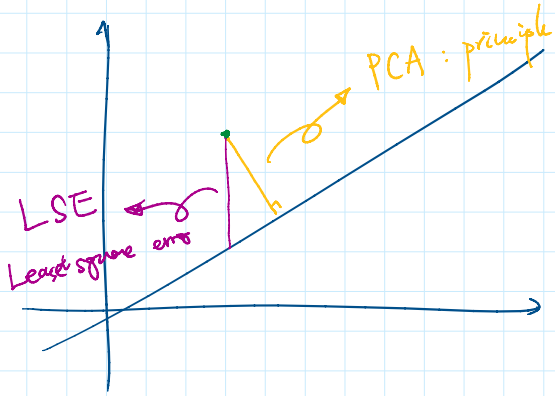
Machine Learning Note

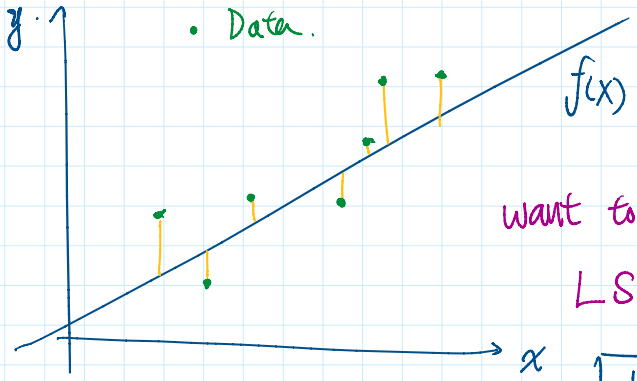
085609 資科工碩一 方偉編

Color Edition is available on GitHub:
https://github.com/Warrenww/Machine-Learning/blob/master/ML%20note.pdf

Statistic & Probability

Thursday, March 5, 2020 10:35 AM





square to eliminate the sign.

want to : $\min \left(\sum_i (f(x) - y_i)^2 \right)$

LSE — error

Linear Regression

* Loss function
* gain function
* objective func.

} try to minimize these functions.

instead to tell computer what to do.

$$L = (ax_0 + b - y_0)^2 + (ax_1 + b - y_1)^2 + \dots$$

to find min a, b $\Rightarrow$ $\frac{\partial L}{\partial a}$, $\frac{\partial L}{\partial b}$ (calculate)

$$\underset{x}{\operatorname{argmin}} \left(\|Ax - b\|^2 \right)$$ — L_2 norm

$\Rightarrow$

$$L = \begin{bmatrix} ax_0 + b - y_0 & ax_1 + b - y_1 & \dots \end{bmatrix} \begin{bmatrix} ax_0 + b - y_0 \\ ax_1 + b - y_1 \\ \vdots \end{bmatrix}$$ (matrix)

$$= (Ax - b)^T (Ax - b)$$

$$= (x^T A^T - b^T)(Ax - b)$$

$$= (x^T A^T - b^T)(Ax - b)$$

$$= x^T A^T A x - x^T A^T b - b^T A x + b^T b$$

$$= x^T A^T A x - 2 x^T A^T b + b^T b$$

x^2

matrix calculus.

∵ b is $n \times 1$, x is $1 \times n$

∴ these are scalar ⇒ they are same.

$$\frac{\partial L}{\partial x} = 2A^T A x - 2A^T b = 0$$

$$\Rightarrow \quad A^T A x = A^T b$$

$$\Rightarrow \quad (A^T A)^{-1}(A^T A)x = (A^T A)^{-1} A^T b$$

$$\Rightarrow \quad x = (A^T A)^{-1} A^T b$$

(?) $A^T A$ always have inverse?

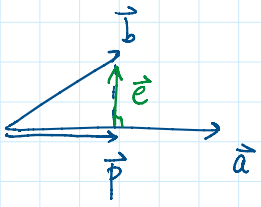
⇒ $A^T A$ is symmetric gram matrix

⇒ $A^T A$ is semi positive definite.

⇒ $A^T A \geq 0$ (not always invertable)

§ orthogonal projection matrix.

1-D



$\vec{p} = \vec{a} \cdot x$ (scalar.)

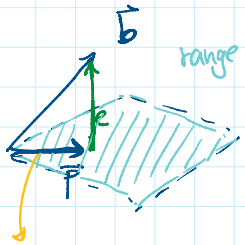
$a^T e = 0$

$\vec{e} = \vec{b} - \vec{p}$

$a^T(b - p) = 0$

$a^T(b - ax) = 0$

2-D



the closest we can get on range of A is the projection.

$A^T(b - Ax) = 0$

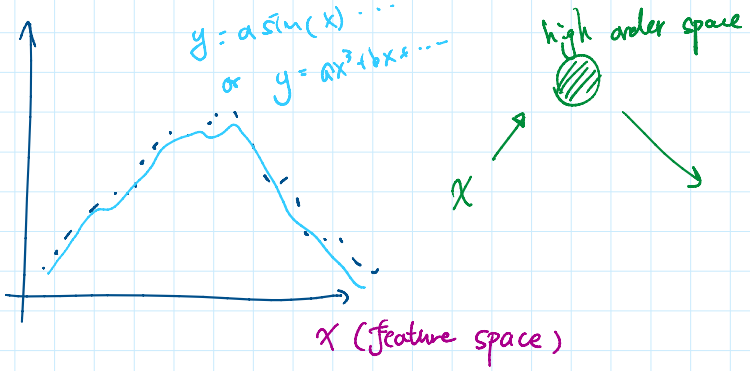
$A^T b = A^T A x$

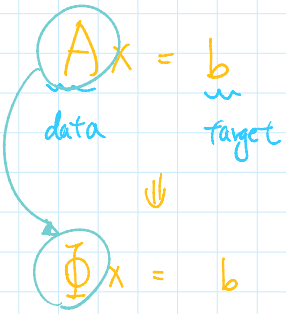
$x = (A^T A)^{-1} A^T b$

$$a^T b - a^T a x = 0$$

$$x = \frac{a^T}{a^T a} b$$

- To solve a non-linear regression.





$$LSE = \|\Phi x - b\|^2$$

$$Ax = \begin{bmatrix} ax_0^3 + bx_0^2 + cx_0 + d & \cdots \end{bmatrix} \begin{bmatrix} ax_0^3 + bx_0^2 + cx_0 + d \\ \vdots \end{bmatrix}$$

basis functions

$$\phi_0(x) = x^3$$
$$\phi_1(x) = x^2$$
$$\phi_2(x) = x$$
$$\phi_3(x) = 1$$

$$\Phi x = \begin{bmatrix} \phi_0(x_0) & \phi_1(x_0) & \phi_2(x_0) & \phi_3(x_0) & \cdots \end{bmatrix} \begin{bmatrix} \vdots \end{bmatrix}$$

Design Matrix

$$L = (\Phi x - b)^T (\Phi x - b)$$

$$\frac{\partial L}{\partial x} = (\Phi^T \Phi)^{-1} \Phi^T b$$

Linear Basis Regression

§ Regularization.

loss

$$LSE = \underset{x}{\operatorname{argmin}} \left(\|Ax - b\|^2 + \|x\|^2 \right)$$

$LSE = \underset{x}{\operatorname{argmin}} \left(\|Ax-b\|^2 + \|x\|^2 \right)$

LSE (penalty term)

regularization term

L_2 norm $= \|x\|^2$ ridge regression

L_1 norm $= |x|$ Lasso regression

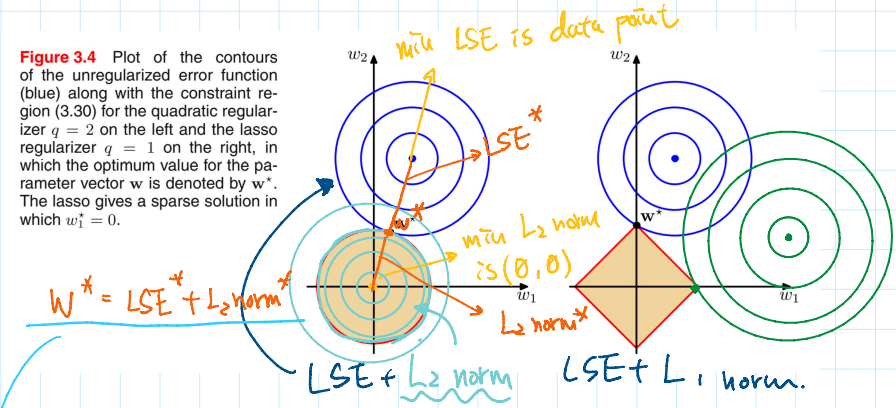
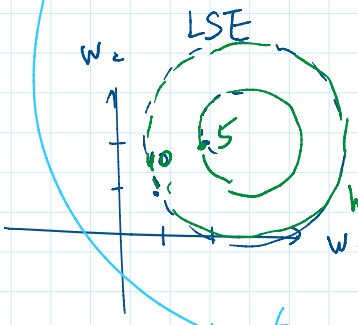


Figure 3.4 Plot of the contours of the unregularized error function (blue) along with the constraint region (3.30) for the quadratic regularizer $q = 2$ on the left and the lasso regularizer $q = 1$ on the right, in which the optimum value for the parameter vector $\mathbf{w}$ is denoted by $\mathbf{w}^\star$. The lasso gives a sparse solution in which $w_1^\star = 0$.

$\|x\|^2 = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$

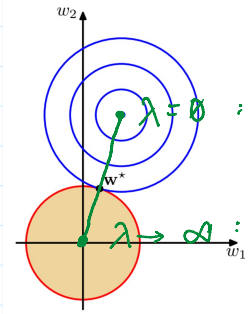


$LSE = \|Ax-b\|^2$

have the same LSE

to control the weight of focusing on LSE or norm.

$\Rightarrow LSE = \underset{x}{\operatorname{argmin}} \left(\|Ax-b\|^2 + \lambda \|x\|^2 \right)$



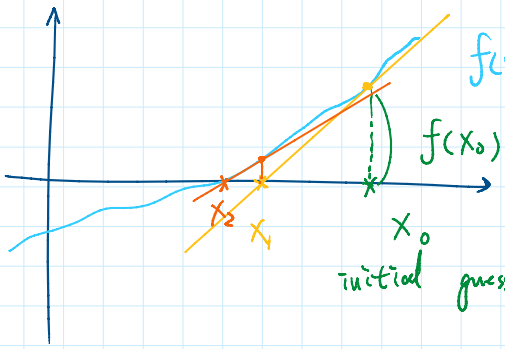
$\lambda = 0$: only LSE term

$\lambda \to \infty$: only L_2 norm term.

§ Non-linear Loss Function.

Newton's Method. : root finding

$f(x) = 0$



$$f'(x_0) = \frac{f(x_0)}{x_0 - x_1}$$

$$\Downarrow$$

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

$$\vdots$$

until converge. $x_n = x_{n-1} - \frac{f(x_n)}{f'(x_n)}$

$\Rightarrow f(x_n) = 0$ root.

§ Taylor's expansion (most useful)

approximate a function.

$\Rightarrow$ In Newton's Method. using tangent line to approximate original function.

the root of tangent line is approximate to the original function.

$$f(x) \approx \underbrace{f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n}_{g(x)} + \cdots$$

$$g(x) \Rightarrow g(x_0) = f(x_0)$$

$$g'(x_0) = f'(x_0)$$

$$\vdots$$

$$g^{(n)}(x_0) = f^{(n)}(x_0)$$

eg: $f(x) = e^x$

$$f(x) \approx 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

eg: $f(x) = e^{ix}$ $\quad f'(x) = e^{ix} \cdot i$, $f''(x) = -e^{ix}$ ...

$$f(x) \approx 1 + ix - \frac{1}{2!}x^2 - \frac{i}{3!}x^3 + \frac{x^4}{4!} + \cdots$$

eg: $f(x) = \sin(x)$ $f'(x) = \cos x$ $f''(x) = \sin x$

$$f(x) \approx \frac{1}{1!}x + \frac{1}{3!}x^3 + \cdots$$

$$f(x) = \cos(x)$$

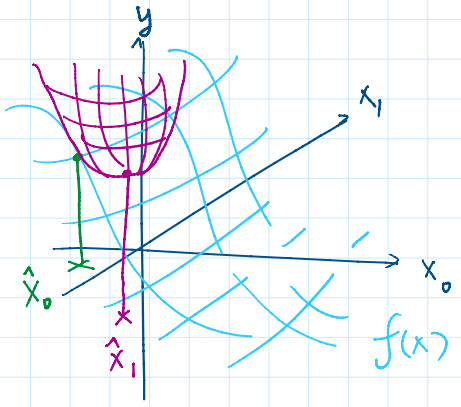
$$f(x) = 1 + \frac{1}{2!}x^2 + \frac{x^4}{4!} + \cdots$$

$$\Rightarrow \quad e^{ix} = \cos x - i\sin x$$

- Nonlinear Loss Function

$$f(\hat{x}) \approx g(x) = f(\hat{x}_0) + f'(\hat{x}_0)\,\Delta x + \frac{1}{2} f''(\hat{x}_0)\cdot \Delta x^2$$

$\Delta x = (\hat{x} - \hat{x}_0)$ quadratic approximate



$$\hat{x} = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} \quad f\left(\begin{bmatrix} x_0 \\ x_1 \end{bmatrix}\right) = y$$

$$g'(x) = f'(\hat{x}_0) + \frac{1}{2} f''(\hat{x}_0)\,\Delta x = 0$$

$$\Rightarrow \quad \Delta x = -\frac{f'(\hat{x}_0)}{f''(\hat{x}_0)}$$

$$a_{n+1} = a_n + \Delta x = x_n + \frac{-f(a_n)}{f''(a_n)}$$

In high dimension: $f''(\hat{x}_n)$ is an matrix

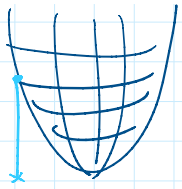
let $f''(\hat{x}_n) = H$

$$\hat{x}_{n+1} = \hat{x}_n + \underbrace{H^{-1} f(\hat{x}_n)}_{\text{Hessian Matrix}} \cdot \underbrace{\nabla f(\hat{x}_n)}_{\text{gradient}}$$

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_0} \\ \frac{\partial f}{\partial x_1} \end{bmatrix}$$

$$Hf(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2}{\partial x_1 x_2} & \cdots \\ \frac{\partial^2}{\partial x_2 \partial x_1} & \frac{\partial f^2}{\partial x_2^2} & \ddots \\ \vdots & & \frac{\partial f^2}{\partial x_n^2} \end{bmatrix}$$

eg: $\|Ax - b\|^2$ LSE



$f(x) = \|Ax - b\|^2 = x^T A^T A x - 2x^T A^T b + b^T b$

$\nabla f(x) = 2A^T A x - 2A^T b$

$Hf(x) = 2A^T A$

$x_0 = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$

$x_{n+1} = x_n - H^{-1} f(x_n) \cdot \nabla f(x_n)$

$= 0 - (2A^T A)^{-1}(2A^T A x_n - 2A^T b)$

$= -\frac{1}{2}(A^T A)^{-1} \cdot 2A^T b$

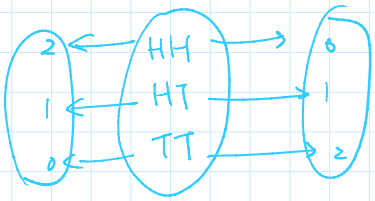
$= (A^T A)^{-1} A^T b$ same result

- Initial Point ⇒ trap in local optimize.
- fastest (least iteration)

  * the time is usually slow ∵ H is hard to calculate.

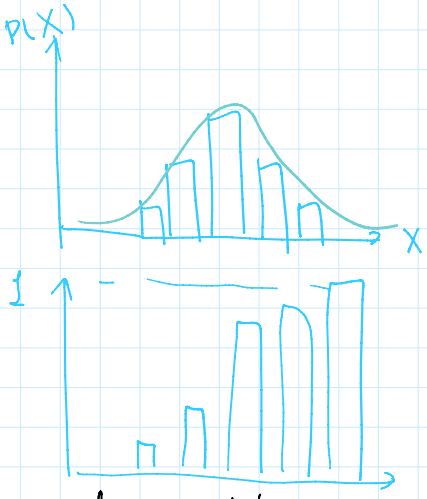
§ Probability.

- trail (試驗) : tossing a coin
- out come : head or tail
- event : set of outcomes.
- sample space ($\mathcal{U}$) all out come
- random variable (X) mapping function.



$P(X = \text{head})$ $P(X=1)$ $P(X=1|\mathcal{U})$

* to get the $P(X=1)$ need infinite times of trails



discrete P.M.f

probability mass function

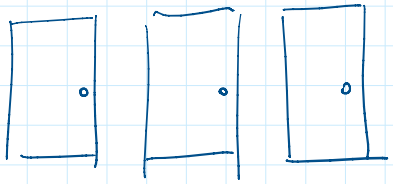
continuous P.D.F

probability density function

C.D.F:

cumulative distribution function

- Conditional Probability
- Joint Probability.
- Monty hall problem.



2 goat 1 car behind the door.

player chose a door.

then host open a wrong door then axle.

the player whether want to change the door or not.

you should change.

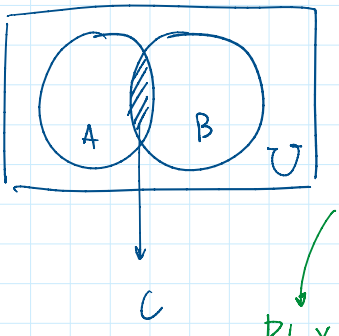
open the door chaing the U

v initial chose

x open door

羊	車	羊	change	not change
V		X	T	F
X	V	X	F	T
X		V	T	F
			2/3	1/3

- Venn diagram.



P(X = A | U)

P(X = B | U)

P(X = C | U)

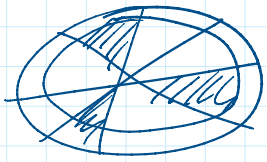
joint probability.

P(X = C | A)

$$P(C|U) = P(A)\cdot P(B|A) = P(B)\cdot P(A|B)$$

$$P(A)P(B|A) = P(B)P(A|B)$$

$$\Rightarrow P(B|A) = \frac{P(B)\cdot P(A|B)}{P(A)}$$ Boyels theorm



Black. Red

50 : 50

B B B B B B B B B □

prior 先验.

Bayesian

prior knowledge

Frequatist

observation

Sum rule: $P(A) = P(A,B) + P(A, \text{not } B)$

product rule: $P(A,B) = P(B|A)\cdot P(A)$

(chain rule) $P(A_4, A_3, A_2, A_1)$

$= P(A_4, A_3 A_2 | A_1)\, P(A_1)$

$= P(A_4 A_3 | A_2, A_1)\, P(A_2|A_1)\, P(A_1)$

prior → parameter

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

posteriness

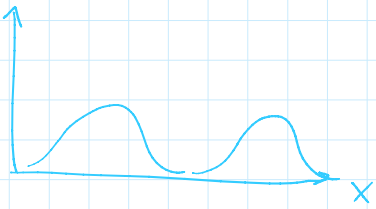
marginal

$$P(\theta|D) = \frac{P(\theta)\,P(D|\theta)}{P(D)}$$

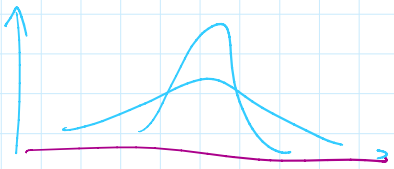
θ: parameter D: data.

§ Distribution.

- Location.



* mean $E_p(X) = \sum p(x_i)\, x_i$ first moment $E(X)$
* mode (max) (most fregent)
* medtan
- dispersion



* variance. $\frac{\sum \| x-\mu \|}{n} = \sum p(x)(x-\mu)^2$

  2nd moment $E((x-\mu)^2)$

  $= E(x^2) - E^2(x)$

* Covariance.

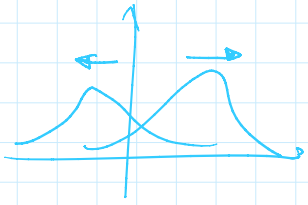
$$Cov(x,y) = E((x-E(x))(y-E(y)))$$

$$= \frac{\sum (x-\mu_x)(y-\mu_y)}{n}$$

important to Principle Componet Analyses.

$$E((x-\mu)^2)$$
$$= \frac{\sum (x-\mu)^2}{n}$$
$$= \frac{1}{n}\sum_i (x^2 - 2x\mu + \mu^2)$$
$$= \frac{1}{n}\sum x^2 - \frac{1}{n} 2\mu \sum x + \frac{1}{n} n\mu^2$$
$$= \frac{1}{n}\sum x^2 - 2\mu \frac{\sum x}{n} + \mu^2$$
$$= \frac{1}{n}\sum x^2 - 2\mu^2 + \mu^2$$
$$= \frac{1}{n}\sum x^2 - \mu^2$$
$$= E(x^2) - E^2(x)$$

- skewness $\sum p(x)(x-\mu)^3$



- Kurtosis. $\sum p(x)(x-\mu)^4$

§ Naive Bayes classifier

Maximum likelihood.

$$P(D \mid \theta) = L(\theta \mid D)$$

$P(H,H,H \mid \theta = 1) = 1$

$P(H,H,H \mid \theta = 0.5) = \frac{1}{8}$

$P(H,H,H \mid \vdots) = \vdots$

D is data obtain

θ is parameter.
here represent the probability of head.

$$P(\theta, D) = \frac{P(\theta)\,P(D \mid \theta)}{P(D)}$$

- Baysian.

$$P(\underbrace{A}_{\theta} \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

$$\underbrace{P(\theta \mid B)}_{\text{posterior}} = \frac{\overbrace{P(B \mid \theta)}^{\text{likelihood}} \cdot \overbrace{P(\theta)}^{\text{prior}}}{\underbrace{P(B)}_{\text{marginal}}}$$

prior: come from assumption / data / experience

marginal $= \sum P(B, \theta)$ ← hard to get.

$$P(\theta \mid B) \propto P(B \mid \theta) \cdot P(\theta)$$

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$$P\left(\begin{array}{l}\text{play} = \text{yes} \\ \text{play} = \text{no}\end{array} \;\middle|\; \begin{array}{l}\text{outlook} = \text{sunny},\ \text{temp} = \text{cool}, \\ \text{Humidity} = \text{high},\ \text{wind} = \text{strong}\end{array}\right) = D$$

$$\left[\begin{array}{l} \dfrac{P(D \mid play=yes)\cdot P(play=yes)}{P(D)} \\ \dfrac{P(D \mid play=No)\cdot P(play=no)}{P(D)} \end{array}\right.$$

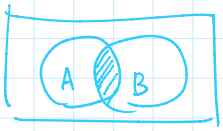
there are no such day satisfy D

$$\Rightarrow \begin{array}{l} P(D \mid play=yes) \\ P(D \mid play=no) \end{array} \doteq 0$$

⇒ Naive Baysian classificatin

⇒ Conditional independent

* Independent



$P(A) \cdot P(B)$

$P(A)\cdot P(B) = P(A,B)$

A, B are independent under some conditions

these conditions are from assumption.

$$P(A,B \mid Y) = P(A \mid Y)\,P(B \mid Y)$$

use chain rule to decompose D

$$P(A_1 A_2 A_3 \underline{A_4}) = P(A_1 \mid A_2 A_3 A_4)\cdot P(A_2 \mid A_3 A_4)\cdot P(A_3 \mid A_4)\,P(A_4)$$

A_4 = play or not

assumption: $P(A_2, A_3 \mid A_4) = P(A_2 \mid A_4)\,P(A_3 \mid A_4)$

$$\Rightarrow \frac{P(A_2 A_3 A_4)}{P(A_4)} \doteq \frac{P(A_2 A_4)}{P(A_4)}\,\frac{P(A_3 A_4)}{P(A_4)}$$

$$\Rightarrow \frac{P(A_2 A_3 A_4)}{} = \frac{P(A_2 A_4)}{}$$

$$\Rightarrow \frac{P(A_2\ A_3\ A_4)}{P(A_3\ A_4)} = \frac{P(A_2\ A_4)}{P(A_4)}$$

$$\Rightarrow P(A_2 \mid A_3 A_4) = P(A_2 \mid A_4)$$

$$\Rightarrow P(A_1\ A_2\ A_3\ A_4) = P(A_1 \mid A_2 A_3 A_4)\cdot P(A_2 \mid A_4) P(A_3 \mid A_4) P(A_4)$$

assumption: $P(A_1\ A_2 A_3 \mid A_4) = P(A_1 \mid A_4) P(A_2 \mid A_4) P(A_3 \mid A_4)$

$$= P(A_1 \mid A_4) \cdot P(A_2 A_3 \mid A_4)$$

$$\frac{P(A_1\ A_2\ A_3\ A_4)}{P(A_4)} = \frac{P(A_2 A_3 A_4)}{P(A_4)} \cdot \frac{P(A_1 A_4)}{P(A_4)}$$

$$\frac{P(A_1\ A_2\ A_3\ A_4)}{P(A_2 A_3 A_4)} = \frac{P(A_1 A_4)}{P(A_4)}$$

$$P(A_1 \mid A_2 A_3 A_4) = P(A_1 \mid A_4)$$

$$\Rightarrow P(A_1\ A_2\ A_3\ A_4) = P(A_1 \mid A_4)\ P(A_2 \mid A_4)\ P(A_3 \mid A_4)\ P(A_4)$$

Now what we need is only:

$P(\text{sunny} \mid \text{Yes}) = 2/9$

$P(\text{cool} \mid \text{Yes}) = 3/9$

$P(\text{high Temp} \mid \text{Yes}) = 3/9$

$P(\text{strong wind} \mid \text{Yes}) = 3/9$

$P(\text{play}) = \frac{9}{14}$

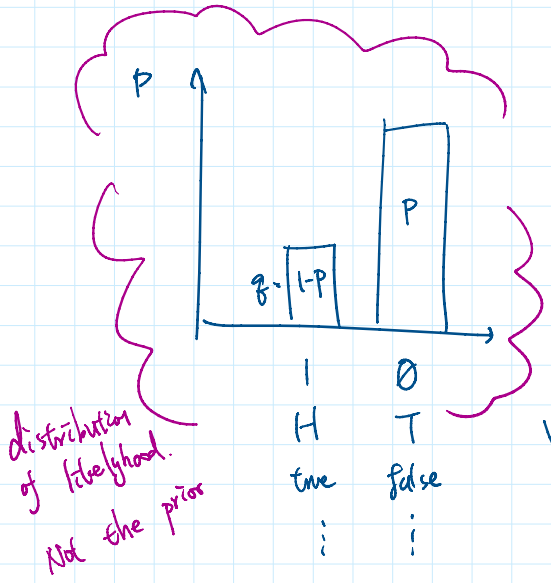
$$P(\text{play} \mid D) = \frac{\left(\frac{2}{9}\cdot\frac{3}{9}\cdot\frac{3}{9}\cdot\frac{3}{9}\right)\cdot\frac{9}{14}}{P(D)} = \frac{0.0053}{P(D)}$$

$$P(\text{not play} \mid D) = \frac{\left(\frac{3}{5}\cdot\frac{1}{5}\cdot\frac{4}{5}\cdot\frac{3}{5}\right)\cdot\frac{5}{14}}{P(D)} = \frac{0.0205}{P(D)}$$

Maximum a posteriori (MAP)

Coin tossing.

- Bernulli Distribution



$$P(X \mid \Theta = p) = p^{x}(1-p)^{1-x}$$

$$E(x) = 0(1-p) + 1(p) = p$$

$$Var(x) = E(x^2) - E(x)^2 = p - p^2 = p(1-p) = pq.$$

Given: $[H, H, T, T, \dots, H]$

$\Downarrow X$

$D = [1, 1, 0, 0, \dots, 1]$ independent.

Maximum Likelyhood (MLE):

$$\prod_i p^{D_i}(1-p)^{(1-D_i)} = Y$$

want to find p maximum the likelyhood.

$\frac{\partial Y}{\partial p}$: hard to calculate.

* $f(a) \le f(b) \iff \log f(a) < \log f(b)$

$$\log Y = \sum_i \log p^{D_i}(1-p)^{1-D_i}$$

$$= \sum \log p^{D_i} + \sum \log (1-p)^{1-D_i}$$

$$= \sum D_i \log p + \sum (1-D_i)\log(1-p)$$

$$\frac{\partial \log Y}{\partial p} = \frac{1}{p}\sum D_i - \frac{1}{1-p}\sum(1-D_i) = 0$$

$$\frac{1}{p}\sum D_i = \frac{1}{1-p}\sum 1 + \frac{1}{1-p}\sum D_i$$

$$\left(\frac{1}{p} - \frac{1}{1-p}\right)\sum D_i = \frac{N}{1-p}$$

$$p = \frac{\sum D_i}{N}$$

$\sum D_i$ → # of head.

N → total.

sample mean.

Information Theory

Tuesday, March 17, 2020 8:18 PM

§ Information Theory.

• Entropy 熵

randomness.

information

$\log_2 P$: unit of randomness (uncertainty)

where P is the probability.

eg: $\log_2 \frac{1}{4}$:= 2 bits to describe the event

$\log_2 \frac{1}{1024}$ = 10 bits to describe the event.

- the expectation of information in event.

$E(X) = \sum X \cdot p(X)$

$E(f(X)) = \sum f(X) \cdot p(X)$

$\Rightarrow H(X) = \sum p(X) \cdot \left(-\log p(X)\right)$

maximum entropy that you can get.

ex: 2 coins.

HH → 2

HT → 1

TH → 1

TT → 0

fair coins

$H(X) = -\sum_{all} \frac{1}{4} \log \frac{1}{4} = 2$ ↑ uniform randomness

unfair coins ($\frac{3}{4}$ H $\frac{1}{4}$ T)

$H(X) = -\left(\frac{9}{16} \cdot \log \frac{9}{16} + 2\frac{3}{16} \cdot \log \frac{3}{16} + \frac{1}{16} \cdot \log \frac{1}{16}\right)$

$= 0.48844$

ex: $-\sum 1 \log 1 = 0 \Rightarrow$ no randomness

- Distance.

• Conditional Entropy.

$$H(X|Y) = -\sum p(x,y) \log \frac{p(x,y)}{p(x)}$$

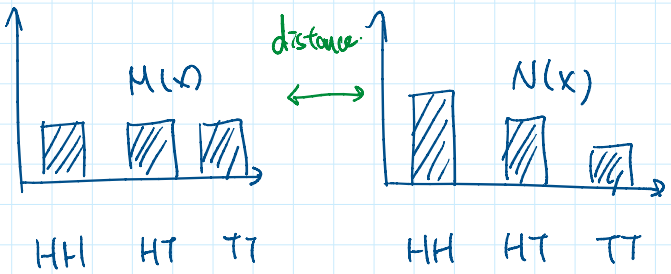
- Joint Entropy.

$$H(X,y) = H(y|x) + H(x)$$

- Relative Entropy.

KL divergence : distance between 2 distributions

[Kullback Leibler] also called information gain.



$$KL(p\|q) = -\sum M(x) \log M(x) - \left(-\sum M(x) \log N(x)\right)$$

where: X : random variable

M, N : distribution.

$$= -\sum M(x) \log \frac{N(x)}{M(x)}$$

* why not N(x)

not work for symmetric distribution KL = 0

- Mutual Information

$$I(X,y) = KL(p(x,y) \| p(x)p(y))$$

$$= H(x) - H(x|y)$$

How independent x,y are.

if x, y are independent $I(x,y) = 0$

Maximum Entropy Principle

cf: Maximum Likelihood.

coin tossing. 10 times

D: H, H, ···· H ← observed (data)

* $L(\theta=0.5 \mid H, H, \cdots H)$ ← Likelihood the coin is fair (0.5 H 0.5 T)

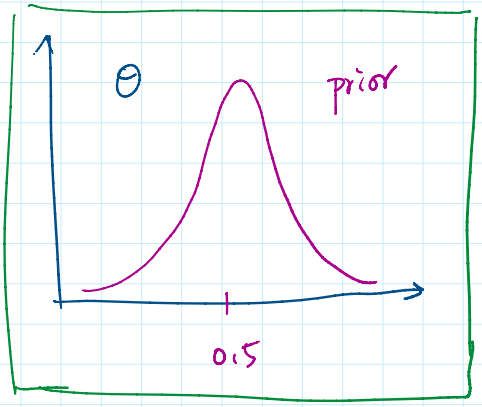
$= P(H, H, \cdots H \mid \theta=0.5)$

$= 1/1024$

* $L(\theta=0.9 \mid D) = P(D \mid \theta=0.9) = (\frac{9}{10})^{10}$

* $L(\theta=1 \mid D) = P(D \mid \theta=1) = 1$

maximum likelihood



Bayesian

frequentist.
⇒ every θ is equal.

Uniform Distribution. ⟶ Maximum Entropy

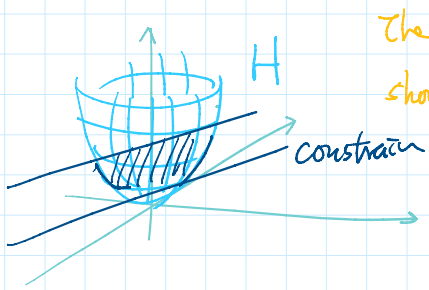
$H = -\sum p(x) \log p(x)$

$= -\int_{-\infty}^{\infty} p(x) \log p(x)\, dx$

, $p(x)$ pdf

- $p(x) \geq 0$
- $\int_a^b p(x) = 1$

- Lagrange multiplier



The gradient of target should parallize the tangent of the constrain.

constrain.

$$L = -\int_a^b p(x)\ln p(x)\,dx + \lambda\left(\int_a^b p(x)\,dx - 1\right)$$

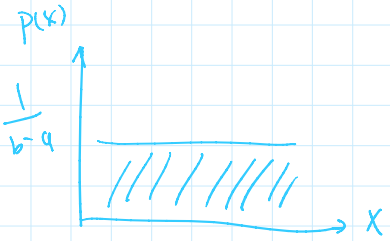
$$\frac{\partial L}{\partial p(x)} = -\left(\frac{\partial p(x)}{\partial p(x)}\cdot \ln p(x) + p(x)\cdot\frac{\partial \ln p(x)}{\partial p(x)}\right) + \lambda\cdot\frac{\partial p(x)}{\partial p(x)}$$

$$= -(\ln p(x) + 1) + \lambda = 0$$

$$\Rightarrow p(x) = e^{\lambda-1}$$

$$\int_a^b e^{\lambda-1}\,dx = 1 \Rightarrow e^{\lambda-1}(b-a) = 1$$

$$\Rightarrow p(x) = e^{\lambda-1} = \frac{1}{b-a}$$



Expectation is known.

$$E(x) = \int_0^\infty x\,p(x)\,dx = \mu$$

new constrain

$$L = -\int_0^\infty p(x)\ln p(x)\,dx + \lambda_0\left(\int_0^\infty p(x)\,dx - 1\right) + \lambda_1\left(\int_0^\infty x\,p(x)\,dx - \mu\right)$$

$$\frac{\partial L}{\partial p(x)} = -(1+\ln p(x)) + \lambda_0 + \lambda_1 x$$

$$\Rightarrow \ln p(x) = \lambda_0 + \lambda_1 x - 1$$

$$p(x) = e^{\lambda_0 + \lambda_1 x - 1}$$

Constraint 1: $\int p(x)\,dx = \int e^{\lambda_0+\lambda_1 x-1}\,dx$

$$= e^{\lambda_0 - 1} \cdot \int e^{\lambda_1 x}\,dx = 1$$

$$\Rightarrow e^{\lambda_0-1} \cdot \frac{1}{\lambda_1} e^{\lambda_1 x}\Big|_0^\infty = 1$$

$\lambda_1 < 0$ or never converge.

$$\Rightarrow e^{\lambda_0-1} \cdot \frac{1}{\lambda_1}\left(\frac{1}{e^{-\lambda_1 \infty}} - \frac{1}{e^0}\right) = 1$$

$$\Rightarrow e^{\lambda_0-1} \cdot \frac{-1}{\lambda_1} = 1$$

$$\Rightarrow \lambda_1 = -e^{\lambda_0 - 1}$$

Constraint 2: $\int_0^\infty x\left(e^{\lambda_0+\lambda_1 x-1}\right)dx = \mu$

$$\Rightarrow e^{\lambda_0-1}\int_0^\infty x e^{\lambda_1 x}\,dx = \mu$$

$$\int_0^\infty x e^{\lambda_1 x}dx = \frac{1}{\lambda_1}\int_0^\infty x\,d e^{\lambda_1 x}$$

$$\int u\,dv = uv - \int v\,du = \frac{1}{\lambda_1}\left(x e^{\lambda_1 x} - \int_0^\infty e^{\lambda x}\,dx\right)$$

$$\Rightarrow e^{\lambda_0-1}\left(x \cdot \frac{1}{\lambda_1} e^{\lambda_1 x}\Big|_0^\infty - \frac{1}{\lambda_1}\int_0^\infty e^{\lambda_1 x}dx\right) = \mu$$

$$\Rightarrow e^{\lambda_0-1}\left(0 - \frac{1}{\lambda_1}\int_0^\infty e^{\lambda x}dx\right) = \mu$$

$$\Rightarrow \mu = \frac{-1}{\lambda_1}\cdot e^{\lambda_0-1}\int_0^\infty e^{\lambda_1 x}\,dx$$

$$= \frac{-1}{\lambda_1}\int_0^\infty e^{\lambda_0+\lambda_1 x-1}\,dx = \int p(x)\,dx = 1$$

$$= \frac{-1}{\lambda_1}$$

$$\Rightarrow \lambda_1 = \frac{-1}{\mu}$$

$$\Rightarrow p(x) = e^{\lambda_0 + \lambda_1 x - 1}$$

$$= e^{\lambda_0 - 1} \cdot e^{\lambda_1 x} = \frac{1}{\mu} \cdot e^{-\frac{x}{\mu}}$$

exponential distribution

- Knowing the 2nd moment.

$$\int x^2 p(x)\, dx = \sigma^2$$

; same as above.

$$-\ln p(x) - 1 + \lambda_0 + \lambda_1 x + \lambda_2 x^2 = 0$$

$$\Rightarrow p(x) = e^{-(1+\lambda_0 + \lambda_1 x + \lambda_2 x^2}$$

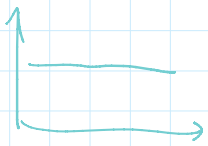
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Gaussian distribution

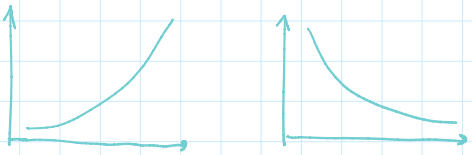
- Conclution

known nothing: Uniform distribution

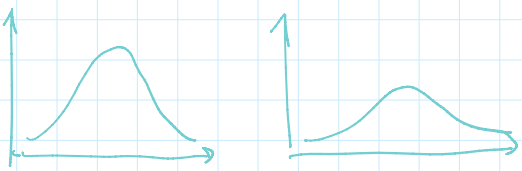
all θ have same value.



given mean: Exponential distribution



given variance: Gaussian distribution



- Location
- dispersion

: moments

▶ toss a coin for one time.

$\theta = \{H, T\} : 1$

Bernoulli Distribution.

$$P\left(\begin{matrix} x=0 \\ x=1 \end{matrix} \middle| \theta\right) = \theta^{x}(1-\theta)^{1-x}$$

$$P(1 \mid \theta) = \theta^{1}(1-\theta)^{0} = \theta$$

$$P(0 \mid \theta) = \theta^{0}(1-\theta)^{1} = 1-\theta$$

$E(X) = \theta \cdot 1 + (1-\theta)\cdot 0 = \theta$ location

$Var(X) = E(X^2) - E^2(X)$

$= \theta - \theta^2 = \theta(1-\theta)$ dispersion

▶ toss a coins for N times.

Binominal Distribution

identical independent (iid)

Given $[1, 0, 1, 1, 0, \cdots] \Rightarrow [X_1, X_2, \cdots, X_n] = D$

$P(D \mid \theta) = \prod_{k=1}^{n} \theta^{x_k}(1-\theta)^{1-x_k}$

likelyhood.

Random variable $X \begin{cases} 0 \\ 1 \end{cases}$

multiple Bernoli trail N times.

X : # of head in N trails.

$D: [x_0, x_1, \cdots x_n]$

$\sum x_i = m.$

$$P(X = m \mid N, \theta) = C_m^N \theta^m (1-\theta)^{N-m}$$

$E(x) = N\theta$

$Var(X) = N\theta(1-\theta)$

MLE : θ frequentist

§ Conjugate Prior

Frequentist → Bayesian

- Prior
- experience

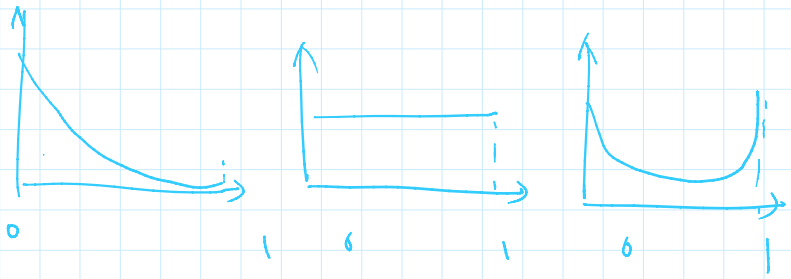
Bernuli / Binominal ⟷ Beta Distribution

Prior & Posterior in the same form

- Beta Distribution

$$\text{Beta}(\theta \mid a,b) = \theta^{a-1}(1-\theta)^{b-1}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

$$= \theta^{a-1}((-\theta)^{b-1} \cdot \frac{1}{\beta(a,b)}$$



Γ(gamma) : factorial $n! = n(n-1)(n-2)\cdots 1$

$$\Gamma(x) = \int_0^\infty p^{x-1} e^{p} dp$$

β (beta) : $\beta(a,b) = \int_0^1 p^{a-1}(1-p)^{b-1} dp$

$$= \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

$$E(x) = \frac{a}{a+b}$$

$$\mathrm{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

$$\int_0^1 x \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}\,dx$$

$$= \frac{1}{B(a,b)} \int_0^1 x \cdot x^{a-1}(1-x)^{b-1}\,dx$$

$$= \frac{1}{B(a,b)} \int_0^1 x^{a}(1-x)^{b-1}\,dx$$

$$= \frac{1}{B(a,b)} B(a+1, b)$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{a\cdot\Gamma(a)\cdot\Gamma(b)}{(a+b)\cdot\Gamma(a+b+1)} = \frac{a}{a+b}$$

- Beta - Binomial conjugation

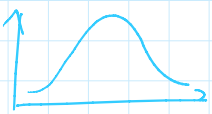
$$\underbrace{P(X \mid N, p)}_{\text{Binomial}} \cdot \underbrace{P(p \mid a, b)}_{\text{prior, Beta}}$$

$$C_M^N p^M (1-p)^{N-M} \cdot p^{a-1}(1-p)^{b-1} \cdot \frac{1}{B(a,b)}$$

$$\propto p^{M+a-1} \cdot (1-p)^{N-M+b-1}$$

- MLE is a point estimation

$$\frac{P(X \mid \theta) P(\theta)}{P(X)} = P(\theta \mid X)$$



$$\left[p^{M+a-1} (1-p)^{N-M+b-1} \right] \cdot \bigcirc$$

$P(X)$

$$\sum_a p(X, a) \quad \text{or} \quad \int p(X, a)\,da.$$

Beta (p | m+a, N-m+b).

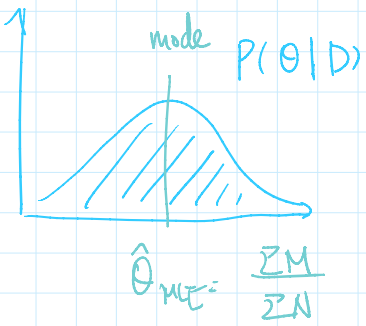
success ≠ fail

prior : Beta (p | a, b)

likelihood : Binomial (X = m | N, P)

posterior : Beta (p | m+a, N-m+b)

most likely θ :



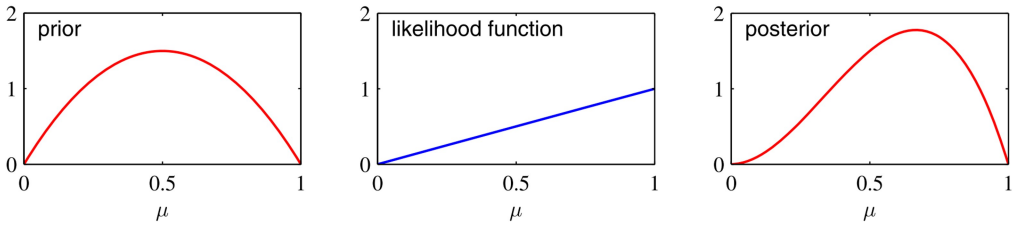
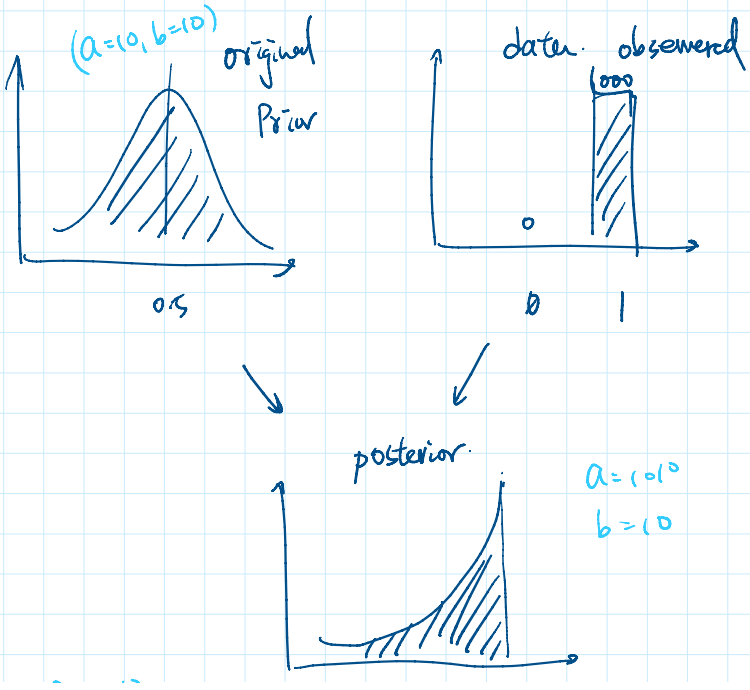
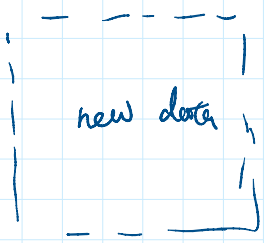


Figure 2.3 Illustration of one step of sequential Bayesian inference. The prior is given by a beta distribution with parameters $a = 2$, $b = 2$, and the likelihood function, given by (2.9) with $N = m = 1$, corresponds to a single observation of $x = 1$, so that the posterior is given by a beta distribution with parameters $a = 3$, $b = 2$.



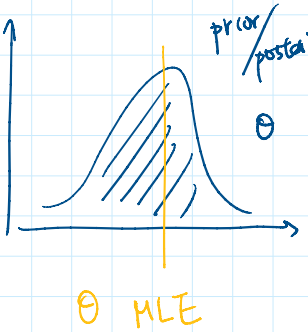
$a = 10$

$b = 10$



update posterior

Sequential learning (online learning)



Frequentist :

likelihood function

→ MLE to the θ

Bayesian : point estimation

- posterior distribution.

  from conjugate prior.

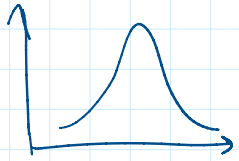
- MLE to the $\hat{\theta}$

  MAP (maximum a posterior)

* $\hat{\theta}_{MAP}$

prior: $\frac{1}{B(a,b)} \cdot p^{a-1}(1-p)^{b-1} \cdot C_m^N \cdot p^m (1-p)^{N-m}$

posterior : $b(p \mid a' = m+a,\ b' = N-m+b)$ update. not re-calculate



mode of $b(a+m,\ N-m+b)$

mode is where they get the maximum likelihood.

$$\frac{1}{B(a+m,\ N-m+b)} \cdot p^{m+a-1} \cdot (1-p)^{N-m+b-1}$$

w.r.t p

$$\frac{db}{dp} = 0 \Rightarrow$$

$$\frac{1}{\beta(a+m, N-m+b)} \left((m+a-1)\, p^{m+a-2} \cdot (1-p)^{N-m+b-1} + (N-m+b-1) \cdot p^{m+a-1} (1-p)^{N-m+b-2} \right) = 0$$

$$\Rightarrow \quad \hat{\theta}_{MLE}$$

Average of $\{x_1, \dots, x_k\}$ $\underline{\underline{x_{k+1}}}$ new data.

$$\frac{\sum x}{k} = \mu$$

$$\text{Avg} = \frac{\sum_{}^{k+1} x}{k+1} \quad (\text{Naive})$$

$$= \frac{k \cdot \mu + x_{k+1}}{k+1}$$

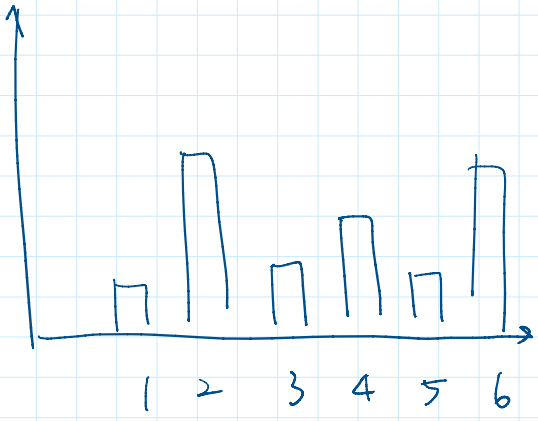
Frequentist : 认为中无 Prior. 又有 data

Bayesian : 利用以前的 knowledge. 累积 Prior.

Prior 就是无随机性. 是惯性论.

§ Multinomial.

v.s. binomial.



dice 100 times = N

$X = [30, 40, \dots\dots]$

$m_1, m_2, \dots m_k$

$$M(x) = \binom{N}{m_1, m_2 \cdots m_k} \pi P_k^{m_k}$$

$$\Downarrow$$

$$\frac{N!}{m_1! \, m_2! \cdots m_k!}$$

conjugate prior

Dirichlet

$$\frac{\Gamma(a_1 + \cdots + a_k)}{\Gamma a_1 \, \Gamma a_2 \cdots \Gamma a_k} \cdot \pi P_k^{a_k - 1}$$

Gaussian distribution

Thursday, April 9, 2020 4:20 PM

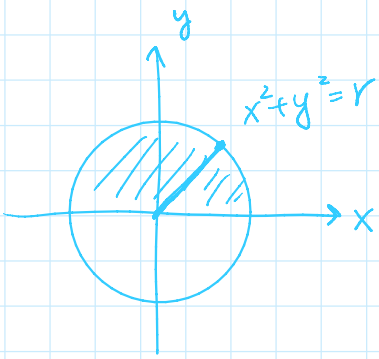
§ Gaussian distribution

- Gaussian integral

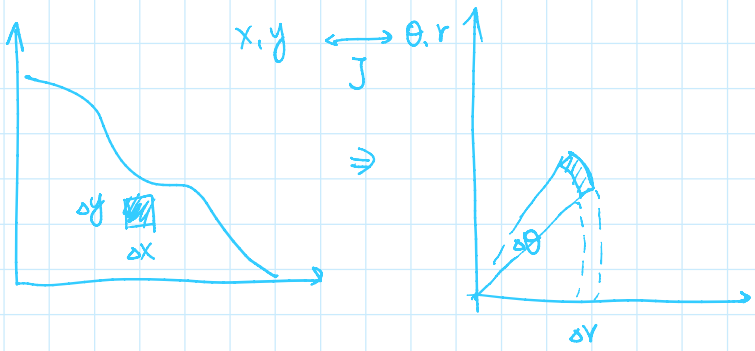
$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

- polar coordinate

$$x = r\cos\theta$$
$$y = r\sin\theta$$



- Jacobian determinant



$$J = \det\left(\begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{bmatrix}\right) = \det\left(\begin{bmatrix} \frac{\partial r\cos\theta}{\partial r} & \frac{\partial r\cos\theta}{\partial \theta} \\ \frac{\partial r\sin\theta}{\partial r} & \frac{\partial r\sin\theta}{\partial \theta} \end{bmatrix}\right)$$

$$= \det\left(\begin{bmatrix} \cos\theta & -r\sin\theta \\ r\cos\theta & \sin\theta \end{bmatrix}\right) = r$$

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\left(\int_{-\infty}^{\infty} e^{-x^2} dx\right)^2}$$

$$= \sqrt{\int_{-\infty}^{\infty} e^{-x^2} dx \cdot \int_{-\infty}^{\infty} e^{-y^2} dy}$$

$$= \sqrt{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)} dx\,dy}$$

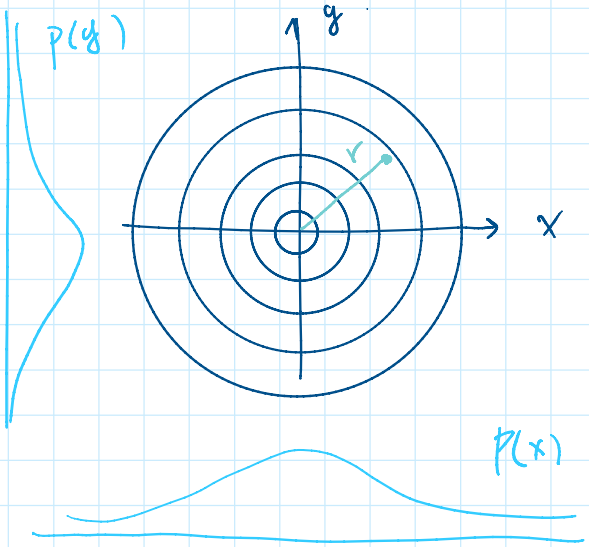
$$= \sqrt{\int_{0}^{2\pi}\int_{0}^{\infty} e^{-(r^2\cos^2\theta + r^2\sin^2\theta)} r\,dr\,d\theta}$$

$$= \sqrt{\int_{0}^{2\pi}\int_{0}^{\infty} e^{-r^2} r\,dr\,d\theta}$$

$$= \sqrt{2\pi \cdot \frac{1}{2}\int_{0}^{\infty} e^{-r^2} dr^2}$$

$$= \sqrt{\pi \cdot (-e^{-r^2})\Big|_0^{\infty}} = \sqrt{\pi}$$

- Gaussian distribution



independent

$p(x,y) = p(x) \cdot p(y)$

$= g(r)$ w.r.t θ.

the probability with same r should be same.

$$\Rightarrow \frac{\partial g(r)}{\partial \theta} = \frac{d\,p(x)\,p(y)}{d\theta} = 0$$

$$p(x) \cdot \frac{dp(y)}{d\theta} + p(y) \cdot \frac{dp(x)}{d\theta} = 0$$

$$p(x) \cdot \frac{dp(y)}{dy} \cdot \frac{dr\sin\theta}{d\theta} + p(y) \cdot \frac{dp(x)}{dx} \cdot \frac{dr\cos\theta}{d\theta} = 0$$

$$p(x)\,p'(y)\cdot r\cos\theta - p(y)\cdot p'(x)\cdot r\sin\theta = 0$$

$$p(x)\cdot p'(y)\cdot x = p(y)\cdot p'(x)\cdot y$$

$$\frac{p(x)\cdot x}{p'(x)} = \frac{p(y)\cdot y}{p'(y)} = C$$

x, y are independent. the only possibility to make these 2 eqs. equal is they converge to a constant.

$$p(x)\cdot x = C\cdot\frac{dp(x)}{dx}$$

$$\frac{1}{C}x\,dx = \frac{1}{p(x)}\,dp(x)$$

$$\int \frac{1}{C}x\,dx = \int \frac{1}{p(x)}\,dp(x)$$

$$\frac{1}{2C}x^2 = \ln p(x)$$

should < 0, otherwise diverge.

$$p(x) = e^{\frac{1}{2c_1}x^2} = e^{-k_1x^2}$$

$$p(y) = e^{\frac{1}{2c_2}y^2} = e^{-k_2y^2}$$

they are pdf $\Rightarrow$ integral to 1.

$$A\cdot\int_{-\infty}^{\infty} e^{-kx^2}\,dx = 1$$

$$\int_{-\infty}^{\infty} e^{-kx^2}\,dx = \frac{1}{A}$$

$$\sqrt{\iint_{-\infty}^{\infty} e^{-k(x^2+y^2)}\,dx\,dy} = \sqrt{\frac{1}{A^2}}$$

$$\sqrt{\int_0^{2\pi}\int_0^{\infty} e^{-kr^2} r\,dr\,d\theta} = \sqrt{2\pi \int_0^{\infty} e^{-kr^2}\frac{dr^2}{2}}$$

$$= \sqrt{\pi\left(\frac{1}{k}e^{-kr^2}\Big|_0^{\infty}\right)}$$

$$= \sqrt{\frac{\pi}{k}} = \sqrt{\frac{1}{A^2}}$$

normalize factor.

$$\Rightarrow A = \sqrt{\frac{k}{\pi}}$$

$$P(x) = \sqrt{\frac{k}{\pi}}\, e^{-kx^2}$$

To find $k \Rightarrow \mu, \sigma$.

Def: $\int_{-\infty}^{\infty} (x-\mu)^2 p(x)\,dx = \sigma^2$

Assume $\mu = 0$

$$\int_{-\infty}^{\infty} x^2 \sqrt{\frac{k}{\pi}}\, e^{-kx^2}\,dx = \sigma^2$$

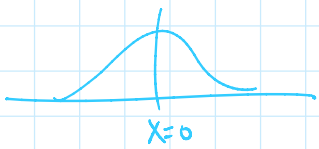
$$\sqrt{\frac{k}{\pi}} \int_{-\infty}^{\infty} x^2 e^{-kx^2}\,dx = \sigma^2$$

$u = x$, $dv = x e^{-kx^2}dx$

$du = dx$ $v = \int x e^{-kx^2}dx = \int e^{-kx^2}\frac{dx^2}{2} = \frac{-1}{2k}e^{-kx^2}$

$$\sqrt{\frac{k}{\pi}}\left(\frac{-x}{2k}e^{-kx^2}\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty}\frac{-1}{2k}e^{-kx^2}dx\right) = \sigma^2$$

↓ it's symmetric



$$2\sqrt{\frac{k}{\pi}}\left(\underbrace{\frac{-x}{2k}e^{-kx^2}\Big|_0^{\infty}}_{=0} - \underbrace{\int_0^{\infty}\frac{-1}{2k}e^{-kx^2}dx}_{=\frac{-1}{2k}\sqrt{\frac{\pi}{k}}}\right) = \sigma^2$$

$$\Rightarrow \quad \sigma^2 = \frac{1}{2k} \quad \Rightarrow \quad k = \frac{1}{2\sigma^2}$$

$$P(x) = \sqrt{\frac{k}{\pi}}\, e^{-kx^2} = \sqrt{\frac{1}{2\pi\sigma^2}}\, e^{\frac{-x^2}{2\sigma^2}}$$

平移至 μ. $\Rightarrow P(x) = \sqrt{\frac{1}{2\pi\sigma^2}}\, e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

In high dimension (multivariable)

$$x \qquad \sigma^2$$

$$\begin{bmatrix} x_0 \\ x_1 \\ \vdots \end{bmatrix} \quad \begin{bmatrix} \sigma_0^2 & cov & \cdots \\ cov & \sigma_1^2 & \\ \vdots & & \ddots \end{bmatrix}$$

$$N(X \mid \mu, \sigma^2) \qquad P(x) = N(X = x \mid \mu, \sigma^2)$$

- Maximum Likelihood Estimation (MLE)

$$\mu_{MLE} = \frac{1}{N} \sum x_n \quad \rightarrow \text{sample mean.}$$

$$\sigma^2_{MLE} = \frac{1}{N} \sum (x_n - \mu_{MLE})^2$$

$$[x_1, x_2 \cdots x_n] \overset{i.i.d}{\sim} N(\mu, \sigma^2)$$

assume independent $\mu_{MLE} \cdot \hat{\mu}$

$$p(x_1 \cdots x_n \mid \mu, \sigma^2) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\left(\frac{-(x_k-\mu)^2}{2\sigma^2}\right)} = L$$

$$\log L = \sum_{k=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(x_k-\mu)^2}{2\sigma^2}}$$

$$= \sum \log \frac{1}{\sqrt{2\pi\sigma^2}} + \sum \frac{-(x_k-\mu)^2}{2\sigma^2}$$

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

log likelihood function

w.r.t μ.

$$\frac{\partial \log L}{\partial \mu} = \frac{-1}{2\sigma^2} \frac{\partial}{\partial \mu} \left(\sum (x-\mu)^2 \right) = 0$$

$$\frac{\partial}{\partial \mu} \left(\sum (x^2 - 2x\mu + \mu^2) \right)$$

$$= \sum_{k=1}^{N} (-2x_k + 2\mu)$$

$$= -2\sum x + 2N\mu = 0$$

$$\Rightarrow \mu_{MLE} = \frac{\sum x}{N}$$

w.r.t. $\sigma^2 = S$

$$\frac{\partial \log L}{\partial S} = \frac{\partial}{\partial S} \left(-\frac{N}{2} \log 2\pi S + \sum \frac{-(x-\mu)^2}{2S} \right)$$

$$= \frac{\partial}{\partial S} \left(-\frac{N}{2} \log 2\pi - \frac{N}{2} \log S + \sum \frac{-(x-\mu)^2}{2} \cdot S^{-1} \right)$$

$$= -\frac{N}{2} \frac{1}{S} - \sum \frac{-(x-\mu)^2}{2} \cdot S^{-2} = 0$$

$$\frac{-N}{2}S^{-1} = \sum \frac{-(x-\mu)^2}{2} S^{-2}$$

$$-NS = \sum -(x-\mu)^2.$$

$$\Rightarrow \sigma^2_{MLE} = S_{MLE} = \frac{\sum (x-\mu)^2}{N}$$

§ Gaussian conjugate

In Bayesian, if posterior and prior are in the same distribution family then called conjugate distribution

* only when σ^2 is known

posterior : Gaussian $\sim N(\mu \mid ?, ?)$

prior : Gaussian $N(\mu \mid \mu_0, \sigma_0^2)$

likelihood : Gaussian $N(D \mid \mu, \sigma^2)$ ← assume fixed. or fit it to another gaussian

$D = \{180, 183, 160, 150, \ldots\}$

$$\mu_{MLE} = \frac{\sum D}{N}$$

$$P(\mu \mid X) = \frac{P(X \mid \mu_0) \cdot P(\mu \mid \mu_0, \sigma_0^2)}{P(X)}$$

likelihood.

$$= \int P(X \mid \mu) \cdot P(\mu \mid \mu_0, \sigma_0^2)\, d\mu.$$

$$\prod_{k=1}^{N} P(X_k \mid \mu, \sigma) \cdot P(\mu \mid \mu_0, \sigma_0^2)$$

prior.

$$= \left(e^{\frac{-(x_1-\mu)^2}{2\sigma^2}} \cdot e^{\frac{-(x_2-\mu)^2}{2\sigma^2}} \cdots \right) \cdot e^{\frac{-1}{2\sigma^2}(\mu-\mu_0)^2} \cdot \text{const}$$

$$= e^{\sum_{k=1}^{n} \frac{-1}{2\sigma^2}(x_k-\mu)^2} \cdot e^{\frac{-1}{2\sigma_0^2}(\mu-\mu_0)^2} \cdot \text{const.}$$

$$= e^{\frac{-1}{2\sigma^2}\sum(x_k^2 - 2x_k\mu + \mu^2) + \frac{-1}{2\sigma_0^2}(\mu^2 - 2\mu\mu_0 + \mu_0^2)} \cdot \text{const.}$$

pf: it is also a Gaussian distribution. ($e^{\frac{-1}{2\sigma^2}(x-\mu)^2}$)

$\Rightarrow N(\mu)$. quadratic completely the square

$$\frac{-1}{2\sigma^2}\sum(x_k^2 - 2x_k\mu + \mu^2) + \frac{-1}{2\sigma_0^2}(\mu^2 - 2\mu\mu_0 + \mu_0^2)$$

$$= \frac{-1}{2\sigma^2}N\mu^2 + \frac{-1}{2\sigma_0^2}\mu^2 + \frac{1}{2\sigma^2}\sum x_i \cdot 2\mu + \frac{1}{2\sigma_0^2}2\mu\mu_0 + \text{const.}$$

$$= -\frac{1}{2}\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)\left(\mu^2 - 2\frac{\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum x_k}{\sigma^2}\right)}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}\mu + m^2\right) + C$$

$$\frac{-1}{2\sigma_n^2}(\mu - m)^2$$

$$\Rightarrow \quad \sigma_n^2 = \frac{1}{\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)}$$

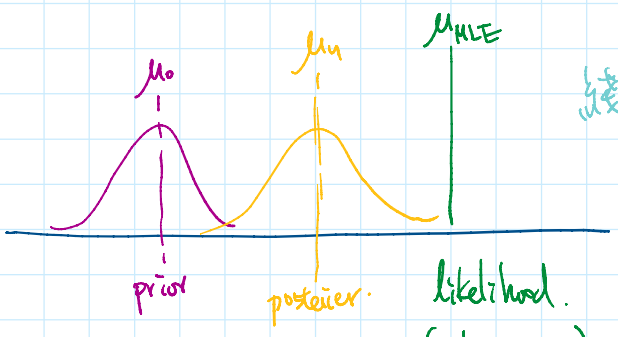
$$\mu_n = m = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{\sum x_k}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}$$

$$P(\mu|D) \sim N(\mu_n, \sigma_n^2)$$

$\mu_{MLE} = \frac{\sum x_k}{N}$

- $\mu_n = \frac{\sigma_n^2}{\sigma_0^2}\mu_0 + \frac{\sigma_n^2}{\sigma^2} \cdot N \cdot \mu_{MLE}$

$$* \quad \frac{\sigma_n^2}{\sigma_0^2} + \frac{\sigma_n^2}{\sigma^2} n = \sigma_n^2 \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right) = \sigma_n^2 \cdot \frac{1}{\sigma_n^2} = 1$$



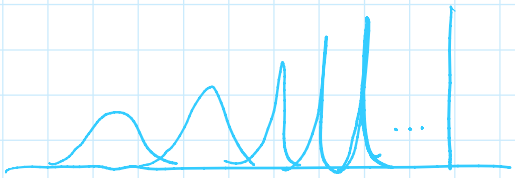
線性組合.

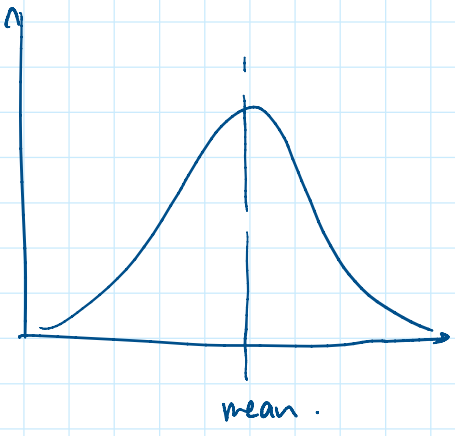
- $(\sigma_n^2)^{-1} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$

$n = 0 \Rightarrow \sigma_n = \sigma_0$ initial value. is prior σ

$n \to \infty$ σ_n^2 determined by σ^2

$\sigma_n^2 \to 0$





- Symetric
- Unimodal
- localization

} distance metric

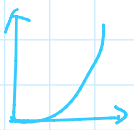
distance ↑ probability ↓

§ Central limit Theorm. 中央極限定理

X — what ever it is like.

random variable.

eg. e^x



its sample mean or sum $\rightarrow y \sim N(\mu, \sigma^2)$

$$[X_{11}\ X_{12}\ \cdots\ X_{1K}] \rightarrow y_1$$
$$[X_{21}\ X_{22}\ \cdots\ X_{2K}] \rightarrow y_2$$
$$\vdots \qquad\qquad \vdots$$

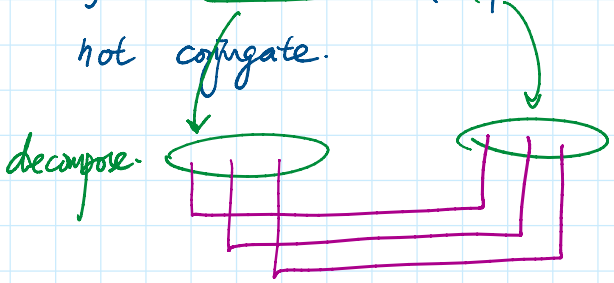
garantee to be Gaussian

- Moment Generating function

$$y = C_1 E(X) + C_2 E(X-\mu)^2 + \cdots$$

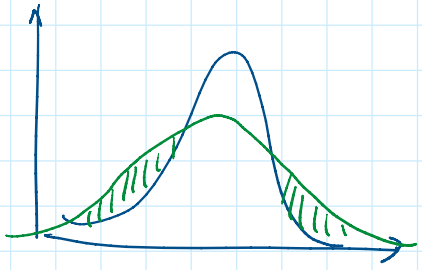
Any variable can be decompose to its moment.

application: if likelihood & prior not conjugate.



Combine the same order moment.

§ Student T distribution



$N(x|\mu, \sigma^2)$

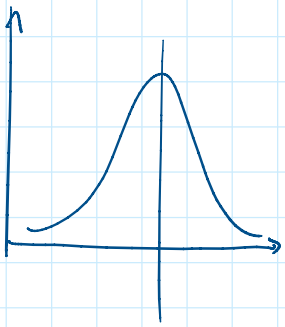
Assume μ is known

σ^2 is unknown.

marginal distribution.

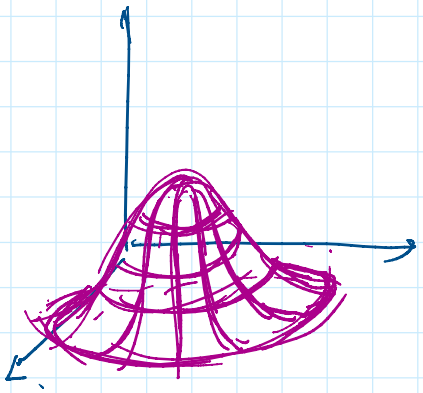
$$T = \int N(x \mid \mu, \sigma^2) \cdot \Gamma(\sigma^2 \mid a, b)\, d\, a, b$$

* more tolerence to high variance.
* Better for { sample size small / variance high.



uni variant Gaussian (単変数)

$$N : \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



K-dim Gaussian.
(multivariant)

$$\frac{1}{2\pi^{\frac{k}{2}}} \cdot \frac{1}{|\Sigma|^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}(x-\mu)^T \cdot \Sigma^{-1} (x-\mu)}$$

$$x : \begin{bmatrix} \\ \\ \end{bmatrix}_{k\times 1} \quad \mu : \begin{bmatrix} \\ \\ \end{bmatrix}_{k\times 1} \quad \Sigma : \begin{bmatrix} & \\ & \end{bmatrix}_{k\times k}$$

Σ : covariance matrix

$$\begin{array}{c|cccc} & x_1 & x_2 & \cdots & x_k \\ x_1 & \sigma_{11} & & & \\ x_2 & & \sigma_{22} & & \\ \vdots & & & \ddots & \\ x_k & & & & \sigma_{kk} \end{array}$$

variance

$$\sigma_{ii} = \frac{\sum (x_i - \mu_i)^2}{n}$$

$$\sigma_{ij} = \frac{\sum (x_i - \mu_i)(x_j - \mu_j)}{N}$$

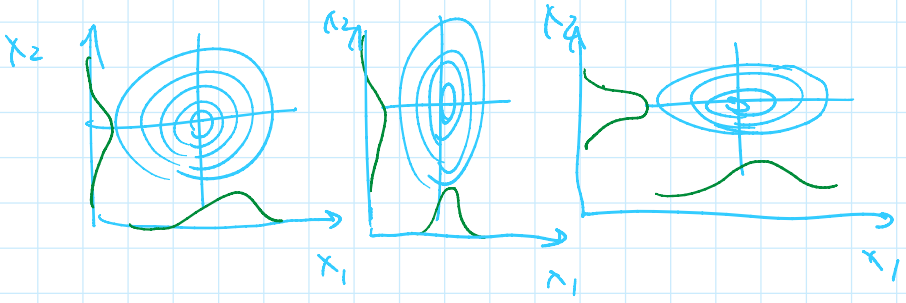
⇒ covariance

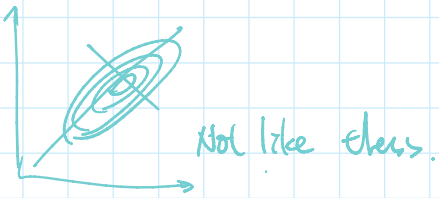
* Covariance show how the x_i change while x_j change.

* $\Lambda = \Sigma^{-1}$ percision matrix

- Orthogonal contour.

$|\Sigma| = \begin{bmatrix} \sigma_{11} & & 0 \\ & \sigma_{22} & \\ 0 & & \sigma_{kk} \end{bmatrix}$ diagnol.





$\Rightarrow \Sigma = \begin{bmatrix} \sigma_{11} & & \text{Non-0} \\ & \sigma_{22} & \\ \text{Non-0} & & \sigma_{kk} \end{bmatrix}$

- if Σ is diagnoal.

$\frac{1}{2\pi^{\frac{k}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}} \cdot e^{\frac{-1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \propto$ Euclidean distance.

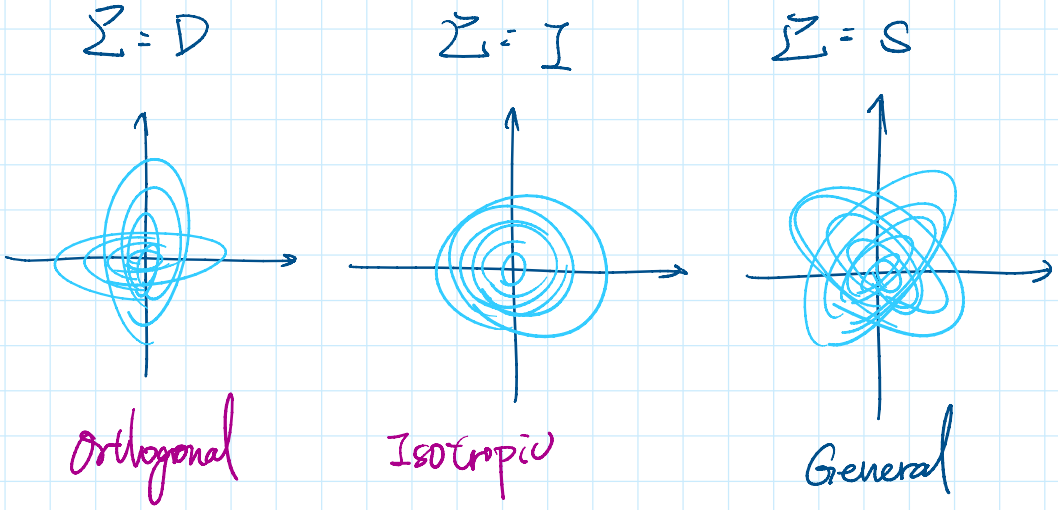
歐幾里德距離 $\sqrt{\Delta x^2 + \Delta y^2}$

$\Rightarrow \sum (x-\mu)^2$

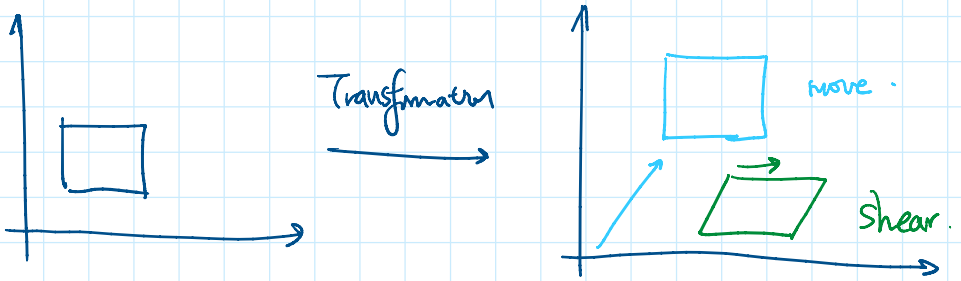
- if Σ is not diagnol.

$(x-\mu)^T \Sigma^{-1} (x-\mu)$ mahalanobis distance

馬氏距離.



§ Affine property.



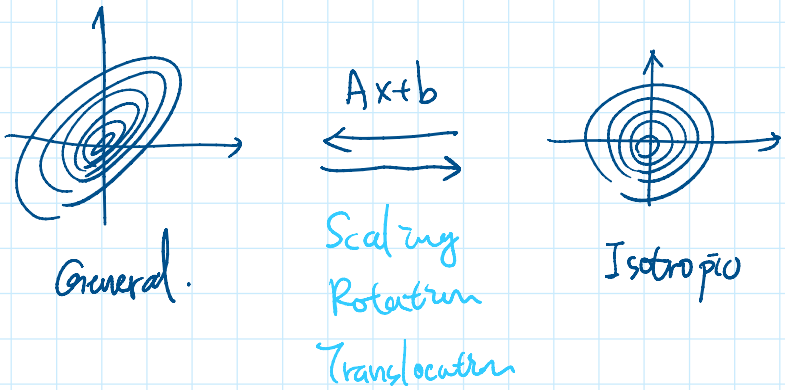
* If you do affine transform to Gaussian it is still Gaussian.

$X = N(\mu, C)$

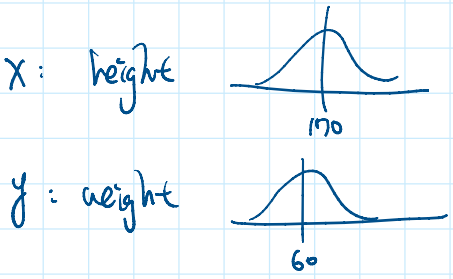
$y = AX + b \sim N(A\mu + b, ACA^T)$

$X \overset{i.i.d}{\sim} N(0,1)$ standard normal

$N(\vec{0}, I) \Leftrightarrow Ax + \mu \sim N(\mu, C)$



§ Linear transformation.



- Sum of 2 independent Gaussian ⇒ Gaussian

$$Y = X_1 + X_2 \Rightarrow E[Y] = \mu_1 + \mu_2$$

$$Var[Y] = \sigma_1^2 + \sigma_2^2$$

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \quad Y = X_1 + X_2 = AX + b$$

$$\Rightarrow A = [1 \;\; 1] \quad b = \vec{0}$$

X_1, X_2 are independent.

$$\Rightarrow E(X) = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad Var(X) = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

$$Y \sim N(A\mu + b, \; AcA^T)$$

$$E(Y) = [1 \;\; 1] \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + 0 = \mu_1 + \mu_2$$

$$Var(Y) = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \sigma_1^2 + \sigma_2^2$$

- Linear Combination.

$$Y = \sum b_i X_i \Rightarrow E[Y] = \sum b_i \mu_i$$

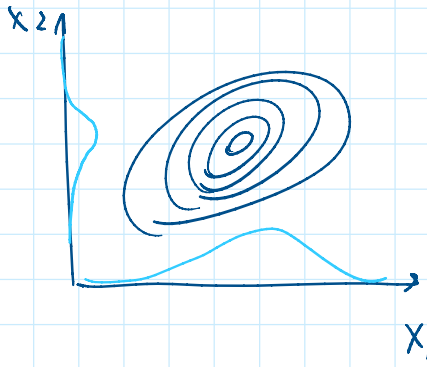
$$Var(Y) = \sum b_i^2 \sigma_i^2$$

- Linear transformation

$$Y = \sum BX \Rightarrow E[Y] = \sum B\mu$$

$$Var[Y] = \sum B^T C B$$

§ Marginal Gaussian



$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, C\right)$$

$$\downarrow$$

$$\begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

$$A = \begin{bmatrix} 1 & 0 \end{bmatrix} \quad b = 0$$

$$AX + b = X_1 \sim N(A\mu + b, ACA^T)$$

from Affine property.

$$A\mu + b = \mu_1 \quad ACA^T = \sigma_1^2$$

$$\Rightarrow X_1 \sim N(\mu_1, \sigma_1^2)$$

$$X = \begin{bmatrix} X_a \\ \hline X_b \end{bmatrix} \quad X_a = \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix}$$

$$X_b = \begin{bmatrix} X_{k+1} \\ \vdots \\ X_n \end{bmatrix}$$

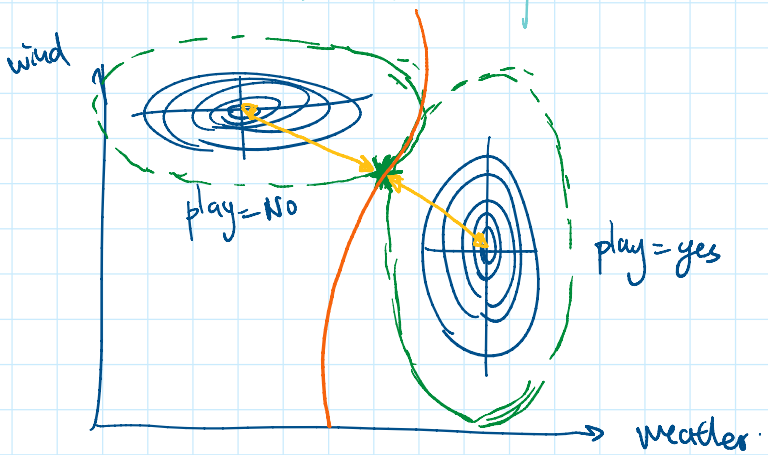
$$\mu = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix} \quad C = \begin{bmatrix} C_{aa} & C_{ab} \\ C_{ba} & C_{bb} \end{bmatrix} \quad C_{aa} = \begin{bmatrix} C_{11} & \cdots & C_{1k} \\ \vdots & \ddots & \vdots \\ C_{k1} & \cdots & C_{kk} \end{bmatrix}$$

$$A = \begin{bmatrix} 1 & & 0 & & \\ & 1 & & & 0 \\ & & \ddots & & \\ 0 & & & 1 & \end{bmatrix} \quad (k \text{ columns}, n-k \text{ columns}) \qquad y = Ax + b = \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix} = x_a$$

$x_a \sim N(\mu_a, C_{aa})$ from Affine Property.

§ Naive Bayes classifier.

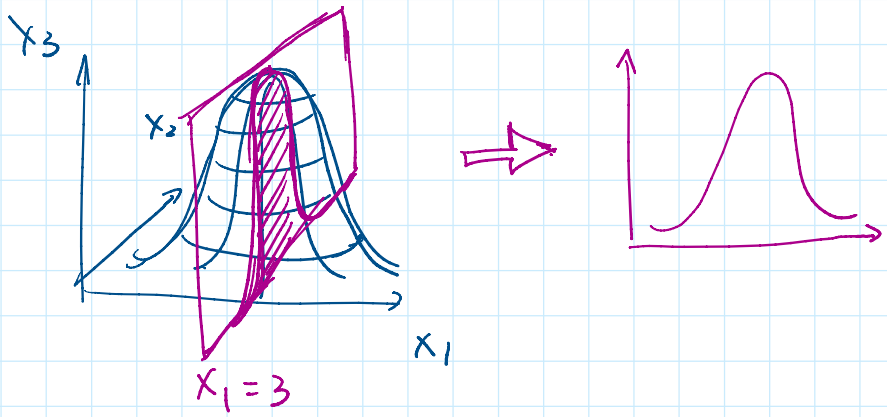
↳ conditional independent.



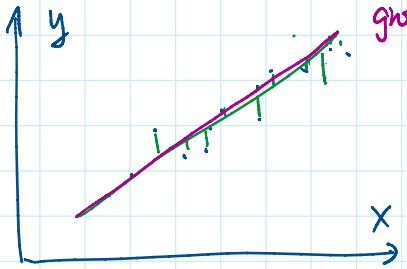
0.5 decision boundary.

distance to the center of both Y/N are the same. (probability is same)

§ Conditional Gaussian.



§ Probabilistic view of linear regression



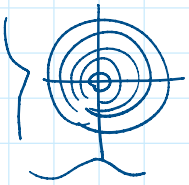
ground truth (model.

$LSE = \| Ax - b \|^2$

↓ line (Ax) — target (b)

↓ MLE

↓ MAP → Fully Bayesian.



random error $\sim N$

⇒ why the points not on line.

minimal error ⇔ maximum probability (likelihood)

$b = Ax + \varepsilon$ (design matrix)

target — Line — error.

$x \xrightarrow{\text{input space}} \Phi \xrightarrow{\text{feature space}} Ax+b \xrightarrow{} \text{output}$

likelihood :

line

$\sim N(y \mid w^T \Phi(x), \sigma^2)$

$\prod_{\text{all } x} N(w^T \Phi(x_n), \sigma^2)$

$\Rightarrow \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_k e^{\frac{1}{2}\left(\frac{y_k - w^T\Phi(x_k)}{\sigma}\right)}$ ⟸ $\begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \\ x_n & y_n \end{bmatrix}$

$\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_k \end{bmatrix}$

MLE : w.r.t w

log transform ⇒ $J = \frac{N}{2}\log 2\pi\sigma^2 + \sum \frac{1}{2\sigma^2}(y - w^T x)^2$

$\| b - Ax \|^2$

same as LSE

$\frac{\partial J}{\partial w} = 0$

Frequentist : MLE = LSE

probability — deterministic.

Bayesian : prior:

$$w \sim N(\vec{0}, b^{-1} I)$$

$$\begin{bmatrix} \sigma^2 & & 0 \\ & \sigma^2 & \\ & \ddots & \\ 0 & & \sigma^2 \end{bmatrix} = \Sigma = \Lambda^{-1}$$

posterior:

$$P(w|D) \propto P(D|w) P(w)$$

likelihood (above)

$$e^{\frac{-1}{2}(x-\mu)^T \Lambda (x-\mu)}$$

$$= e^{\frac{-b}{2} w^T w} = e^{\frac{-b}{2}} \cdot e^{w^T w}$$

$$\log P(w|D) = \| \vec{y} - w^T \vec{x} \|^2 + \bigcirc\, w^T w$$

same as $\| Ax - b \|^2 + \lambda x^2$ from of regularized LSE.

- LSE → MLE
  RLSE → MAP

let $\sigma^{-1} = a$ (covariance^{-1} = precision)

$$\Rightarrow J = e^{\frac{-a}{2}(w^T x - y)^T (w^T x - y) - \frac{b}{2} w^T w}$$

$$\Rightarrow \log J = \frac{-a}{2}(w^T x - y)^T (w^T x - y) - \frac{b}{2} w^T w$$

$w_0 + w_1 x^1 + \cdots + w_k x^k$

w : weight $\begin{bmatrix} w_0 \\ \vdots \\ w_k \end{bmatrix}$ $k \times 1$

$A = \Phi$ = design Matrix $\begin{bmatrix} x_0 \cdots x_0^k \\ \vdots \\ x_n \cdots x_n^k \end{bmatrix}$ $n \times k$

y : target $\begin{bmatrix} y_0 \\ \vdots \\ y_n \end{bmatrix}$ $n \times 1$

rewrite $w^T x$ as Aw (Both represent the line)

$$\Rightarrow \log J : \frac{-a}{2}(w^T A^T A w - w^T A^T y - y^T A w + y^T y) - \frac{b}{2} w^T w$$

$$= \frac{-a}{2}(w^T A^T A w - 2 w^T A^T y + y^T y) - \frac{b}{2} w^T w$$

$w^T A^T y$
$1 \times k \times k \times n \times n \times 1 = 1 \times 1$

$y^T A w$
$1 \times n \times n \times k \times k \times 1 = 1 \times 1$

want to turn into the form $(x-\mu)^T \Lambda (x-\mu)$

$$= x^T \Lambda x \quad - 2 x^T \Lambda \mu + \bigcirc$$

$$= \frac{-1}{2}(a w^T A^T A w + b w^T w - 2a w^T A^T y + \bigcirc)$$

$$= \frac{-1}{2}(w^T (a A^T A + b I) w - 2 (a \Lambda^{-1} A^T y) w^T + \bigcirc)$$

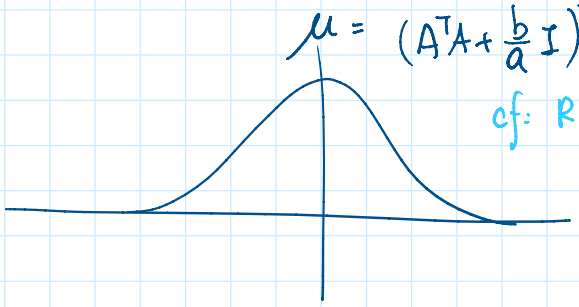
$$\Rightarrow P(w|D) \sim N\left(a\Lambda^{-1}A^T y,\ (aA^TA+bI)^{-1}\right)$$

MAP $\sim N\left(a(aA^TA+bI)^{-1}A^T y,\ (aA^TA+bI)^{-1}\right)$

$\sim N\left(\left(A^TA+\frac{b}{a}I\right)^{-1}A^T y,\ (aA^TA+bI)^{-1}\right)$

$\mu = \left(A^TA+\frac{b}{a}I\right)^{-1}A^T y$

cf: RLSE: $(A^TA+\lambda I)^{-1}A^T b$



§ Bayesian Linear Regression.

◉ Fully Bayesian

want to find.

MLE → LSE → likelihood.

MAP → RLSE → posterior.

Fully Bayesian → predictive distribution

(no $\vec{w}$, Φ)

# of basis → ∞

Kernal method.

* every $\vec{w}$ is a distribution

▥ Marginalized w

$$P(y|D) = \int P(y|w,D)\cdot P(w|D)\,dw$$

$$\int N(y|w^Tx,\sigma^2)\cdot N(w|\mu,b^{-1}I)\,dw$$

Λ^{-1}

univariant gaussian.

$$= \int \mathcal{N} \; e^{\frac{a}{2}(y-w^T x)^2} \cdot e^{\frac{-1}{2}(w-\mu)^T \Lambda (w-\mu)} \, dw$$

$$= \int \mathcal{N} \; e^{\frac{-1}{2}(ay^2 - 2w^T x y a + a w^T x x^T w + w^T \Lambda w - 2w^T \Lambda \mu + \mu^T \Lambda \mu)} \, dw$$

* Marginal is still Gaussian $\left(x^T \Lambda x - 2 x^T \Lambda \mu + \mu^T \Lambda \mu \right)$

$$= \int \mathcal{N} \; e^{\frac{-1}{2}\left(w^T (a x x^T + \Lambda) w - 2 w^T (x y a + \Lambda \mu) + a y^2 + \mu^T \Lambda \mu \right)} \, dw$$

$$\Rightarrow C = a x x^T + \Lambda \;,\; \mu' = C^{-1}(x y a + \Lambda \mu)$$

$$= \mathcal{N} \int e^{\frac{-1}{2}\left((w^T C w - 2 w^T C \mu' + \mu'^T C \mu') - \mu'^T C \mu' + a y^2 - \mu^T \Lambda \mu \right)} \, dw$$

$$= \mathcal{N} \int e^{\frac{-1}{2}(w-\mu')^T C (w-\mu')} \cdot e^{\frac{-1}{2}\mu'^T C \mu' - \frac{1}{2} a y^2 - \frac{1}{2}\mu^T \Lambda \mu} \, dw$$

$$= \mathcal{N} \; e^{\frac{-1}{2}\mu'^T C \mu' - \frac{1}{2} a y^2 - \frac{1}{2}\mu^T \Lambda \mu} \int e^{\frac{-1}{2}(w-\mu')^T C (w-\mu')} \, dw$$

it's a gaussian

integral all w over gaussian = 1

$$= \mathcal{N} \; e^{\frac{1}{2}(a y^2 - \mu'^T C \mu' + \mu^T \Lambda \mu)}$$

$$\left[C^{-1}(x y a + \Lambda \mu) \right]^T \cdot C \cdot \left[C^{-1}(x y a + \Lambda \mu) \right]$$

$$= (x y a + \Lambda \mu)^T C^{-1T} C \cdot C^{-1} \cdot (x y a + \Lambda \mu)$$

$$= (x y a + \Lambda \mu)^T \cdot C^{-1T} \cdot (x y a + \Lambda \mu)$$

$$= (a^2 x^T C^{-1} x) y^2 + 2 (a x^T C^{-1} \Lambda \mu) y + \mu^T \Lambda^T C^{-1} \Lambda \mu$$

$$= \exp\left\{ \frac{1}{2} \left((a - a^2 x^T C^{-1} x) y^2 - 2 (a x^T C^{-1} \Lambda \mu) y + (\mu^T \Lambda \mu - \mu^T \Lambda^T C^{-1} \Lambda \mu) \right) \right\}$$

let $\lambda = a - a^2 x^T C^{-1} x$ → $C = aXX^T + \Lambda$

$m = \frac{1}{\lambda}(aX^T C^{-1}\Lambda\mu)$

$P(y|D) = N(m, \lambda) = N\left(\frac{1}{\lambda}aX^T C^{-1}\Lambda\mu,\ a - a^2 x^T C^{-1} x\right)$

§ Sherman Morrison Formula

$$C^{-1} = \Lambda^{-1} - \frac{\Lambda^{-1}a\Phi\Phi^T\Lambda^{-1}}{1+a\Phi^T\Lambda^{-1}\Phi}$$

$$(\Lambda + a\Phi\Phi^T)\left(\Lambda^{-1} - \frac{\Lambda^{-1}a\Phi\Phi^T\Lambda^{-1}}{1+a\Phi^T\Lambda^{-1}\Phi}\right)$$

$$= I + a\Phi\Phi^T\Lambda^{-1} - \frac{a\Phi\Phi^T\cdot\Lambda^{-1}a\Phi\Phi^T\Lambda^{-1}}{1+a\Phi^T\Lambda^{-1}\Phi} - \frac{1\Lambda^{-1}a\Phi\Phi^T\Lambda^{-1}}{1+a\Phi^T\Lambda^{-1}\Phi}$$

$$= I + a\Phi\Phi^T\Lambda^{-1} - \frac{a\Phi(\Phi^T\Lambda^{-1}a\Phi+I)\Phi^T\Lambda^{-1}}{1+a\Phi^T\Lambda^{-1}\Phi}$$

$$= I + a\Phi\Phi^T\Lambda^{-1} - a\Phi\Phi^T\Lambda^{-1} = I$$

$\lambda = a - a^2\Phi^T C^{-1}\Phi$

$= a - a^2\cdot\Phi^T\left(\Lambda^{-1} - \frac{\Lambda^{-1}a\Phi\Phi^T\Lambda^{-1}}{1+a\Phi^T\Lambda^{-1}\Phi}\right)\Phi$

let $\alpha = \Phi^T\Lambda^{-1}\Phi$ $\quad a\cdot\Phi\cdot\alpha\cdot\Phi^T = a\cdot\alpha$

$= a - a^2\cdot\Phi^T\left(\Lambda^{-1} - \frac{\Lambda^{-1}a\alpha}{1+a\alpha}\right)\Phi$

$= a - a^2\left(\Phi^T\Lambda^{-1}\Phi - \frac{\Phi^T\Lambda^{-1}a\alpha\Phi}{1+a\alpha}\right)$

$= a - a^2\left(\alpha - \frac{a\alpha^2}{1+a\alpha}\right) = a - a^2\left(\frac{\alpha+a\alpha^2-a\alpha^2}{1+a\alpha}\right)$

$= a - \frac{a^2\alpha}{1+a\alpha} = \frac{a+a^2\alpha-a^2\alpha}{1+a\alpha}$

$= \frac{a}{1+a\alpha}, \quad \alpha = \Phi^T\Lambda^{-1}\Phi$

$\sigma^2 = \frac{1}{\lambda} = \frac{1}{a} + \alpha$

$m = \frac{1}{\lambda}a\Phi^T C^{-1}\Lambda\mu$

$m = m^T = \left[\frac{1}{\lambda}a\Phi^T C^{-1}\Lambda\mu\right]^T$ * m is scalar

$$= \mu^T \left(\frac{1}{\lambda} a \lambda^T C^{-1^T} \Phi \right) \quad * \ \Lambda \cdot C \text{ is symmetric}$$

$$= \mu^T \left(\frac{1+a\alpha}{a} \cdot a \cdot \Lambda \cdot \left(\Lambda^{-1} - \frac{\Lambda^T a \alpha}{1+a\alpha} \right) \cdot \Phi \right)$$

$$= \mu^T (1+a\alpha) \cdot \left(I - \frac{a\alpha}{1+a\alpha} \right) \cdot \Phi$$

$$= \mu^T \left(I - \frac{a\alpha}{1+a\alpha} + a\alpha - \frac{a^2\alpha^2}{1+a\alpha} \right) \cdot \Phi$$

$$= \mu^T \left(I + a\alpha - \frac{a\alpha(1+a\alpha)}{1+a\alpha} \right) \cdot \Phi$$

$$= \mu^T (I + a\alpha - a\alpha) \cdot \Phi$$

$$= \mu^T \Phi$$

$$\Rightarrow N \left(\mu^T \Phi, \ \frac{1}{a} + \Phi^T \Lambda^{-1} \Phi \right)$$

μ: mean of prior

Fully Bayesian: Maximum every possible of a line.

Kernel $\Phi \Phi^T$

§ Decision Theory.

- Regression → classification
  (*Logistic regression)
- EM.
- clustering.
  ⇒ non-supervising. (No label)
- Dimension Reduction.

supervising learning
+ every data has label

Estimator. (bias/variance)

A statistic to approximate the property of a distribution.

Eg: $\hat{\mu}_{MLE} = \frac{1}{n}\sum x_i$

$D = [x_1, \dots x_n] \overset{iid.}{\sim} N(\mu, \sigma^2)$

$\mu_{MLE} = \hat{\mu} = \bar{x} = \frac{1}{n}\sum x_i$

$\sigma^2_{MLE} = \hat{\sigma} = \frac{1}{n}\sum (x_i - \bar{x})^2$ ⇒ bias

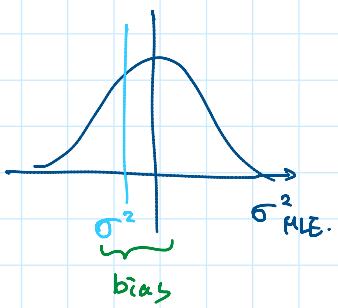
$\sigma^2_{Alt} = \frac{1}{n-1}\sum (x_i - \bar{x})^2$ * ⇒ unbias

- Bias

  Mean of estimator diffacult to real.

  $E(\hat{\theta}) - \theta$

  Eg: $E(\sigma^2_{MLE}) - \sigma^2_{MLE}$



▲ $E(\hat{\mu}) = E\left(\frac{1}{n}\sum X_i\right)$

$= \frac{1}{n}\sum \boxed{E(X_i)} \to \mu.$

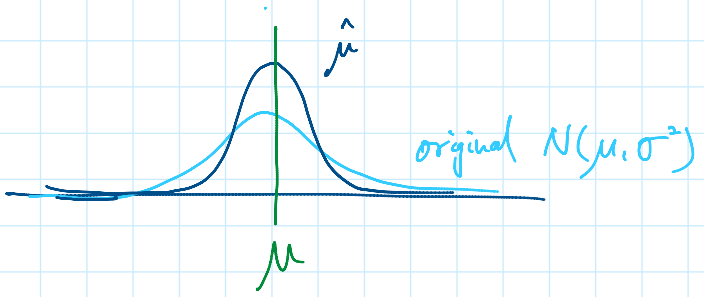
$= \frac{1}{n}\cdot n\mu = \mu.$

$E(\hat{\mu}) - \mu = 0 \Rightarrow$ No bias.

$Var(\hat{\mu}) = Var\left(\frac{X_1 + X_2 \cdots X_n}{n}\right)$

$= \frac{1}{n^2} Var(X_1 \cdots X_n)$

$= \frac{1}{n^2}(\sigma^2 \cdot n) = \frac{\sigma^2}{n}$



* $n \to \infty$ $Var(\hat{\mu}) \longrightarrow 0$

$\Rightarrow \hat{\mu}$ is a single number.

▲ $E(\sigma^2_{MLE})$

$= E\left(\frac{1}{n}\sum (X_i - \hat{\mu})^2\right)$ dummy variable.

$= E\left(\frac{1}{n}\sum \left((X_i - \mu) - (\hat{\mu} - \mu)\right)^2\right)$

$= E\left(\frac{1}{n}\sum \left((X_i - \mu)^2 - 2(X_i - \mu)(\hat{\mu} - \mu) + (\hat{\mu} - \mu)^2\right)\right)$

$= E\left(\frac{1}{n}\sum (X_i - \mu)^2 - \frac{2}{n}\sum (X_i - \mu)(\hat{\mu} - \mu) + \frac{1}{n}\sum (\hat{\mu} - \mu)^2\right)$

$= E\left(\frac{1}{n}\sum (X_i - \mu)^2\right) - E\left(\frac{2(\hat{\mu} - \mu)}{n}\sum (X_i - \mu)\right) + E\left(\frac{n}{n}(\hat{\mu} - \mu)^2\right)$

$= \frac{1}{n}E\left((X_1 - \mu)^2 + (X_2 - \mu)^2 + \cdots\right) - E\left(2(\hat{\mu} - \mu)\cdot\frac{1}{n}\left(\sum X_i - \sum \mu\right)\right) + E(\hat{\mu} - \mu)^2$

$= \frac{1}{n}\left(Var(X_1) + Var(X_2) + \cdots\right) - E\left(2(\hat{\mu} - \mu)\cdot\left(\frac{\sum X_i}{n} - \frac{n}{n}\mu\right)\right) + E((\hat{\mu} - \mu)^2)$

$= \frac{1}{n}\cdot n \cdot Var(X_i) - E\left(2(\hat{\mu} - \mu)(\hat{\mu} - \mu)\right) + E((\hat{\mu} - \mu)^2)$

$= \sigma^2 - 2E(\hat{\mu} - \mu)^2 + E(\hat{\mu} - \mu)^2$

$= \sigma^2 - \boxed{E(\hat{\mu} - \mu)^2} \to Var(\hat{\mu}) = \frac{\sigma^2}{n}$

$= \sigma^2 - \frac{\sigma^2}{n}$

$= \frac{n-1}{n}\sigma^2 \neq \sigma^2$ Bias !

- $E(\sigma_{MLE}^2)$

$= E\left(\frac{1}{n}\sum(x_i-\hat{\mu})^2\right)$

$= E\left(\frac{1}{n}\sum(x_i-\mu-(\hat{\mu}-\mu)^2\right)$

$= E\left(\frac{1}{n}\sum a^2\right) - E\left(\frac{2n}{n}b^2\right) + E\left(\frac{n}{n}b^2\right)$

$= E\left(\frac{1}{n}\sum a^2\right) - E\left(\frac{n}{n}b^2\right)$ $\frac{\sum a^2}{n} = \sigma^2$

$\frac{1}{n}\sum a^2 = \frac{n}{n}\sigma^2$

$\frac{n}{n}\cdot\frac{\sigma^2}{n} = \frac{\sigma^2}{n}$

$= \frac{n}{n}\sigma^2 - \frac{1}{n}\sigma^2 = \sigma^2$ No bias !

Mean Square Error (MSE)

$E((\hat{\theta}-\theta)^2)$ ← property of data (ground truth) ⇒ usually don't know

↑ estimator ⇒ random variable.

- MSE is error from ground truth
  LSE is error from datas

$bias = E(\hat{\theta})-\theta = E(\hat{\theta}-\theta)$

$MSE(\hat{\theta}) = bias^2(\hat{\theta}) + Var(\hat{\theta})$

$E((\hat{\theta}-\theta)^2)$

$= E(((\hat{\theta}-\mu)-(\theta-\mu))^2)$

$= E((\hat{\theta}-\mu)^2) - 2E((\hat{\theta}-\mu)(\theta-\mu)) + E((\theta-\mu)^2)$ (const.)

$2(\theta-\mu)E(\hat{\theta}-\mu)$

$= 2(\theta-\mu)(E(\hat{\theta})-\mu)$

* for unbias estimator

$E(\hat{\theta}) = \mu$

$= E((\hat{\theta}-\mu)^2) + E((\theta-\mu)^2)$

$= E((\hat{\theta}-\mu)^2) + (\theta-\mu)^2$

$= Var(\hat{\theta}) + (\theta - E(\hat{\theta}))^2$

$= Var(\hat{\theta}) + bias^2(\hat{\theta})$

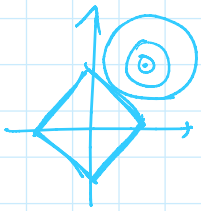
* if bias is reduced ↓
  ⇒ all error come from Variance.
  ⇒ trade-off

- Bias ∝ flexibility / Generalisity.
- Vars ∝ sensitivity to traning data

① Complexity of model

$\Phi : \phi_1(x) \dots \phi_K(x)$ ← feature.

⇒ ① dimension Reduction.
② Regularization - EX: L1 norm



* low bias ⇔ overfitting. (more complicated)
  — usually when.:
  - non-linear.
  - non parametric (no assumption)
  - k-nn.

* low variance ⇔ underfitting (simple)
  - Regression. (with regularization)

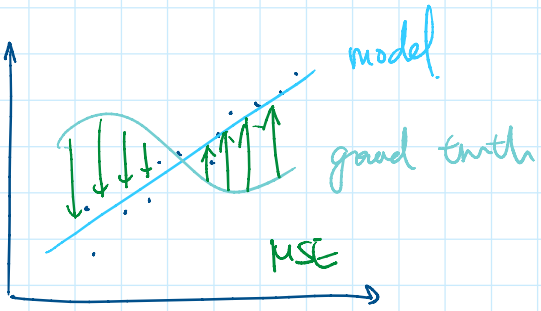
- Naive Bayes
- Linear model

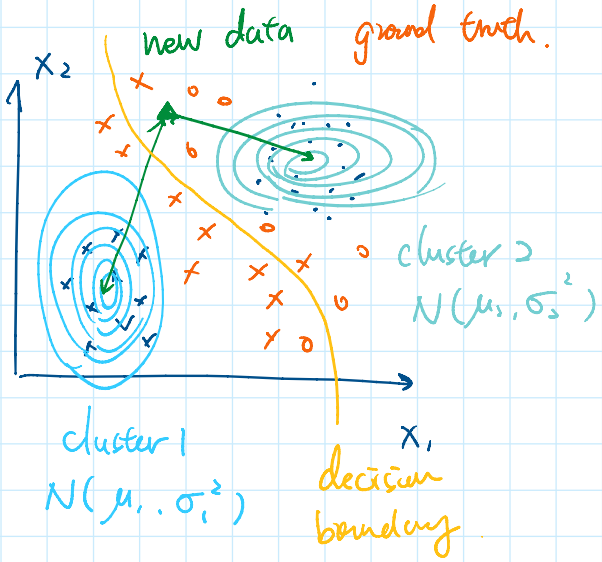
Bias - variance trade off

MSE

How much error that a model makes.

{ continous target
  discrete target





Contigerty table (confusion matrix)

ground truth \ prediction	$\theta = Y$	$\theta = N$	sum
θ = yes	True positive	False Negative	Yes
θ = No	False positive	True Negative	No
sum	positive	negative	total

$$\text{Specificity} = \frac{TP + TN}{\text{total}}$$

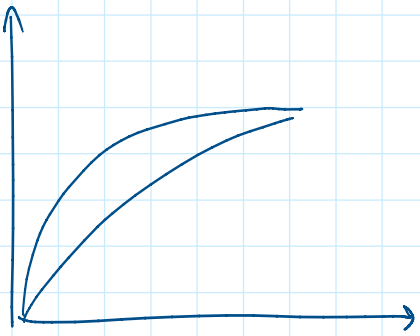
Sensitivity (Recall / Power) $= \frac{TN}{N_0} = \frac{TN}{FP+TN}$

Accuracy. =

False positive rate. =

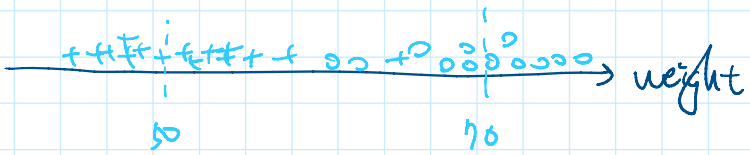
Precission (PPV) :

ROC (A.U.C)



§ Regression v.s. Classification.

Supervised Learning

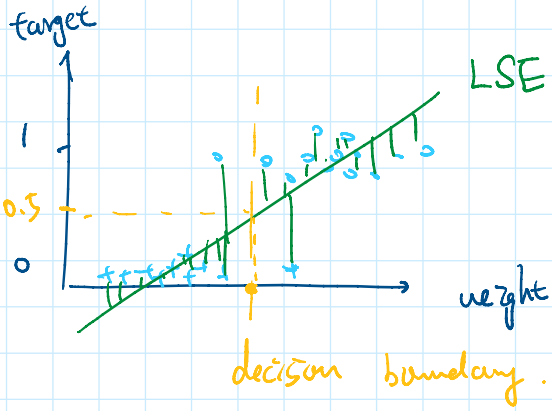


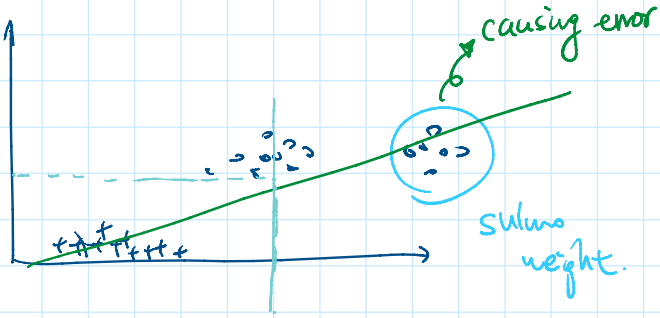
+ : female o : male.

let male : 1 ⇒ $f(0) = 1$

female : 0 $f(+) = 0$

indicator function.





To solve ⇒ using other loss function

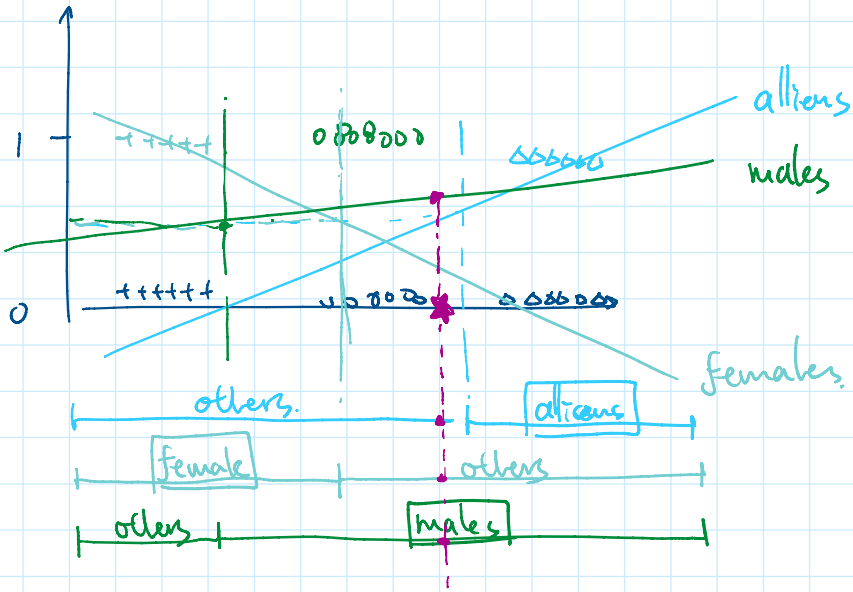
§ Loss Function.

- multi class

  * one of k coding.

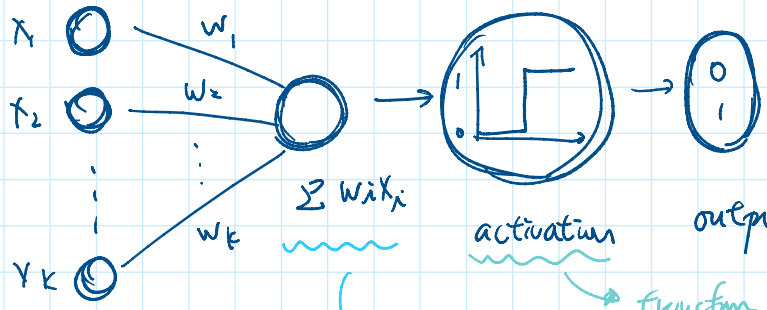
  - o male
  - + female
  - △ allien

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$



- Fisher's linear discriminat.

- Perceptron algorithm (simple neural network)

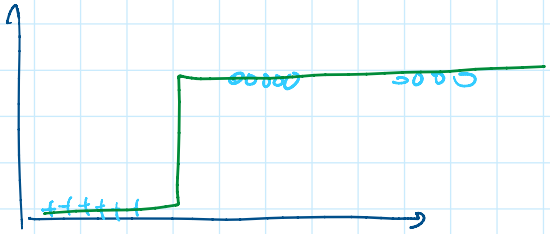


input neural

$w^T \Phi \rightarrow f(w^T \Phi)$ transform the line into another space.

* step function could tolerance extream value.

* But it's not a smooth function.



§ Gradient desent.

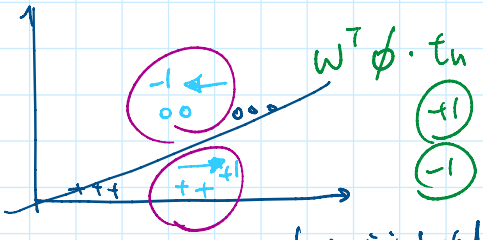
Newton's method (Taylor's expension) — 2nd order approximation

Gradient desent — 1st order.

$$X_{n+1} = x_n - \gamma \cdot \nabla f$$

γ: learning rate.

- Perceptron criterion
  - correct prediction gives no error
  - use misclassified data to update the line



epoch : mini batch

$X_0 = W_0 \rightarrow X_1 = W_1$

error made by the misclassified data $= w\phi t$

$$\frac{\partial}{\partial w} w\phi t = \phi t$$

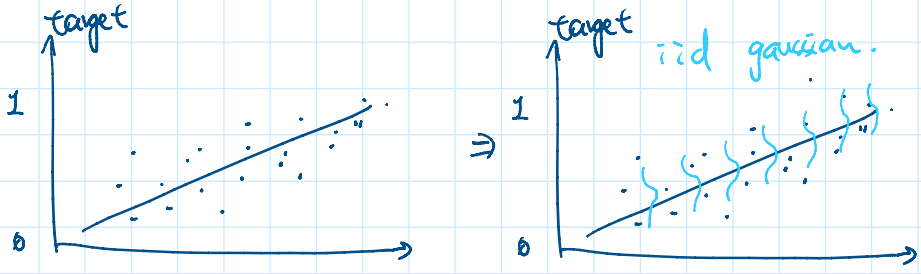
§ Logistic Regression

$$f(x) = \frac{1}{1+e^{-x}}$$

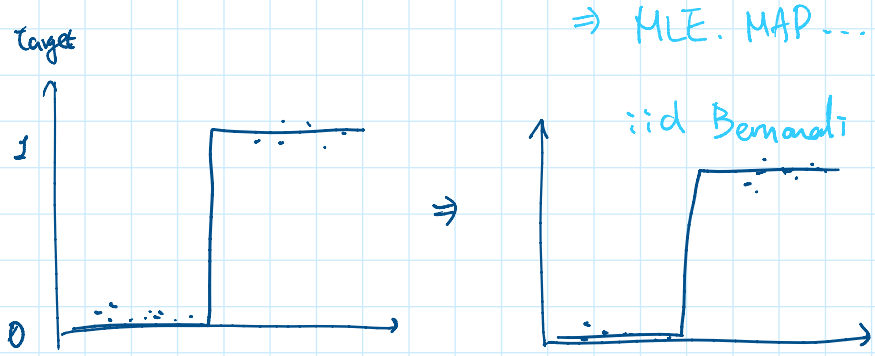
probit function

CDF of gaussian.

$$f(w^T\phi) = \frac{1}{1+e^{-w^T\phi}}$$



⇒ MLE. MAP ...



Given $D = (x_i, y_i)$ → $y_i \in \{0, 1\}$

$y_i \sim$ Bernulli $(f(w^T \phi))$ → Bernulli

Bernulli: $B(X|p) = p^X (1-p)^{1-X}$

MLE:

likelihood $:= \prod$ Bernulli$(y_i \mid f(w_i \phi))$

$$= \prod \left(\frac{1}{1+e^{-w^T x}} \right)^{y_i} \left(\frac{e^{-w^T x}}{1+e^{-w^T x}} \right)^{1-y_i}$$

$$-\log P(D|w)$$

$$= -\sum y_i \boxed{\log \frac{1}{1+e^{-w^T x_i}}} + (1-y_i) \log \frac{e^{-w^T x_i}}{1+e^{-w^T x_i}}$$

$$\frac{\partial}{\partial w_j} -\log P(D|w)$$

$$\frac{\partial}{\partial w_j} \left(-\log 1 + e^{\overbrace{-w^T x_i}^{[w_0 \cdots w_k] \begin{bmatrix} x_0 \\ \vdots \\ x_k \end{bmatrix}}} \right)$$

$= w_0 x_0 + \cdots + w_k x_k$

let $u = 1 + e^v$, $v = -w^T x_i$

$$\frac{\partial}{\partial w_j} \left(\log \frac{1}{1+e^{-w^T x_i}} \right) = \frac{\partial (-\log u)}{\partial u} \cdot \frac{\partial u}{\partial v} \cdot \frac{\partial v}{\partial w_j}$$

$$= \frac{1}{1+e^{-w^T x_i}} \cdot e^{-w^T x_i} \cdot x_{ij} \quad \text{—①}$$

$$\frac{\partial}{\partial w_j} \left(\log \frac{e^{-w^T x_i}}{1+e^{-w^T x_i}} \right) = \frac{\partial}{\partial w_j} \left(\log e^{-w^T x_i} - \log 1 + e^{-w^T x_i} \right)$$

$$= -x_{ij} + x_{ij} \frac{e^{-w^T x_i}}{1+e^{-w^T x_i}}$$

$$= -x_{ij} \cdot \left(\frac{1}{1+e^{-w^T x_i}} \right) \quad \text{—②}$$

from ①, ②

$$\frac{\partial}{\partial w_j} -\ln P(D|w)$$

$$= -\sum \left[y_i x_{ij} \left(\frac{e^{-w^T x_i}}{1+e^{-w^T x_i}} \right) - ((-y_i) x_{ij} \left(\frac{1}{1+e^{-w^T x_i}} \right) \right]$$

$$= -\sum \left[y_i x_{ij} \left(1 - \frac{1}{1+e^{-w^T x_i}} \right) - ((-y_i) x_{ij} \left(\frac{1}{1+e^{-w^T x_i}} \right) \right]$$

$$= -\sum \left[y_i x_{ij} - \frac{y_i x_{ij}}{1+e^{w^T x_i}} - \frac{x_{ij}}{1+e^{w^T x_i}} + \frac{y_i x_{ij}}{1+e^{w^T x_i}} \right]$$

$$= \sum x_{ij} \left(\frac{1}{1+e^{w^T x_i}} - y_i \right) = 0$$

* Recall: Newtons Method

$$X_{n+1} = X_n - H^{-1} f(X_n) \cdot \nabla f(X_n)$$

$$\nabla f = \begin{bmatrix} \frac{\partial J}{\partial w_1} \\ \vdots \\ \frac{\partial J}{\partial w_k} \end{bmatrix} \quad \sum x_{ij} \left(\frac{1}{1+e^{w^T x_i}} - y_i \right)$$

$$\Phi = \begin{bmatrix} x_{11} \cdots x_{1d} \\ \vdots \\ x_{n1} \cdots x_{nd} \end{bmatrix}_{n \times d} \quad \phi_1 \cdots \phi_d$$

$$\Rightarrow \nabla f = \Phi^T \left(\frac{1}{1+e^{w^T x}} - y \right) \begin{bmatrix} y_0 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1}$$

$$H = \begin{bmatrix} \frac{\partial^2 J}{\partial w_1 \partial w_1} & \frac{\partial^2 J}{\partial w_1 \partial w_2} & \cdots \\ \frac{\partial^2 J}{\partial w_2 \partial w_1} & \frac{\partial^2 J}{\partial w_2 \partial w_2} & \\ \vdots & & \ddots \end{bmatrix}$$

$$\frac{\partial}{\partial w_k} \left[\sum x_{ij} \left(\frac{1}{1+e^{-w^T x_i}} - y_i \right) \right]$$

$$\frac{\partial}{\partial w_k}\left(1+e^{-w^T x_i}\right)^{-1}$$

$u = 1+e^{-w_k x_i}$, $v = w_k x_{ik}$.

$$= \frac{\partial u^{-1}}{\partial u}\cdot\frac{\partial u}{\partial v}\cdot\frac{\partial v}{\partial w_k}$$

$$= -u^{-2}\cdot\left(-e^{-w_k x_{ik}}\right)\cdot x_{ik}$$

$$= x_{ik}\,\frac{e^{-w_k\cdot x_{ik}}}{\left(1+e^{-w_k x_{ik}}\right)^2}$$

$$\Rightarrow \sum x_{ij}\cdot\left(x_{ik}\cdot\frac{e^{-w_k x_{ik}}}{\left(1+e^{-w_k x_{ik}}\right)^2}\right)$$

$A^T \cdot D \cdot A$

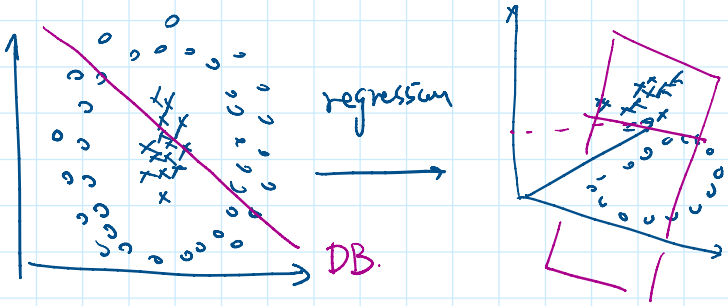
$$\begin{bmatrix} & \end{bmatrix}\begin{bmatrix} w_1 & 0 \\ & \ddots \\ 0 & w_n \end{bmatrix}\begin{bmatrix} \\ \\ \end{bmatrix}$$

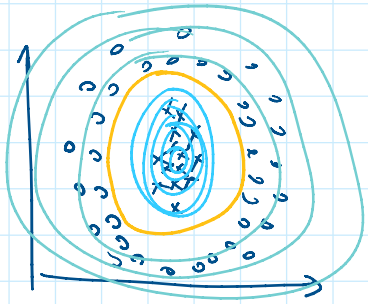
§ Knn : k - nearest neighbor

Knn : classification — [low bias ; high variant] ⇒ overfitting.

Kmeans : clustering.

§ Decision boundary.





Naive Bayes classifier

$N(\vec{\mu}_1, \Sigma_1) = N(\vec{\mu}_2, \Sigma_2)$

§ Clustering. (unsupervised learning)

Regression

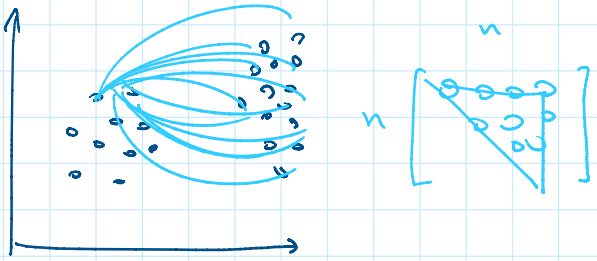
classification

supervised learning.

complete data

⇒ ~~target~~ labels known

- hierachical clustering. (knn)



§ EM algorithm.

(E) expectation

(M) maximization

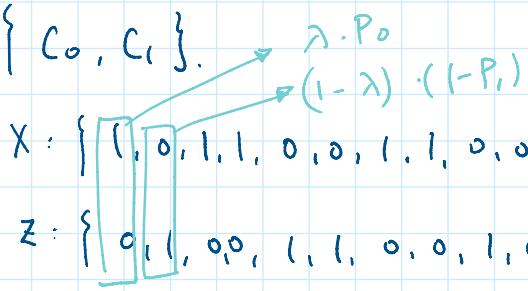
For complete data:

— 1 coin flipping $X: \{1,1,0,0,1,1\}$.

$$\theta_{MLE} = \frac{\sum x_i}{n} = \frac{4}{6} \quad B(\theta)$$

— 2 coins flipping.

$\{C_0, C_1\}$.



$X: \{1,0,1,1,0,0,1,1,0,0\}$.

$Z: \{0,1,0,0,1,1,0,0,1,0\}$ chance to pick C_0

likelihood: $p(X, Z \mid \theta)$, $\theta = P_0, P_1, \lambda$.

P_0: C_0 head; P_1: C_1 head

$$\prod \left[\lambda P_0^{x_i}(1-P_0)^{1-x_i}\right]^{1-z_i} \left[(1-\lambda)\cdot P_1^{x_i}(1-P_1)^{1-x_i}\right]^{z_i}$$

log likelyhood:

$$J = \sum\left[(1-z_i)\log\lambda + z_i\log(1-\lambda)\right] + \sum\left[(1-z_i)x_i\log P_0\right] + \dots$$

$$\frac{\partial J}{\partial \lambda} = 0 \Rightarrow \frac{1}{\lambda}\sum(1-z_i) = \frac{1}{1-\lambda}\sum z_i$$

$$\lambda_{MLE} = \frac{n - \sum z_i}{n} = \frac{10-4}{10} = \frac{6}{10}$$

$$\frac{\partial J}{\partial P_1} = 0 \Rightarrow \frac{1}{P_1}\sum z_i x_i = \frac{1}{1-P_1}\sum z_i(1-x_i)$$

$$P_1 = \frac{\sum z_i x_i}{\sum z_i} = \frac{0}{4}$$

$$\frac{\partial J}{\partial P_0} = 0 \Rightarrow \frac{1}{P_0}\sum(1-z_i)x_i = \frac{1}{1-P_0}\sum(1-z_i)(1-x_i)$$

$$P_0 = \frac{\sum x_i - \sum x_i z_i}{n - \sum z_i} = \frac{5}{6}$$

- Incomplete data.

$$X: \{1, 0, 1, 1, 0, 0, 1, 0, 0, 0\}$$

$$Z: \{ \text{missing} \}$$

EM algorithm * responsibility -

$$Z = \{0, 1\}$$

$$W = \{0.5, 0.5\}$$

approximate.

W ~ weight

$$P(z_i = C_0, x_i | \theta) = \lambda P_0^{x_i} (1-P_0)^{1-x_i}$$

$$P(z_i = C_1, x_i | \theta) = (1-\lambda) P_1^{x_i} (1-P_1)^{1-x_i}$$

Step 1 initialize all parameters.

$$\lambda^{(0)}, P_0^{(0)}, P_1^{(0)}$$

Step E calculate responsibility.

$$W_0 = \frac{P(z_i=0, x_i|\theta)}{P(z_i=0, x_i|\theta) + P(z_i=1, x_i|\theta)}$$

$$W_1 = \frac{P(z_i=1, x_i|\theta)}{P(z_i=0, x_i|\theta) + P(z_i=1, x_i|\theta)}$$

$\Rightarrow \quad X = \{1, 0, 1, 1, 0, 0, 1, 0, 0, 0\}$

$w_0 = \{ \cdots \cdots \}$

$[w_1 = \{ \underset{1-w_0}{\odot} \cdots \cdots \}]$

$\Rightarrow$ using w_0 instead of Z.

* w is a posterior.

$$w = \frac{P(z_i, x|\theta)}{\sum P(z_i, x|\theta)} = \frac{P(z, x|\theta)}{P(x|\theta)} = P(z|x, \theta)$$

step M update parameters.

$$J = \sum w_i \log\left(\lambda P_0^{x_i}(1-P_0)^{1-x_i}\right) + \sum (1-w_i) \log\left((1-\lambda) P_1^{x_i}(1-P_1)^{1-x_i}\right)$$

$$\frac{\partial J}{\partial \lambda} = 0 \Rightarrow \lambda_{MLE} = \frac{\sum w_i}{n}$$

$$\frac{\partial J}{\partial P_0} = 0 \Rightarrow P_{0\,MLE} = \frac{\sum w_i x_i}{\sum w_i}$$

$$\frac{\partial J}{\partial P_1} = 0 \Rightarrow P_{1\,MLE} = \frac{\sum x_i - \sum x_i w_i}{n - \sum w_i}$$

use these updated parameters to do the E step. repeat until converge.

$X = \{1, 0, 1, 1, 0, 0, 1, 0, 0, 0\}$

step1 : $\lambda^{(0)} = \frac{1}{3}$, $P_0^{(0)} = \frac{1}{3}$, $P_1^{(0)} = \frac{2}{3}$

$$P(x_i, z_i|\theta) = \left[\lambda P_0^{x_i}(1-P_0)^{1-x_i}\right]^{1-z_i}\left[(1-\lambda) P_1^{x_i}(1-P_1)^{1-x_i}\right]^{z_i}$$

E step: $P(z=0, \underset{\sim}{x}=1 \mid \theta) = \frac{1}{3}\times\frac{1}{3} = \frac{1}{9}$

$P(z=0, x_i=0 \mid \theta) = \frac{1}{3}\times\frac{2}{3} = \frac{2}{9}$

$P(z=1, \underset{\sim}{x}=1 \mid \theta) = \frac{2}{3}\times\frac{2}{3} = \frac{4}{9}$

$P(z=1, x_i=0 \mid \theta) = \frac{2}{3}\times\frac{1}{3} = \frac{2}{9}$

$\Rightarrow$ when $x_i = 1$

$$w_0 = \frac{\frac{1}{9}}{\frac{1}{9}+\frac{4}{9}} = \frac{1}{5} \qquad w_1 = \frac{4}{5}$$

when $x_i = 0$

$$w_0 = \frac{\frac{2}{9}}{\frac{2}{9}+\frac{2}{9}} = \frac{1}{2} \qquad w_1 = \frac{1}{2}$$

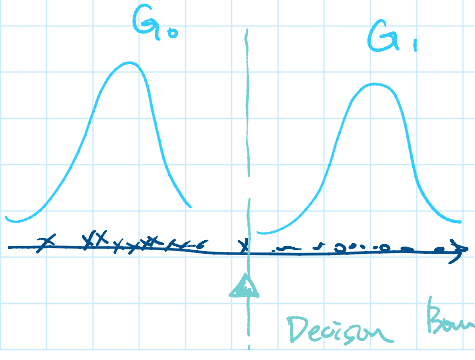
M step:

$$\lambda^{(1)} = \frac{\sum w_i}{n} = \frac{4\times\frac{1}{5}+6\times\frac{1}{2}}{10} = 0.38$$

$$p_0^{(1)} = \frac{\sum w_i x_i}{\sum w_i} = \frac{4\times\frac{1}{5}}{4\times\frac{1}{5}+6\times\frac{1}{2}} = \frac{0.8}{3.8} = \frac{4}{19}$$

$$p_1^{(1)} = \frac{\sum x_i - \sum x_i w_i}{n-\sum w_i} = \frac{4-0.8}{10-3.8} = \frac{3.2}{6.2} = \frac{16}{31}$$

§ Gaussian Mixture Model (GMM)



$\Rightarrow$

- $\lambda \Rightarrow G_0$
- μ_0 ($\vec{\mu}_0$)
- μ_1 ($\vec{\mu}_1$)
- σ_0^2 (Σ_0)
- σ_1^2 (Σ_1)

} assume to be 1 (I)

likelihood

$$\prod \left(\lambda e^{(x_i - \mu_0)^2} \right)^{1-z_i} \left((1-\lambda) e^{(x_i - \mu_1)^2} \right)^{z_i}$$

log likelihood

$$J = \sum (1-z_i)\left(\log\lambda + (x_i - \mu_0)^2 \right) + \sum z_i \left(\log(1-\lambda) + (x_i - \mu_1)^2 \right)$$

$$\frac{\partial J}{\partial \lambda} = 0 \Rightarrow \frac{1}{\lambda} \sum (1 - z_i) = \frac{1}{1-\lambda} \sum z_i$$

$$\lambda_{MLE} = \frac{n - \sum z_i}{n}$$

$$\frac{\partial J}{\partial \mu_0} = 0 \Rightarrow \mu_{0MLE} = \frac{\sum x_i - \sum x_i z_i}{\sum (1 - z_i)}$$

$$\frac{\partial J}{\partial \mu_1} = 0 \Rightarrow \mu_{1MLE} = \frac{\sum x_i z_i}{\sum z_i}$$

mean of x_i

Complete ⇒ Incomplete

$$w = \frac{e^{(x-\mu_1)^2}}{e^{(x-\mu_0)^2} + e^{(x-\mu_1)^2}}$$

$w_0 > w_1$ $w_1 > w_0$

$Z = \{ 0, 1, \dots \}$

$w_0 = \{ \dots \}$

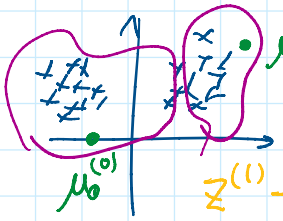
$w_1 = \{ \dots \}$

using w_0, w_1 generate new Z.

and update the parameters with new Z.

⇒ **K means clustering**

# of cluster. , here $k = 2$



① initial the arbitary parameters.

② calculate the responsibility.

③ use new z to update the parameters.