac{1}{2}} e}$$

$$= \sum_1 -\frac{1}{2}\log 2\pi + \sum \frac{1}{2}\log\beta + \sum -\frac{\beta}{2}(x_i-\mu)^2.$$

$$= \frac{N}{2}\log 2\pi + \frac{N}{2}\log\beta - \frac{\beta}{2}\sum(x_i-\mu)^2$$

$$\frac{\partial}{\partial \mu} \log L = 0 + 0 - \frac{\beta}{2} \sum_i 2(x_i - \mu)(-1) = 0$$

$$\sum_i x_i = \sum_i \mu = N\mu$$

$$\hat{\mu}_{ML} = \frac{\sum x_i}{N}$$ || max likelihood estimator

Estimator is a function of random variable(s)

$\Rightarrow$ it is a random variable

① unbiased $E[\hat{\mu}] = \mu.$
② Prefer low variance

} properties of good estimator

$$\frac{\partial \log L}{\partial \beta} = \frac{N}{2}\frac{1}{\beta} - \frac{1}{2} \sum (x_i - \mu)^2 = 0$$

$$\frac{1}{\hat{\beta}_{ML}} = \frac{1}{N} \sum (x_i - \mu)^2$$ ← replace with $\hat{\mu}_{ML}$

$$\frac{1}{\hat{\beta}_{ML}} = \frac{1}{N} \sum (x_i - \mu)^2$$ } ← biased estimator

$$E[\hat{\sigma}^2_{ML}] = \frac{N-1}{N}\sigma^2.$$

$$\frac{1}{N-1}\sum (x_i - \hat{\mu}_{ML})$$ ← corrected unbiased

$$E[\hat{\mu}_{ML}] = E\left[\frac{1}{N}\sum x_i\right]$$

$$= \frac{1}{N}\sum E[x_i]$$

$$= \frac{1}{N}\cdot\sum \mu.$$

$$= \frac{1}{\cancel{N}}\cdot\cancel{N}\mu = \mu$$

$\therefore$ it is an unbiased estimator

Machine Learning

① A model explains how data is generated

② We estimate the concrete model (parameters) using data

① Maximum Likelihood - non bayesian

② Prior + data → posterior → MAP

→ Predictive distribution } bayesian

all information

prior protects against overfitting

coin model : p

$$L = p^{\#H}(1-p)^{\#T}$$

$$\text{prior } \beta(a,b) = \frac{p^{a-1}\,p^{b-1}}{\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}}$$

Used Beta(3,3) in example

$$\text{posterior} \propto \text{prior} \times L \propto p^{\#H+a-1}(1-p)^{\#T+b-1}$$

$$\text{dist.} \propto p^2(1-p)^2$$

must be

$$\Pr(p) = \frac{\Gamma(3+3)}{\Gamma(3)\Gamma(3)}\,p^2(1-p)^2 = \frac{6!}{3!\,3!}\,p^2(1-p)^2$$

$$= \frac{\cancel{3 \cdot 5 \cdot 4^2}}{\cancel{2 \cdot 2}}\,p^2(1-p)^2$$

$$= 30\,p^2(1-p)^2$$

$$\int p^3 (1-p)^7 \, dp \times \frac{\Gamma(4+8)}{\Gamma(4)\Gamma(8)} \times \frac{\Gamma(4)\Gamma(8)}{\Gamma(4+8)}$$

$$\frac{\Gamma(4)\Gamma(8)}{\Gamma(4+8)} \int \beta(4,8) \, dp$$

$$\frac{4!\,8!}{12!}$$

Normal distribution

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2\sigma^2}x^2} e^{\frac{+x\mu}{\sigma^2}} e^{\frac{-\mu^2}{2\sigma}}$$

$$\propto e^{\frac{2\mu x - x^2}{2\sigma^2}}$$

"completing the squares"

$$p(x) \propto e^{-16x^2+8x}$$ ← normal distribution

$$\frac{2\mu x - x^2}{2\sigma^2} = -16x^2 + 8x$$

$$\frac{+1}{2\sigma^2} = +16 \quad \text{and} \quad \frac{+\mu x}{2\sigma^2} = 8$$

$8/32 = \frac{1}{4}$

$\sigma^2 = 32$ $\quad$ $8\mu = \cancel{8 \times 32}$

Model 2: independant samples from Normal$(\mu, \frac{1}{\beta})$

$$L = \prod_i \sqrt{\frac{\beta}{2\pi}}\, e^{-\frac{\beta}{2}(x-\mu)^2}$$

$$\hat{\mu}_{ML} = \frac{\sum x_i}{N}$$

What would a conjugate prior for μ look like?

prior $\times L \propto$ posterior

$$L = \prod_i \sqrt{\frac{\beta}{2\pi}}\, e^{-\frac{\beta}{2}(x_i-\mu)^2}$$

$$= \prod_i \sqrt{\frac{\beta}{2\pi}}\, e^{-\frac{\beta}{2}x_i^2}\, e^{-\beta x_i \mu}\, e^{\frac{\beta}{2}\mu^2}$$

exponential quadratic μ.

$$\text{Prior} \quad p(\mu \mid \mu_0\, \beta_0) = \sqrt{\frac{\beta}{2\pi}}\, e^{-\frac{\beta_0}{2}(\mu-\mu_0)}$$

$$= \sqrt{\frac{\beta}{2\pi}}\, e^{-\frac{\beta_0}{2}\mu^2}\, e^{+\beta_0 \mu\mu_0}\, e^{-\frac{\beta_0 \mu_0^2}{2}}$$

$$= \sqrt{\frac{\beta}{2\pi}}\, e^{-\frac{\beta \mu_0}{2}} \left(e^{-\frac{\beta_0}{2}\mu^2 + \beta_0 \mu_0 \mu}\right.$$

What are mean & precision ($= \frac{1}{\sigma^2}$) for posterior?

(Todo)

House

	Age	Size	Condition	Centrality	Price
House 1	2	1000	10	-1000	10
House 2	100	500	8	1	15
					? output

$\Phi = \left(\quad \right)$ ← i^{th} example description $\phi(x_i)$

j^{th} attribute

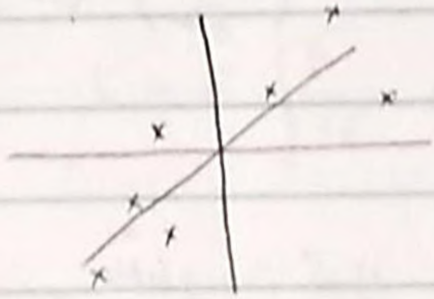
t = vector of values we want to predict, one for each example

NOTE: All vectors are column vectors

x_i = i^{th} example

$\phi(x_i)$ = representation of i^{th} example

Linear regression



x_i's are arbitrary

$$y_i = \bar{w}^T \phi(x_i)$$

true

$$y_i = \sum_j w_j \phi_j(x_i)$$

observed

$$t_i \sim N\left(y_i, \frac{1}{\beta}\right)$$

assume variance of all examples is same.

t_i's are independant of each other

Likelihood $$L = \prod_i \sqrt{\frac{\beta}{2\pi}}\, e^{-\frac{\beta}{2}(y_i - t_i)^2}$$

$$L = \prod_i \sqrt{\frac{\beta}{2\pi}}\, e^{-\frac{\beta}{2}(w^T\phi(x_i) - t_i)^2}$$

$$\log L = \frac{N}{2}\log\frac{\beta}{2\pi} - \frac{\beta}{2}\sum_{i=1}^{N}(w^T\phi(x_i) - t_i)^2$$

Max. likelihood for w

max log L is same as max $-\sum_{i=1}^{N}(w^T\phi(x_i) - t_i)^2$

or min $\sum_{i=1}^{N}(w^T\phi(x_i) - t_i)^2$

A matrix vector product is the Linear combination of the column vectors of matrix.

$\min \sum (w^T \phi(x_i) - t_i)^2$ least squares

$$\begin{bmatrix} A \end{bmatrix} \begin{bmatrix} b \end{bmatrix} = \begin{bmatrix} \sum_i a_i b_i \end{bmatrix}$$

a_i

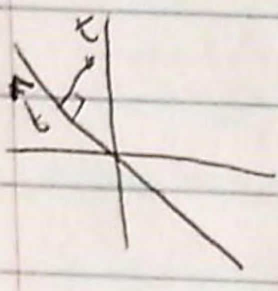
$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} a\alpha \\ d\alpha \\ h\alpha \end{pmatrix} + \begin{pmatrix} b\beta \\ e\beta \\ i\beta \end{pmatrix} + \begin{pmatrix} c\gamma \\ f\gamma \\ j\gamma \end{pmatrix}$$

$w^T \phi(x_i)$
$= \phi(x_i)^T w$

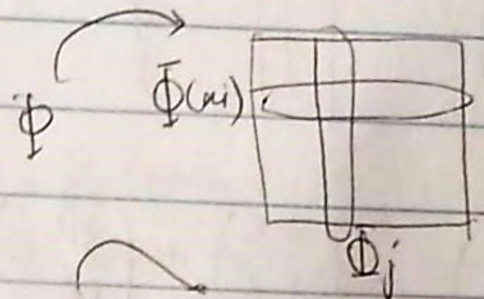
Min $\sum (w^T \phi(x_i) - t_i)^2$
min $\| \Phi w - t \|^2$
Find $\hat{w}$ such that $\| \hat{t} - t \|^2$ is minimised where $\hat{t} = \Phi \hat{w}$ parameters

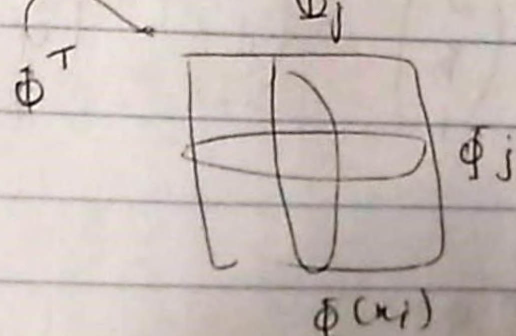
prediction

$\hat{t}$ is a linear combination of columns of Φ

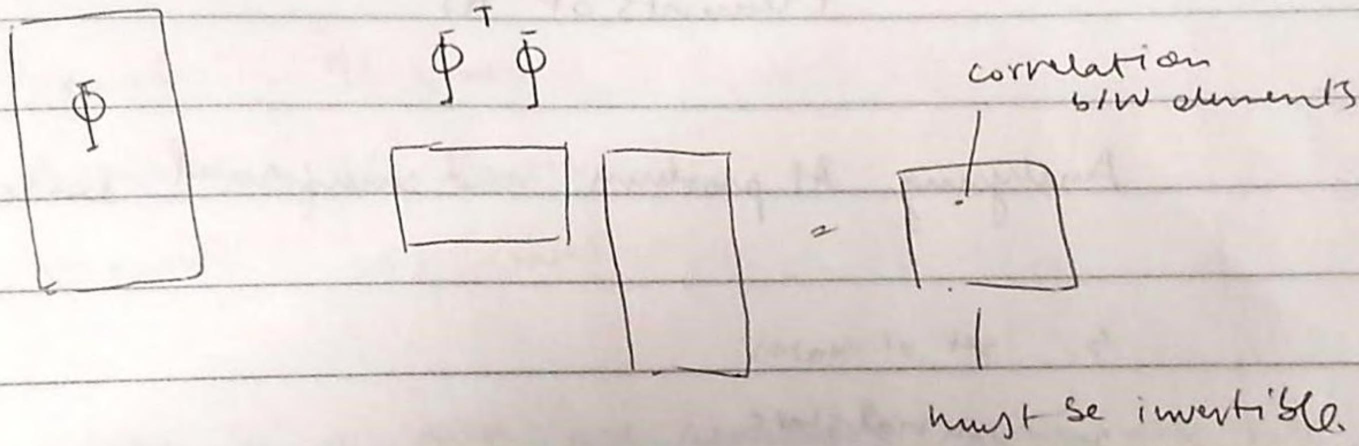


$\cancel{t - \hat{t}}$ $t - \hat{t}$ is perpendicular to columns of Φ





$$\Phi^T (t - \hat{t}) = 0$$
$$\Phi^T (t - \Phi \hat{w}) = 0$$
$$\cancel{\Phi^T t = \Phi^T \Phi \hat{w}}$$
$$\Phi^T t = \Phi^T \Phi \hat{w}$$
$$\boxed{\hat{w} = (\Phi^T \Phi)^{-1} \Phi^T t}$$

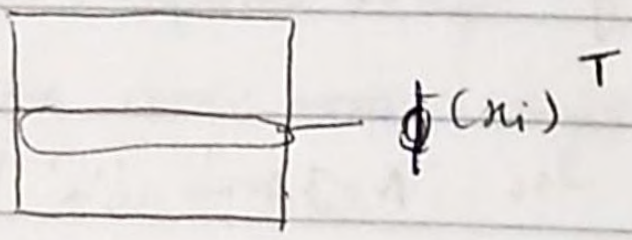


For any A

$A^T A$ has an inverse

iff col of A are independent

Machine Learning.



x_i example

$\Phi(x_i)$ column vector, represents the example

true hidden value

$$y_i = w^T \phi(x_i)$$

observe

$$t_i \sim N(y_i, \frac{1}{\beta})$$

Every example is drawn independently

$$L = \prod_i \frac{\sqrt{\beta}}{\sqrt{2\pi}} e^{-\frac{\beta}{2}(w^T\phi(x) - t_i)^2}$$

$\vdots$

Maximum likelihood :

$$\text{Max} \quad \frac{N}{2} \log \beta - \frac{\beta}{2} \sum_{i=1}^{N} w^T \Phi(x_i) - t_i^2$$

max L for w

is same as

$$\min \sum_i (w^T \phi(x_i) - t_i)^T$$

$$\min \| \Phi w - t \|^2$$

Soln #1 Geometric

$$\hat{\omega}_{ML} = (\Phi^T \Phi)^{-1} \Phi^T t$$

Soln #2 Using Derivatives

$$\min \sum (w^T \phi(x_i) - t_i)^2$$

$$w^T \phi(x_i) = \sum_{k=1}^{K} w_k \phi_k(x_k)$$

$$\frac{\partial}{\partial w} \underbrace{(w^T \phi(x_i) - t_i)^2}_{= \text{let } A} = \sum_i 2(w^T \phi(x_i) - t_i) \phi_k(x_i)$$

$$\frac{\partial A}{\partial w} = \begin{pmatrix} \frac{\partial A}{\partial w_1} \\ \frac{\partial A}{\partial w_2} \\ \vdots \\ \frac{\partial A}{\partial w_K} \end{pmatrix} = 2 \sum (w^T \phi(x_i) - t_i) \phi(x_i)$$

$$= \bar{0} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

what is the vector whose entries are $\phi_k(x_i)$: $\begin{pmatrix} \phi_1(x_i) \\ \vdots \\ \phi_K(x_i) \end{pmatrix}$

$\phi(x_i)$

$$\underbrace{\sum (w^T \phi(x_i)) \phi(x_i)}_{\Phi^T \Phi w} = \underbrace{\sum t_i \phi(x_i)}_{\Phi^T t}$$

vector whose entries are $\sum w^T \phi(x_i) = \sum \phi(x_i)^T w = \Phi w$

59

$$a^T b = \sum_k a_k b_k$$

$$\Rightarrow \Phi^T \Phi \hat{W} = \Phi^T t$$

$$\Rightarrow \boxed{W_{ML} = (\Phi^T \Phi)^T \Phi^T t}$$

Soln #3 Vector derivatives

$$\min \|\Phi w - t\|^2$$

$$\|\Phi w - t\|^2 = (\Phi w - t)^T (\Phi w - t)$$ scalar

$$= w^T \Phi^T \Phi w - w^T \Phi^T t - t^T \Phi w + t^T t$$

$$= w^T \Phi^T \Phi w - 2w^T \Phi^T t + t^T t$$

(Assuming A is symmetric)

$$\frac{\partial w^T A w}{\partial w} = 2Aw$$

$$\frac{\partial w^T b}{\partial w} = b$$

$$\frac{\partial \|\Phi w - t\|^2}{\partial w} = 2\Phi^T \Phi w - 2\Phi^T t = 0$$

$$\Rightarrow \Phi^T \Phi w = \Phi^T t$$

$$\Rightarrow \hat{w} = (\Phi^T \Phi)^{-1} \Phi^T t$$

Regularization : trick to restrict W from getting arbitrary values & therefore overfitting the data.

New objective : $\min \|\Phi w - t\|^2 + \lambda \|w\|^2$

regularization parameter

penalty for increase in norm of w.

$\lambda \|w\|^2 = \lambda w^T w$

$\frac{d}{dw} \lambda \|w\|^2 = 2\lambda w$

$\therefore \frac{d}{dw} \|\Phi w - t\|^2 + \lambda \|w\|^2$

$= 2\Phi^T \Phi w - 2\Phi^T t + 2\lambda w$

$\frac{d}{dw}$ objective $= 0$

$\Phi^T \Phi w + \lambda w - 2\Phi^T t = 0$

$\Phi^T \Phi w + \lambda I w - 2\Phi^T t = 0$

$(\Phi^T \Phi + \lambda I) w = \Phi^T t$

$\hat{w}_R = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T t$

$$\log L = -\frac{N}{2}\log 2\pi + \frac{N}{2}\log\beta - \frac{\beta}{2}\|\Phi w - t\|^2$$

compute MLE for β

$$\frac{d\log L}{d\beta} = \cancel{\quad} 0 + \frac{N}{2\beta} - \frac{1}{2}\|\Phi W - t\|^2$$

$$\sigma^2 = \frac{1}{\beta} = \frac{1}{N}\|\Phi w - t\|^2$$

Conjugate prior for W?

Prior $\times$ Likelihood $\propto$ posterior

$$L = \prod\left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{\beta}{2}(w^T\phi(x_i)-t)^2}$$

$$\propto e^{-\frac{\beta}{2}\|\Phi w - t\|^2}$$

$$\propto e^{-\frac{\beta}{2}\cancel{(\Phi w - t)^T}(w^T\Phi^T w - 2w^T\Phi^T t + t^T t)}$$

$$\propto e^{-\frac{\beta}{2}w^T\Phi^T w}\; e^{\frac{\beta}{2}2w^T\Phi^T t}\; \underbrace{e^{-\frac{\beta}{2}t^T t}}_{\text{ignore}}$$

$$\propto e^{-\frac{\beta}{2}w^T\Phi^T w}\; e^{\beta w^T\Phi^T t}$$

$e^{-ve\ \text{Quadratic}}$ e^{linear}

~~Mult~~ Multivariate Gaussian Distribution

* looks like a gaussian in n-dimensions

* Read Linear Algebra review

Machine Learning

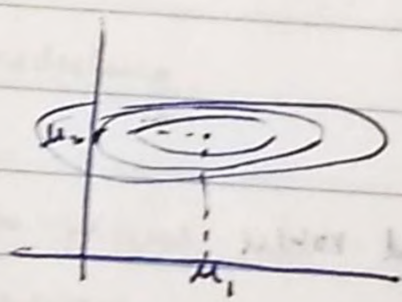
Independant x_1 x_2

$x_1 \sim N(\mu_1, 100)$

$x_2 \sim N(\mu_2, 100)$

$$P(x_1, x_2) = \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{1}{2\sigma_1^2}(x_1-\mu_1)^2} \times$$

$$\frac{1}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{1}{2\sigma_2^2}(x_2-\mu_2)^2}$$

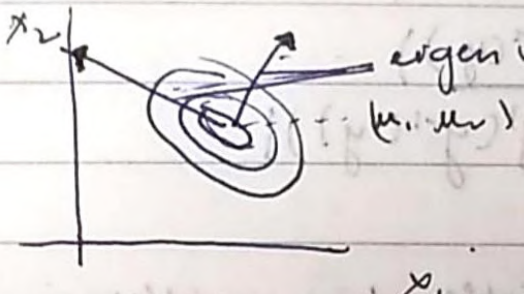


$$P(x_1 x_2) = \left(\frac{1}{\sqrt{2\pi}}\right)^2 \frac{1}{\sigma_1\sigma_2} e^{-\frac{1}{2}(x_1-\mu_1)\frac{1}{\sigma_1^2}(x_1-\mu_1)} e^{-\frac{1}{2}(x_2-\mu_2)\frac{1}{\sigma_2^2}(x_2-\mu_2)}$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^2 \frac{1}{\sigma_1\sigma_2} e^{-\frac{1}{2}(x_1-\mu_1)(x_2-\mu_2)\begin{pmatrix}\frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2}\end{pmatrix}\begin{pmatrix}x_1-\mu_1 \\ x_2-\mu_2\end{pmatrix}}$$

$$P(x) = \left(\frac{1}{2\pi}\right)^{m/2} \frac{1}{\sigma_1 \dots \sigma_m} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

Direction and width of ellipse is determined by the eigen decomposition of Σ



eigenvalues = variances

$$\Sigma = V \Lambda V^T$$
$$y = V^T (x - \mu)$$

① $X \in \{0, 1\}$ win

$Y = 3 + X$

$P(X=1) = 0.7$

$P(X=0) = 0.3$

$Y \in \{3, 4\}$

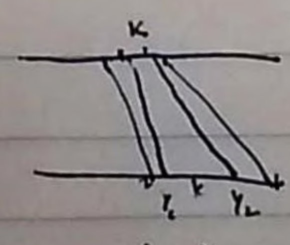
$P(Y=4) \; 0.7$

$P(Y=3) = 0.3$

② $X \in [1, 2]$ $\quad Y = 2X$

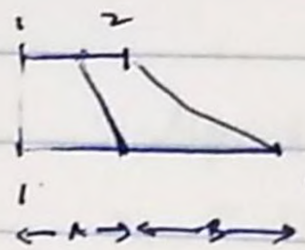
$P_x(x) = 1$

$$\int_0^1 \underset{P_x(x)}{1} \, dx = x \Big|_0^1 = 1$$



$\frac{Y_1}{2} = X_1 \qquad \frac{Y_2}{2} = X_2$

③ $x \in [1,2]$ $p_x(x) = 1$ $y = x^2$



$\sum_A p_y(y) > \sum_B p_y(y)$

$y = g(x)$ one one & invertible.

$$p_y(y) = \frac{p_x(g^{-1}(y))}{|g'(g^{-1}(y))|}$$

change of variable for variables in 1D

④ $$p_y(y) = \frac{p_x(x)}{g'(g^{-1}(y))} = \frac{1}{2}$$

slope of transformation at its image at y.

⑤ $$p_y(y) = \frac{1}{2\sqrt{y}}$$

$y = g(x) = x^2$

$g'(x) = 2x$

$g^{-1}(y) = \sqrt{y}$

$g'(g(y)) = 2\sqrt{y}$

④ $x \in [1, 2]$

$y = g(x) = x^2$

$y \in [1, 4]$

$p_y(y) = \text{uniform } [1, 4]$

$= \frac{1}{3}$

$p_x(x) = ?$

$x = f(y) = \sqrt{y}$

$$p_x(x) = \frac{p_y(f^{-1}(x))}{|f'(f^{-1}(x))|}$$

$$= \frac{p_y(x^2)}{\left|\frac{-1}{2x}\right|}$$ — uniform $[1, 4]$ $= \frac{1}{3}$

$$= \frac{1/3}{\frac{1}{2}\left|\frac{1}{x}\right|}$$

$$= \frac{2}{3}|x| = \frac{2}{3}x$$

$f^{-1}(x) = x^2$

$f'(y) = \frac{-1}{2\sqrt{y}}$

$f'(f^{-1}(x))$

$= \frac{-1}{2\sqrt{x^2}}$

$= \frac{-1}{2x}$

$$P_x(g^{-1}(y)) \cdot |J_{g^{-1}}(y)| = \frac{P_x(g^{-1}(y))}{|J_{g^{-1}}(g^{-1}(y))|}$$

$$P_x(g^{-1}(y))|(g^{-1})'(y)| = \frac{P_x(g^{-1}(y))}{|g'(g^{-1}(y))|}$$

Multivariate

$y = g(x)$ $\quad Y = g(X)$ $\quad X \in \mathbb{R}^k$ $\quad y \in \mathbb{R}^k$

$$P_Y(Y) = \frac{P_x(g^{-1}(Y))}{|J_g(g^{-1}(y))|}$$

Jacobian

output

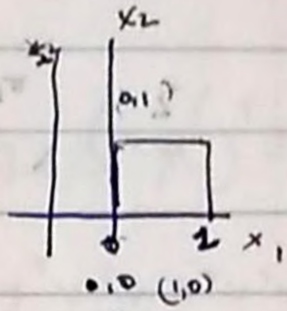
$$\begin{pmatrix} y_1 \\ \vdots \\ y_k \end{pmatrix}$$

input $(x_1 \ldots x_k)$

$$J = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_k}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \ddots & \frac{\partial y_k}{\partial x_2} \\ \vdots & & \\ \frac{\partial y_1}{\partial x_k} & \cdots & \frac{\partial y_k}{\partial x_k} \end{pmatrix}$$

— linear transform.

(5) $Y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = g(x)$

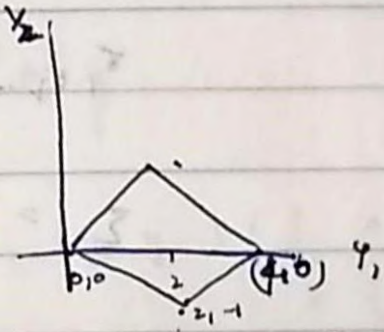


$X \sim \text{Uniform } [0,1]^2$

$p_X(X) = 1$

$p_Y(Y) = ?$

$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = g^{-1}(y) = \frac{1}{4} \begin{pmatrix} 1 & 2 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$



$$J = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 2 & -1 \end{pmatrix}^T = (\text{L. transform})$$

$\text{abs}(|J|) = -2 - 2 = |-4| = 4$

$$p_Y(y) = \frac{p_X(g^{-1}(y))}{|J[g^{-1}(y)]|} = \frac{1}{4}$$

is diagonal.

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_n^2 \end{pmatrix}$$

$$|\Sigma| = \prod_{i=1}^{m} \sigma_i$$

$$\sigma_1 \sigma_2 \dots \sigma_m = |\Sigma|^{1/2}$$

$$\Sigma = V \Lambda V^T \qquad (|||)(\ddots)(\equiv)$$

capital lambda.

special linear transform

$$Y = V^T(x - \mu) = g(x)$$

$$x = g^{-1}(Y) = Vy + \mu$$

orthonormal.

$V = (V^T)^{-1}$

or $V^{-1} = V^T$

$$P_Y(Y) = \frac{P_X(g^{-1}(y))}{|J|}$$

$\Sigma = V \Lambda V^T$

$\Sigma^{-1} = V \Lambda^{-1} V^T$ (see Linear Algebra notes)

$$p_x(x) = (2\pi)^{-m/2} \, \cancel{\Sigma^{-1/2}} \, |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

$$p_x(g^{-1}(y)) = (2\pi)^{-m/2} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(\underline{Vy + \cancel{\mu}} - \cancel{\mu})^T V \Lambda^{-1} V^T (Vy + \cancel{\mu} - \cancel{\mu})}$$

$$p_x(g^{-1}(y)) = (2\pi)^{-m/2} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2} \cancel{(Vy)^T V \Lambda^{-1}}} \quad y^T \underbrace{V^T V}\Lambda^{-1} \underbrace{V^T V} y$$

$\cancel{-\frac{1}{2} y^T \Lambda^{-1}}$

exponent $= -\frac{1}{2} y^T \Lambda^{-1} y$

var of y is the eigenvalues of Σ $\Longrightarrow$ $\frac{1}{2} (y_1 \; y_2) \begin{pmatrix} \frac{1}{\lambda_1} & 0 \\ 0 & \frac{1}{\lambda_2} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$

$E[x] = \int p_x(x) \, x \, dx$

$z = x - \mu = Vy$

$dz = dx$

$$E[x] = \int \left(\frac{1}{2\pi}\right)^{m/2} \frac{1}{|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2} \overset{z}{\widetilde{(x-\mu)}}^T \Sigma^{-1} \overset{z}{\widetilde{(x-\mu)}}} \quad \frac{d\widetilde{z} \; \widetilde{(z+\mu)}}{dx \; (x)}$$

$$= c \int \underbrace{e^{-\frac{1}{2} z^T \Sigma^{-1} z}}_{\text{symmetric}} \underbrace{(z+\mu)}_{\text{asymmetric}} dz$$

$\int$ sym $\times$ Asym $= 0$

$\int p(z) \mu \, dz = \mu$

$\int$ symmetric $\times$ asymmetric $\to 0$.

$I = \int f(x)\, g(x)\, dx$

$x = -z$

$dx = -dz$

$I = \int f(-z)\, g(-z)\, dz$

$= \int f(z)\, g(-z)\, dz$

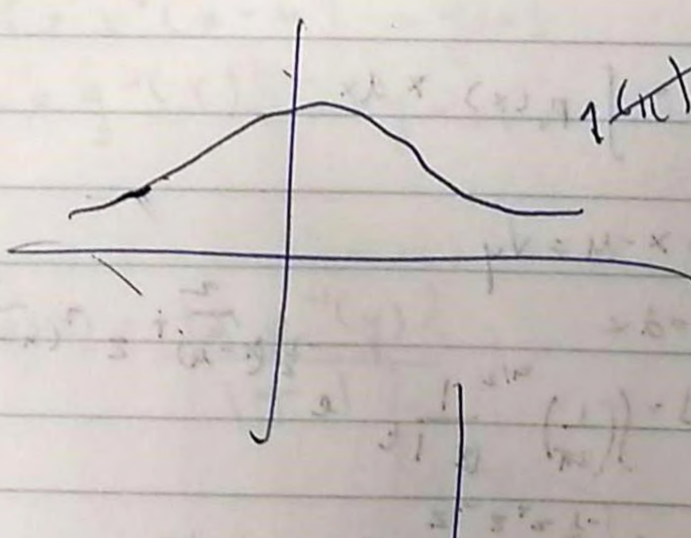
$= -\int f(z)\, g(z)\, dz$

$2I = \int f(x)\, g(x)\, dx - \int f(x)\, g(x)\, dx$

$= 0$

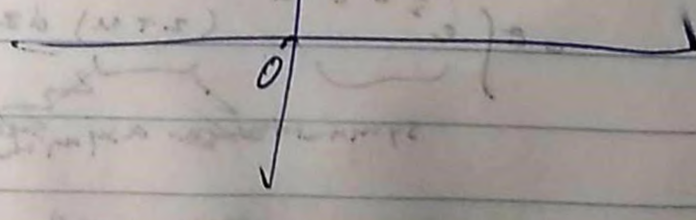
$g(-x) = -g(x)$

$g(x)$



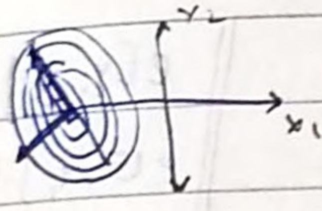
$\cancel{g(x) = -g}$

$g(-x) = -g(x)$



Machine Learning.

$$p(x) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$



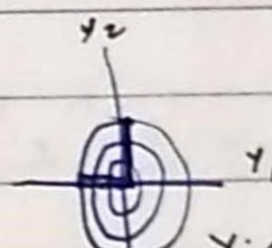
$Y = V^T(x-\mu)$

$X = VY + \mu \Leftrightarrow$

where

$\Sigma = V \Lambda V^T$

$\Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{pmatrix}$



$Y_i \sim N(0, \lambda_i)$

Y_i's are independent

mean ~~zero~~ (0)

Variance λ_i

$$E[x] = \int_x \frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \times dx$$

let $x - \mu = z \Rightarrow x = z + \mu$

$$= \int_z \frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(z)^T \Sigma^{-1}(z)} (z+\mu)\, dz$$

$$= \int_z \underbrace{P(z+\mu)\, z\, dz}_{} + \int_z \underbrace{P(z+\mu)\, \mu\, dz}_{1 \times \mu}$$

$\because f(-z) = -f(z)$

$I = 0$

$\Rightarrow E[x] = \mu.$

$$\mathrm{Cov}(x) = E[(x-\mu)(x-\mu)^T]$$

$$= E[z z^T] = E[VY(VY)^T]$$

$$= E[V Y Y^T V^T]$$

$$= V E[Y Y^T] V^T$$

$$= V \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_d \end{bmatrix} V^T$$

$$= V \Lambda V^T = \Sigma$$

$E[y_1 y_3] = ?$

$= E[y_1] E[y_3]$

$\quad 0 \qquad 0$

$= 0$

$E[y_i^2] = \cancel{E} ?$

$\mathrm{var} = \lambda_i = E[y_i]^2 - [E[y_i]]^2$

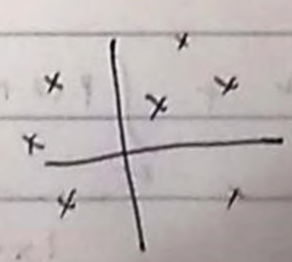
$\qquad 0$

$E[y_i] = \lambda_i$

diagonals

Exercise

Data is generated independently by a multivariate normal distribution. Estimate parameters μ, Σ using Maximum likelihood



data : ~~$x_1, x_2, \ldots$~~ $z^{(1)}, z^{(2)}, \ldots z^{(N)}$

Likelihood =

$$\prod (2\pi)^{-d/2} |\Sigma|^{-1/2} e^{-\frac{1}{2}(z^i-\mu)^T \Sigma^{T} (z^i-\mu)}$$

$$\log L = \sum_i -\frac{d}{2}\log 2\pi + \sum_i \left(-\frac{1}{2}\right)\log \Sigma - \frac{1}{2}\sum_i (z^i-\mu)^T \Sigma^{-1} (z^i-\mu)$$

$$\frac{\partial a^T B a}{\partial a} = ? \quad \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

B symmetric

$$a^T B a = \Sigma_i \Sigma_j a_i a_j B_{ij}$$

$$\frac{\partial (a^T B a)}{\partial a_k} = \sum_j a_j B_{kj} + \sum_i a_i B_{ik}$$

$$\cancel{= \sum_j a_j B_{jk} + \Sigma_i a_i B_{ik}}$$

$$= \sum_j B_{kj} a_j + \sum_i B_{ki} a_i$$

$$= 2\,(k^{th} \text{ row of } B)\, a$$

$$\frac{\partial (a^T B a)}{\partial a_k} = 2 b_k^T a_k$$

$b_k = k^{th}$ col of B.

$$\frac{\partial}{\partial a}(a^T B a) = 2Ba.$$

$$\frac{\partial}{\partial \mu}\left(\frac{-1}{2}\sum_i (z^i-\mu)^T \Sigma^{-1}(z^i-\mu)\right)$$

$$\frac{\partial}{\partial a}(a^T B a) = 2Ba$$

$$= \frac{\partial}{\partial \mu}\left[\frac{-1}{2}\left(\sum_i \underbrace{z^{iT}\Sigma^{-1}z^i}_{0} - 2\mu^T\Sigma^{-1}z^i + \mu^T\Sigma^{-1}\mu\right)\right]$$

$$= 0 - \sum_i 2\Sigma^{-1}z^i + \sum_i 2\Sigma^{-1}\mu = 0$$

$$-\sum_i z^i + \Sigma\mu = 0$$

$$\mu = \frac{\sum_i z^i}{N}$$

$$\boxed{\begin{aligned} &\frac{\partial}{\partial A}\log|A| = (A^{-1})^T \\ &\frac{\partial}{\partial A}(a^T A^{-1} b) = -A^{-1T} a b^T A^{-1T} \end{aligned}}$$

$$\log L = \underbrace{\sum_i -\frac{d}{2}\log 2\pi}_{0} - \frac{1}{2}\sum_i \log|\Sigma| - \frac{1}{2}\sum_i (z^i-\mu)^T\Sigma^{-1}(z^i-\mu)$$

$$\frac{\partial \log L}{\partial \Sigma} = -\frac{1}{2}\cdot N(\Sigma^{-1})^T \quad -\frac{1}{2}\sum_i (-1)(\Sigma^{-1})^T(z^i-\mu)(z^i-\mu)^T(\Sigma^{-1})^T$$

$$\frac{\partial}{\partial \Sigma} \log L = \frac{N}{2}(\Sigma^{-1})^T + \frac{1}{2}\sum_i (\Sigma^{-1})^T (z^i-\mu)(z^i-\mu)^T(\Sigma^{-1})^T = 0$$

multiply on right by Σ^T

$$\frac{N}{2} I = \frac{1}{2}\sum_i (\Sigma^{-1})^T \sum_i (z^i-\mu)(z^i-\mu)^T$$

$$\Sigma^T = \frac{1}{N}\sum_i \underbrace{(z^i-\mu)(z^i-\mu)^T}_{\text{symmetric}}$$

$$\Rightarrow \Sigma^T = \Sigma = \frac{1}{N}\sum_i (z^i-\mu)(z^i-\mu)^T$$

Develop derivative identities

ex. 1

$$\underset{\text{scalar}}{\frac{\partial}{\partial x}} \underset{}{\overset{\text{matrices}}{(AB)_{ij}}} \qquad C = AB \qquad C_{ij} = \sum_k A_{ik} B_{kj} \qquad \begin{pmatrix} i \; — \end{pmatrix}\begin{pmatrix} | \; j \end{pmatrix}$$

$$\frac{\partial}{\partial x} C_{ij} = \sum_k \frac{\partial A_{ik}}{\partial x} B_{kj} + \frac{\partial}{\partial x} B_{kj} A_{ik}$$

$$\left(\frac{\partial C}{\partial x}\right)_{ij} = \left(\frac{\partial A}{\partial x} B\right)_{ij} + \left(A \frac{\partial B}{\partial x}\right)_{ij}$$

$$\frac{\partial C}{\partial x} = \frac{\partial A}{\partial x} B + A \frac{\partial B}{\partial x}$$ (matrix product rule)

ex. 2

$$\frac{\partial}{\partial x}(A^{-1}A) = \frac{\partial}{\partial x}(I)$$

$$= \frac{\partial A^{-1}}{\partial x} A + A^{-1}\frac{\partial A}{\partial x} = 0$$

$$\frac{\partial A^{-1}}{\partial x} = -A^{-1}\frac{\partial A}{\partial x} A^{-1}$$

$$\frac{\partial \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}}{\partial d} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

ex. 3

$$\frac{\partial A^{-1}}{\partial A_{kl}} = -A^{-1}\frac{\partial A}{\partial A_{kl}} A^{-1}$$

$$= A^{-1}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} A^{-1}_{kl}$$

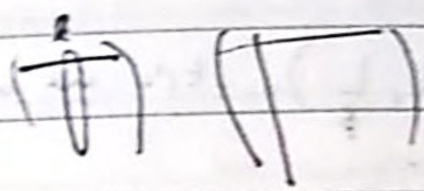
ex. 4

$$\frac{\partial\, a^T A^{-1} b}{\partial A_{kl}} \begin{matrix} \}\text{scalar} \\ \}\text{scalar} \end{matrix}$$

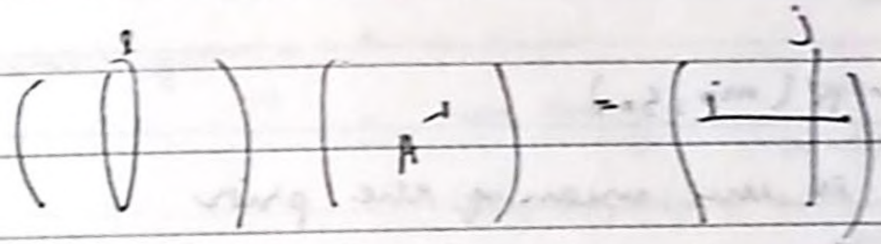
$$= a^T \frac{\partial A^{-1}}{\partial A_{kl}} b$$

$$= -a^T A^{-1}\begin{pmatrix} 0 & \cdots & 0 \\ 0 & 1 & 0 \\ \vdots & \ddots & 0 \\ 0 & \cdots & 0 \end{pmatrix}_{kl} A^{-1} b$$

$$A^{-1}\begin{pmatrix} \text{all zeros} \\ \text{except the } k\ell \\ \text{which is } 1 \end{pmatrix} = \begin{pmatrix} \text{zeros except} \\ \ell^{th} \text{ column} \end{pmatrix}$$



k^{th} column of A^{-1} moves to ℓ^{th} column of result



$A^{-1}_{ik} A^{-1}_{\ell j}$

$\hat{A}$ = Matrix whose ij entry is $A^{-1}_{ik} A^{-1}_{\ell j}$

$$\frac{\partial a^T A^{-1} b}{\partial A_{k\ell}} = -a^T \hat{A} b$$

$$= -\sum_i \sum_j a_i b_j \hat{A}_{ij}$$

$$= -\sum_i \sum_j a_i b_j A^{-1}_{ik} A^{-1}_{\ell j}$$

$$\Rightarrow \frac{\partial a^T A^{-1} b}{\partial A} = \hat{\hat{A}}$$

where entries in $\hat{\hat{A}}$ are

$$\frac{\partial}{\partial A}(a^T A^{-1} b) = -A^{-1T} a b^T A^{-1T}$$

Linear Regression

$L = p(t|w)$

$t|w \sim N(\Phi w, \frac{1}{\beta} I)$

$\Rightarrow t_i|w \sim N(w^T \phi(x_i), \frac{1}{\beta})$ t_i's are independent

Normal prior for w.

$w \sim N(m_0, S_0)$

m_0 is the mean of the prior

= mean after seeing 0 observation

S_0 is the covariance matrix of prior

= cov. after seeing 0 observation.

$\rightarrow w|t \propto P(t|w) p(w)$

$$e^{-\frac{1}{2}(t-\Phi w)^T(\frac{1}{\beta} I)^{-1}(t-\Phi w)} \; e^{-\frac{1}{2}(w-m_0)^T S_0^{-1}(w-m_0)}$$

multiply & see what this looks like in terms of w

Machine Learning

Linear Regression

$t_i \sim \mathcal{N}(w^T\phi(x_i), \frac{1}{\beta})$

t_i independent

$t|w \sim \mathcal{N}(\Phi w, \frac{1}{\beta} I)$ $\quad \begin{bmatrix} \frac{1}{\beta} & & \\ & \frac{1}{\beta} & \\ & & \ddots \end{bmatrix}$

$w \sim \mathcal{N}(m_0, S_0)$ (gaussian prior)

$x \sim \mathcal{N}(m_x, S_x)$

$y|x \sim \mathcal{N}(Ax+b, S_y)$

observe y & want $P(x \mid y = y)$

$t \equiv y$

$w \equiv x$

Generic MVN, $z \sim \mathcal{N}(\mu, \Sigma)$

$= \left(\frac{1}{2\pi}\right)^{d/2} \frac{1}{|\Sigma|^{1/2}} e^{-(z-\mu)^T\Sigma^{-1}(z-\mu)}$

$e^{-\frac{1}{2}z^T\Sigma^{-1}z} \underbrace{e^{-\frac{1}{2}\mu^T\Sigma^{-1}\mu}}_{\text{does not depend on } z \Rightarrow \text{ignore}} e^{z^T\Sigma^{-1}\mu}$

$P(z) \propto e^{-\frac{1}{2}z^T\Sigma^{-1}z} e^{z^T\Sigma^{-1}\mu}$

$P(X,Y) = P(x)\, P(Y|X) = ?$

$$= \left(\underset{①}{e^{-\frac{1}{2}x^T S_x x}} \; \underset{②}{e^{x^T S_x^{-1} m_x}} \right) \left(e^{-\frac{1}{2}(y-(Ax+b))^T S_y^{-1} (y-(Ax+b))} \; \cancel{S_y^{-1}(y-(Ax+b))} \right)$$

consider exponents of second term

$$-\frac{1}{2} y^T S_y^{-1} y - \underbrace{\frac{1}{2}(Ax+b)^T S_y^{-1}(Ax+b)} + (Ax+b)^T S_y^{-1} y$$

$-\frac{1}{2} x^T A^T S_y^{-1} A x$ ③

$- x^T A^T S_y^{-1} b$ ④

$-\frac{1}{2} b^T S_y^{-1} b$

$x^T A^T S_y^{-1} y$ ⑤

$+ b^T S^{-1} y$

ignore

①&③ quadratic terms

$$-\frac{1}{2} x^T (S_x^{-1} + A^T S_y^{-1} A) x \Rightarrow \Sigma^{-1} = S_x^{-1} + A^T S_y^{-1} A$$

on comparision

②, ④ & ⑤

$$x^T (S_x^{-1} m_x - A^T S_y^{-1} b + A^T S_y^{-1} y)$$

$$= x^T (S_x^{-1} m_x + A^T S_y^{-1} (y-b))$$

$$\Rightarrow \Sigma^{-1} \mu = S_x^{-1} m_x + A^T S_y^{-1} (y-b)$$

on comparision

$$\Rightarrow \mu = (S_x^{-1} + A^T S_y^{-1} A)^{-1} (S_x^{-1} m_x + A^T S_y^{-1} (y-b))$$

$$\left(\frac{1}{\beta}I\right)^{-1} = \beta I$$

$m_x = m_0, S_x = S_0$

$A = \Phi, b = 0$

$S_y = \frac{1}{\beta} I$

Posterior

$$w|t \sim N(\mu^*, \Sigma^*)$$

$$\Sigma^* = \left(S_0^{-1} + \Phi^T \left(\frac{1}{\beta} I\right)^{-1} \Phi\right)^{-1}$$

$$\mu^* = \Sigma^* \left(S_0^{-1} m_0 + \Phi^T \beta I \Phi\right)$$
$$= \Sigma^* \left(S_0^{-1} m_0 + \beta \Phi^T \Phi\right)$$

MAP = ~~maximum~~ input that maximizes posterior
(in normal dist., it is same as mean)
$= \mu^*$.

$$w|t \sim N(m_N, S_N)$$
$$S_N = (S_0^{-1} + \beta \Phi^T \Phi)^{-1}$$
$$m_N = S_N(\beta \Phi^T t + S_0^{-1} m_0)$$

"no information prior" (fallback to MLE)

variance $\to \infty$ $\quad S_0 = \begin{pmatrix} \infty & & \\ & \infty & \\ & & \ddots \end{pmatrix} \quad S_0^{-1} = \begin{pmatrix} 0 & & \\ & 0 & \\ & & \ddots \end{pmatrix}$

$$S_N = (\beta \Phi^T \Phi)^{-1}$$
$$m_N = (\beta \Phi^T \Phi)^{-1} (\beta \Phi^T t + 0)$$
$$= (\beta \Phi^T \Phi)^{-1} (\beta \Phi^{-1} t)$$
$$= \frac{1}{\beta} (\Phi^T \Phi)^{-1} \beta (\Phi^{-1} t) \quad = (\Phi^T \Phi)^{-1} (\Phi^{-1} t)$$

$$w \sim N(0, \frac{1}{\alpha} I)$$

alpha

$$S_N = \left(\alpha I + \beta \phi^T \phi\right)^{-1}$$

$$m_N = \left(\alpha I + \beta \phi^T \phi\right)^{-1} \left(\beta \phi^T t + 0\right)$$

$$= \left(\frac{\alpha}{\beta} I + \phi^T \phi\right)^{-1} (\phi^T t)$$

$$= \left(\frac{\alpha}{\beta} I + \phi^T \phi\right)^{-1} (\phi^T t)$$

let $\lambda = \frac{\alpha}{\beta}$

$$m_N = (\lambda I + \phi^T \phi)^{-1} (\phi^T t)$$

regularized linear regression

Predictive distribution

$$P(t_{N+1} | w) \sim N(\phi(x_N + 1)^T w, \frac{1}{\beta})$$

$$P(t_N + 1 | t) \sim ?$$

$w \sim N(m_N, S_N)$

$t_{N+1} | w \sim N(\Phi(x_{N+1})^T w, \frac{1}{\beta})$

$A = \phi^T$

$b = 0$

$t_{N+1} \sim N(\Phi(x_{N+1})^T m_N,$

$\frac{1}{\beta} + \Phi(x_{N+1})^T S_N \Phi(x_{N+1}))$

What if β is not known?

conjugate prior/posterior for $(w, beta)$

Likelihood

$\left(\frac{1}{2\pi}\right)^{N/2} \beta^{N/2} e^{-\frac{1}{2}(\phi w - t)^T \beta I (\phi w - t)}$

$\beta^{N/2} e^{-\frac{\beta}{2}(\phi w - t)^T(\phi w - t)}$

$\Sigma = \frac{1}{\beta} I$

$|\Sigma| = \left(\frac{1}{\beta}\right)^N$

$\frac{1}{|\Sigma|} = \beta^{N/2}$

What would be a conjugate prior for β?

$\beta^{\square} e^{-\beta \square}$

Gamma dist.

$\text{Gamma}(\lambda | a, b)$

$= \frac{1}{\Gamma(a)} b^a \lambda^{a-1} e^{-b\lambda}$

(ay :

$$\left.\begin{array}{l} w \sim N(w_0, S_0) \\ \beta \sim \text{Gamma}(\beta, a, b) \end{array}\right\} \text{independent}$$

$p(\beta, w \mid t)$

$S_N = (S_0^{-1} + \beta \Phi^T \Phi)^{-1}$ – dependent

would $S_N = \frac{1}{\beta} S_N$

Machine Learning

Goal: use bayesian solution for both w, β.

problem: { even if we start with independent $p(w)$ & $P(\beta)$ in prior, posterior will have $p(w, \beta)$ with w, β correlated

$$t_i \sim N(w^T \phi(x_i), \frac{1}{\beta})$$

$$t \sim N(\Phi w, \frac{1}{\beta} I)$$

Gamma Distribution

$$P(X | a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} e^{-b\lambda} \qquad a, b, \lambda > 0$$

Distribution over positive random variable

Prior

$P(\beta | a_0, b_0) = \text{Gamma}(\beta, a_0, b_0)$ ... marginal

$P(w | m_0, S_0, \beta) = N(m_0, \frac{1}{\beta} S_0)$ --- conditional

Hope for posterior

$P(\beta | a_N, b_N) = \text{Gamma}(\beta | a_N, b_N)$

$P(w | m_N, S_N, \beta) = N(m_N, \frac{1}{\beta} S_N)$

Posterior $\propto$ Prior $\times$ Likelihood

β prior . . . $\frac{1}{\Gamma(a_0)} b_0^{a_0} \beta^{a_0-1} e^{-b_0\beta}$

w prior . . . $\left(\frac{1}{2\pi}\right)^{d/2} \frac{1}{|\frac{1}{\beta}S_0|^{1/2}} e^{-\frac{1}{2}(w-m_0)^T(\frac{1}{\beta}S_0)^{-1}(w-m_0)}$

Likelihood . . $\left(\frac{1}{2\pi}\right)^{N/2} \frac{1}{\sqrt{|\frac{1}{\beta}I|}} e^{-\frac{1}{2}(t-\Phi w)^T\beta I(t-\Phi w)}$

$\sqrt{|\frac{1}{\beta}I|} \longrightarrow \sqrt{\left(\frac{1}{\beta}\right)^N} = \left(\frac{1}{\beta}\right)^{N/2}$

param	(w,β)	prior(w,β)
data	t	
posterior	$p(w\mid t)$	

$|\frac{1}{\beta}S_0|^{1/2}$ $= \left(\frac{1}{\beta}\right)^d |S_0|$ $\Rightarrow \frac{1}{\sqrt{|\frac{1}{\beta}S_0|}} = \beta^{d/2}\frac{1}{\sqrt{|S_0|}}$

ignoring constant terms

$\beta^{a_0-1} e^{-b_0\beta} e^{-\frac{\beta}{2}w^TS_0^{-1}w} e^{-\frac{\beta}{2}m_0^TS_0^{-1}m_0} e^{\beta w^TS_0^{-1}m_0}$

$\beta^{N/2} e^{-\frac{\beta}{2}t^Tt} e^{-\frac{\beta}{2}w^T\Phi^T\Phi w} e^{\beta w^T\Phi^T t}$

Quadratic term

$$-\frac{\beta}{2} w^T(S_0^{-1} + \Phi^T\Phi) w \quad \text{v/s} \quad -\frac{1}{2} z^T \Sigma^{-1} z$$

$$\Sigma^{-1} = \beta(S_0^{-1} + \Phi^T\Phi)$$

$$\Sigma = \frac{1}{\beta}\underbrace{(S_0^{-1}+\Phi^T\Phi)^{-1}}_{S_N} \qquad S_N = (S_0^{-1}+\Phi^T\Phi)^{-1}$$

linear term

$$w^T(\beta S_0^{-1} m_0 + \beta \Phi^T t) \quad \text{v/s} \quad z^T\Sigma^{-1}\mu$$

$$\cancel{w^T}\,[\;\;] = \cancel{z^T}\,\Sigma^{-1}\mu$$

$$\Sigma^{-1}\mu = [\;\;]$$

$$\mu = \Sigma\,[\;\;]$$

$$\mu = \Sigma(\beta)(S_0^{-1}m_0 + \Phi^T t)$$

$$\mu = \frac{1}{\cancel{\beta}}(S_0^{-1}+\Phi^T\Phi)^{-1}\cancel{\beta}(S_0^{-1}m_0+\Phi^T t)$$

$$\boxed{\mu = (S_0^{-1}+\Phi^T\Phi)^{-1}(S_0^{-1}m_0+\Phi^T t)}$$

next work on β

But

look at posterior on w.

$$\frac{1}{(2\pi)^{d/2}}\frac{1}{\sqrt{\frac{1}{\beta}S_N}}\, e^{-\frac{1}{2}(w^T(\frac{1}{\beta}S_N)^{-1}w} \quad \frac{e^{-\frac{1}{2}m_N^T(\frac{1}{\beta}S_N)^{-1}m_N}}{} \quad e^{w^T(\frac{1}{\beta}S_N)^{-1}m_N}$$

$$\underline{\frac{1}{\beta}^{d/2}} \quad \frac{1}{\sqrt{|S_0|}}$$

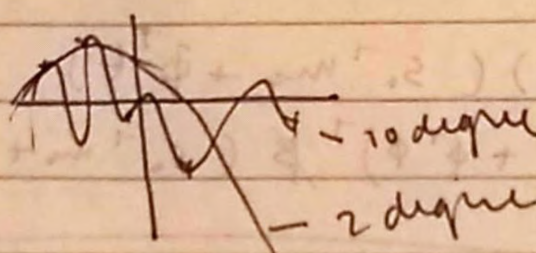
not accounted for

$$\text{Gamma}(x \mid a, b) \propto x^{a-1} e^{-bx}$$

$$\beta^{a_0 + \frac{N}{2} - 1} \Rightarrow a_N = a_0 + \frac{N}{2}$$

$$e^{-\beta \underbrace{\left(b_0 + \frac{1}{2} m_0^T S_0^{-1} m_0 + \frac{1}{2} t^T t - \frac{1}{2} m_N^T S_0^{-1} m_N\right)}_{b_N}}$$

∴ no term remains, by completing the squares the form of the posterior is Normal × Gamma



$$\Phi = \begin{pmatrix} 1 & x & x^2 & \cdots & x^{10} \\ \vdots & \vdots & \vdots & & \end{pmatrix}$$

$$\Phi = \begin{pmatrix} 1 & x & x^2 \\ \vdots & \vdots & \end{pmatrix}$$

$$\begin{pmatrix} x \\ \vdots \\ t \end{pmatrix}$$

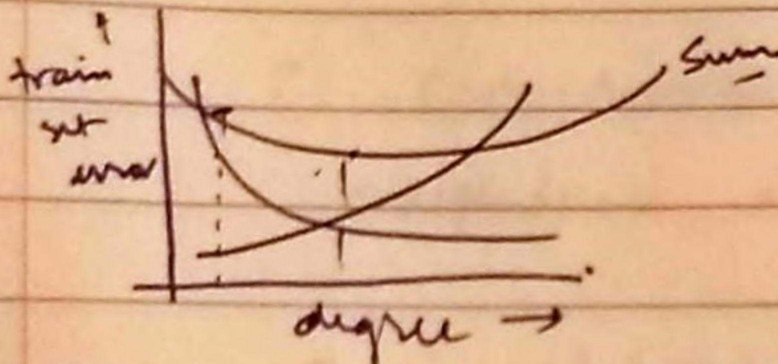
Model Selection

How to pick extra parameters

& param of prior

Simple intuition

Bayesian information criteria



Size of confidence interval grows with complexity of model class

Evidence maximization.

Machine Learning

Model Selection

$p(w) = N(0, \frac{1}{\alpha} I)$

$p(t|w) = N(\Phi w, \frac{1}{\beta} I)$

Posterior $\quad p(w|t) \sim N(m_N, S_N)$

$m_N = \beta S_N \Phi^T t$

$S_N = (\alpha I + \beta \Phi^T \Phi)^{-1}$

what values of α, β ?

hyper-parameters

Polynomial regression

$\phi(x) = \begin{pmatrix} 1 \\ x \\ \vdots \\ x^d \end{pmatrix}$

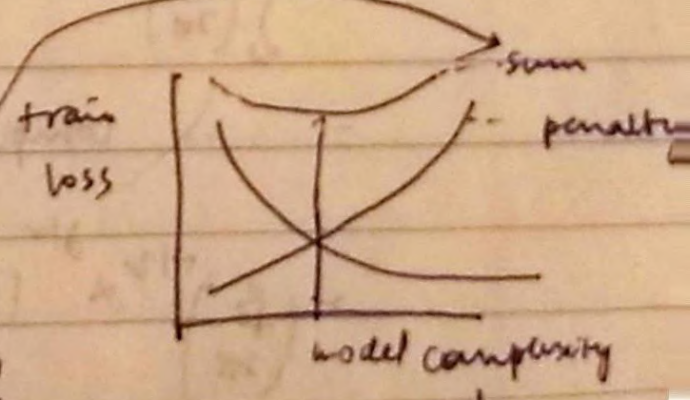
what value of d?

Problem 1	Problem 2	problem
fixed d	pick α, β, d	θ parameters
pick α, β	$\vdots$	γ hyperparameters
so as to		Prior $p(\theta\|\gamma)$
get "best"		Likelihood $p(t\|\theta)$
posterior		find $p(\theta\|t, \gamma)$
		what value for γ

Standard trick

min $\text{Loss}(\gamma) + \text{Penalty}(\gamma)$

to pick γ.



BIC,
AIC,
MDL, SRM

Pick r using evidence function

Evidence $P(t \mid r)$

$\equiv$ Max. likelihood on hyper parameters

$\equiv$ 2nd level Max. L

$\equiv$ Emperical bayes

$\equiv$ Evidence approximation.

Bishop: ① intuitive arguments that this works

② rough calculation shows that we get BIC as a special case.

$$\text{Evidence}(r) = \int_\theta \underbrace{P(\theta \mid r)}_{\text{prior}} \; \underbrace{P(t \mid \theta)}_{L} \, d\theta$$

pick hyper param to max

Model 1

$$Ev = \int_w \left(\frac{1}{2\pi}\right)^{d/2} \alpha^{d/2} e^{-\frac{\alpha}{2} w^T w} \cdot \left(\frac{1}{2\pi}\right)^{N/2} \beta^{N/2} e^{-\frac{\beta}{2}(\phi w - t)^T(\phi w - t)}$$

$= \ldots$ completing the square to normal dist

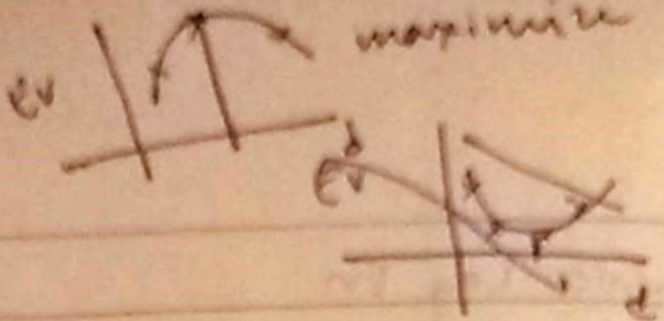
$$= \left(\frac{\beta}{2\pi}\right)^{N/2} \alpha^{d/2} |S_N|^{1/2} e^{-\frac{\beta}{2}\|\phi m_N - t\|^2} e^{-\frac{\alpha}{2}\|m_N\|^2}$$

1. Calculate Log Fn
2. Take derivative
3. Solve for α, β.

Solution gives an iterative algorithm
1. Calc. λ_i eigenvalues of $\beta \phi^T \phi$
2. Calc. $\gamma = \sum_{i=1}^{d} \frac{\lambda_i}{\lambda_i + \alpha}$
3. Update $\alpha = \frac{\gamma}{\|m_N\|^2}$

$$\frac{1}{\beta} = \frac{1}{N-\gamma} \|\phi m_N - t\|^2$$

Model Selection Algorithm
1. Initialize α, β
2. Repeat until convergence
   - Calculate m_N, S_N
   - Update α, β



Model 2,

alg. to select α, β, d

for $d = 1, \ldots D$

run. alg. for Model Selection #1 to pick α_d, β_d

Calculate evidence using d, α_d, β_d

Pick d, α_d, β_d which max. Evidence.

	eigenvalues	
$\Phi^T\Phi$	$\hat{\lambda}_i$	
$\beta\Phi^T\Phi$	$\lambda_i = \beta\hat{\lambda}_i$	
$\alpha I + \beta\Phi^T\Phi$	$\alpha + \beta\hat{\lambda}_i$	
S_N	$\frac{1}{\lambda_i + \alpha}$	$= \frac{1}{\beta\hat{\lambda}_i + \alpha}$

$\lambda_i \to \infty \Rightarrow$ var in direction ≈ 0

$\lambda_i \to 0 \Rightarrow$ var $\approx \infty$

$\frac{\lambda_i}{\lambda_i + \alpha}$ $\xrightarrow{(\lambda_i \to \infty)} 1$

$\xrightarrow{(\lambda_i \to 0)} 0$

$\gamma = \sum \frac{\lambda_i}{\lambda_i + \alpha} \approx$ no. of determined dimensions

Machine Learning

Linear Regression

Limitations: model is linear

predicting $t \in \mathbb{R}$

Classification : when t is discrete

$$\begin{pmatrix} \Phi \end{pmatrix} \quad \begin{array}{c} t \\ \begin{pmatrix} t \\ s \\ l \end{pmatrix} \end{array} \quad \begin{array}{c} t \\ \begin{pmatrix} Y \\ N \\ Y \\ Y \end{pmatrix} \end{array} \quad \begin{array}{c} t \\ \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \end{pmatrix} \end{array}$$

multi-class   binary

target value = label.

Simplest soln ~~fo~~ for 2 class

use linear regression

$Y \longleftrightarrow 1$

$N \longleftrightarrow -1$

use max $L \rightarrow w$

For new example x,

compute $a = w^T x$

if $a \geq 0 \Rightarrow$ say yes

otherwise $\Rightarrow$ no

Methodology

Specify a model that explains how data is generated

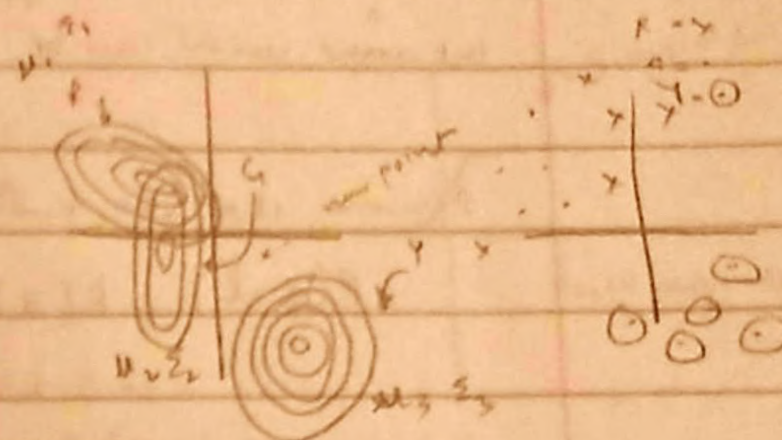
Then given data learn parameters of that model. (or a posterior over parameters)

Example

$p(c=1)$ $P(R) = 0.5$

$p(c=2)$ $P(G) = 0.4$

$p(c=3)$ $P(Y) = 0.1$



$P(x|R)$

$P(x|G)$

$P(x|Y)$

To generate data

for each i

pick class $l_i \in \{1, \ldots, K\}$

pick x from $P(x \mid c = l_i)$

We also want to include cases where features are discrete

$$\phi = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

we will need a different $P(x \mid C = l)$

but same model can be used

Assume that we know

$P(C = l) \quad P(x \mid C = l)$

How is a new example classified.

compute $P(C = j \mid x)$

$$P(C = j \mid x) = \frac{P(x \,\&\, C = j)}{P(x)}$$

$$= \frac{P(C = j)\, P(x \mid C = j)}{\sum_i P(x \mid C = i)\, P(C = i)}$$

Define $a_j = \log \left[P(C = j)\, P(x \mid C = j) \right]$

$$\frac{e^{a_j}}{\sum_i e^{a_i}} \leftarrow \text{Softmax}$$

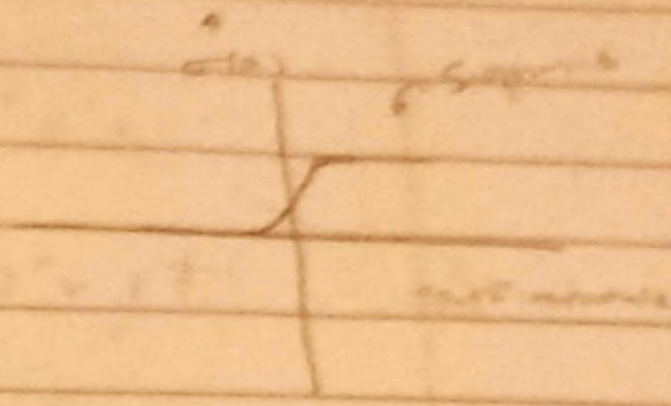
In 2 class case

$$P(C=1) = \frac{e^{a_1}}{e^{a_1} + e^{a_2}}$$

$$= \frac{1}{1 + e^{a_2 - a_1}}$$

Define a to be $a_1 - a_2$

$$\Rightarrow \frac{1}{1 + e^{-a}}$$



with equal costs, predict $C=1$

$\iff P(C=1) \geq 1/2$

$\iff a \geq 0$

What does prediction look like in geometric space

when $P(X \mid C=j) = N(\mu_j, \Sigma_j)$

Consider case with $K=2$, $\Sigma_1 = \Sigma_2 = \Sigma$

When do we predict class $=1$

$a = a_1 - a_2$

$$= \log \frac{P(C=1)\, P(X \mid C=1)}{P(C=2)\, P(X \mid C=2)}$$

$a = \log P(c=1) - \log P(C=2)$

$$- \cancel{\frac{d}{2}\log 2\pi} - \cancel{\frac{1}{2}\log|\Sigma|} - \frac{1}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)$$

$$+ \cancel{\frac{d}{2}\log(2\pi)} + \cancel{\frac{1}{2}\log|\Sigma|} + \frac{1}{2}(x-\mu_2)^T \Sigma^{-1} (x-\mu_2)$$

$$\geq \frac{1}{2}$$

linear in x

$$-\frac{1}{2} x^T \Sigma^{-1} x - \frac{1}{2}\mu_1^T \Sigma^{-1}\mu_1 + x^T \Sigma^{-1}\mu_1$$

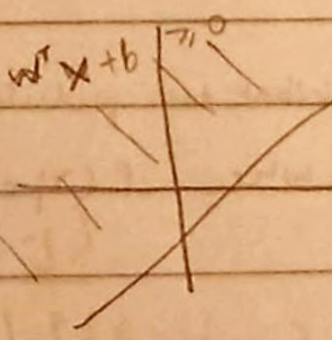
cancel out

$$+\frac{1}{2} x^T \Sigma^{-1} x + \frac{1}{2}\mu_2^T \Sigma^{-1}\mu_2 + x^T \Sigma^{-1}\mu_2 \geq \frac{1}{2}$$

$$W = \Sigma^{-1}(\mu_2 - \mu_1)$$

$$b = -\frac{1}{2} + \log\left(\frac{P(c=1)}{P(c=2)}\right) - \frac{1}{2}\mu_1^T \Sigma^{-1}\mu_1 - \frac{1}{2}\mu_2^T \Sigma^{-1}\mu_2$$

$$x^T w + b \geq 0$$

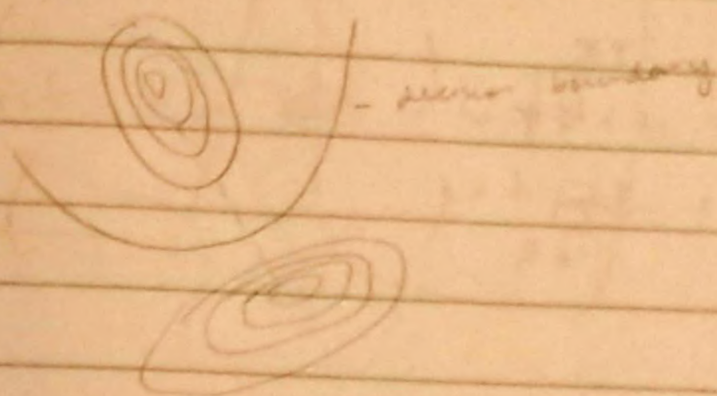


When $\Sigma_1 \neq \Sigma_2$

only one step in the derivation changes

$$\frac{1}{2} x^T \Sigma_2^{-1} x - \frac{1}{2} x^T \Sigma_1^{-1} x = x^T \left(\frac{1}{2} \left(\Sigma_2^{-1} - \Sigma_1^{-1} \right) \right) x$$

Quadratic



ex

4 classes

Max likelihood estimate

$\{ P_i, \mu_i, \Sigma_i \}$

$P_i = P(C = i)$

Likelihood $= \left[\prod_{i \in G} P_G N(x_i | \mu_G, \Sigma_G) \right]$

$\left[\prod_{i \in R} P_R N(x_i | \mu_R, \Sigma_R) \right]$

$\left[\prod_{i \in Y} P_Y N(x_i | \mu_Y, \Sigma_Y) \right]$

$\left[\prod_{i \in B} P_B N(x_i | \mu_B, \Sigma_B) \right]$

each μ_i is independent of μ_j

$$\log L = \left(\sum_{i \in G} \log P_G + \sum_{i \in G} \log N(x_i \mid \mu_G, \Sigma_G) \right)$$

$$+ \sum_{i \in R} (\quad \cdot \quad + \quad)$$

$$+ \sum_{i \in Y} (\quad + \quad)$$

$$+ \sum_{i \in B} (\quad \cdot \quad + \quad)$$

For μ_i Σ_i same as max L for MVN params

For $P_G + P_R + P_B + P_Y = 1$

$$P_G = \frac{\#\text{ points } G}{\text{total }\#\text{ of points}}$$

$$\Sigma = \frac{1}{N} \overbrace{\Sigma}^{\text{sum}} (x_i - \mu)(x_i - \mu)^T$$

Shared covariance

$$\Sigma = \frac{1}{N} \left(\sum_G (x_i - \mu_G)(x_i - \mu_G)^T \right.$$

$$+ \sum_B (x_i - \mu_B)(x_i - \mu_B)^T$$

$$\left. + \ \ldots \right)$$

data $\rightsquigarrow$ lot of params

$\{p_i, \mu_i, \Sigma_i\}_{i=1..k}$

$k + k \cdot d + k \cdot d^2$ params

param $\rightsquigarrow$ w, b

Predict C_1 (x) $w^T x + b > 0$

$d + 1$ params

Machine Learning

$P(C=c)$

$p(X|C=c)$

observed $x_1, \dots, x_N$

$\Rightarrow$ Max L for param of $P(C=c)\,P(X|C=c)$

$\{P_i, \mu_i, \Sigma_i\}_{i=1\dots k}$

To predict on new example

$w = f(\{P_i, \mu_i, \Sigma_i\})$

$b = g(\dots)$

z is yes $\Leftrightarrow$ $w^T \Phi(z) + b \geq 0$

instead of writing $w^T \Phi(x) + b$

$w = (w_1 \dots w_d)$

$v = (\underbrace{w_0}_{\text{represents } b}, w_1 \dots w_d)$

$\hat{\Phi}(x) = (1, \phi_1(x) \dots \phi_d(x))$

$v^T \hat{\Phi}(x) = w_0 + w^T \Phi(x)$

set $w_0 = b$

v:-

~~logistic~~ linear regression

$a_i = w^T \phi(x_i)$

$y_i = a_i$

$t_i \sim N(y_i, \frac{1}{\beta})$

logistic regression

$a_i = w^T \phi(x_i)$

$y_i = \sigma(a_i)$

sigmoid

$t_i \sim \text{Bernoulli}(y_i)$

drawn independently

random variable

Prob that label of ith example is 1 $= y_i$

$\Phi = \begin{pmatrix} \equiv \end{pmatrix} \qquad t = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}$

$\Rightarrow w_{\max L}$

$$L = \prod_i y_i^{t_i} (1-y_i)^{1-t_i}$$

take log

$$\log L = \sum_i t_i \log y_i + (1-t_i)\log(1-y_i)$$

diff wrt w,

$$\frac{d \log L}{dw} = \sum_i t_i \frac{1}{y_i}\frac{dy_i}{dw} + (1-t_i)\left(\frac{-1}{1-y_i}\right)\frac{dy_i}{dw}$$

$$\frac{\partial y_i}{\partial w} = \frac{\partial y_i}{\partial a_i} \frac{\partial a_i}{\partial w}$$

$$y = \sigma(a)$$

$$\frac{dy}{da} = \frac{d}{da}\left(\frac{1}{1+e^{-a}}\right)$$

$$= (-1)\frac{1}{(1+e^{-a})^2}(e^{-a})(-1)$$

$$= \cancel{e^{-a}} \frac{1}{1+e^{-a}} \cdot \frac{e^{-a}}{1+e^{-a}}$$

$$= \sigma(a)\left(1-\sigma(a)\right)$$

$$\frac{\partial a_i}{\partial w} = \frac{d}{dw} w^T \phi(x_i) = \phi(x_i)$$

$$\frac{d}{dw}\log L = \sum_i t_i \frac{1}{y_i} \frac{dy_i}{\partial a_i}\frac{\partial a_i}{\partial w} + \sum_i (1-t_i)\frac{1}{1-y_i}\frac{\partial y_i}{da_i}\frac{da_i}{dw}(-1)$$

$$= \sum_i \frac{t_i}{\cancel{y_i}} y_i(1-y_i)\,\phi(x_i) - \sum\left(\frac{1-t_i}{\cancel{1-y_i}}\right) y_i \cancel{(1-y_i)}\,\phi(x_i)$$

$$= \sum_i \left[t_i\phi(x_i) - \cancel{t_i y_i \phi(x_i)} - y_i \phi(x_i) + \cancel{t_i y_i \phi(x_i)}\right]$$

$$= \sum_i \left[\underbrace{(t_i - y_i)}_{\text{vector elements}} \underbrace{\phi(x_i)}_{\text{matrix columns}}\right] = \Phi^T(t-y)$$

$$\frac{d}{dw}\log L = 0$$

$$\cancel{\sum_i t_i \phi(x_i) = \sum_i y_i \phi(x_i)}$$

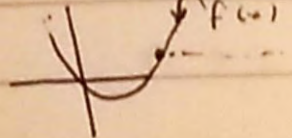
$$\Phi^T(t-y) = 0$$

log L is concave and has only one maxima. So, GD will converge.

$$\sum t_i \, \Phi(x_i) = \sum y_i \, \Phi(x_i)$$
$$= \sum \sigma(w^T \Phi(x_i)) \, \phi(x_i)$$

No simple "closed form solution" for w.

optimize directly.



where to go from here?

if f is increasing go left
if f is decreasing go right

To maximize, go with gradient
To minimize, go against gradient

Gradient Descent.

initialize x

Repeat

$$x \leftarrow x - \eta \cdot f'(x)$$

Gradient Descent for Logistic Regression

$$w \leftarrow w + \eta \, \underbrace{\Phi^T (t-y)} \;\text{.....}\; \frac{d}{dw} \log L$$

$$\equiv w \leftarrow w + \eta \frac{d}{dw} \log L$$

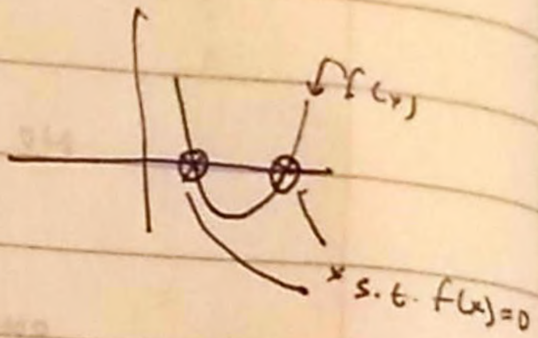
Maximize log likelihood.

if fn. is linear
this method terminates in
1 step and is exact

Newton's method.

To find a <u>zero</u> of $f(x)$

$$x_{n+1} \Leftarrow x_n - \frac{f(x_n)}{f'(x_n)}$$



ex. $f(x) = (x-2)(x-6) = x^2 - 8x + 12$

$x_0 = 8$

$$x_1 = 8 - \frac{12}{8}$$

why?

$$f(x_0+h) = f(x_0) + f'(x_0)h + \frac{1}{2} f''(x_0)h^2 + \ldots$$

Taylor expansion

higher order terms are small; ignore them

$$\approx f(x_0) + f'(x_0)h$$

if $f(x_0+h) = 0$

$$\Rightarrow f(x_0) + f'(x_0)h = 0$$

$$\Rightarrow h = \frac{-f(x_0)}{f'(x_0)}$$

if fn is quadratic
this method terminates in 1 step
and is exact

Newton's method for finding extremum of a function
[find zero of $f'(x)$]

$$x_{n+1} \leftarrow x_n - \frac{f'(x_n)}{f''(x_n)}$$

ex. $f(x) = (x-2)(x-6) = x^2 - 8x + 12$

$x_0 = 8$

$x_1 = 8 - \frac{8}{2} = 8 - 4 = 4 \leftarrow$ minimum at point this

consider $F : \mathbb{R}^k \longrightarrow \mathbb{R}$

$$F(x_0+h) = F(x_0) + \left(\frac{\partial F}{\partial x}\Big|_{x_0}\right)^T h + \underbrace{\dots}_{\text{ignore higher order terms}}$$

$F \quad \mathbb{R}^k \longrightarrow \mathbb{R}^k$

$$J = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_k} \\ \vdots & & & \vdots \\ \frac{\partial y_k}{\partial x_1} & \frac{\partial y_k}{\partial x_2} & & \frac{\partial y_k}{\partial x_k} \end{pmatrix}$$

J is different from change of R.V.

$$h = \begin{pmatrix} h_1 \\ \vdots \\ h_k \end{pmatrix}$$

To find zero of $F(\cdot)$

$$0 = F(x_0+h) \approx F(x_0) + J|_{x_0} \cdot h$$

$$h = -J|_{x_0}^{-1} F(x_0)$$

To find min of $G : \mathbb{R}^n \to \mathbb{R}$

find zero of $f = \frac{\partial G}{\partial x}$

In this case $J(F)$ = Matrix of 2nd derivatives

$$\text{Hessian} = \begin{pmatrix} \frac{\partial^2 G}{\partial x_1^2} & \frac{\partial G}{\partial x_1 \partial x_2} & \cdots & \frac{\partial G}{\partial x_1 x_n} \\ \vdots & & & \\ \frac{\partial G}{\partial x_2 \partial x_1} & & \cdots & \frac{\partial G}{\partial x_n \partial x_n} \end{pmatrix}$$

$$x_{n+1} \leftarrow x_n - H|_{x_n}^{-1} \frac{\partial G}{\partial x}\Big|_{x_n}$$

$g(x) = 5x_1^2 + 6x_1x_2 + 3x_2^2$

find minimum of g

$$\frac{dg}{dx} = \begin{pmatrix} 10x_1 + 6x_2 \\ 6x_1 + 6x_2 \end{pmatrix}$$

$$\begin{array}{ccc} & \frac{\partial^2}{\partial x_1^2} & \frac{\partial\partial}{\partial x_1 x_2} \\ H = & \begin{pmatrix} 10 & 6 \\ 6 & 6 \end{pmatrix} & \\ & \frac{\partial\partial}{\partial x_2 x_1} & \frac{\partial\partial}{\partial x_2 x_2} \end{array}$$

$$H^{-1} = \frac{1}{24} \begin{pmatrix} 6 & -6 \\ -6 & 10 \end{pmatrix}$$

$h(x) = 5x_1^3 + 6x_1x_2 + 3x_2^2$

$$\frac{\partial h}{\partial x} = \begin{pmatrix} 15x_1^2 + 6x_2 \\ 6x_1 + 6x_2 \end{pmatrix}$$

$$H = \begin{pmatrix} 30x_1 & 6 \\ 6 & 6. \end{pmatrix}$$

Logistic Regression

Also called iterative reweighted least squares (IRLS)

$$\frac{\partial}{\partial w} \log L = \sum_i (t_i - y_i)\, \phi(x_i)$$

$$\frac{\partial \log L}{\partial w \partial w^T} = -\sum_i \cancel{\left(\frac{\partial t}{\partial w}\right)} \phi(x) \frac{\partial y_i}{\partial w^T}$$

$$= -\sum_i \phi(x_i) \frac{\partial y_i}{\partial w^T}$$

$A_1 \left(\diagdown \right)$

$\sum a_i b_i^T = AB^T$

$$= -\sum_i \phi(x_i)\, \underbrace{y_i(1-y_i)\, \phi(x_i)^T}_{\text{almost same as calculated earlier}}$$

$$= -\sum_i \underbrace{\phi(x_i)\, y_i(1-y_i)}\, \underbrace{\phi(x_i)^T}$$

$$= -\Phi^T R \Phi \qquad R = \operatorname{diag}\{y_i(1-y_i)\}$$

$R \equiv$ positive (or 1)

$$w \leftarrow w + (\Phi^T R \Phi)^{-1} \Phi^T (\cancel{t} - y)$$

$$\text{or } w \leftarrow w - (\Phi^T R \Phi)^{-1} \Phi^T (y - t)$$

Proof that our function is concave (i.e. has 1 maxim

~~or Guarantee that~~

we need $H \preceq 0$ ~~positive~~ negative definite

$c^T H c \leq 0$ $\quad$ c - arbitrary vector

$-c^T \Phi^T R \Phi c < 0$ $\quad$ $c \neq 0$

$-(c^T \Phi^T R^{1/2})(R^{1/2} \Phi c) < 0$

$-\underbrace{(R^{1/2} \Phi^T c)^T (R^{1/2} \Phi c)}_{\text{norm}} < 0$

this hold if $R^{1/2}\Phi$ is full rank.

~~ie~~ if columns in data matrix are linearly independent

Trade off : Newton $\quad$ optimal step size but compute hessian

GD $\quad$ fixed non optimal step size but ~~need~~ no additional computation

Machine Learning

Generative model

K classes $c=1, \dots c=K$

Prob. of class k : C_k

Prob of generating data $P(X|C_k)$

Given model.

prediction

Softmax $= \dfrac{e^{a_k}}{\sum_j e^{a_j}}$

sigmoid, $\sigma = \dfrac{1}{1+e^{-(a_1 - a_2)}}$

$a_j = \ln P(C_k) \, P(X|C_k)$

When $X|C_j \sim N(\mu_j, \Sigma_j)$

Shared $\Sigma = \Sigma_j \ \forall j$

Distinct Σ_j

non-shared is easy

① Shared Σ, 2 class

Prediction $P(C=1) = \sigma(w^T X + w_0)$

(from data dist. to model)

let $w = \Sigma^{-1}(\mu_1 - \mu_2)$

$w_0 = \ln \dfrac{P(C_1)}{P(C_2)} - \frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2}\mu_2^T \Sigma^{-1} \mu_2$ 39

-ve definite hessian
$\Rightarrow$ concave fn.

Single maxm ma.

Discriminative model

Logistic Regression

$$P(C=1 \mid x) = \sigma(w^T \phi(x_i))$$

absorbed bias term w_0 into extra dimension

for analogy $\phi(x) = x$

$$a_i = w^T \phi(x_i)$$
$$y_i = \sigma(a_i)$$
$$t_i \sim \text{Bernoulli}(y_i)$$

Max^m likelihood is hard

No closed form solution

$\Rightarrow$ Optimization: Gradient ascent or Newton's method

$$\frac{\partial \log L}{\partial w} = \sum_i \phi(x_i)\left[t_i - y_i\right]$$

$$= \Phi^T (t - y)$$

$$\frac{\partial^2 \log L}{\partial w^2} = -\Phi^T R \, \Phi$$

$$R = \text{diag}\{y_i(1-y_i)\}$$

Naive Bayes

Generative model for discrete data.

Say Binary features.

$$\bar{\Phi} = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & \\ 0 & \cdots & \end{pmatrix}$$

$$P(x \mid C_k) = ?$$

how to write a compact prob. dist?

Given C_k

$$\left.\begin{array}{l} P(0\,0\,0 \mid C_k) \\ P(0\,0\,1 \mid C_k) \\ P(0\,1\,0) \\ P(0\,1\,1) \\ P(1\,0\,0) \\ P(1\,0\,1) \\ P(1\,1\,0) \\ P(1\,1\,1) \end{array}\right\}$$

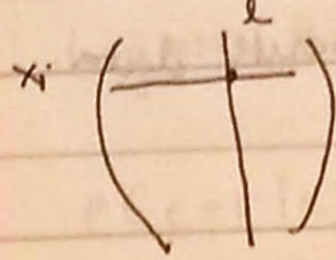
2^n - exponential for general dist.

Specifying a general dist. requires too much time / space

A very simple restriction: assume that features are conditionally independent given the label.

$E_i \perp E_j$ $\quad x_i \perp x_j \mid c$

Notation



x_{il} = lth bit of ith example.

$$P(x_i \mid C_k) = \prod_{l=1}^{d} \mu_{kl}^{x_{il}} (1-\mu_{kl})^{1-x_{il}}$$

μ_{kl} = parameter for l^{th} bit given that label = k^{th} class

Let first ~~label~~ example have label 3

$$P(C_3)\, p(x_1 \mid C_3)$$
$$= P(C_3) \prod \mu_{3l}^{x_{1l}} (1-\mu_{3l})^{1-x_{il}}$$

boolean

$$L = \prod_i \prod_k \left[P(C_k) \prod_l \mu_{kl}^{x_{il}} (1-\mu_{kl})^{1-x_{il}} \right]^{t_{ik}}$$

MLE for $P(C_k)$ is same as general case

$$\log L = \sum_i \sum_k \left[\ln P(C_k) + \sum_l x_{il} \ln \mu_{kl} + \sum_l (1-x_{il}) \ln(1-\mu_{kl}) \right] \cdot t_{ik}$$

$$\frac{\partial \log L}{\partial \mu_{kl}} = \sum_{\substack{i \\ \text{such that} \\ t_{ik}}} \left[\frac{x_{il}}{\mu_{kl}} - \frac{1-x_{il}}{1-\mu_{kl}} \right] = 0$$

$$\Rightarrow \sum_i x_{il}(1-\mu_{kl}) - \sum_i \mu_{kl}(1-x_{il}) = 0$$

$$\sum x_{il} - \cancel{\sum x_{il}\,\mu_{kl}} - \sum_i \mu_{kl} + \cancel{\sum \mu_{kl}}\, x_{il} = 0$$

$$\sum_{t_i=k} x_{il} - N_k\mu_{kl} = 0$$

$$\sum_{t_i=k} x_{il} = N_k\mu_{kl}$$

$$\mu_{kl} = \frac{\sum_{t_i=k} x_{il}}{N_k} = \frac{\text{no. of examples in } l^{th} \text{ bit set}}{\text{no. of examples of class } k}$$

It can be shown that

$$P(C_k|x) = \frac{e^{a_k}}{\sum e^{a_j}}$$

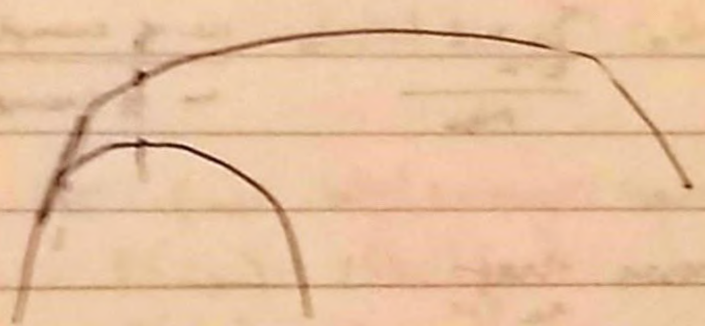
$a_j = w^T x + w_0$ $\qquad \exists\ w, w_0$ for some

1st order	gradient ascent	gradient descent
move with gradient	move against gradient
$w \leftarrow w + \eta \frac{\partial \log L}{\partial w}$	$w \leftarrow w - \eta \frac{\partial \log L}{\partial w}$

Q η = ?

Newton's method.

assume that the function is quadratic and jump to the location that gives the maxima of the quadratic fn.



2nd order

$$w \leftarrow w + H^{-1} \frac{\partial \log L}{\partial w}$$

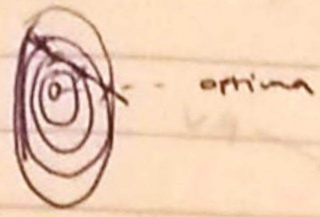
① compute hessian

② invert hessian

$$H = \frac{\partial^2 \log L}{\partial w \partial w^T}$$

alternative : approximate hessian in linear time

Line search. to choose η



want to find ideal η

1. Brute force search (expensive)
2. backtracking line search

Backtracking line search
for minimization

initialize $\eta = 1$

while $f\left(w - \eta \frac{\partial f}{\partial w}\right) > f(w) - \frac{\eta}{2} \left\| \frac{\partial f}{\partial w} \right\|^2$

$$\eta \leftarrow \frac{\eta}{2}$$

Compromise backtracking line search w/ gradient ascent/descent

What if computing derivative is expensive?

$$\frac{\partial L}{\partial w} = \sum_{i=1}^{N} \phi(x_i)(t_i - y_i)$$

sum over all data points expensive

45

Stochastic Gradient Descent

GD: $w \leftarrow w - \eta G$

$$G = \frac{\partial f}{\partial w}$$

if w can get $\hat{G}$ $\leftarrow$ RV

s.t. $E[\hat{G}] = G$ // unbiased estimator

then we can use $\hat{G}$ instead of G

$\Rightarrow w \leftarrow w - \eta_t \hat{G}$ converges to the min^m of fn.

where η_t must satisfy some conditions

$$\eta_t = \frac{1}{t}$$

* Cheap random estimate of gradient for logistic regression

To get $\hat{G}$

pick $i \in 1 .. N$ at random (uniformly)

$\hat{G} = N\,\phi(x_i)(t_i - y_i)$

$E(\hat{G}) = \frac{1}{N} \sum_i N(\phi_i)(t_i - y_i) = \sum \phi_i (t_i - y_i)$

Alternative minibatch SGD-

take k examples and take mean

instead of 1 example

Machine Learning

Exponential family of distributions

Any distribution that can be written in form $p(x) = h(x)\, g(\eta)\, e^{\eta^T u(x)}$ is a member of exponential family

* constrains the form of PDF.

$\cancel{\eta^T u(x) - A(\eta)}$

$A(\eta) = -\log g(\eta) \Rightarrow p(x) = h(x)\, e^{\eta^T u(x) - A(\eta)}$

Bernoulli $\in$ exp. family

$$p(x) = \mu^x (1-\mu)^{1-x}$$
$$= e^{(\log \mu)x + (\log(1-\mu))(1-x)}$$
$$= e^{(\log \mu)x + \log 1-\mu - \log(1-\mu)x}$$
$$\cancel{= e^{(\log \mu)x}}$$
$$= (1-\mu)\, e^{x\left(\log \frac{\mu}{1+\mu}\right)}$$

$h(x)$ = base measure

$g(\eta)$ = log normalizer

η = Natural parameters
canonical parameters

$u(x)$ = Sufficient Statistics

dim. $\eta = 1$ $\qquad \eta = \log \frac{\mu}{1-\mu}$ $\qquad h(x) = 1$

$g(\eta) = ? = 1-\mu$

$\eta = \log \frac{\mu}{1-\mu} \Rightarrow e^{\eta} = \frac{\mu}{1-\mu} \Rightarrow e^{\eta} - \mu e^{\eta} - \mu = 0$

$\Rightarrow \boxed{\frac{e^{\eta}}{1+e^{\eta}} = \mu}$

$g(\eta) = 1 - \frac{e^{\eta}}{1+e^{\eta}} = \frac{1}{1+e^{\eta}}$ $\qquad \mu = \frac{1}{1+e^{-\eta}} = \sigma(\eta)$ sigmoid

$P(x) = \underbrace{\frac{1}{1+e^{\eta}}}_{g(\eta)} \cdot e^{\eta x}$

$U(x) = x$

$h(x) = 1$

$E[U(x)]$ = mean parameter $= \theta$ $\qquad = \mu$ for bernoulli

if entries of $U(x)$ are linearly independent

then $\theta \overset{1-1}{\longleftrightarrow} \eta$

Normal

$$p(x) = (2\pi)^{-1/2}\sigma^{-1}\exp\{-(2\sigma^2)^{-1}(x-\mu)^2\}$$

$$= (2\pi)^{-1/2}\sigma^{-1}\exp\left\{\frac{-\mu^2}{2\sigma^2}\right\}\exp\left\{\frac{-1}{2\sigma^2}x^2 - \frac{\mu}{\sigma^2}x\right\}$$

$$h(x) = 1 \quad u(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \quad \eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix}$$

$$g(\eta) = (2\pi)^{-1/2}\sigma^{-1}\exp\left\{\frac{-\mu^2}{2\sigma^2}\right\}$$

$$= \frac{\sqrt{\eta_2}}{\sqrt{\pi}}\exp\{-\eta_1^2/4\eta_2\}$$

Poisson

$$p(x) = \lambda^x \underbrace{e^{-\lambda}}_{g(\eta)} \underbrace{\frac{1}{x!}}_{h(x)}$$

$$\frac{1}{x!}e^{-\lambda}e^{(\log\lambda)x}$$

dim. $= 1$ $\quad \eta = \log\lambda \quad \lambda = e^{\eta}$

$h(x) = \frac{1}{x!} \quad U(x) = x$

Fact 1 If $x_1, \ldots x_N$ are iid sampled from an exp. family distribution then

$$p(x_1 \ldots x_N) = \left(\prod_i h(x_i)\right)\left(g(\eta)^N\right) e^{\eta^T \sum_{i=1}^{N} u(x_i)}$$

u: sufficient stats since only sum is needed for MLE and not each x_i.

Fact 2: $E[u(x)] = -\frac{\partial}{\partial \eta} \log g(\eta)$

$$\text{cov}(u(x)) = E\left[\left(u(x) - E[u(x)]\right)\left(u(x) - E[u(x)]\right)^T\right]$$

$$= -\frac{\partial^2}{\partial \eta \, \partial \eta^T} \log g(\eta)$$

Fact 3: (3.1) If we have one sample from exp. fam. dist. Max likelihood Solution is obtained when

$$u(x) = -\frac{\partial}{\partial \eta} \log g(\eta)$$

(3.2) For iid samples

$$\frac{1}{N}\sum_i u(x_i) = -\frac{\partial}{\partial \eta} \log(g(\eta))$$

$$\rightarrow \text{Max } L \cdot \frac{1}{N}\sum u(x_i) = E[u(x)]$$

Max L for Bernoulli dist

$\bar{x}$ part 3

~~$p(x) = \lambda^x e^{-\lambda} / x!$~~

~~$h(x) = \frac{1}{x!}$ $\eta = \log \lambda$ $\lambda = e^{\eta}$ $u(x) = x$~~

~~max L~~

$p(x) = \mu^x (1-\mu)^{1-x}$

$\eta = \log \frac{\mu}{1-\mu}$ $h(x) = 1$

$g(\eta) = 1-\mu = 1 - \frac{e^{\eta}}{1+e^{\eta}}$

$= \frac{1}{1+e^{\eta}}$

$= \frac{\partial}{\partial \eta} g(\eta)$

$= \frac{\partial}{\partial \eta} \left(\frac{1}{1+e^{\eta}} \right)$

$= \frac{-1}{(1+e^{\eta})^2} \cdot e^{\eta}$

$\frac{\partial}{\partial \eta} \log g(\eta)$

$\bar{x}_{ML} = -\frac{1}{g(\eta)} \frac{\partial}{\partial \eta} g(\eta)$

$\bar{x}_{ML} = + \frac{e^{\eta}}{(1+e^{\eta})^2} (1+e^{\eta})$

$\bar{x} = + \frac{e^{\eta}}{1+e^{\eta}}$ $\Rightarrow \eta = \log \frac{\bar{x}}{1-\bar{x}}$

Sanity check.

$\bar{y}$ = mean of y

$$\mu = \frac{1}{1+e^{-\eta}} = \frac{1}{1+\frac{1-\bar{y}}{\bar{y}}} = \bar{y}$$

Solving using fact 2 + fact 3 together.

$$E[\mu(x)] = \bar{x}$$

$$\Rightarrow \mu = \bar{x}$$

$$E[U(x)] = \frac{1}{N}\sum U(x_i)$$

$$\begin{pmatrix} E[x] \\ E[x^2] \end{pmatrix} = \cancel{\frac{1}{N}\sum} \frac{1}{N}\begin{pmatrix} \sum x_i \\ \sum x_i^2 \end{pmatrix}$$

$$\begin{pmatrix} E[x] \\ E[x^2] \end{pmatrix} = \frac{1}{N}\begin{pmatrix} \sum x_i \\ \sum x_i^2 \end{pmatrix}$$

$$\mu = \frac{1}{N}\sum x_i$$

$$\mu^2 + \sigma^2 = \frac{1}{N}\sum x_i^2$$

$$\sigma^2 = \frac{1}{N}\sum (x_i - \hat{\mu}_{ML})^2$$

Proof of Fact 3.1

$$L = h(x)\, g(\eta)\, e^{\eta^T U(x)}$$

$$\log L = \log h(x) + \log g(\eta) + \eta^T U(x)$$

$$\frac{\partial \log L}{\partial \eta} = \frac{\partial}{\partial \eta} \log g(\eta) + U(x) = 0$$

$$U(x) = -\frac{\partial}{\partial \eta} \log g(\eta)$$

Fact 3.2

$$L = \prod_i h(x_i)\, g(\eta)^N e^{\eta^T \sum_i U(x_i)}$$

$$\log L = \sum \log h(x_i) + N \log g(\eta) + \eta^T \sum U(x_i)$$

$$\frac{\partial \log L}{\partial \eta} = N \frac{\partial}{\partial \eta} \log g(\eta) + \sum U(x_i) = 0$$

Proof of fact 2

$$\int h(x)\, g(\eta)\, e^{\eta^T U(x)}\, dx = 1$$

$$g(\eta) \int h(x)\, e^{\eta^T U(x)}\, dx = 1$$

$$\frac{\partial}{\partial \eta} \left(g(\eta)\right) \left(\int h(x)\, e^{\eta^T U(x)}\, dx\right) = 0$$

product rule

$$\frac{\partial}{\partial \eta} g(\eta) \int h(x)\, e^{\eta^T U(x)}\, dx + g(\eta) \frac{\partial}{\partial \eta} \int h(x)\, e^{\eta^T U(x)} U(x) = 0$$

$$\int h(x)\, e^{\eta^T U(x)}\, dx = \frac{1}{g(\eta)} \quad \text{and} \quad \int h(x) e^{\eta^T U(x)}\, U(x)\, dx = \frac{E[U(x)]}{g(\eta)}$$

$$\frac{\partial}{\partial \eta} g(\eta) \frac{1}{g(\eta)} + \cancel{g(\eta)} \frac{E[U(x)]}{\cancel{g(\eta)}} = 0$$

$$\underline{\frac{1}{g(\eta)} \frac{\partial g(\eta)}{\partial \eta}} + E[U(x)] = 0$$

$$\frac{\partial}{\partial \eta} \log g(\eta) + E[U(x)] = 0 \Rightarrow E[U(x)] = -\frac{\partial}{\partial \eta} \log g(\eta)$$

$$\text{cov}(U(x_i), U(x_j)) = E\left[\left(U(x_i) - E[U(x_j)]\right)\left(U(x_j) - E[U(x_j)]\right)\right]$$

$$\frac{\partial}{\partial \eta} \log g(\eta) + g(\eta) \int h(x) e^{\eta^T U(x)} U_2(x)\, dx = 0$$

$$\frac{\partial}{\partial \eta_i} \log g(\eta) + \underbrace{g(\eta) \int h(x)\, e^{\eta^T U(x)}\, U_i(x)\, dx}_{\text{product rule}} = 0$$

$$\frac{\partial^2}{\partial \eta_i \partial \eta_k} \log g(\eta) + g(\eta) \int h(x)\, e^{\eta^T U(x)}\, U_i(x)\, U_k(x)\, dx$$

$$+ \frac{\partial}{\partial \eta_k} g(\eta) \int h(x)\, e^{\eta^T U(x)}\, U_i(x)\, dx = 0$$

$$\frac{\partial}{\partial \eta_k} \log g(\eta) + E(U_k(x)) = 0$$

$$E(U_k(x)) = -\frac{\partial}{\partial \eta_k} \log g(\eta)$$

$$= \cancel{-g(\eta)} \; \frac{+1}{g(\eta)} \frac{\partial}{\partial \eta_k} g(\eta)$$

$$\cancel{\frac{\partial}{\partial \eta_k}} \Rightarrow \frac{\partial}{\partial \eta_k} g(\eta) = E[U_k(x)]\, g(\eta)$$

$$\int h(x) e^{\eta^T U(x)} U_i(x)\, dx = \frac{E[U_i(x)]}{g(\eta)}$$

$$\int h(x) e^{\eta^T U(x)} U_i(x) U_k(x)\, dx = \frac{1}{g(\eta)} E[U_i(x) U_k(x)]$$

$$\frac{\partial^2 \log g(\eta)}{\partial \eta_i \partial \eta_k} - E[U_i(x)]\, E[U_k(x)] + E[U_i(x) U_k(x)] = 0$$

$$-E[U_i(x)]\, E[U_k(x)] + E[U_i(x) U_k(x)] = -\frac{\partial^2}{\partial \eta_i \partial \eta_k} [\log g(\eta)]$$

no. of obs.

$$L \propto g(\eta)^N e^{\eta^T \sum_i u(x_i)}$$

→ sum of sufficient stats

Conjugate prior should have

$\nu \equiv$ pretend to have seen ν no. of examples

$\bar{g} \equiv$ mean of such pseudo observations

$$p(\eta) \propto g(\eta)^\nu e^{\eta^T g}$$

added counts

$$= g(\eta)^\nu e^{g^T \eta} = g(\eta)^\nu e^{(\nu \bar{g})^T \eta}$$

mean for newer obs.

$p(x|\lambda) = \lambda^x e^{-\lambda} \frac{1}{x!}$

$\lambda > 0$

Machine Learning

Quiz 3 model selection - GLM, HW3.

Generalized Linear Models.

Linear Regression

$a_i = w^T \phi(x_i)$

$t_i \sim N(a_i, \frac{1}{\beta})$

Logistic Regression

$a_i = w^T \phi(x_i)$

$y_i = \sigma(a_i)$

$t_i \sim \text{Bernoulli}(y_i)$

$\frac{\partial \log L}{\partial w} = \sum (t_i - y_i) \phi(x_i)$

$= \Phi^T (t - y)$

$\frac{\partial^2 \log L}{\partial w} = -\sum y_i(1-y_i) \phi(x_i) \phi(x_i)^T$

$= -\Phi^T R \Phi$

$R = \text{diag}(y_i(1-y_i))$

Poisson Regression

$a_i = w^T \phi(x_i)$

$y_i = e^{a_i}$

$t_i \sim \text{Poisson}(y_i)$

Poisson Regression

$L = \prod_i y^{t_i} e^{-y_i} \frac{1}{t_i!}$

same form.

$\log L = \sum t_i \log y_i - y_i - \log t_i!$

$= \sum_i t_i (\log e^{a_i}) - y_i - \log t_i!$

$= \sum t_i a_i - y_i - \log t_i!$

$\frac{\partial \log L}{\partial w} = \sum_i t_i \phi(x_i) - e^{a_i} \phi(x_i)$

$= \sum_i (t_i - y_i) \phi(x_i)$

$$\frac{\partial \log L}{\partial w \partial w^T} = \sum_i \phi(x_i) \frac{\partial y_i}{\partial w^T}$$

$$= \sum_i \phi(x_i) e^{a_i} \phi(x_i)^T$$

$$= \sum_i e^{a_i} \phi(x_i) \phi(x_i)^T$$

$$= \sum_i y_i \phi(x_i) \phi(x_i)^T$$

$$= \Phi^T R \Phi \qquad R = \mathrm{diag}(y_i)$$

$$w \leftarrow w - (\Phi^T R \Phi)^{-1} \Phi^T (y - t)$$

Exponential Family of Distributions

$$P(x|\eta) = h(x) g(\eta) e^{\eta^T U(x)}$$

η: natural parameter

g: normalizer

U: sufficient statistics

$\mu = E[U(x)]$: mean parameter

If U is linearly independent then we can write PDF in 2 ways (given η with μ)

let $\eta = \psi(\mu)$

$\mu = \psi^{-1}(\eta)$

Bernoulli

$$P(x|\mu) = \mu^x (1-\mu)^{1-x}$$

$$P(x|\eta) = \frac{1}{1+e^\eta} e^{\eta x}$$

$$\mu = \frac{1}{1+e^{-\eta}} = \sigma(\eta)$$

$$\eta = \log \frac{\mu}{1-\mu}$$

Poisson

$$P(x|\eta) = \frac{1}{x!} e^{-e^\eta} e^{\eta x}$$

$$\eta = \log \lambda$$

$$\lambda = e^\eta$$

Any PDF $f(x)$ can be changed to add a scale parameter.

① $P(x) = \frac{1}{s} f(\frac{x}{s})$

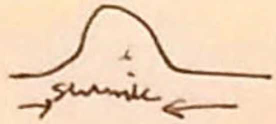
Apply to a 1D exp. family dist. $u(x) = x$

$$P(x) = \frac{1}{s} h(\frac{x}{s}) g(\eta) e^{\frac{1}{s}\eta \cdot x}$$

For any exp family dist.

$$E[u(x)] = -\frac{\partial}{\partial\eta} \log g(\eta)$$

① $E[x] = -s \frac{\partial}{\partial\eta} \log g(\eta)$

scale(∼ , s) → 

$$\mathrm{cov}(U(x)) = -\frac{\partial^2}{\partial \eta \, \partial \eta^T} \log g(\eta)$$

③ $$\mathrm{var}(x) = -s^2 \frac{\partial^2}{\partial \eta^2} \log g(\eta)$$

$$p(x) = (2\pi)^{-1/2} \sigma^{-1/2} \exp\{-(2\sigma^2)^{-1}(x^2 - 2\mu x + \mu^2)\}$$

$\sigma^2 = s$

$$= \underbrace{(2\pi\sigma^2)^{-1/2} e^{-(2\sigma^2)^{-1}\mu^2}}_{g(\eta)} \; \underbrace{e^{-(2\sigma^2)^{-1}x^2}}_{h(\frac{x}{s})} \; \underbrace{e^{\cancel{2}(\sigma^2)^{-1}\mu x}}_{e^{\frac{1}{s}\mu x}}$$

in this representation $\eta = \mu$ $\quad \psi =$ identity

Generalized Linear Model

$a_i = w^T \phi(x_i)$

$y_i = f(a_i)$ ← mean parameter and mean of t_i

f: activation $\qquad \eta_i = \psi(y_i)$ natural parameter

f^{-1}: link function.

$t_i \sim P(t_i | \eta_i)$ (exp. family dist.)

1D exp family dist.

Canonical link function picks $f(a) = \Psi^{-1}(a)$

$$\eta_i = \Psi(y_i) = \psi(\Psi^{-1}(a_i)) = a_i$$

for linear regression, Ψ = identity.

for logistic regression

$t_i \sim \text{Bernoulli}(y_i)$

$\sim \text{Ber.}^{\text{natural}}(t_i \mid a_i)$

for poisson

$t_i \sim \text{Poisson}(y_i)$

$\sim {}^{\text{natural}}\ \text{poisson}(a_i)$

GLM

$$L = \prod_i \frac{1}{s} h\left(\frac{t_i}{s}\right) g(\eta_i) e^{\frac{1}{s}\eta_i t_i}$$

$u(x) = x$

$t_i \sim$ 1D exp. family dist.

s scale

$$\log L = \sum_i \log s + \log h\left(\frac{t_i}{s}\right) + \log g(\eta_i) + \frac{1}{s}\eta_i t_i$$

$$\frac{\partial \log L}{\partial w} = \sum_i \frac{1}{g(\eta_i)} \frac{\partial g(\eta_i)}{\partial w} + \frac{1}{s}\frac{\partial \eta_i}{\partial w} t_i$$

$$= \sum_i \frac{1}{g(\eta_i)} \frac{\partial g(\eta_i)}{\partial \eta} \cdot \frac{\partial (\eta_i)}{\partial y} \frac{\partial y_i}{\partial a_i} \frac{\partial a_i}{\partial w}$$

$$+ \frac{1}{s} \frac{\partial \eta_i}{\partial y} \frac{\partial y_i}{\partial a_i} \frac{\partial a_i}{\partial w}$$

$$= \sum \frac{\partial}{\partial \eta_i} \log g(\eta_i)\, \psi'(y_i) f'(a_i) \phi(x_i) + \frac{1}{s} t_i \psi(y_i) f'(a_i) \phi(x_i)$$

Canonical link

$\eta_i = a_i$

$\frac{\partial \eta_i}{\partial w} = \phi(x_i)$

$\Rightarrow \frac{\partial}{\partial w} \log L = \underline{\frac{\partial}{\partial \eta_i} \log g(\eta_i)} \, \phi(x_i) + \frac{1}{s} t_i \phi(x_i)$

$E[u(x)] = -\frac{\partial}{\partial \eta} \log g(\eta)$

$E[x] = -s \frac{\partial}{\partial \eta} \log g(\eta)$

$\frac{\partial}{\partial \eta} \log g(\eta) = -\frac{1}{s} E[x]$

$\Rightarrow \frac{\partial}{\partial w} \log L = -\frac{1}{s} E[t] \, \phi(x_i) + \frac{1}{s} t_i \phi(x_i)$

$= -\frac{1}{s} y_i \phi(x_i) + \frac{1}{s} t_i \phi(x_i)$

$= \frac{1}{s} \sum_i (t_i - y_i) \phi(x_i)$

choose f

ψ is determined by exp. family dist. type

t_i describes data generation

$\frac{\partial^2 \log L}{\partial w \partial w^T} = -\frac{1}{s} \sum f'(a_i) \phi(x_i) \phi(x_i)^T$

82 $= -\frac{1}{s} \phi^T R \phi \qquad R = \mathrm{diag}(r_i) \quad r_i = f'(a_i)$

$$\frac{\partial}{\partial w}\log L = \sum_i \frac{\partial}{\partial \eta_i}\log g(\eta_i)\, \psi'(y_i)\, f'(a_i)\, \phi(x_i)$$

$$+ \frac{1}{s} t_i\, \psi'(y_i)\, f'(a_i)\, \phi(x_i)$$

$\frac{1}{s} y_i$

$$= \sum_i \frac{1}{s}(t_i - y_i)\, \psi'(y_i)\, f'(a_i)\, \phi(x_i)$$

$\frac{1}{s}\Phi^T$

$$\frac{\partial}{\partial w \partial w^T}\log L = \sum_i \frac{1}{s}(t_i - y_i)\Big[\psi'(y_i) f''(a_i)\, \phi(x_i)\phi(x_i)^T + f'(a_i)^2 \psi''(y_i)\, \phi(x_i)\phi(x_i)^T\Big]$$

$$-\frac{1}{s}\, \psi'(y_i)\, f'(a)^2 \quad \phi(x_i)\, \phi(x_i)^T$$

let $r_i = (t_i - y_i)\left[\psi'(y_i) f''(a_i) + f'(a_i)^2 \psi''(y_i) - \psi'(y_i) f'(a_i)^2\right]$

$$\Rightarrow \sum_i \frac{1}{s} r_i\, \phi(x_i)\, \phi(x_i)^T = -\frac{1}{s}\Phi^T R\, \Phi$$

$$w \leftarrow w + s(\Phi^T R \Phi)^{-1} \frac{1}{s} \Phi^T (t_i - y_i)$$

$$R = \mathrm{diag}(f'(a_i))$$

Canonical link

Normal $f(a) = a$ $\quad f'(a) = 1$ $\quad R = I$

Bernoulli $f(a) = \sigma(a)$ $\quad f'(a) = y_i(1-y_i)$

Poisson $f(a) = e^a$ $\quad f'(a) = e^a$

Machine Learning.

Logistic Regression

$a_i = w^T\phi(x_i)$

$y_i = \sigma(a_i)$

$$L = \prod_i \sigma(w^T\Phi(x_i))^{t_i} (1-\sigma(w^T\Phi(x_i)))^{1-t_i}$$

$$\frac{\partial \log L}{\partial w} = \Phi^T(t-y)$$

$$\frac{\partial^2 \log L}{\partial w \partial w^T} = -\Phi^T R \Phi \qquad R = \text{diag}(y_i(1-y_i))$$

$$w \leftarrow w - (\Phi^T R \Phi)^{-1} \Phi^T(t-y)$$

iterative method

what would be a conjugate prior for w?

prior $\times$ likelihood $\propto$ Posterior

same form

No known PDF satisfies conjugate requirement

Let's use a normal prior $w \sim N(m_0, S_0)$
(Since it is most convenient)

This model is not conjugate

$t_i | w \sim \text{Bernoulli}(w^T \phi(x_i), \frac{1}{\beta})$

Posterior $\propto$ prior $\times$ likelihood

also, posterior $\propto \prod y_i^{t_i} (1-y_i)^{1-t_i} \left(\frac{1}{2\pi}\right)^{d/2} \frac{1}{|S_0|^{1/2}} e^{-\frac{1}{2}(w-m_0)^T S_0^{-1} (w-m_0)}$

No known form for posterior

Sol 1: Maybe we can represent it "indirectly"
(typically via sampling)

Sol 2: Approximate posterior "as well as we can"

Before Posterior, let's compute MAP.

$$\log \text{Posterior} = \text{const} + \sum t_i \ln y_i + \sum (1-t_i) \ln(1-y_i) - \frac{1}{2}(w-m_0)^T S_0^{-1} (w-m_0)$$

diff wrt w.

$$\underbrace{\frac{\partial}{\partial w} \log \text{Posterior} = 0 + \phi^T (t-y) - S_0^{-1}(w-m_0)}_{G}$$

(no closed form solution form w)

$$\underbrace{\frac{\partial^2}{\partial w \partial w^T} \log \text{Posterior} = -\phi^T R \phi - S_0^{-1}}_{H}$$

$w \leftarrow w - H^{-1} G$

$= w - (-H^{-1})(-G)$ — finds w_{MAP}

$= w - (\phi^T R \phi + S_0^{-1})^{-1} (\phi^T (y-t) + S_0^{-1}(w-m_0))$

Laplace Approximation

1D Case :

want to approximate $f(x)$ as $\mathcal{N}$

$f(x) \propto e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$

So, approx. $g(x) = \log f(x)$ as a quadratic fn.

$x = x_0 + h$

taylor expansion $g(x) \cong g(x_0) + \boxed{g'(x_0) \cdot h} + \frac{1}{2} g''(x_0) h^2 + \dots$

$h = x - x_0$ $\qquad$ $\boxed{g'(x_0) \cdot h} \to 0$

$= g(x_0) + \frac{1}{2} g''(x_0)(x-x_0)^2$ $\qquad$ ignore higher order terms

$g'(x_0) = 0 \Rightarrow x_0 = \mu$

$A = \frac{1}{\sigma^2} = -g''(x_0)$

i.e. mean of normal = point of maxima of g

curvature at mean / maxima is same.

$f(x) \propto e^{\frac{A}{2}(x-x_0)^2} = e^{\frac{1}{2\sigma^2}(x-\mu)^2}$

Multivariate Case

$A = -H^{-1}$

$f(x) \propto e^{\frac{1}{2}(x-x_0)^T H (x-x_0)}$

$= e^{-\frac{1}{2}(x-x_0)^T A^{-1}(x-x_0)}$

A is inverse covariance matrix

Coming back to Logistic Regression

$x_0 = w_{MAP}$

$H = -[\Phi^T R \Phi + S_0^{-1}]$

$\Rightarrow m_N = w_{MAP}$

$S_N = [\Phi^T R \Phi + S_0^{-1}]^{-1}$

$w_{LAPLACE} \sim N(m_N, S_N)$

How to predict?

Using w_{MAP}: Given x_{N+1} predict $p(C=1)$.

$\sigma(w_{MAP}^T \Phi(x_{N+1}))$

For bayesian solution

$$p(C=1) = \int_w p(w \mid m_N, S_N) \, p(C=1 \mid w) \, dw$$

$$\int_w N(w \mid m_N, S_N) \, \sigma(w^T \Phi(x_{N+1})) \, dw$$

Approximate $\sigma(a)$ as $\Phi(\lambda a)$ Φ = CDF of normal dist (mean = 0, variance = 1)

$\lambda = \sqrt{\frac{\pi}{8}}$

$\sigma(a) \approx \Phi(\lambda a)$

$a = w^T \Phi(x_{n+1})$ $\quad a \sim N(\Phi(x_{n+1})^T m_N, \Phi(x_{n+1})^T S_N \Phi(x_{n+1}))$

if w is gaussian, a linear transform of w is also gaussian

$$p(c=1) = \int N(a \mid \mu_a, \sigma_a^2)\, \sigma(a)\, da$$

$$\int N(a \mid \mu_a, \sigma_a^2)\, \Phi(\lambda a)\, da$$

$\vdots$

Final result $\quad \sigma\left(\dfrac{\mu_a}{\sqrt{1 + \frac{\pi}{8}\sigma_a^2}}\right)$

σ: $x > 0 \Rightarrow 1$

$x < 0 \rightarrow 0$

if same cost of +ve & -ve labels

$p(c=1)$ depends only on sign of μ_a

Machine Learning

Applying ML Algorithms.

Nearest neighbours and decision trees.

knn algorithm for classification

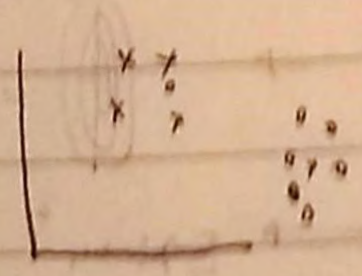
- store all examples
- find k nearest neighbours
- predict label based on voting for each neighbour

knn for regression

- take mean of k nearest neighbours.

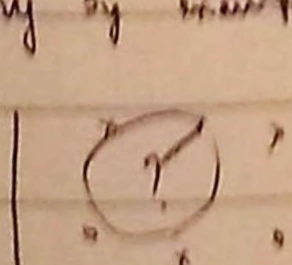
* NON PARAMETRIC METHOD. — no prior commitment to hyp.
  complexity increases w/ more data

!noise



increase value of k for random noise

!query by example — search is expensive



with Δ inequality
geometric D.S. (k-d tree)

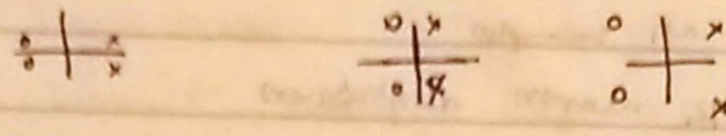
! k is free parameter

how to choose k?

model selection ← use validation data to evaluate how each k performs

! sensitivity to how data is presented (scale)

since it depends on distance → normalize
- linear
- z scaling



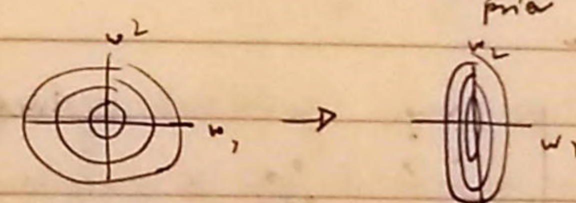
Sensitive to irrelevant features

irrelevant features dominate relevant features → apply dimensionality reduction

Bayes. LR

$$\sum (t_i - w^T \Phi(x_i))^2 + \underbrace{\lambda w^T w}_{\text{prior}}$$

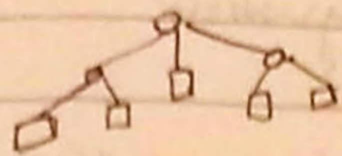
modify prior :

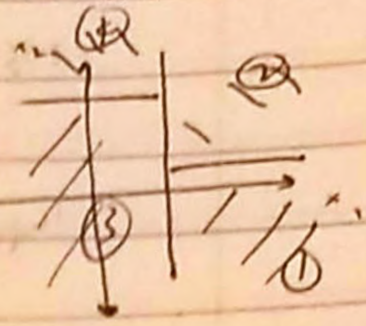


$$\lambda w^T w \longrightarrow \sum \lambda_i w_i^2$$

Decision Trees.

non-parametric and non-probabilistic classifier.





recursively split feature space.

! tree can be very large (exponential)

! it may overfit

how to build a good tree (best tree = np hard)

Splitting criteria.

algorithm:

if data has pure class

make leaf

else:

pick feature to split on

split data into subsets

apply algorithm recursively for each subtree

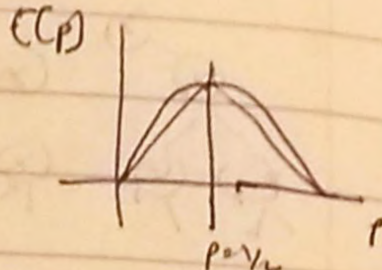
Measure uncertainity

entropy - classification (accuracy does not work as well)

MSE - regression

Information gain = reduction in uncertainity due to split.

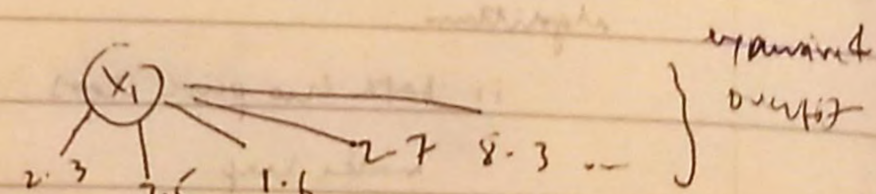
$$\text{Entropy}(p_1 \dots p_n) = -\sum p_i \log p_i$$



$$\text{Gain}(\text{split}) = \text{Ent}(S) - \sum_i \frac{|S_i|}{|S|} \text{Ent}(S_i)$$

Real valued Attributes.

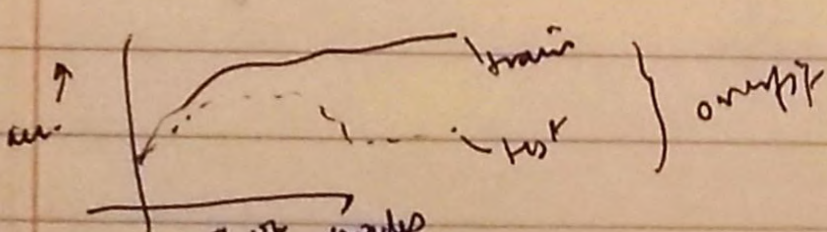
x_1	
2.3	+
3.5	−
1.6	−
2.7	+
8.3	−
6.5	+



−	+	+	−	+	−
1.6	2.3	2.7	3.5	6.5	8.3

① only finite # points make difference

② no need to test values in [2.3, 2.7]



solutions to overfitting

prevent -

min # points at any level

min. information gain

grow tree to full size. } best.

Then prune [use valid. set]

Ex. Reducing error Pruning

App Machine Learning

Applying ML.

individual feature preprocessing

linear scaling to [0,1]

$$x \leftarrow \frac{x - x_{min}}{x_{max} - x_{m}}$$

normal scaling

$$x \leftarrow \frac{x - \mu}{\sigma}$$

Discretizing features

unsupervised

1. equal bin size predetermined by range
2. equal frequency adapts to data
3. cluster

Supervised.

use labels. Run decision tree on single feature

most helpful after pruning

discrete to numerical

unit vector 1000, 0100, 0010, 0001

increasing weight vectors

1000, 1100, 1110, 1111

Manifold methods

data resides on a "manifold"

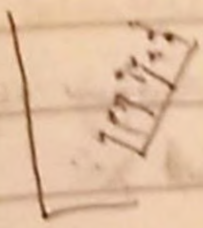
embed data in low dim space, preserving local distances.

produces "embedding"

PCA principal Component Analysis

linear dim. reduction

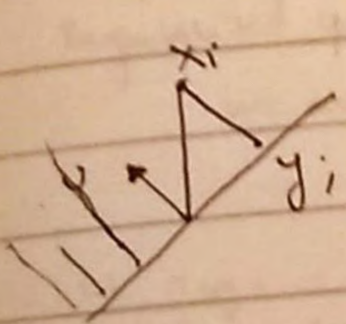
project data onto k dim with max variation



center data matrix

$\Phi^T \Phi = V \Lambda V^T$ eigendecomposition

take top k eigenvectors.



Data $= x_1 \dots x_N$

$y = u^T x_i$

$\bar{x} = \frac{1}{N} \Sigma x_i$ vector

$\bar{y} = \frac{1}{N} \Sigma y_i$ scalar

$\bar{y} = \frac{1}{N} \sum u^T x_i = \frac{1}{N} u^T \bar{x}$

Projected variance

$J = \frac{1}{N} \Sigma (y_i - \bar{y})^2$

want to max

also, $\bar{J} = \frac{1}{N}\sum_i [U^T(x_i - \bar{x})]^2$

$= \frac{1}{N}\sum_i U^T(x_i-\bar{x})(x_i-\bar{x})^T U$

$= U^T\left[\frac{1}{N}\sum(x_i-\bar{x})(x_i-\bar{x})^T\right]U$

$= U^T \Sigma_x U$

constraint norm $U = 1$

Use Lagrange Multipliers

objective + λ(constraint) = new objective

solve for original vars. and λ

max $U^T S U$ s.t. $U^T U = 1$

$\equiv U^T S U \; \lambda(U^T U - 1)$

$\frac{\partial \bar{J}(U,\lambda)}{\partial \lambda} = 1 - U^T U = 0 \Rightarrow \|U\| = 1$

$\frac{\partial \bar{J}(U,\lambda)}{\partial U} = 2SU - 2\lambda U = 0$

$SU = \lambda U$ — eigenvector $\|U\| = 1$

eigen value

which eigenvector?

$\int_u u^T S u = u^T \lambda u = \lambda \|u\|^2$

$= \lambda$

pick max eigen value

Projection $= \phi U$

Feature Selection.

Filter method

calculate score (ex. info gain, correlation w/ label)

for each feature

the pick top k

! issue - may end up choosing same/similar duplicate

features

L1 Regularization

Regularized linear regression

$w = \text{argmin} \frac{1}{2} \underbrace{\sum (w^T \phi(x_i) - t_i)^2}_{\text{loss}} + \underbrace{\lambda w^T w}_{\text{regularization}} \rightarrow \lambda \sum_k w^2$

L1 reg.

$w = \text{argmin} \frac{1}{2} \text{LOSS} + \lambda \sum_k |w_k|$

use Laplace distribution

instead of gaussian for prior

Evaluating ML outcomes.

what to measure? Regression, classification

MSE acc.

confusion matrix

	+	-	← classified as
+	TP	FN	
-	FP	TN	

$$acc = \frac{TP + TN}{TP + TN + FN + FP}$$

IR terminology

ignore true negative

Precision $= \frac{TP}{TP+FP}$ of those that I predicted, how many are POSITIVE

Recall $= \frac{TP}{TP+FN}$ of the ones that I should've found how many did I find

$$F = \frac{2RP}{R+P}$$

medical ~~ter~~ terminology

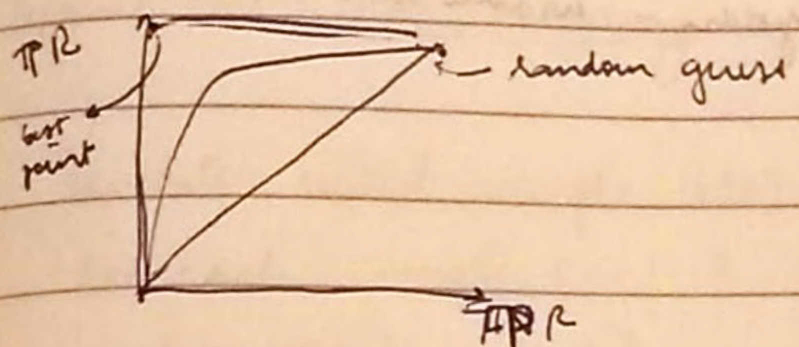
sensitivity = recall accuracy in + class

Specificity $= \frac{TN}{TN+FP}$ accuracy in − class signal detection

TPrate = Recall

FPrate = 1- Specificity

ROC (receiver operator characteristic) curve



get points by changing threshold

area under ROC curve (< 1 ~ probability that random example from test set is classified correctly)

How to measure?

validation set method

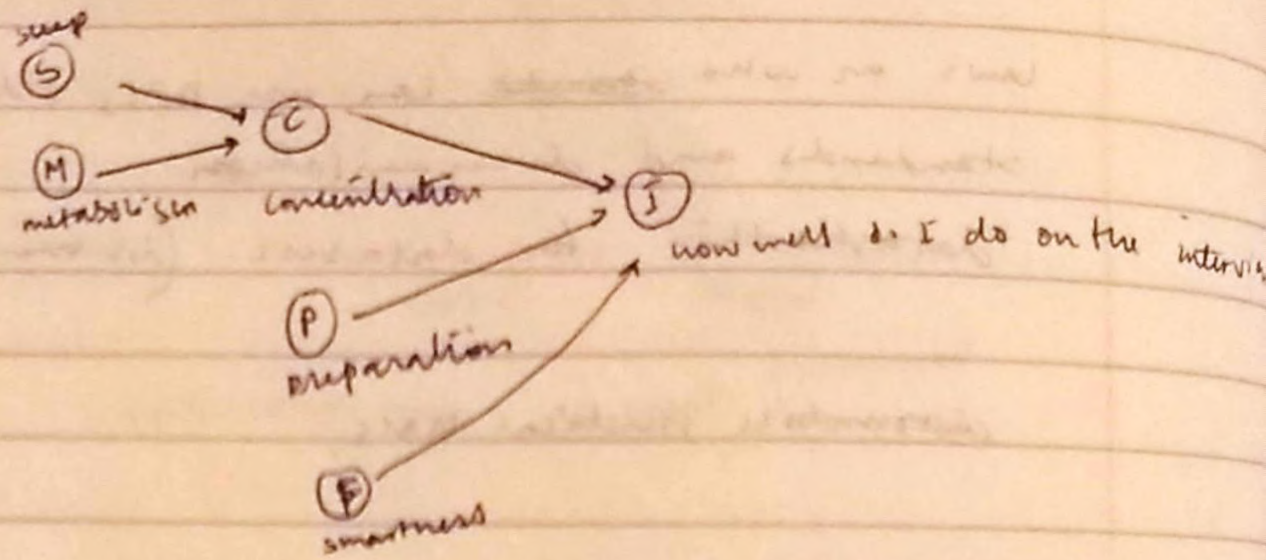
+ unbiased estimate of data
- variance due to choice of valid set
- wastes data

do many times and take average

(-introduces bias) → k fold (disjoint) cross validation (- class label dist bias) → stratified CV

Machine Learning

Graphical Models.



Bayesian network

a directed acyclic graph

probablistic relationship b/w variables

for every node v: $P(V \mid \text{parents}(V))$

General form

Nodes are $X_1 \ldots X_N$

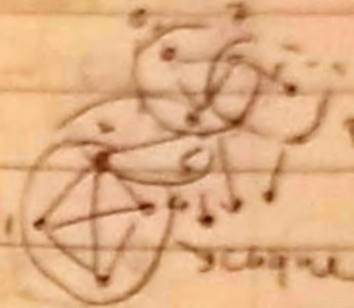
$P(X_i \mid Pa(X_i))$

Joint Dist

$$P(X_1 \ldots X_N) = \prod_i P(X_i \mid Pa(X_i))$$

12/6

Undirected Graphical models
(Markov Random Field)
Graph with no edges
For every clique in graph
potential function $\psi(x_c)$



$$P(x_1 \ldots x_n) \propto \prod_c \psi_c(x_c)$$

$C_1 = 1,2,3,4$

$C_2 = 5$

$C_3 = 5,6,7$

$C_4 = 7,8$

ex.

$C_1 = x_1 x_2$

$C_2 = x_2 x_3$

ψ_{C_1},

x_1	x_2	ψ_c
0	0	5
0	1	1
1	0	3
1	1	100

ψ_{C_2}

x_2	x_3	
0	0	1
0	1	1
1	0	2
1	1	2

$$P(x_1 \ldots x_n) = \frac{1}{Z} \prod_c \psi_c(x_c)$$

$$Z = \sum_{x_1 \ldots x_n} \prod_c \psi_c(x_c)$$

• generalization of Bayes net
c_i = i and parents

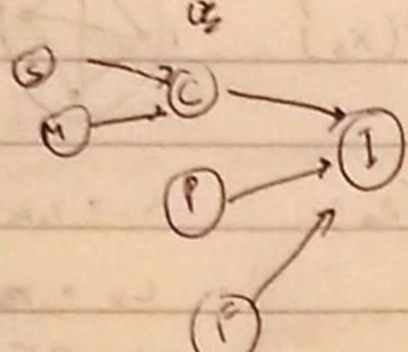
$\psi(x_i, P_a(x_i))$

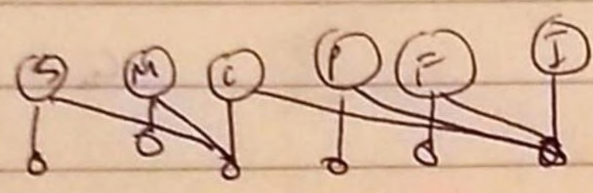
129

More general → factor graph.

Random variables $x_1, \dots x_N$

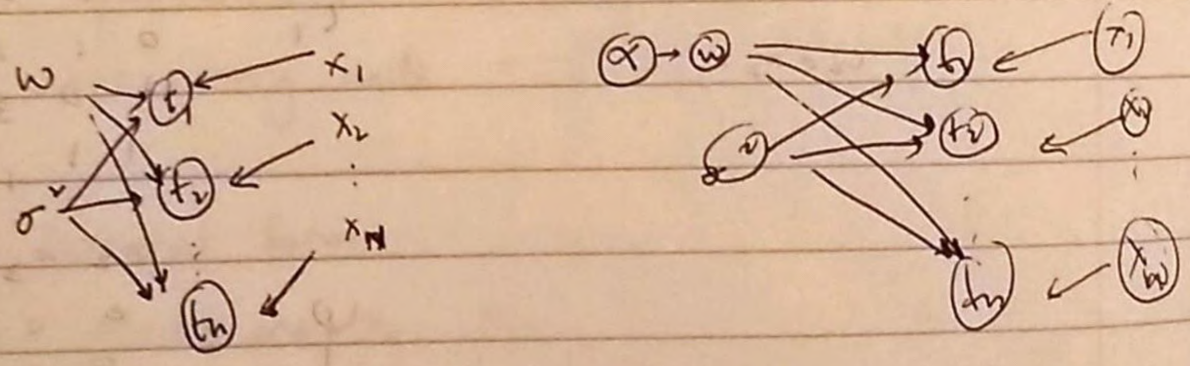
and functions of RVs.





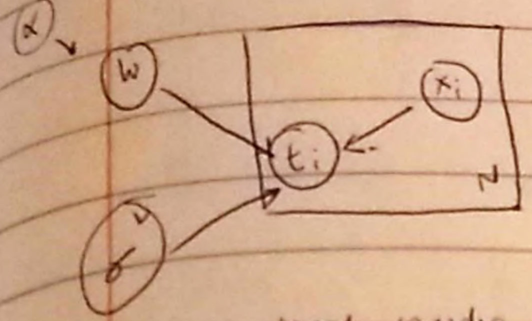
In linear regression

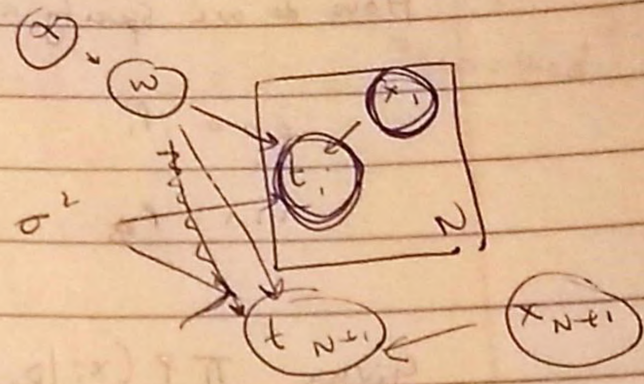
in Bayesian linear regression



140

Plate notation.





Observe some nodes,

ML, MAP :

find assignment to some variable(W)

s.t P(Evidence| W) is Max.

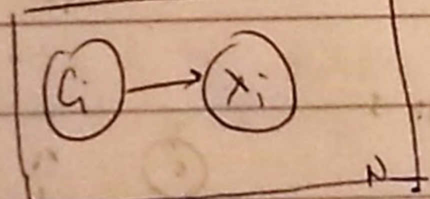
PA

Predictive distribution

find P(t_{N+1} | Evidence)

Generative model.

$P(C_i)$ $P(X_i | C_i)$



How do we specify a dist. over 3 binary r.v. ?

000 p_1

$\vdots$ $\qquad \sum p_i = 1$

111 p_8

Given $\prod p(x_i \mid p_a(x_i))$

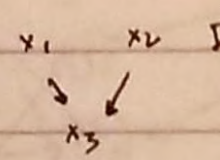
& obs ~~$x_3=1$~~ $x_3 = 1$

$p(x_2 = 1) = ?$

$$p(x_2 = 1) = \sum_{x_1, x_3} p(x_1, x_2 = 1, x_3)..$$

Can ~~we~~ first write a full table of joint dist and marginalize variables not of interest (x_1, x_3)

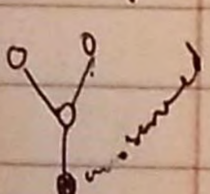
I ? $x_1 \perp x_2$

(A)  I

$$p(x_1 x_2 x_3) = p(x_1)\, p(x_2)\, p(x_3 \mid x_1 x_2)$$

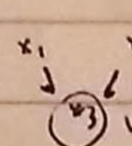
$$p(x_1 x_2) = \sum p(x_1)\, p(x_2)\, p(x_3 \mid x_1 x_2)$$

$$= p(x_1)\, p(x_2) \underbrace{\sum p(x_3 \mid x_1 x_2)}_{=1}$$

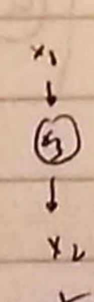
(B') 

not I

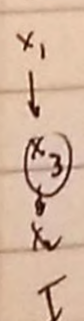
unobserved

(B)  unobserved

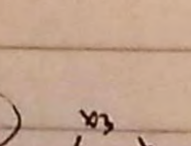
not I

(C) 

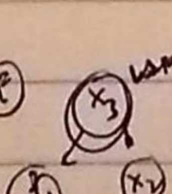
not I

(D)  unobserved

I

(E) 

not I

(F)  unobserved

I

142

head to head observed
incoming outgoing unobserved
tail to tail unobserved

Theorem: V is independent of U

$\Leftrightarrow$ every path from $v \in V$ to $u \in U$ is "blocked"

$$x_1 \to x_2 \to x_3 \to x_4 \to x_5$$

$P(x_1 = 1) = 0.7$

$P(x_i = 1 \mid x_{i-1} = 1) = 0.9$

$P(x_i = 1 \mid x_{i-1} = 0) = 0.3$

① $P(x_3 = 1)$ — no evidence.

② $P(x_3 = 1 \mid x_1 = 1)$ — parent to children

③ $P(x_3 = 1 \mid x_5 = 1)$ — children to parent

① $P(x_3 = 1) = \sum_{x_1 x_2 x_4 x_5} P(x_1) P(x_2 \mid x_1) P(x_3 = 1 \mid x_2) P(x_4 \mid x_3 = 1) P(x_5 \mid x_4)$

$P(x_1 = 1) = .7$

$P(x_2 = 1) = \sum_{v = \{0,1\}} P(x_1 = v) \, P(x_2 = 1 \mid x_1 = v)$

$= .72$

$P(x_3 = 1) = \sum_{x_2 = v} P(x_2 = v) \, P(x_3 = 1 \mid x_2 = v)$

$$p(x_3=1) = \sum_{x_1 x_2 x_4 x_5} p(x_1)\, p(x_2|x_1)\, p(x_3=1|x_2)\, p(x_4|x_3=1)\, p(x_5|x_4)$$

$$= \sum_{x_1 x_2 x_4} p(x_1)\, p(x_2|x_1)\, p(x_3=1|x_2)\, p(x_4|x_3=1) \underbrace{\sum p(x_5|x_4)}_{=1}$$

$$= \sum_{x_1 x_2} p(x_1)\, p(x_2|x_1)\, p(x_3=1|x_2) \underbrace{\sum p(x_4|x_3=1)}_{=1}$$

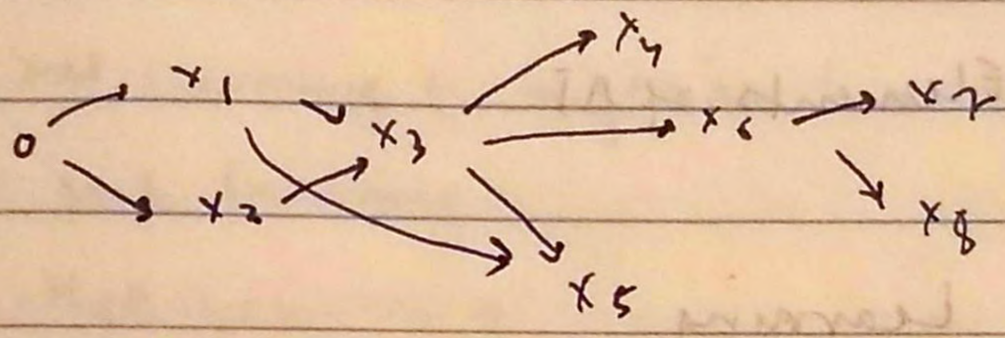
$$= \sum_{x_2} p(x_3=1|x_2) \sum_{x_1} \underbrace{p(x_1)\, p(x_2|x_1)}_{=p(x_2)}$$

(2) $p(x_1=1 | x_3=1) = \dfrac{p(x_1=1, x_3=1)}{p(x_1=1)}$

(3) $p(x_3=1 | x_5=1) = \dfrac{p(x_3=1, x_5=1)}{p(x_3=1)}$

$$p(x_3=1, x_5=1) = \sum_{x_1 x_2 x_4} p(x_1)\, p(x_2|x_1)\, p(x_3=1|x_2)\, p(x_4|x_3=1)\, p(x_5=1|x_4)$$

$$= \sum_{x_1}\sum_{x_2} p(x_1)\, p(x_2|x_1)\, p(x_3=1|x_2) \sum p(x_4|x_3=1)\, p(x_5=1|x_4)$$

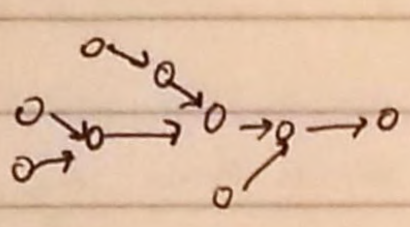


sum x_3 => result depends on

x_1 x_2 x_4 x_5 x_6

Machine learning

If bayesian net is a polytree
then, var. elimination produces tables
which are $\leq$ tables in original B.N.



if bayes net is not polytree,
tables could be $>$ than size of tables in BN.

$$\sum_{x_j} f(x_1, x_2, x_3)\, g(x_1, x_5, x_6)$$

→ function of $(x_2\ x_3\ x_5\ x_6)$

for continuous R.V.,

$$P(x_1) = \int_{x_2} p(x_1 \mid x_2)\, p(x_2)\, dx$$

integrations should be simple

var. elim. is NP hard. for non polytrees

Belief Propagation (loopy B.P.)
assume graph is polytree
approximate inference.

Alternate idea:
instead of computing marginals exactly, try to sample from the marginal distribution.

ex:
logistic regression
$p(w \mid \text{data})$
Previously, we computed a wrong posterior $\hat{p}(w \mid \text{data})$
instead, try to sample $w_1, w_2 \dots w_k$
$w_i \sim p(w \mid \text{data})$

- To predict instead of
$\int p(w \mid \text{data})\, p(y_{next} \mid w)\, dw$ pred dist.
compute $\frac{1}{k} \sum_i p(y_{next} \mid w_i)$

$$x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow x_4 \rightarrow x_5$$

$p(x_3)$ sample x_1 then $x_2 \mid x_1$ then $x_3 \mid x_2$ ..
$p(x_3 \mid x_1 = 1)$ fix $x_1 = 1$ ..
$p(x_3 \mid x_5 = 1)$ @ rejection sampling

sampling ≡ monte-carlo.

(a) Rejection sampling

take sample, discard if $x_5 \neq 1$

correct but slow (wastes sampling time)

(b) Likelihood weighting

instead of sampling and rejecting 11000

stop after 1100 and force $x_5 = 1$

and use with weight of $0.3. = \sum P(x_5=1 | x_4 = \ldots$

11001	.3
10101	.3
00101	.3
	0.9 total sample

does not waste samples. Still slow.

For these algorithms, we must be able to sample from $P(x_i \mid Pa(x_i))$

Easy for discrete. Various methods & tricks for continuous variables.

Monte Carlo Markov Chain

Sample entire string at once

1101 → 1011 → 00101 → ...

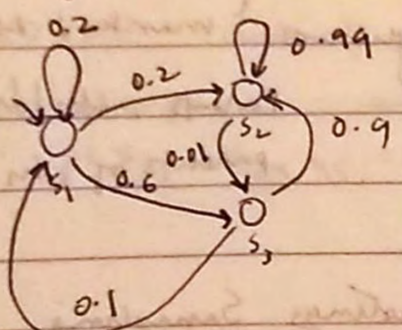
define a random process

in limit, the distribution is same as exactly what we want.

Markov Chain

Nodes are states.

Edges are probabilistic transitions.

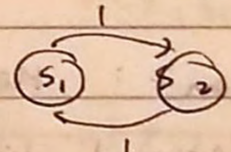


Stationary distribution over states

$\forall s \quad P(s) = P(\text{arrive in } s \text{ in next step})$

$\forall i \quad P(s_i) = \sum_j P(s_j) P(s_i \mid s_j)$

ex.



does not have s.p.

We will build a markov chain s.t. states are value configurations of bayes net and its stationary distribution is P(unobserved vars. | observed vars.)

Transitions in the MC go from $x_1 \ldots x_n$ to
(11001)

another valuation (01001)

Option 1 : Gibbs Sampling

Pick $i \in \{1 \ldots N\}$ at random (uniformly)

Pick x_i from dist $P(x_i \mid x_1 \ldots x_{i-1}\, x_{i+1} \ldots x_n)$

$$0 \to \boxed{0} \to 0 \to 0 \to 0$$

need only neighbours "markov blanket"
parents, children,
parents of children

Option 2 : Metropolis Hastings Sampling

* Proposal distribution $q(y \mid x)$

repeat

① Pick y from $q(y \mid x)$

② compute accept. probability $A = \min\left(1, \dfrac{p(y)\, q(x \mid y)}{p(x)\, q(y \mid x)}\right)$

③ Accept y i.e. $x \leftarrow y$ with prob A

stationary dist

Detailed Balance

Markov Chain with Transition prob T has Detailed Balance relative to P_T

$P(a)\ T(b|a) == P(b)\ T(a|b)$

Fact 1 DB -> P is stationary for MC

Fact 2 MH has DB for P.

Fact 3 Gibbs is special case of MH where A=1.

00100 current state

resample x_2 using Gibbs.

$$P(x_2 \mid x_{1345} = 0100) = \frac{P(x_2 = v \text{ and } x_{1345} = 0100)}{P(x_{1345} = 0100)}$$

~~$P(x_1=0)\ P(x_2=v|x_1=0)\ P(x_2=1)\ P(x_2=v$~~

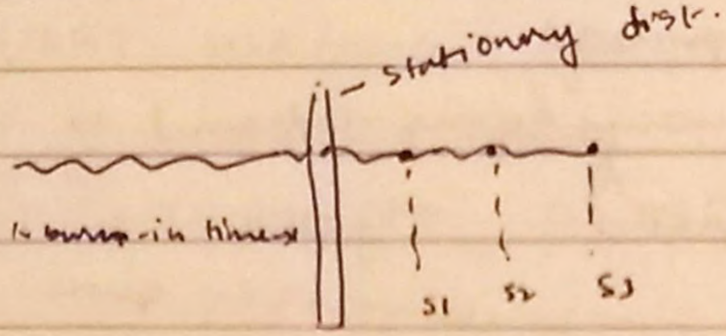
numerator: $P(x_1=0)\ P(x_2=v|x_1=0)\ P(x_3=1|x_2=v)$
$P(x_4=0|x_3=1)\ P(x_5=0|x_4=0)$

$v = \{0, 1\}$

$$\frac{P(x_2=0 \text{ and } x_{1345} = 0100)}{P(x_2=1 \text{ and } x_{1345} = 0100)} = \frac{P(x_2=0|x_1=0)\ P(x_3=1|x_2=0)}{P(x_2=1|x_1=0)\ P(x_3=1|x_2=1)}$$

Machine Learning

Markov Chain Monte Carlo.



s1, s2, s3 are not independent

but we assume so.

Gibbs sampling $\quad p(v_i \mid v_1 \ldots v_{i-1}, v_{i+1} \ldots v_N)$ – can't compute for undirected models

MH sampling $\quad \min\left(1, \dfrac{p(v')\, q(v \mid v')}{p(v)\, q(v' \mid v)}\right)$

use MH

Fact1: DB implies P() is stationary.

Fact2: MH satisfies DB for P().

Fact3: Gibbs is MH & A=1.

proof in slides

Latent Dirichlet Allocation.

completely unsupervised text learning.

Dirichlet

constraint $\sum \mu_i = 1$

$\alpha = (\alpha_0 \ldots \alpha_k)^T \quad \alpha_0 = \sum \alpha_i$

$$p(\mu|\alpha) = Dir(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\prod \Gamma(\alpha_i)} \mu_i^{\alpha_i - 1}$$

Discrete dist

Posterior $= Dir(\mu | \alpha + m)$

"topic" : dist over words.

each doc is about multiple topics.

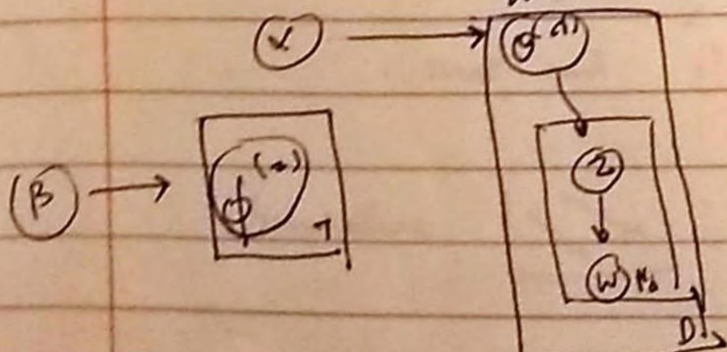
~~each~~ each doc draws topic from dir

each topic draws word from dir

For each doc, for each location

decide topic

draw word from topic



LSA - matrix of ~~real~~ nos.

Topic model - matrix of probs.

N : total # of words

N_d : # of words in doc. d.

N_k : # of ~~words~~ times a topic occured.

$N_{k|d}$: # of times topic k appeared in d.

$N_{i|k}$: # of times word i occured in topic k.

prior : $\prod_d Dir(\theta_d | \alpha)$ $\quad \prod_k Dir(\Phi_k | \beta)$

Posterior . . -

but we do not know topics (z's)

Gibbs sampling

remove word.

Learning α and β.

Evidence Maximization

direct → exponential

use sum of log of samples

Machine Learning

Y observed variables
Z hidden variables
want to max. $p(Y|\theta) = \int_z p(Y, Z|\theta)\, dz$

$Q(\theta^{new}, \theta^{old}) = E_{p(Z|Y,\theta^{old})}[\ln p(Y, Z|\theta^{new})]$

EM Algorithm
init θ^{old}
repeat
* E step: calculate $Q(\theta^{new}, \theta^{old})$
* M step: pick θ^{new} to max $Q(\theta^{new}, \theta^{old})$

what does this mean?
Mixture of coin models
for each i
pick $z_i \in \{1 \dots k\}$ from discrete$\{p_1 \dots p_k\}$
ex. $k = 3$ $z = \{1, 2, 3\}$
$p = \{0.7, 0.1, 0.2\}$
$z_i = 1 \Leftrightarrow 100$
$2 \Leftrightarrow 010$
$3 \Leftrightarrow 001$

flip coin z_i T times to get y_i

ex. $T=6$ $Y_i = HHTHTH$

Coin i has $p(H) = \mu_i$

likelihood $= \mu_i^{\#H} (1-\mu_i)^{\#T}$

Complete data likelihood

$$L = \prod_i \prod_j \left[p_j \prod_t \mu_j^{y_{it}} (1-\mu_j)^{y_{it}} \right]^{z_{ij}}$$

coins

labels

$\theta = (\{p_j\}\{\mu_j\})$

$\ln(P\, Y, Z \mid \theta^{new}) =$

$$\sum_i \sum_j z_{ij} \underbrace{\left[\ln p_j^{new} + \sum_t y_{it} \ln \mu_j^{new} + (1-y_{it}) \ln (1-\mu_j^{new}) \right]}$$

Next figure out Q.

$$Q(\theta^{new}, \theta^{old}) = \sum_i \sum_j E_{P(z \mid Y, \theta^{old})} [z_{ij}] (\ldots)$$

fn of new params

Let $\alpha_{ij} = p_j\, P(Y_i \mid \theta_j) = p_j \prod_t \mu_j^{y_{it}} (1-\mu_j)^{1-y_{it}}$

fn. of old params

Let $\gamma_{ij} = E_{P(z \mid Y, \theta^{old})} [z_{ij}] = \gamma_{ij}$

$$Q(\theta^{new}, \theta^{old}) = \sum_i \sum_j \gamma_{ij} \ln \alpha_{ij}$$

$r_{ij} = E_{P(Z|Y,\theta^{old})}[z_{ij}] = P(z_{ij}=1 \mid Y, \theta^{old})$

$$= \frac{P(z_{ij}=1, y_i \mid Y^{-i}, \theta^{old})}{P(y_i \mid Y^{-i}, \theta^{old})}$$

$$= \frac{P(z_{ij}=1, y_i \mid \theta^{old})}{P(y_i \mid \theta^{old})}$$ because $y_i \perp y^{-i} \mid \theta^{old}$.

Numerator $= P(z_i = j)\, P(y_i \mid z_i = j)$

$= P_j\, P(y_i \mid \theta_j)$

$= \alpha_{ij}$

binomial

assumed

$$\Rightarrow \boxed{r_{ij} = \frac{\alpha_{ij}}{\sum_l \alpha_{il}}} \quad \text{I}$$

$$Q = \sum_i \sum_j r_{ij} \left[\ln P_j + \sum_t y_{it} \ln \mu_j + (1 - y_{it}) \ln (1-\mu_j)\right]$$

P_j, μ_j are new.

Must satisfy $\sum_j P_j = 1$ || use Lagrange multipliers

$$Q = \text{Const}(\{P_j\}) + \sum_i \sum_j r_{ij} \ln P_j$$

$$L = \sum_i \sum_j r_{ij} \ln P_j + \lambda \left(\sum_j P_j - 1\right)$$

No of examples in class j

$$\frac{\partial}{\partial \lambda} = \sum p_j - 1 = 0 \; ; \quad \frac{\partial}{\partial p_j} = \sum_i \gamma_{ij} \cdot \frac{1}{p_j} + \lambda = 0$$

$$\Rightarrow p_j = -\frac{1}{\lambda} \sum_i r_{ij} \qquad \boxed{N_j = \sum_i r_{ij}}$$

$$= -\frac{1}{\lambda} N_j$$

$$\sum_j p_j = 1 \Rightarrow \sum_j -\frac{1}{\lambda} N_j = 1 \Rightarrow N = -\lambda$$

$$\boxed{\therefore p_j = \frac{N_j}{N}} \quad \text{II}$$

$$\frac{\partial}{\partial \mu_j} = \sum_i \sum_t \left[\frac{y_{it}}{\mu_j} - \frac{1-y_{it}}{1-\mu_j}\right] r_{ij} = 0$$

$$\sum_i \sum_t \left(y_{it}(1-\mu_j) - \mu_j (1-y_{it})\right) r_{ij} = 0$$

$$\sum_i \sum_t r_{ij} \left(y_{it} - \cancel{y_{it}\mu_j} - \mu_j + \cancel{\mu_j y_{it}}\right) = 0$$

$$\sum_i \sum_t r_{ij} \mu_j = \sum_i \sum_t r_{ij} y_{it}$$

$$\Rightarrow \mu_j \sum_t \sum_i r_{ij} = \sum_i r_{ij} \left(\sum_t y_{it}\right)$$

\# heads

$$\Rightarrow \mu_j T N_j = \sum_i r_{ij} \left(\sum_t y_{it}\right)$$

$$\Rightarrow \boxed{\mu_j = \frac{\sum_i r_{ij} \left(\sum_t y_{it}\right)}{T N_j}} \quad \text{III}$$

182

EM for Mixture of Coins

init $\{P_j, \mu_j\}$

Repeat

Calculate γ_{ij} using I

Calculate $\{P_j, \mu_j\}$ using II, III

EM does marginal likelihood estimation when one or more parameters are not observed.

Why does it work?

Fact if $Q(\theta^{new}, \theta^{old}) > Q(\theta^{old}, \theta^{old})$

then $P(Y|\theta^{new}) > P(Y|\theta^{old})$

Proof $0 < Q(\theta^{new}, \theta^{old}) - Q(\theta^{old}, \theta^{old})$

$= E_{P(Z|Y,\theta^{old})}[\ln P(Y, Z|\theta^{new}) - \ln P(Y, Z|\theta^{old})]$

$= E_{P(Z|Y,\theta^{old})}\left[\ln \frac{P(Y|\theta^{new})\, P(Z|Y, \theta^{new})}{P(Y|\theta^{old})\, P(Z|Y, \theta^{old})}\right]$

$= E_{P(\ldots)} \ln \frac{P(Y|\theta^{new})}{P(Y|\theta^{old})} + E_{P(\ldots)} \ln \frac{P(Z|Y, \theta^{new})}{P(Z|Y, \theta^{old})}$

no z term

less than 0 claim

$0 \le \ln \frac{P(Y|\theta^{new})}{P(Y|\theta^{old})} \Rightarrow \frac{P(Y|\theta^{new})}{P(Y|\theta^{old})} \ge 1$

Jensen's inequality : concave f

$f[E[x]] \geq E[f($

ex. variance $E[x]^2 - E[x^2] \geq 0$

$P_1(v) \quad P_2(v)$

KL divergence
not symmetric

$$d_{KL}(P_1 \| P_2) = \int P_1(v) \ln \frac{P_1(v)}{P_2(v)} \, dv$$

$$= E_{P_1(v)} \left[\ln \frac{P_1(v)}{P_2(v)} \right] \geq 0$$

$$-d_{KL}(P_1 \| P_2) = E_{P_1(v)} \ln \frac{P_2(v)}{P_1(v)}$$

$$\leq \ln E_{P_1(v)} \frac{P_2(v)}{P_1(v)}$$

$$= \ln \int P_1(v) \frac{P_2(v)}{P_1(v)} \, dV = \ln 1 = 0$$

Machine Learning

Kernel Function : Fast way to compute inner product in some feature space

$$k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

For ex. euclidean $(x_i^T x_j)$, quadratic $(x_i^T x_j + 1)^2$, polynomial $(x_i^T x_j + 1)^d$

inf. polynomial → RBF $\left(e^{-\frac{1}{2}\|x_i - x_j\|^2 / s^2} \right)$ $\boxed{Alg} \leftarrow \boxed{kernel} \leftarrow \boxed{Data}$

Kernel Method : Learning algorithm that works with kernels

For ex. Perceptron, kNN, Regularized Linear Regression.

Mercer's Theorem: $K(\cdot, \cdot)$ is a kernel $\Leftrightarrow$

① K is symmetric ② Kernel matrix is Positive Semi Definite $\forall$ finite set of points

All eigenvalues ≥ 0 $\Leftrightarrow$ $\forall c \; c^T K c \geq 0$

Proof one dim. if $K(a,b) = \phi(a)^T \phi(b)$ then $c^T K c \geq 0$

$$\begin{aligned} c^T K c &= \sum_i \sum_j c_i c_j K(x_i, x_j) \\ &= \sum_i \sum_j c_i c_j \phi(x_i)^T \phi(x_j) \\ &= \left[\sum_i c_i \phi(x_i)\right]^T \left[\sum_j c_j \phi(x_j)\right] \\ &= \|\sum c_i \phi(x_i)\|^2 \geq 0 \end{aligned}$$

Fact if K_1 is a kernel, K_2 is a kernel

then ⓐ $K_3 = K_1 + K_2$ is a kernel

ⓑ $K_4 = K_1 \times K_2$ is a kernel

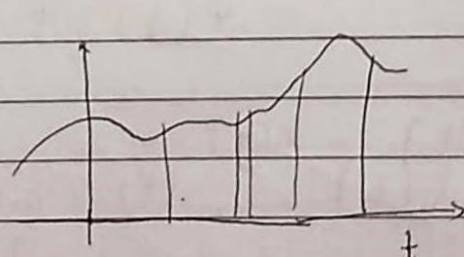
Proof of ⓐ $c^T K_3 c = c^T (K_1 + K_2) c = c^T K_1 c + c^T K_2 c \geq 0$

Aliter $K_3(x_i, x_j) = K_1(x_i, x_j) + K_2(x_i, x_j)$

$$= \phi_1^T(x_i)\phi_1(x_j) + \phi_2^T(x_i)\phi_2(x_j)$$

$$= \hat{\phi}(a)^T \hat{\phi}(b) \quad \Big| \quad \hat{\phi}(a) = \text{concat.} \begin{matrix} \phi_1(a) \\ \phi_2(a) \end{matrix}$$

Gaussian Process



A distribution over functions such that for any finite set of inputs $x_1 \dots x_N$ the vector of function values

mean of $f(x_i)$

$$f = (f(x_1) \dots f(x_N))^T \sim N((m(x_1) \dots m(x_N))^T, C)$$

$$C = \begin{array}{c} \quad j \\ i \begin{pmatrix} & | & \\ - & + & - \\ & | & \end{pmatrix} \end{array} C(x_i, x_j)$$

is distributed normally

It is specified by mean fn. and covariance new point is first in covariance matrix

Given $x_1 \dots x_N$

$f_1 \dots f_N$

x_{N+1}

Predict f_{N+1}

$\bar{f}_{N+1} = (f_1 \dots f_N)$

$$C_{N+1} = \left(\begin{array}{c|c} c & v^T \\ \hline v & C_N \end{array}\right) \quad c = C(x_{N+1}, x_{N+1}) \quad v = \begin{pmatrix} c(x_1, x_{N+1}) \\ \vdots \\ c(x_N, x_{N+1}) \end{pmatrix}$$

C_N = c applied to $x_1 \dots x_N$

Assume $m(x) = 0 \ \forall x$

$\bar{f}_{N+1} = (f_{N+1}, (\bar{f}_N))^T \sim N(0, C_{N+1})$

observe $\bar{f}_N$

$p(f_{N+1})$

Application of MVN conditional template

$$\begin{pmatrix} x_a \\ x_b \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}\right)$$

$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab} \Sigma_{bb}^{-1} \Sigma_{ba}$$

$$\mu_{a|b} = \mu_a + \Sigma_{ab} \Sigma_{bb}^{-1} (x_b - \mu_b)$$

$$\Sigma = c - v^T C_N^{-1} v$$

$$\mu = 0 + v^T C_N^{-1} (\bar{f}_N)$$

Bayes LR.

$p(w) = N(0, \frac{1}{\alpha} I)$, $y = \Phi w$; $t \sim N(y, \frac{1}{\beta} I)$

$t_i \sim N(w^T \phi(x_i), \frac{1}{\beta})$

Claim: t is sampled from gaussian process

$E[y] = \Phi E[w] = 0$

$$E[yy^T] = E[\Phi w w^T \Phi^T]$$

$$= \Phi E[ww^T] \Phi^T$$

$$= \Phi (\tfrac{1}{\alpha} I) \Phi^T = \tfrac{1}{\alpha} \Phi \Phi^T$$

$\Phi = \begin{pmatrix} \vdots \end{pmatrix}_{N \times D}$

$\Phi^T \Phi = (\cdots)\begin{pmatrix} \vdots \end{pmatrix}$

$= (\ )_{D \times D}$

$\Phi \Phi^T = \begin{pmatrix} \vdots \end{pmatrix} (\cdots)$

$x_i^T x_j$

$= \left(\ \right)_{N \times N}$

inner product

$E[t] = E[y] = 0$

$$E[tt^T] = cov(Y) + cov(t|y)$$

$$= \frac{1}{\alpha} K + \frac{1}{\beta} I$$

For any $x_1 \dots x_N$

$t = (t(x_1) \dots t(x_N)) \sim N(0, C)$

$$C = \frac{1}{\alpha} K + \frac{1}{\beta} I$$

$C = \frac{1}{\alpha} K + \frac{1}{\beta} I$ $\quad k_{RBF} = e^{-\frac{1}{2}\|x_i - x_j\|^2 / s^2}$

Evidence $= P(t \mid \alpha, \beta) = N(0, C_N)$

$\log E = -\frac{N}{2} \log 2\pi - \frac{1}{2} \log |C_N| - \frac{1}{2} t^T C_N^{-1} t$

= Marginal Likelihood.

$\frac{\partial}{\partial \alpha} \quad \frac{\partial}{\partial \beta} \quad \frac{\partial}{\partial s}$ $\Longleftrightarrow$ Gradient Descent

Logistic - Bernoulli ?

if $m \neq 0$ then ?