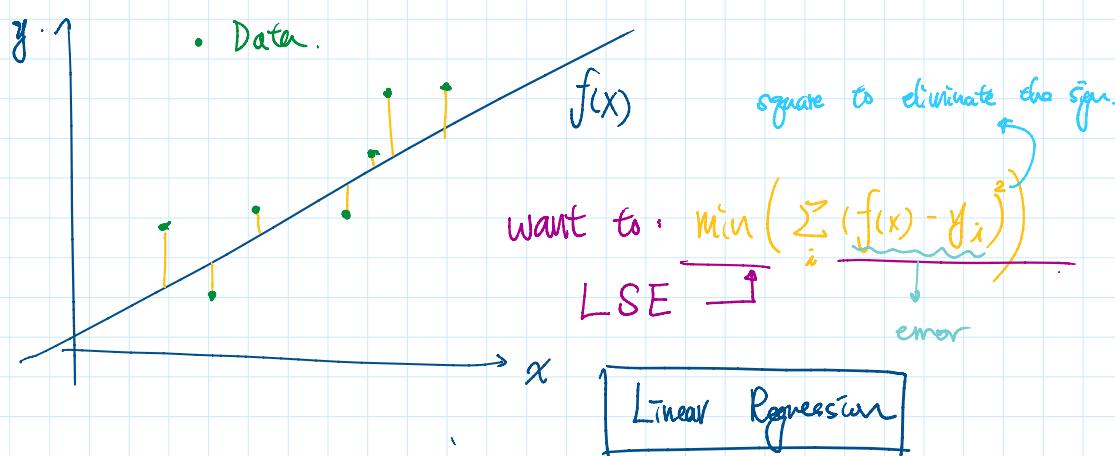
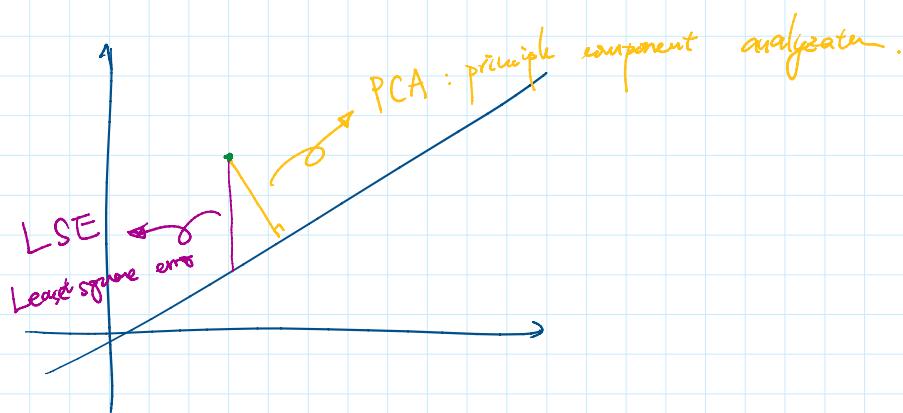


Machine Learning Note

0856(09) 資科工碩一 方偉綸



- * Loss function
- * gain function
- * objective func.

try to minimize these function.
instead to tell computer what to do.

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

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

$$\arg \min_x (\|Ax - b\|^2) \quad \xrightarrow{\text{L}_2 \text{ norm}}$$

$$\begin{aligned} L &= [ax_0 + b - y_0 \ ax_1 + b - y_1 \ \dots] \begin{bmatrix} ax_0 + b - y_0 \\ ax_1 + b - y_1 \\ \vdots \end{bmatrix} \quad (\text{matrix}) \\ &= (Ax - b)^T (Ax - b) \end{aligned}$$

$$\begin{aligned}
 &= (x^T A^T - b^T)(Ax - b) \\
 &= x^T A^T A x - x^T A^T b - b^T A x + b^T b \quad \because b \text{ is } n \times 1 \\
 &= x^T A^T A x - 2x^T A^T b + b^T b \quad \because x \text{ is } 1 \times n \\
 &\quad \xrightarrow{x^2} \text{matrix} \quad \text{calculus.} \\
 &\quad \therefore \text{these are scalar} \\
 &\quad \Rightarrow \text{they are same.}
 \end{aligned}$$

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

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

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

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

① $A^T A$ always have inverse?

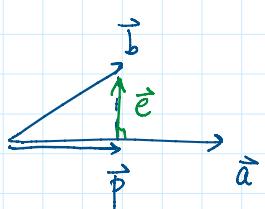
$\Rightarrow A^T A$ is symmetric gram matrix

$\Rightarrow A^T A$ is semi positive definite.

$\Rightarrow A^T A \geq 0$ (not always invertible)

§ orthogonal projection matrix.

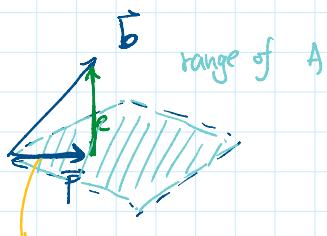
1-D



$$\vec{p} = \vec{a} \cdot \vec{x}$$

scalar.

2-D



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

$$\vec{a}^T \vec{e} = 0$$

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

$$\vec{a}^T (\vec{b} - \vec{p}) = 0$$

$$\vec{a}^T (\vec{b} - \vec{a}x) = 0$$

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

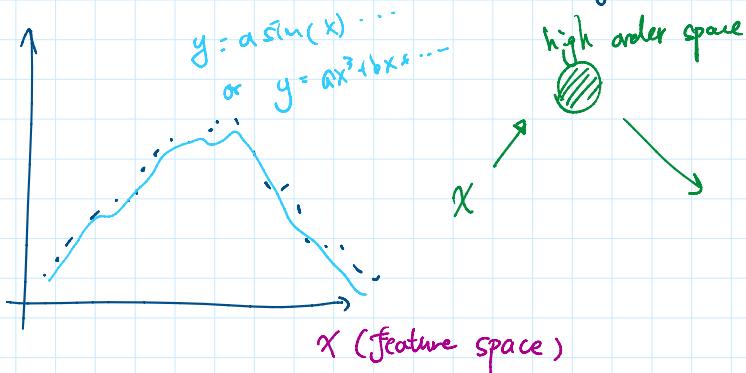
$$A^T b = A^T A x$$

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

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

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

- To solve a non-linear regression.



$$\begin{aligned} Ax &= b \\ \text{data} &\quad \text{target} \\ \Phi x &= b \\ \Phi &= \begin{bmatrix} \phi_0(x) & \phi_1(x) & \phi_2(x) & \dots \end{bmatrix} \quad \begin{bmatrix} \phi_0(x_0) & \phi_1(x_0) & \phi_2(x_0) & \dots \end{bmatrix} \\ \text{basis functions} & \quad \begin{bmatrix} \phi_0(x) = x^3 \\ \phi_1(x) = x^2 \\ \phi_2(x) = x \\ \phi_3(x) = 1 \end{bmatrix} \\ \text{LSE} &= \| \Phi x - b \| \\ Ax &= \begin{bmatrix} ax_0^3 + bx_0^2 + cx_0 + d & \dots \end{bmatrix} \quad \begin{bmatrix} ax_0^3 + bx_0^2 + cx_0 + d \\ \vdots \end{bmatrix} \\ \Phi x &= \begin{bmatrix} \phi_0(x_0) & \phi_1(x_0) & \phi_2(x_0) & \phi_3(x_0) & \dots \end{bmatrix} \quad \begin{bmatrix} \phi_0(x_0) & \phi_1(x_0) & \phi_2(x_0) & \phi_3(x_0) & \dots \end{bmatrix} \\ \boxed{\text{Design Matrix}} & \quad \begin{bmatrix} \vdots \\ \vdots \end{bmatrix} \end{aligned}$$

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

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

Linear Basis Regression

§ Regularization -

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

loss

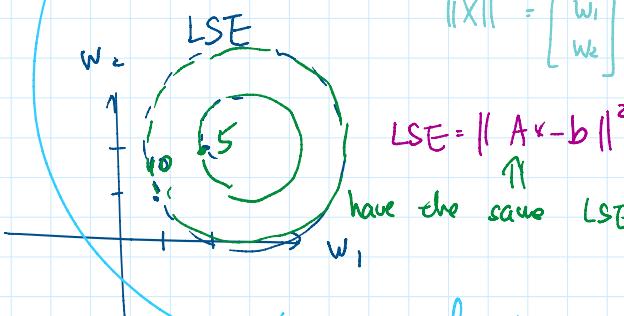
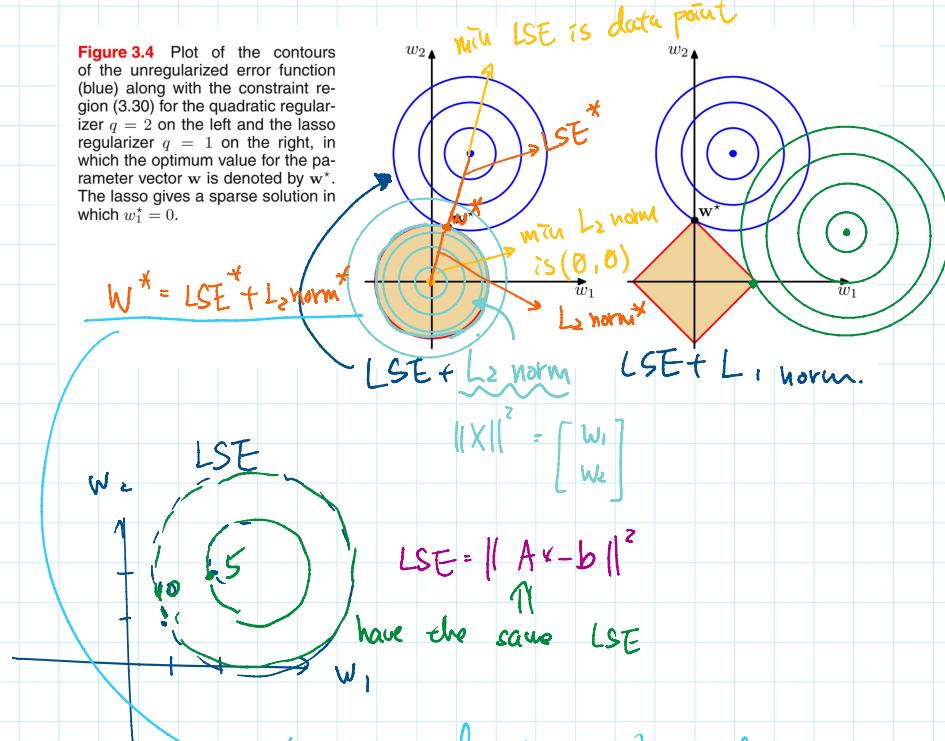
$$\text{LSE} = \underset{x}{\operatorname{argmin}} \left(\underbrace{\|Ax - b\|^2}_{\text{LSE}} + \underbrace{\lambda \|x\|^2}_{\text{(penalty term)}} \right)$$

regularization term

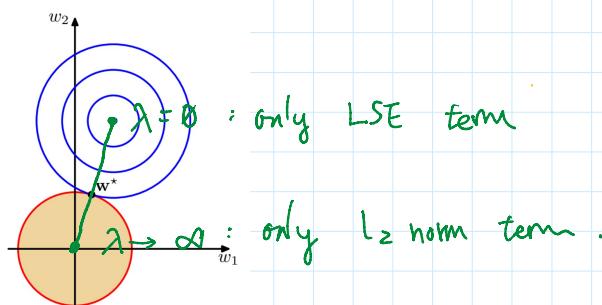
$$L_2 \text{ norm} = \|x\|^2 \quad \text{ridge regression}$$

$$L_1 \text{ norm} = |x| \quad \text{Lasso regression}$$

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 w is denoted by w^* . The lasso gives a sparse solution in which $w_1^* = 0$.

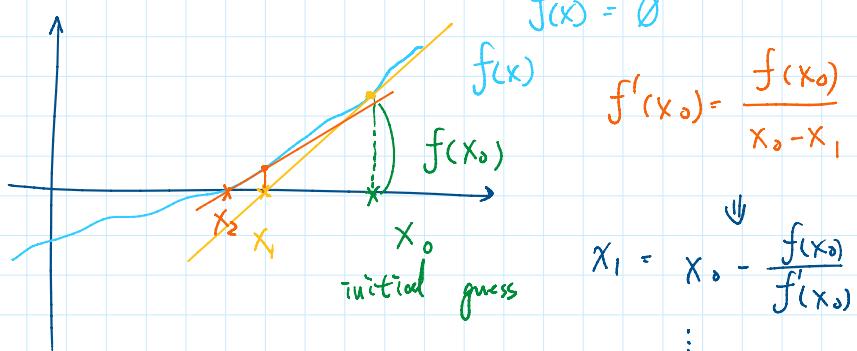


$$\Rightarrow \text{LSE} = \underset{x}{\operatorname{argmin}} \left(\|Ax - b\|^2 + \lambda \|x\|^2 \right)$$



§ Non-linear Loss function.

Newton's Method. : root finding



$$f(x) = 0$$

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

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

⋮

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

$$\Rightarrow f(x_n) = 0 \text{ root.}$$

§ Taylor's expansion (most useful)

approximate a function.

⇒ In Newton's Method, using tangent line to approximate original function.

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

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

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

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

⋮

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

$$\text{eg: } f(x) = e^x$$

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

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

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

$$\text{eg: } f(x) = \sin(x) \quad f'(x) = \cos x \quad f''(x) = -\sin x$$

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

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

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

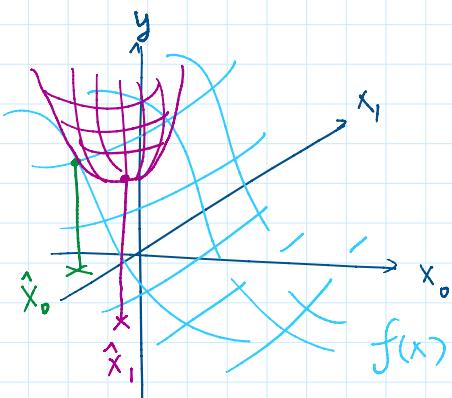
$$\Rightarrow e^{ix} = \cos x - i \sin x$$

• Nonlinear Loss Functions

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

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

quadratic approximate



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

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

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

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

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

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

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

Hessian Matrix gradient: $\begin{bmatrix} \frac{\partial f}{\partial x_0} \\ \frac{\partial f}{\partial x_1} \end{bmatrix}$

$$H f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_0^2} & \frac{\partial^2 f}{\partial x_0 \partial x_1} & \dots \\ \frac{\partial^2 f}{\partial x_1 \partial x_0} & \frac{\partial^2 f}{\partial x_1^2} & \dots \\ \vdots & \vdots & \ddots \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

$$\text{eg: } \|Ax - b\|^2 \text{ LSE}$$

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

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

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

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

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

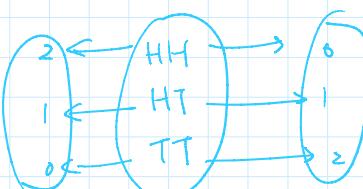
$$= - \frac{1}{2} (A^T A)^{-1} \cdot 2A^T b$$

$$= (A^T A)^{-1} A^T b \quad \boxed{\text{same result}}$$

- Initial Point \Rightarrow trap in local optimize.
- fastest (least iteration)
 - * the time is usually slow $\because H$ is hard to calculate.

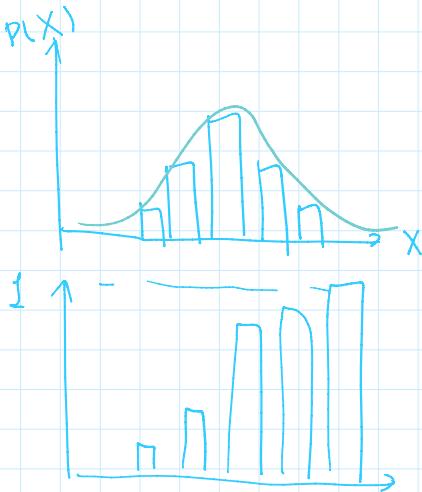
§ Probability.

- trial (試験) : tossing a coin
- out come : head or tail
- event : set of outcomes.
- sample space (Ω) : all outcome
- random variable (X) : mapping function.



$$P(X = \text{head}) \quad P(X = 1) \quad P(X = 1 | \Omega)$$

* To get the $P(X = 1)$ need infinite times of trials



discrete P.M.F

probability mass function
continuous P.D.F

probability density function

C.D.F :

cumulative distribution function

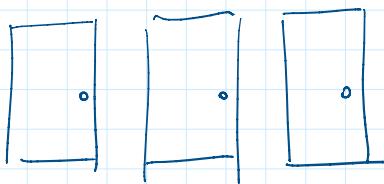
⇒ Conditional Probability

⇒ Joint Probability.

⇒ Monty Hall problem.

= goat / car.

behind the door.



player chose a door.

then host open a wrong door then ask.

the player whether want to change the door.

or not.

open the door choosing the TJ

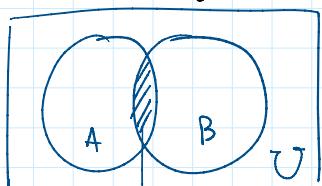
	initial choice	λ	≠	λ	change	not change.
✓	✓	✗	✗	✗	T	F
✗	✗	✓	✓	✓	F	T
					T	F

⇒ Venn diagram.

$$P(X = A | \cup)$$

$$P(X = B | \cup)$$

$$P(X = C | \cup)$$



Joint probability.
 $P(X = C | A)$

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

$$\frac{1}{\square} = \frac{\circ}{\square} \cdot \frac{\circ}{\circ}$$

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

$$\Rightarrow P(B|A) = \frac{P(B) \cdot P(A|B)}{P(A)} \quad \text{Bayes' theorem}$$



$$\begin{array}{ccccccccc} B & B & B & B & B & B & B & B & B \\ \hline & & & & & & & & \square \\ & & & & & & & \uparrow & \\ \text{prior 先驗.} & & & & & & & ? & \end{array}$$

<u>Bayesian</u>	<u>frequentist</u>
prior knowledge	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) $P(A_1, A_2, A_3, A_4)$

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

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

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

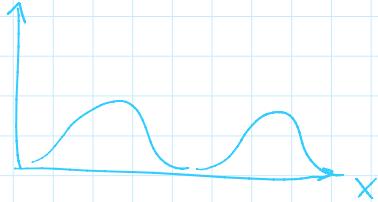
posterior prior parameter likelihood
marginal

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

θ : parameter D: data.

§ Distribution.

- Location.

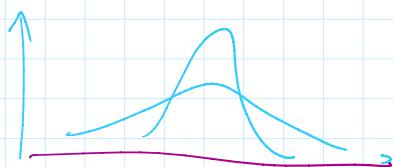


* mean $E(x) = \sum p(x_i) x_i$ first moment $E(x)$

* mode (max) (most frequent)

* median

- dispersion



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

$$\begin{aligned} & \text{2nd moment } E((x-\mu)^2) \\ & = E(x^2) - E^2(x) \end{aligned}$$

* Covariance.

$$\text{Cov}(x, y) = E((x-E(x))(y-E(y)))$$

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

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

$$= \frac{\sum (x-\mu)^2}{n}$$

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

$$= \frac{1}{n} \sum x_i^2 - \frac{1}{n} 2\mu \sum x_i + \frac{1}{n} n\mu^2$$

$$= \frac{1}{n} \sum x_i^2 - 2\mu \frac{\sum x_i}{n} + \mu^2$$

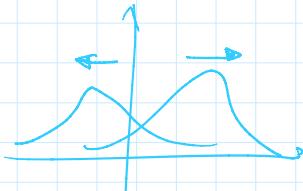
$$= \frac{1}{n} \sum x_i^2 - 2\mu^2 + \mu^2$$

$$= \frac{1}{n} \sum x_i^2 - \mu^2$$

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

- Skewness

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



- Kurtosis.

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

§ Naive Bayes classifier

Maximum likelihood.

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

$$P(H, H, H | \theta = 1) = P(D \text{ is data obtain})$$

$$P(H, H, H | \theta = 0.5) = \frac{1}{8} \quad \theta \text{ is parameter.}$$

$$P(H, H, H | \dots) = \dots \quad \text{here represent the probability of head.}$$

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

• Bayesian.

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

$$\frac{P(\theta | B)}{\text{posterior}} = \frac{P(B | \theta) \cdot P(\theta)}{P(B)} \quad \begin{array}{l} \text{likelihood.} \\ \text{prior} \\ \text{come from} \\ \text{assumption / data / experience} \end{array}$$

$$= \sum p(B, \theta) \quad \leftarrow \text{hard to get.}$$

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

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

$$P(\text{play} = \text{yes} | \text{outlook} = \text{sunny}, \text{temp} = \text{cool}, \text{humidity} = \text{high}, \text{wind} = \text{strong}) = D$$

$$P \left(\begin{array}{c} \text{play} = \text{yes} \\ \text{play} = \text{no} \end{array} \mid \begin{array}{l} \text{outlook} = \text{sunny}, \text{temp} = \text{cool} \\ \text{humidity} = \text{high}, \text{wind} = \text{strong} \end{array} \right) = D$$

$$\left[\frac{P(D \mid \text{play} = \text{yes}) \cdot P(\text{play} = \text{yes})}{P(D)} \right] - \left[\frac{P(D \mid \text{play} = \text{no}) \cdot P(\text{play} = \text{no})}{P(D)} \right]$$

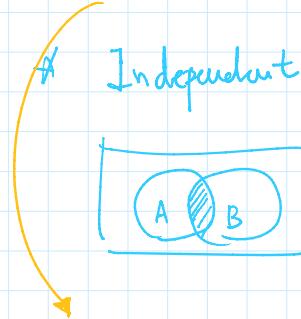
there are no such day satisfy D

$$\Rightarrow P(D \mid \text{play} = \text{yes}) = 0$$

$$\Rightarrow P(D \mid \text{play} = \text{no}) = 0$$

\Rightarrow Naive Bayesian classification

\Rightarrow Conditional independent



$$P(A), P(B)$$

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

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

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

play or not

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

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

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

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

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

$$\Rightarrow P(A_1 A_2 A_3 A_4) = \underbrace{P(A_1 | A_2 A_3 A_4)}_{\text{assumption}} \cdot P(A_2 | A_4) P(A_3 | A_4) P(A_4)$$

$$\begin{aligned} \text{assumption: } P(A_1 A_2 A_3 | A_4) &= P(A_1 | A_4) P(A_2 | A_4) P(A_3 | A_4) \\ &= P(A_1 | A_4) \cdot P(A_2 A_3 | A_4) \end{aligned}$$

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

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

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

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

Now what we need is only:

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

$$P(\text{cold} | \text{Yes}) = 3/9 \quad P(\text{play}) = \frac{9}{14}$$

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

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

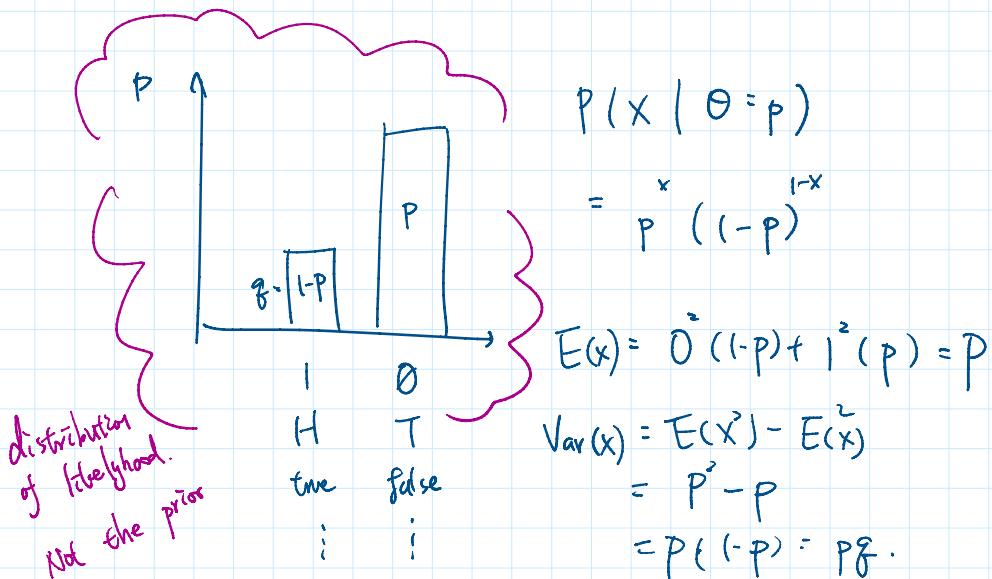
$$P(\text{play} | D) = \frac{\left(\frac{2}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{3}{9}\right) \cdot \frac{9}{14}}{P(D)} = \frac{0.0053}{P(D)}$$

$$P(\text{not play} | D) = \frac{\left(\frac{3}{9} \cdot \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{3}{5}\right) \cdot \frac{5}{14}}{P(D)} = \frac{0.0205}{P(D)}$$

Maximum a posteriori (MAP)

④ Coin tossing.

• Bernoulli Distribution



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

$\Downarrow X$

$$D = [\underbrace{1, 1, 0, 0, \dots, 1}_{\text{independent}}]$$

Maximum Likelihood (MLE) :

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

want to find p maximum the likelihood.

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

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

$$\begin{aligned} \log Y &= \sum_i \log p^{D_i} (1-p)^{1-D_i} \\ &= \sum_i D_i \log p + \sum_i (1-D_i) \log (1-p) \\ &= \sum D_i \log p + \sum (1-D_i) \log (1-p) \end{aligned}$$

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

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

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

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

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

sample mean.

↑ of head.
total.

§ Information Theory.

① Entropy $H(X)$

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.

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

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

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

maximum entropy that you can get.

Ex: 2 coins.

$$HH \rightarrow 2$$

$$HT \rightarrow 1$$

$$TH \rightarrow 1$$

$$TT \rightarrow 0$$

fair coins

$$H(x) = - \sum_{\text{all}} \frac{1}{4} \log \frac{1}{4} = 2 \quad \text{uniform randomness}$$

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

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

$$= 0.48844$$

$$\text{Ex: } - \sum | \log | = 0 \Rightarrow \text{no randomness}$$

• Distance.

② Conditional Entropy.

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

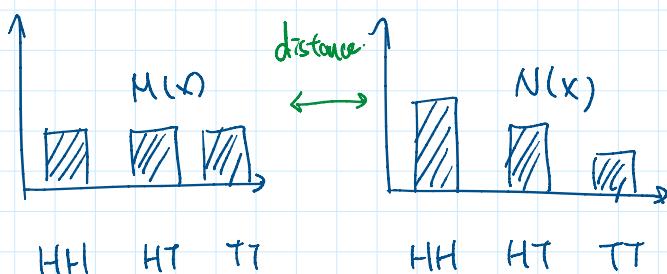
- Joint Entropy.

$$H(X,Y) = H(Y|X) + H(X)$$

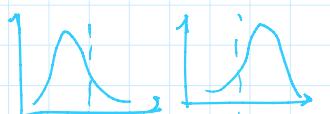
- Relative Entropy.

KL divergence: distance between 2 distributions

[Kullback Leibler] also called information gain.



* why not $N(x)$



not work for symmetric distribution $KL=0$

$$KL(p||q) = - \sum_{x \in X} p(x) \log q(x) - \left(\sum_{x \in X} p(x) \log p(x) \right)$$

where: X : random variable

M, N : distribution.

$$= - \sum_{x \in X} p(x) \log \frac{N(x)}{M(x)}$$

- 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

c.f. Maximum Likelihood.

coin tossing. 10 times

D: H, H, ... H \leftarrow observed (data)

$$* L(\theta=0.5 \mid H, H, \dots H) \leftarrow \text{Likelihood the coin}$$

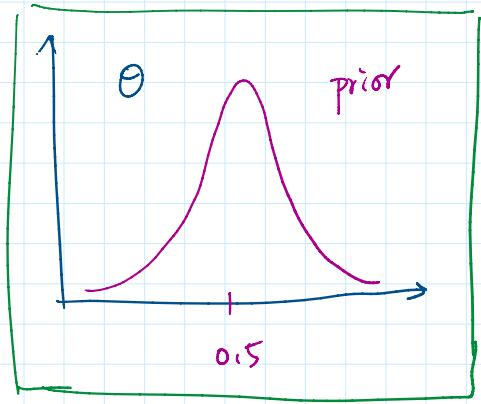
$$= P(H, H, \dots H \mid \theta=0.5) \quad \text{is fair (0.5H or T)}$$

$$= 1/1024$$

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

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

maximum likelihood.



frequentist.

\Rightarrow every θ is equal.

■ Uniform Distribution. \rightarrow Maximum Entropy

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

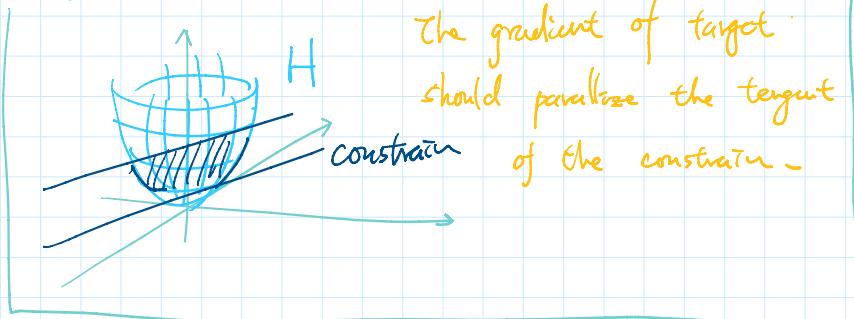
$$= - \int_{-\infty}^{\infty} p(x) \log p(x) dx$$

, $p(x)$ pdf

- $p(x) \geq 0$

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

Lagrange multiplier



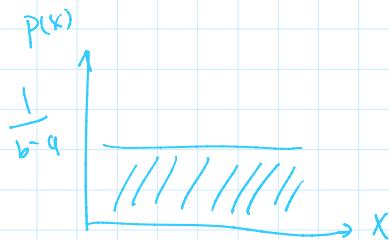
The gradient of target
should parallelize the tangent
of the constraint.

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

$$\begin{aligned} \frac{\partial L}{\partial p(x)} &= - \left(\frac{\partial p(x)}{\partial p(x)} \cdot \ln p(x) + p(x) \cdot \frac{\partial \ln p(x)}{\partial p(x)} \right) + \lambda \cdot \frac{\partial p(x)}{\partial p(x)} \\ &= -(\ln p(x) + 1) + \lambda = 0 \\ \Rightarrow p(x) &= e^{\lambda-1} \end{aligned}$$

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

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



■ Expectation is known.

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

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

new constraint

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

exponential distribution

④ Knowing the 2nd moment.

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

; same as above.

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

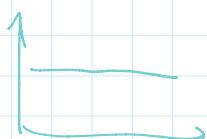
$$\Rightarrow p(x) = e^{-1 + \lambda_0 + \lambda_1 x + \lambda_2 x^2}$$

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

Gaussian distribution

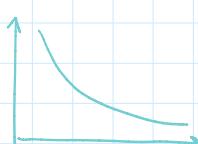
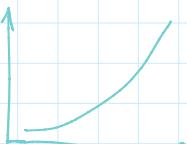
⇒ Conclusion

known nothing: Uniform distribution

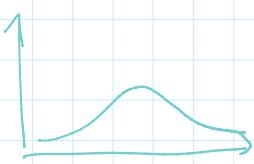
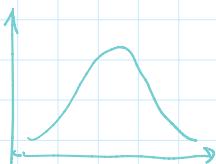


all θ have same value.

given mean: Exponential distribution



given variance: Gaussian distribution



- Location
- dispersion;
moments.

→ toss a coin for one time.

$$\theta = \text{Head} = 1$$

Bernoulli Distribution:

$$P(X=0 | \theta) = \theta^0 (1-\theta)^{1-x}$$

$$P(1 | \theta) = \theta^1 (1-\theta)^0 = \theta$$

$$P(0 | \theta) = \theta^0 (1-\theta)^1 = 1-\theta$$

$$E(x) = \theta \cdot 1 + (1-\theta) \cdot 0 = \theta \quad \text{location}$$

$$\begin{aligned} \text{Var}(x) &= E(x^2) - E(x)^2 \\ &= \theta - \theta^2 = \theta(1-\theta) \quad \text{dispersion} \end{aligned}$$

→ toss a coin for N times.

Binomial Distribution

identical independent (iid)

Given $[1, 0, 1, 1, 0, \dots] \Rightarrow \underbrace{[x_1, x_2, \dots, x_n]}_{\text{multiple Bernoulli trials}} = D$

$$\underbrace{P(D|\theta)}_{\text{likelihood.}} = \prod_{k=1}^N \theta^{x_k} (1-\theta)^{1-x_k}$$

Random variable x $\begin{cases} 0 \\ 1 \end{cases}$
 multiple Bernoulli trial N times.

X : # of head in N trials.

$D = [x_1, x_2, \dots, x_n]$

$$\sum x_i = m.$$

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

$$E(x) = N\theta$$

$$\text{Var}(x) = N\theta(1-\theta)$$

MLE : θ
 frequentist

§ Conjugate Prior

Frequentist \rightarrow Bayesian

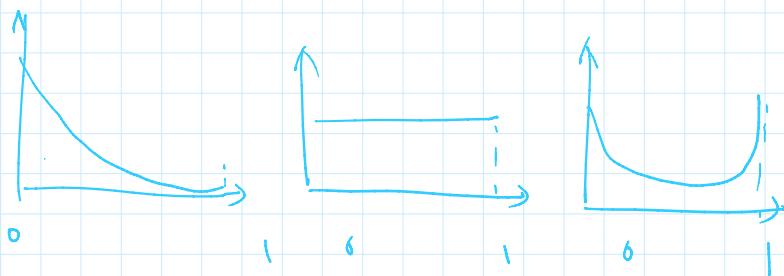
- Prior
- Experience

Bernoulli
Binomial \longleftrightarrow Beta Distribution

Prior & Posterior in the same form

- Beta Distribution

$$\begin{aligned}\text{Beta}(\theta | a, b) &= \theta^{a-1} (1-\theta)^{b-1} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \\ &= \theta^{a-1} (1-\theta)^{b-1} \cdot \frac{1}{\beta(a, b)}\end{aligned}$$



$\Gamma(\text{gamma})$: factorial $n! = n(n-1)(n-2)\dots 1$

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

$$\begin{aligned}\beta(\text{beta}) : \beta(a, b) &= \int_0^1 p^{a-1} (1-p)^{b-1} dp \\ &= \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}\end{aligned}$$

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

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

$$\begin{aligned} & \int_0^1 x \frac{1}{\beta(a,b)} x^{a-1} (1-x)^{b-1} dx \\ &= \frac{1}{\beta(a,b)} \int_0^1 x \cdot x^{a-1} (1-x)^{b-1} dx \\ &= \frac{1}{\beta(a,b)} \int_0^1 x^a (1-x)^{b-1} dx \\ &= \frac{1}{\beta(a,b)} \beta(a+1, b) \end{aligned}$$

$$\begin{aligned} &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{a \cdot \Gamma(a) \cdot \Gamma(b)}{(a+b) \cdot \Gamma(a+b+1)} = \frac{a}{a+b} \end{aligned}$$

• Beta-Binomial conjugation

$$P(X|N, P) \cdot P(P|\alpha, \beta)$$

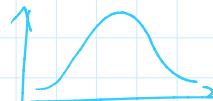
Binomial prior, Beta

$$\binom{N}{m} P^m (1-P)^{N-m} \cdot P^{\alpha-1} (1-P)^{b-1} \cdot \frac{1}{\beta(\alpha, \beta)}$$

$$\propto P^{m+\alpha-1} \cdot (1-P)^{N-m+b-1}$$

- MLE is a point estimation

$$\frac{P(x|\theta)p(\theta)}{p(x)} = p(\theta|x)$$



$$\left[P^{m+\alpha-1} (1-P)^{N-m+b-1} \right] \cdot \text{O}$$

$$\sum_a p(x,a) \cdot \int p(x,a) da.$$

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 :

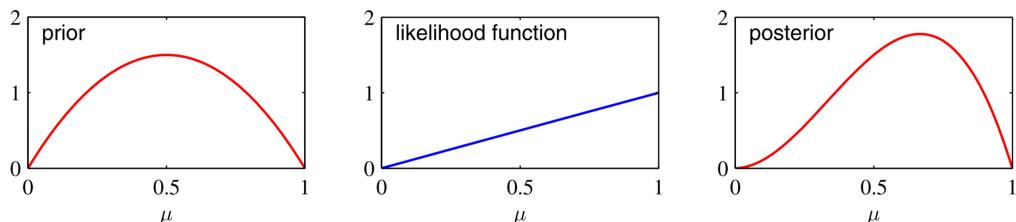
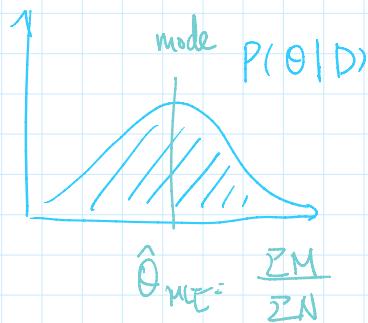
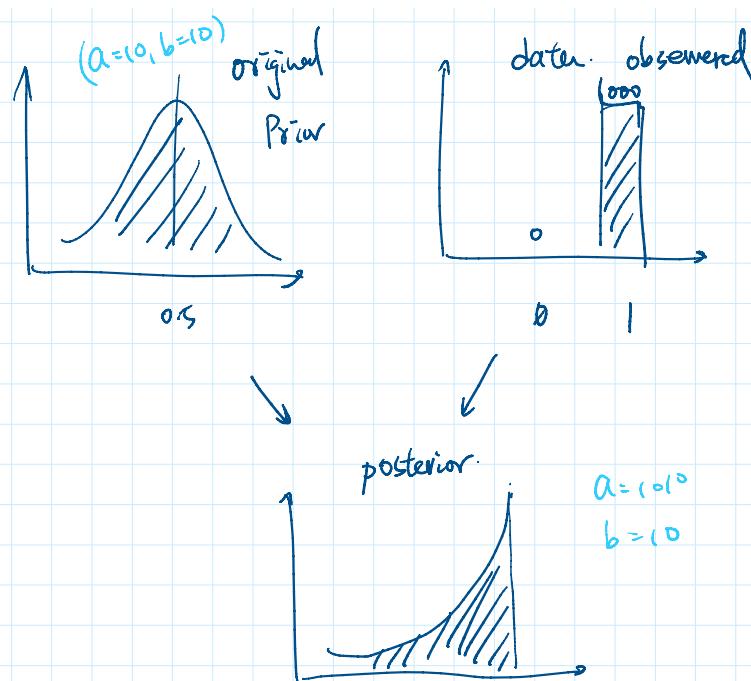
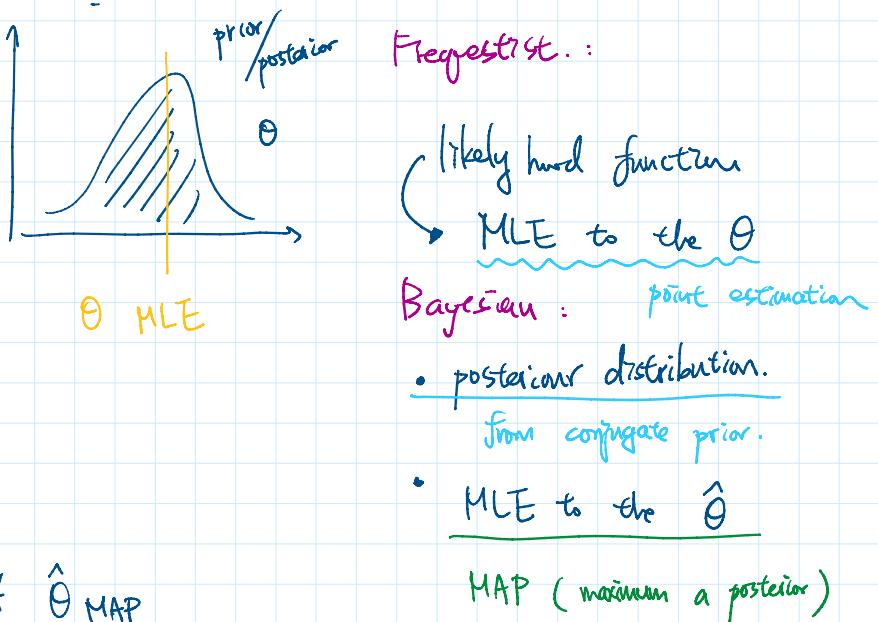
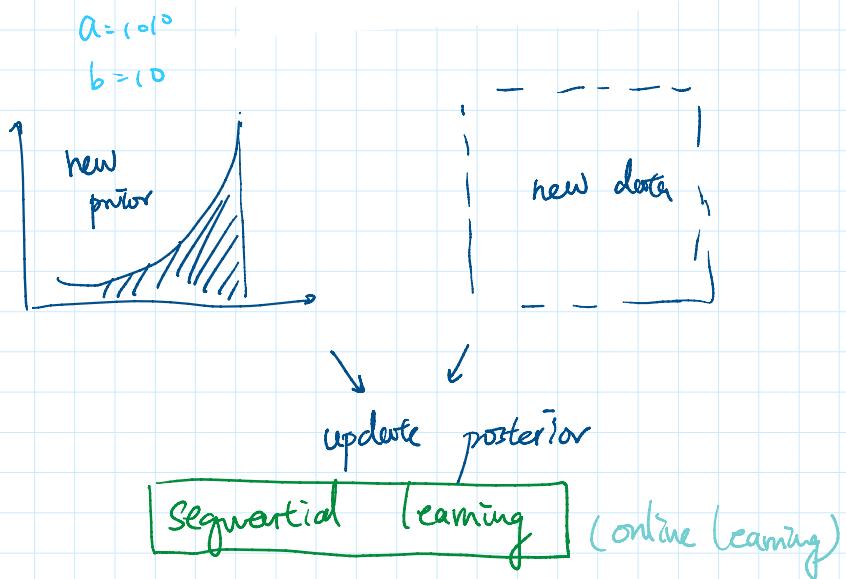


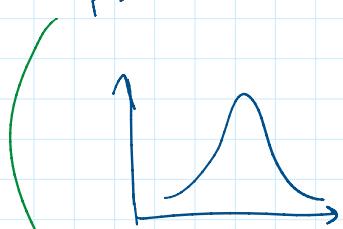
Figure 2.3 Illustration of one step of sequential Bayesian inference. The prior is given by a beta distribution with parameters $a = 2, b = 2$, and the likelihood function, given by (2.9) with $N = m = 1$, corresponds to a single observation of $x = 1$, so that the posterior is given by a beta distribution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





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

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.

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

w.r.t p

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

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

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

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

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

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

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

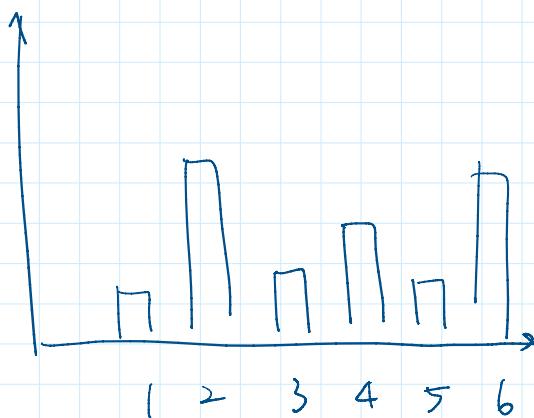
Frequentist : 以中立 Prior. 及有 data

Bayesian : 利用以前的 knowledge. 設置 Prior.

Prior 手中隨機性. 是個性質.

§ Multinomial.

v.s. binomial.



dice 100 times = N

$$X = [\underbrace{30, 40, \dots}_{m_1, m_2, \dots, m_k}, \dots]$$

$$M(X) = \binom{N}{m_1, m_2, \dots, m_k} \pi p_k^{m_k}$$

\Downarrow

$$\frac{N!}{m_1! m_2! \dots m_k!}$$

Conjugate prior

Dirichlet

$$\frac{\Gamma(a_1 + \dots + a_k)}{\Gamma(a_1) \Gamma(a_2) \dots \Gamma(a_k)} \cdot \pi p_k^{a_k - 1}$$

Gaussian distribution

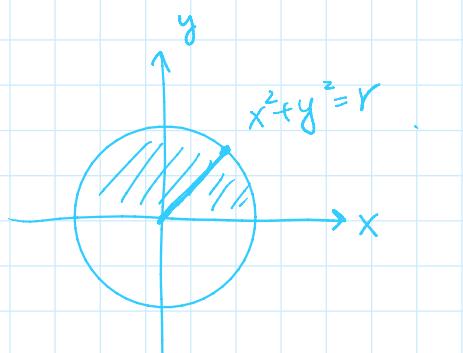
Thursday, April 9, 2020 4:20 PM

§ Gaussian distribution

- Gaussian integral

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

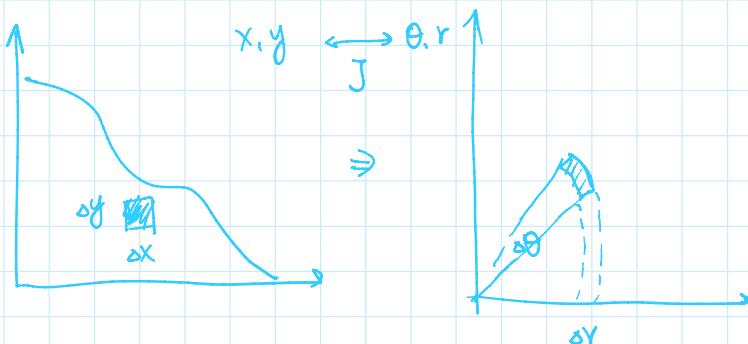
- polar coordinate



$$x = r \cos \theta$$

$$y = r \sin \theta$$

- Jacobian determinant



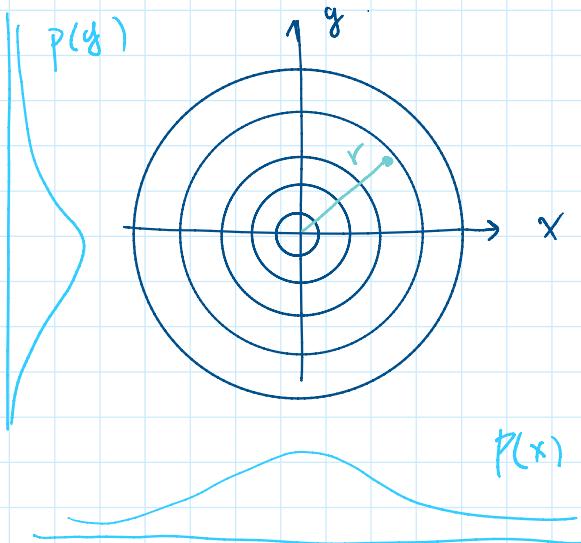
$$J = \det \begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{bmatrix} = \det \begin{bmatrix} \frac{\partial \cos \theta}{\partial r} & \frac{\partial \cos \theta}{\partial \theta} \\ \frac{\partial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

$$= \det \begin{bmatrix} -r \sin \theta & -r \cos \theta \\ r \cos \theta & r \sin \theta \end{bmatrix} = r$$

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

$$\begin{aligned}
 &= \sqrt{\int_{-\infty}^{\infty} e^{-x^2} dx \cdot \int_{-\infty}^{\infty} e^{-y^2} dy} \\
 &= \sqrt{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)} dx dy} \\
 &= \sqrt{\int_0^{2\pi} \int_0^{\infty} e^{-(r^2(\cos^2\theta + \sin^2\theta))} r dr d\theta} \\
 &= \sqrt{\int_0^{2\pi} \int_0^{\infty} e^{-r^2} r dr d\theta} \\
 &= \sqrt{2\pi \cdot \frac{1}{2} \int_0^{\infty} e^{-r^2} dr^2} \\
 &= \sqrt{\pi \cdot (-e^{-r^2}) \Big|_0^{\infty}} = \sqrt{\pi}
 \end{aligned}$$

• Gaussian distribution



independent

$$\begin{aligned}
 p(x, y) &= \underbrace{p(x)}_{= g(r)} \cdot \underbrace{p(y)}_{\text{w.r.t. } \theta} \\
 &= g(r)
 \end{aligned}$$

the probability with same r should be same,

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

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

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

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

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

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

x, y are independent.

The only possibility to make these 2 eqs. equal is they converge to a constant.

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

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

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

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

should < 0 , otherwise diverge.

$$p(x) = e^{\frac{1}{2C_1} x^2} = e^{-k_1 x^2}$$

$$p(y) = e^{\frac{1}{2C_2} y^2} = e^{-k_2 y^2}$$

\sim they are pdf \Rightarrow integral to 1.

$$A \cdot \int_{-\infty}^{\infty} e^{-k_1 x^2} dx = 1$$

$$\int_{-\infty}^{\infty} e^{-k_1 x^2} dx = \frac{1}{\sqrt{A}}$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-k_1 x^2 - k_2 y^2} dx dy = \sqrt{\frac{1}{A}}$$

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

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

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

normalize factor.

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

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

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

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

$$\text{Assume } \mu = 0$$

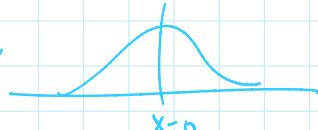
$$\int_{-\infty}^{\infty} x^2 \sqrt{\frac{k}{\pi}} e^{-kx^2} dx = \sigma^2$$

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

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

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

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

\downarrow it's symmetric 

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

$$= 0 \qquad \qquad = \frac{-1}{2k} \sqrt{\frac{\pi}{k}}$$

$$\Rightarrow \sigma^2 = \frac{1}{2K} \Rightarrow \sigma = \frac{1}{\sqrt{2}\sigma}$$

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

平均至 μ $\Rightarrow P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

In high dimension (multivariable)

$$\begin{bmatrix} x \\ x_0 \\ x_1 \\ \vdots \end{bmatrix} \quad \begin{bmatrix} \sigma^2 \\ \sigma_0^2 & \text{cor} \dots \\ \text{cor} & \sigma_1^2 & \ddots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

$$N(X | \mu, \sigma^2) \quad P(x) = N(x | \mu, \sigma^2)$$

- Maximum Likelihood Estimation (MLE)

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

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

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

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

$$P(x_1, \dots, x_n | \mu, \sigma^2) = \prod_{k=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_k - \mu)^2}{2\sigma^2}} = L$$

$$\begin{aligned}
 \log L &= \sum_{k=1}^n \log \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x_k-\mu)^2}{2\sigma^2}} \\
 &= \sum \log \frac{1}{\sqrt{2\pi\sigma^2}} + \sum \frac{-(x_k-\mu)^2}{2\sigma^2} \\
 &= -\frac{N}{2} \log 2\pi\sigma^2 + \sum \frac{-(x_k-\mu)^2}{2\sigma^2}
 \end{aligned}$$

log likelihood function

w.r.t μ .

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

$$\begin{aligned}
 &\frac{\partial}{\partial \mu} \left(\sum (x^2 - 2x\mu + \mu^2) \right) \\
 &= \sum_{k=1}^N (-2x_k + 2\mu) \\
 &= -2 \sum x + 2N\mu = 0
 \end{aligned}$$

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

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

$$\begin{aligned}
 \frac{\partial \log L}{\partial S} &= \frac{\partial}{\partial S} \left(-\frac{N}{2} \log 2\pi S + \sum \frac{-(x-\mu)^2}{2S} \right) \\
 &= \frac{\partial}{\partial S} \left(-\frac{N}{2} \log 2\pi - \frac{N}{2} \log S + \sum \frac{-(x-\mu)^2}{2} \cdot S^{-1} \right) \\
 &= -\frac{N}{2} \frac{1}{S} - \sum \frac{-(x-\mu)^2}{2} \cdot S^{-2} = 0
 \end{aligned}$$

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

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

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

§ 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 $\sim N(\mu | \textcircled{1}, \textcircled{2})$

[prior]

Gaussian $N(\mu | \mu_0, \sigma_0^2)$

[likelihood]

Gaussian

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

assume

fixed or

fit it to another gaussian

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

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

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

P(X)

$$= \int p(x | \mu) \cdot p(\mu | \mu_0, \sigma_0^2) d\mu$$

$$\prod_{k=1}^N \underbrace{p(x_k | \mu, \sigma^2)}_{\text{Likelihood}} \cdot \underbrace{p(\mu | \mu_0, \sigma_0^2)}_{\text{prior}}$$

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

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

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

\ddagger : it's also a Gaussian distribution. ($e^{\frac{-1}{2\sigma^2} (x-\mu)^2}$)
 $\Rightarrow N(\mu)$. quadratio complete the square

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

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

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

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

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

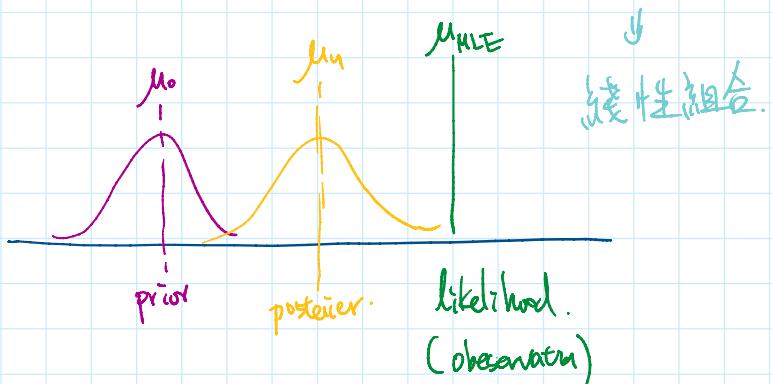
$$\mu_n = m = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{\sum x_k}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}$$

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

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

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

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

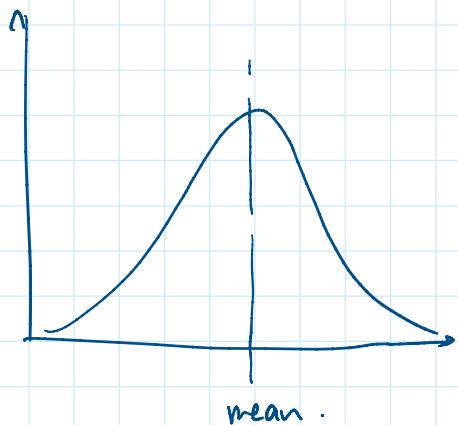
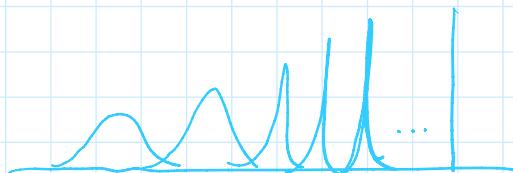


$$\bullet (\sigma_n^2)^{-1} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$$

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

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

$$\sigma_n^2 \rightarrow 0$$



- Symmetric
 - Unimodal
 - Localization
- } distance metric

distance ↑ probability ↓

§ Central Limit Theorem. 中央極限定理

X what ever it is like.
 random variable.
 e.g. e^X 

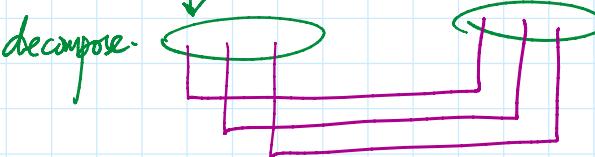
its sample mean or sum $y \sim N(\mu, \sigma^2)$
 $[X_{11} X_{12} \dots X_{1k}] \rightarrow y_1$
 $[X_{21} X_{22} \dots X_{2k}] \rightarrow y_2$
 : :
 guarantee to be Gaussian

- Moment Generating function

$$y = C_0 E(x) + C_1 E(x-\mu)^1 + \dots$$

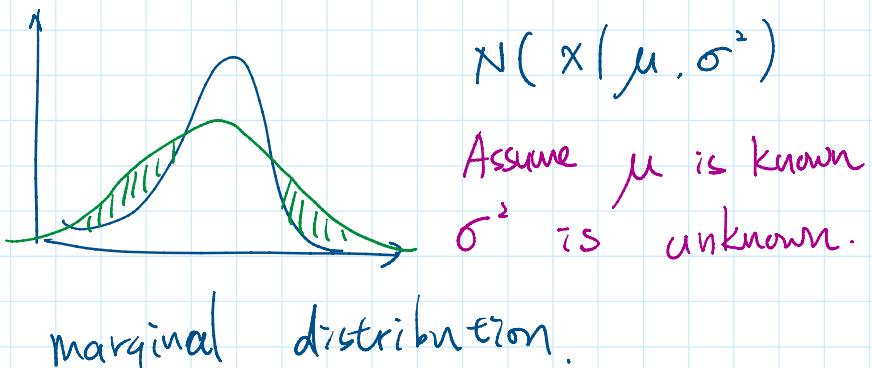
Any variable can be decompose to its moment.

application .. if Likelihood & prior
 not conjugate.

decompose. 

Combine the same order moment.

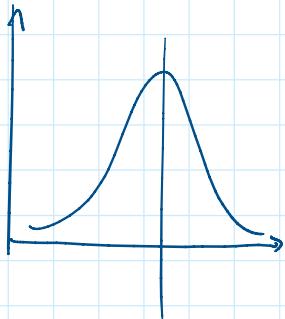
§ Student T distribution



$$T = \int N(x | \mu, \sigma^2) \cdot P(\sigma^2 | a, b) da, db$$

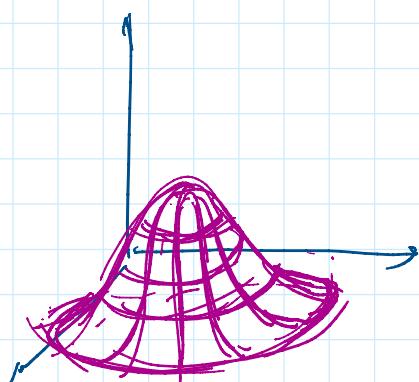
* more tolerance to high variance.

* Better for {sample size small
variance high.}



uni variant Gaussian (单变量)

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

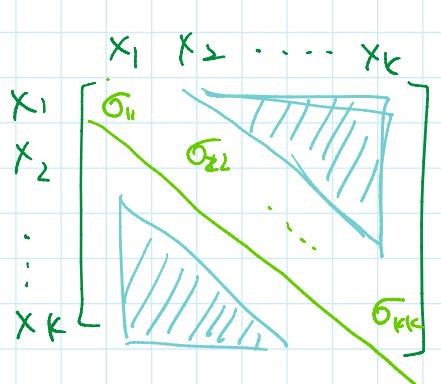


K-dim Gaussian.
(multivariant)

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

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

Σ : covariance matrix



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

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

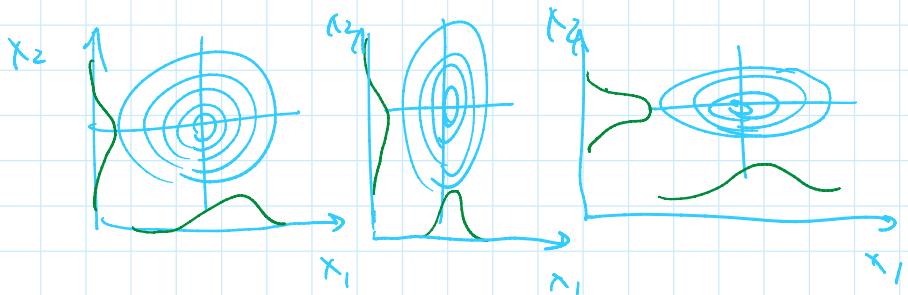
⇒ covariance

* Covariance show how the x_i change while x_j change.

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

* Orthogonal transform.

$$|\Sigma| = \begin{bmatrix} \sigma_1 & & \\ & \sigma_2 & 0 \\ 0 & & \ddots & 0_{n \times n} \end{bmatrix} \text{ diagonal.}$$



$$\Rightarrow \Sigma = \begin{bmatrix} \sigma_1 & \text{Non-0} \\ & \sigma_2 \\ \text{Non-0} & \ddots & \sigma_{kk} \end{bmatrix}$$

• if Σ is diagonal.

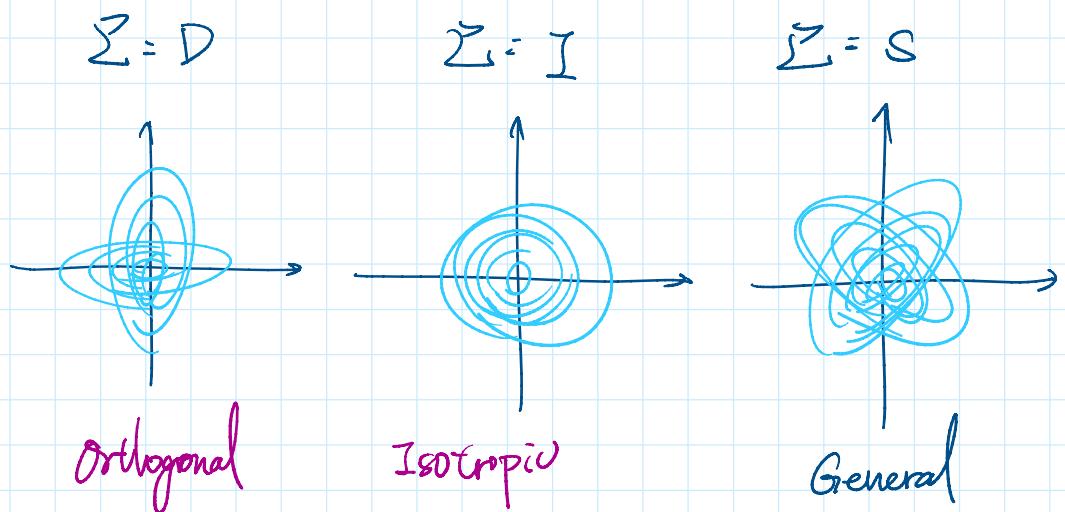
$$\frac{1}{2\pi\sqrt{\det(\Sigma)}} \cdot e^{\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \propto \text{Euclidean distance.}$$

歐幾里得距離 $\sqrt{(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2}$

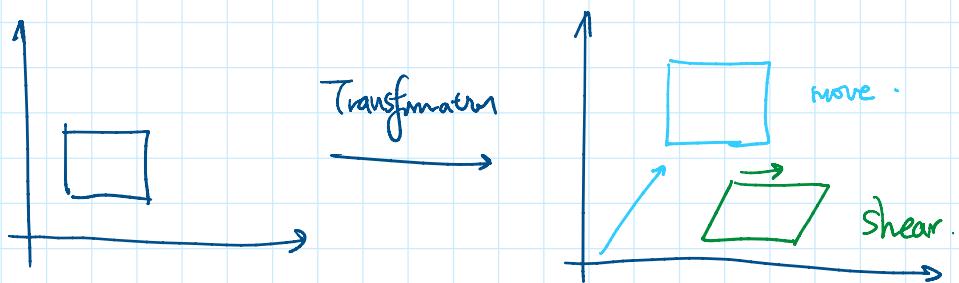
• if Σ is not diagonal.

$$(x-\mu)^T \Sigma^{-1} (x-\mu) \propto \text{Mahalanobis distance}$$

馬氏距離.



Σ Affine property.



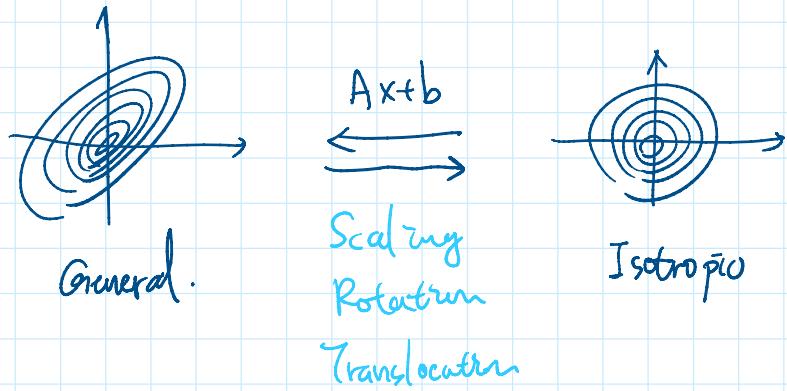
* If you do affine transform to Gaussian,
it is still Gaussian.

$$X \sim N(\mu, \Sigma)$$

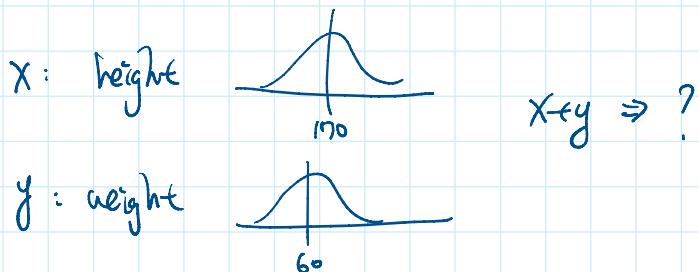
$$y = Ax + b \sim N(A\mu + b, A\Sigma A^T)$$

$X \stackrel{\text{i.i.d}}{\sim} N(0, I)$ standard normal

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



§ Linear transformation.



- Sum of 2 independent Gaussian \Rightarrow Gaussian

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

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

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

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

X_1, X_2 are independent.

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

$$Y \sim N(A\mu + b, A\sigma^2 A^T)$$

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

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

- Linear Combination.

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

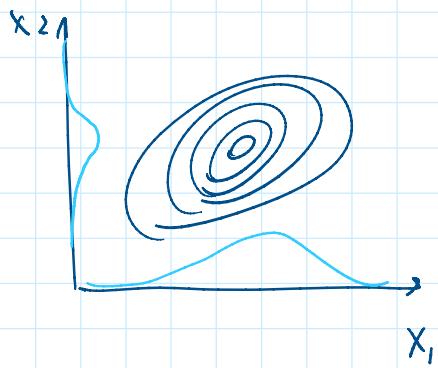
$$\text{Var}(Y) = \sum b_i^2 \sigma_i^2$$

- Linear transformation

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

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

§ Marginal Gaussian



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

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

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

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

$$A\mu + b = \mu_1 \quad AC A^T = \sigma_1^2 \quad \text{from Affine property.}$$

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

$$X = \begin{bmatrix} X_a \\ X_b \end{bmatrix} \quad X_a = \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix} \quad X_b = \begin{bmatrix} X_{k+1} \\ \vdots \\ X_n \end{bmatrix}$$

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

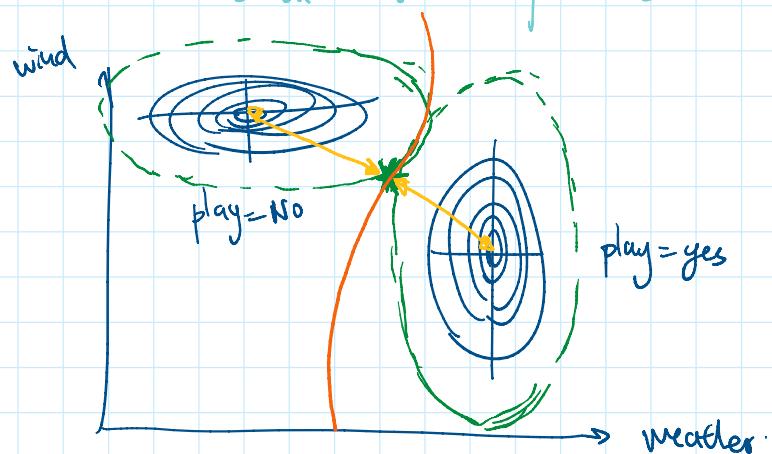
k $n-k$

$$A = \begin{bmatrix} I_k & 0 \\ 0 & I_{n-k} \end{bmatrix} \quad y = Ax + b = \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix} = x_a$$

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

§ Naive Bayes classifier

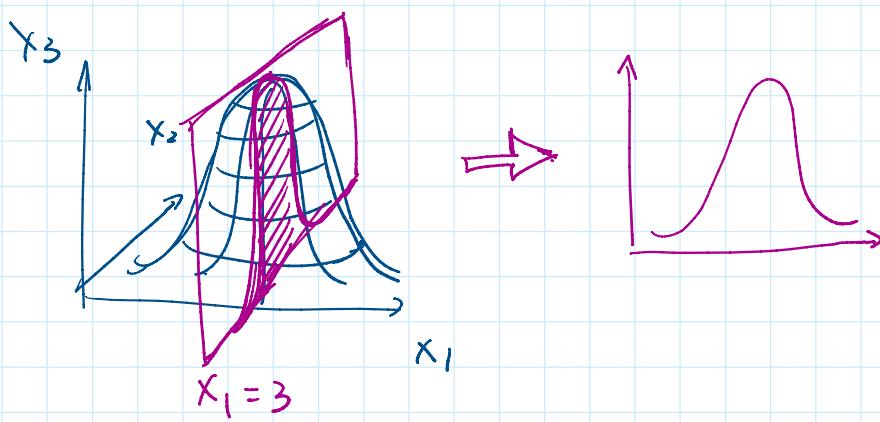
↳ conditional independent.



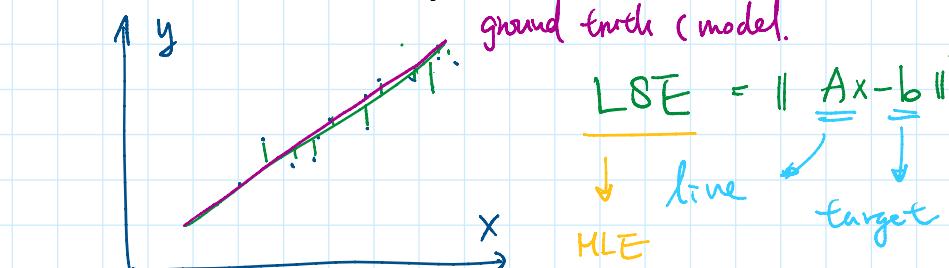
0.5 decision boundary

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

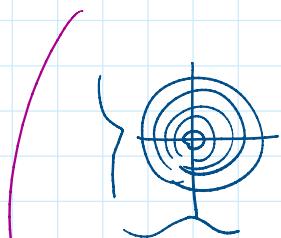
§ Conditional Gaussian



§ Probabilistic view of linear regression



$\text{MAP} \rightarrow \text{Fully Bayesian}$



random error $\sim N$

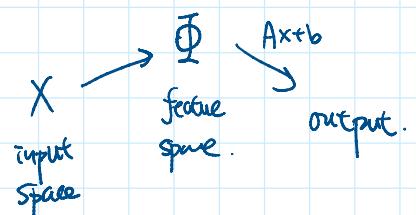
\Rightarrow why the points not on line.

minimal error \Leftrightarrow maximum probability (likelihood)

$$b = Ax + \varepsilon$$

target line error.

design matrix

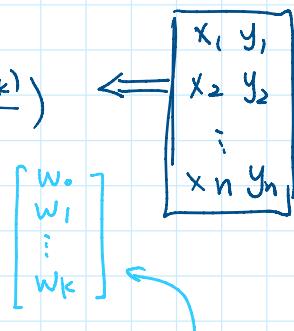


likelihood:

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

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

$$\Rightarrow \left(\frac{1}{\sigma \sqrt{2\pi}} \right)^n \prod_k e^{-\frac{1}{2} \left(\frac{y_k - w^T \Phi(x_k)}{\sigma} \right)^2}$$



MLE: w.r.t w

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

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

$\|b - Ax\|^2$
same as LSE

Frequentist: $\frac{\text{MLE}}{\text{probability}} = \frac{\text{LSE}}{\text{determinist.}}$

Bayesian: prior

$$\begin{bmatrix} \sigma^2 & & & \\ & \sigma^2 & & \\ & & \ddots & \\ & & & \sigma^2 \end{bmatrix}$$

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

posterior:

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

likelihood (above)

$$= \Sigma = \Lambda^{-1}$$

$$e^{\frac{-1}{2}(x-w)^T \Lambda (x-w)} = e^{\frac{-b}{2} w^T w} = e^{\frac{-b}{2} \cdot w^T w}$$

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

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

regularized LSE

- LSE \rightarrow MLE
- RLSE \rightarrow MAP

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

$$w + a(x_1 + \dots + x_n) \Rightarrow J = e^{\frac{-a}{2} (w^T x - y)^T (w^T x - y) - \frac{b}{2} w^T w}$$

$$w: \text{weight} \begin{bmatrix} w_0 \\ \vdots \\ w_k \end{bmatrix} \Rightarrow \log J = \frac{-a}{2} (w^T x - y)^T (w^T x - y) - \frac{b}{2} w^T w$$

$A = \Phi$ = design Matrix
rewrite $w^T x$ as Aw (Both represent the line)

$$\begin{bmatrix} x_0 & \dots & x_k \end{bmatrix}_{n \times k} \Rightarrow \log J: \frac{-a}{2} (w^T A^T Aw - \underbrace{w^T A^T y - y^T A w + y^T y}_{\text{want to turn into the form } (x-w)^T \Lambda (x-w)}) - \frac{b}{2} w^T w$$

$$g: \text{target} \begin{bmatrix} y_0 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1} = \frac{-a}{2} (w^T A^T Aw - 2w^T A^T y + y^T y) - \frac{b}{2} w^T w$$

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

$$= \frac{x^T \Lambda x}{\cancel{-T}} - \frac{?x^T \Lambda y}{\cancel{-T}} + \bigcirc$$

$$\Rightarrow \frac{1}{2} (\cancel{aw^T A^T Aw} + \cancel{bw^T w} - \cancel{2aw^T A^T y} + \bigcirc)$$

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

$$\left\{ \begin{array}{l} w^T A^T y \\ |x|k \times k \times n \times n \times 1 = 1 \times 1 \end{array} \right.$$

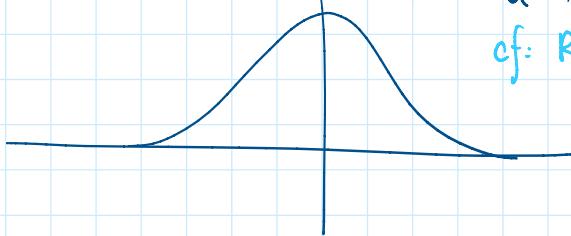
$$\left. \begin{array}{l} y^T A w \\ |x|n \times m \times k \times k \times 1 = 1 \times 1 \end{array} \right.$$

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

MAP

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

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

$$u = (A^T A + \frac{b}{a} I)^{-1} A^T y$$


cf: RLSE
 $(A^T A + 2I)^{-1} A^T b$

§ Bayesian Linear Regression.

• Fully Bayesian

want to find.

MLE \rightarrow LSE \rightarrow likelihood.

MAP \rightarrow RLSE \rightarrow posterior.

Fully Bayesian \rightarrow predictive distribution

(no \vec{w} , $\vec{\Phi}$)

Kernal method.

$\#$ of basis $\rightarrow \infty$

* every \vec{w} is a distribution

■ Marginalized w \downarrow

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

$\int N(y|w^T x, \sigma^2) \cdot N(w|\mu, \Sigma^{-1}) dw$

univariate gaussian.

$$= \int \textcircled{1} e^{\frac{-\alpha}{2}(y-w^T x)^2} \cdot e^{-\frac{1}{2}(w-\mu)^T \Lambda (w-\mu)} dw$$

$$= \int \textcircled{1} e^{-\frac{1}{2}(\alpha y^2 - 2w^T x y + w^T x^T x + w^T \Lambda w - 2w^T \Lambda \mu + \mu^T \Lambda \mu)} dw$$

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

$$= \int \textcircled{1} e^{-\frac{1}{2}(\underline{w^T (\alpha x x^T + \Lambda) w} - \underline{2w^T (x y + \Lambda \mu)} + \underline{ay^2 + \mu^T \Lambda \mu})} dw$$

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

$$= \textcircled{2} \int e^{-\frac{1}{2}((w^T C w - 2w^T C \mu' + \mu'^T C \mu') - \mu'^T C \mu' + ay^2 - \mu^T \Lambda \mu)} dw$$

$$= \textcircled{2} \int e^{-\frac{1}{2}(w-\mu)^T C (w-\mu')} \cdot e^{-\frac{1}{2}\mu'^T C \mu' - \frac{1}{2}ay^2 - \frac{1}{2}\mu^T \Lambda \mu} dw$$

$$= \textcircled{2} e^{-\frac{1}{2}\mu'^T C \mu' - \frac{1}{2}ay^2} \int e^{-\frac{1}{2}(w-\mu)^T C (w-\mu')} dw$$

it's a gaussian

integral all w over gaussian = 1

$$= \textcircled{2} e^{\frac{1}{2}(ay^2 - \underline{\mu'^T C \mu'} + \mu^T \Lambda \mu)}$$

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

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

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

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

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

$$\text{let } \lambda = a - a^T C^{-1} x \quad \xrightarrow{*} C = a x x^T + \lambda I$$

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

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

S Sherman Morrison Formula

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

$$\begin{aligned} & (\lambda + a \Phi \Phi^T) \left(I - \frac{\lambda^T a \Phi \Phi^T \lambda^{-1}}{I + a \Phi \lambda^{-1} \Phi} \right) \\ &= I + a \Phi \Phi^T \lambda^{-1} - \frac{a \Phi \Phi^T \lambda^T a \Phi \Phi^T \lambda^{-1}}{I + a \Phi \lambda^{-1} \Phi} - \frac{\lambda \lambda^T a \Phi \Phi^T \lambda^{-1}}{I + a \Phi \lambda^{-1} \Phi} \\ &= I + a \Phi \Phi^T \lambda^{-1} - \frac{a \Phi (\cancel{\lambda^T \lambda} + I) \Phi^T \lambda^{-1}}{\cancel{I + a \Phi \lambda^{-1} \Phi}} \\ &= I + a \Phi \Phi^T \lambda^{-1} - a \Phi \Phi^T \lambda^{-1} = I \end{aligned}$$

$$\begin{aligned} \lambda &= a - a^T \Phi^T C^{-1} \Phi \\ &= a - a^T \cdot \Phi^T \left(I - \frac{\lambda^T a \Phi \Phi^T \lambda^{-1}}{I + a \Phi \lambda^{-1} \Phi} \right) \Phi \\ \text{let } \alpha &= \Phi^T \lambda^{-1} \Phi \quad a \cdot \Phi \cdot \alpha \cdot \Phi^T = a \cdot \alpha \\ &= a - a^T \cdot \Phi^T \left(I - \frac{\lambda^T a \alpha}{I + a \alpha} \right) \Phi \\ &= a - a^T \left(\Phi^T \lambda^{-1} \Phi - \frac{\Phi^T \lambda^T a \alpha \Phi}{I + a \alpha} \right) \\ &= a - a^T \left(\alpha - \frac{a \alpha^2}{I + a \alpha} \right) - a - a^T \left(\frac{\alpha + a \alpha^2 - a \alpha^2}{I + a \alpha} \right) \\ &= a - \frac{a^T \alpha}{I + a \alpha} = \frac{a + a \alpha^2 - a \alpha^2}{I + a \alpha} \\ &= \frac{a}{I + a \alpha}, \quad \alpha = \Phi^T \lambda^{-1} \Phi \end{aligned}$$

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

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

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

$$\begin{aligned}
&= \mu^T \left(\frac{1}{\lambda} \alpha \lambda^T C^{-1} \Phi \right) * \text{A. } C \text{ is symmetric} \\
&= \mu^T \left(\frac{1+\alpha\lambda}{\lambda} \cdot \alpha \cdot \lambda \cdot \left(I - \frac{\lambda^T \alpha \alpha}{1+\alpha\lambda} \right) \cdot \Phi \right) \\
&= \mu^T \left((1+\alpha\lambda) \cdot \left(I - \frac{\alpha \lambda}{1+\alpha\lambda} \right) \cdot \Phi \right) \\
&= \mu^T \left(I - \frac{\alpha \lambda}{1+\alpha\lambda} + \alpha \lambda - \frac{\alpha^2 \lambda^2}{1+\alpha\lambda} \right) \cdot \Phi \\
&= \mu^T \left(I + \alpha \lambda - \frac{\alpha \lambda (1+\alpha\lambda)}{1+\alpha\lambda} \right) \cdot \Phi \\
&= \mu^T (I + \alpha \lambda - \alpha \lambda) \cdot \Phi \\
&= \mu^T \Phi
\end{aligned}$$

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

μ : mean of prior

Fully Bayesian: Maximum every possible of a line.

Kernel $\Phi \Phi^T$

§ Decision Theory.

- Regression → classification
 $(+ \text{Logistic regression}) \rightarrow$
- EM.
 supervised learning
 + every data has label
- clustering
 \Rightarrow non-supervised. (No label)
- Dimension Reduction.

■ Estimator. (bias/variance)

A statistic to approximate the property of a distribution.

$$\text{Eq. } \hat{\mu}_{\text{MLE}} = \frac{1}{n} \sum x_i$$

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

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

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

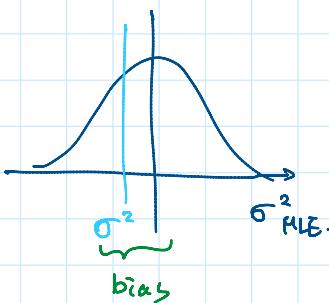
$$\hat{\sigma}^2_{\text{Alt}} = \frac{1}{n-1} \sum (x_i - \bar{x})^2 \Rightarrow \text{unbias}$$

• Bias

Mean of estimator different to real.

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

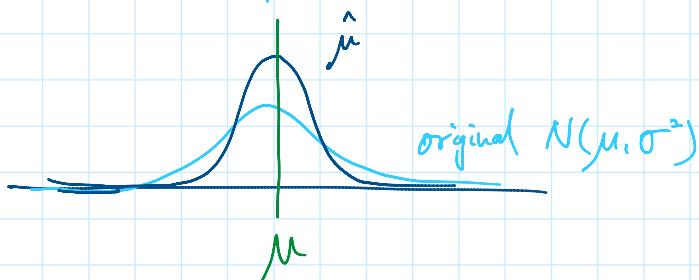
$$\text{Eq. } E(\hat{\sigma}^2_{\text{MLE}}) - \sigma^2_{\text{MLE}}$$



$$\begin{aligned} E(\hat{\mu}) &= E\left(\frac{1}{n} \sum x_i\right) \\ &= \frac{1}{n} \sum E(x_i) \xrightarrow{\text{as } n \rightarrow \infty} \mu \\ &= \frac{1}{n} \cdot n\mu = \mu. \end{aligned}$$

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

$$\begin{aligned} \text{Var}(\hat{\mu}) &= \text{Var}\left(\frac{x_1 + x_2 + \dots + x_n}{n}\right) \\ &= \frac{1}{n^2} \text{Var}(x_1, \dots, x_n) \\ &= \frac{1}{n^2} (\sigma^2 \cdot n) = \frac{\sigma^2}{n} \end{aligned}$$



* $n \rightarrow \infty \quad \text{Var}(\hat{\mu}) \rightarrow 0$
 $\Rightarrow \hat{\mu}$ is a single number.

$$\begin{aligned} &E(\sigma_{\text{ME}}^2) \\ &= E\left(\frac{1}{n} \sum (x_i - \hat{\mu})^2\right) \quad \text{dummy variable...} \\ &= E\left(\frac{1}{n} \sum ((x_i - \mu) - (\hat{\mu} - \mu))^2\right) \\ &= E\left(\frac{1}{n} \sum ((x_i - \mu)^2 - 2(x_i - \mu)(\hat{\mu} - \mu) + (\hat{\mu} - \mu)^2)\right) \\ &= E\left(\frac{1}{n} \sum (x_i - \mu)^2 - \frac{2}{n} \sum (x_i - \mu)(\hat{\mu} - \mu) + \frac{1}{n} \sum (\hat{\mu} - \mu)^2\right) \\ &= E\left(\underbrace{\frac{1}{n} \sum (x_i - \mu)^2}_{\text{Var}(x_i)} - \underbrace{\frac{2}{n} \sum (x_i - \mu)(\hat{\mu} - \mu)}_{E((\hat{\mu} - \mu)(x_i - \mu))} + E\left(\frac{n}{n} (\hat{\mu} - \mu)^2\right)\right) \\ &= \underbrace{\frac{1}{n} E((x_i - \mu)^2 + (x_2 - \mu)^2 + \dots)}_{\frac{1}{n} \sum (x_i - \mu)^2} - E(2(\hat{\mu} - \mu) \cdot \underbrace{\frac{1}{n} \sum (x_i - \mu)}_{\frac{\sum x_i}{n}}) + E((\hat{\mu} - \mu)^2) \\ &= \underbrace{\frac{1}{n} (\text{Var}(x_1) + \text{Var}(x_2) + \dots)}_{\frac{1}{n} \cdot n \cdot \text{Var}(x_i)} - E(2(\hat{\mu} - \mu) \cdot \left(\frac{\sum x_i}{n} - \frac{n}{n} \mu\right)) + E((\hat{\mu} - \mu)^2) \\ &= \underbrace{\frac{1}{n} \cdot n \cdot \text{Var}(x_i)}_{\sigma^2} - E(2(\hat{\mu} - \mu)(\hat{\mu} - \mu)) + E((\hat{\mu} - \mu)^2) \\ &= \underbrace{\sigma^2}_{\sigma^2} - \underbrace{2E((\hat{\mu} - \mu)^2)}_{E((\hat{\mu} - \mu)^2)} + E((\hat{\mu} - \mu)^2) \\ &= \sigma^2 - \boxed{E((\hat{\mu} - \mu)^2)} \xrightarrow{\text{as } n \rightarrow \infty} \text{Var}(\hat{\mu}) = \frac{\sigma^2}{n} \\ &= \sigma^2 - \frac{\sigma^2}{n} \end{aligned}$$

$$= \frac{n}{N} \sigma^2 \neq \sigma^2 \text{ Bias!}$$

■ $E(\hat{\sigma}_{\text{att}}^2)$

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

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

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

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

$$\begin{aligned} & \text{brace under } E\left(\frac{1}{n-1} \sum x_i^2\right) \quad \frac{1}{n-1} \sum x_i^2 = \frac{n}{n-1} \sigma^2 \\ & \frac{n}{n-1} \cdot \frac{\sigma^2}{n} = \frac{\sigma^2}{n-1} \end{aligned}$$

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

■ Mean Square Error (MSE)

$$E((\hat{\theta} - \theta)^2) \xrightarrow{\substack{\text{property of data (general truth)} \\ \rightarrow \text{usually don't know}}}$$

$\hat{\theta}$ $\xrightarrow{\substack{\text{estimator} \\ \rightarrow \text{random variable.}}}$

- MSE is error from general truth
- LSE is error from data

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

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

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

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

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

!!

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

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

* for unbiased estimator

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

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

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

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

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

* if bias is indeed 0
 ⇒ all error come from Variance.
 ⇒ trade-off

- Bias \propto flexibility / Generalisity.
- Vars \propto sensitivity to training data

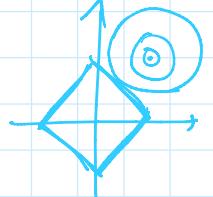
② Complexity of model

$$\Phi = \phi_1(x) \dots \phi_k(x)$$

feature .

⇒ ① dimension Reduction .

② Regularization - Ex: L1 norm



* low bias \Leftrightarrow overfitting. (more complicated)

— usually when:

- non-linear .
- non parametric (no assumption)
- k-nn .

* low Variance \Leftrightarrow underfitting (simple)

- Regression. (with regularization)

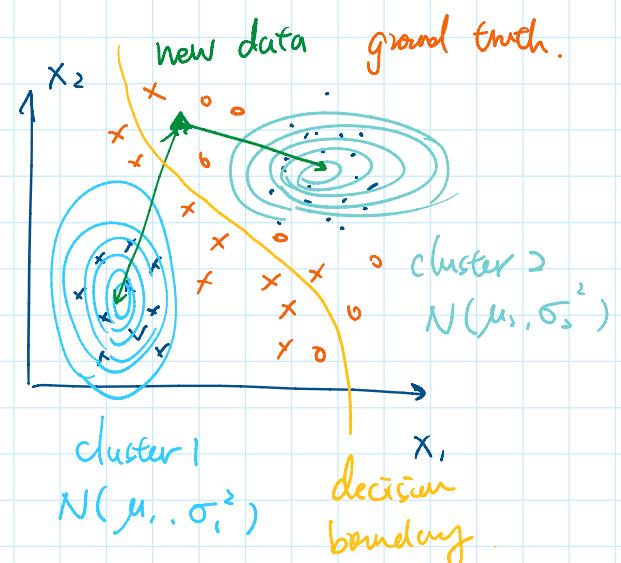
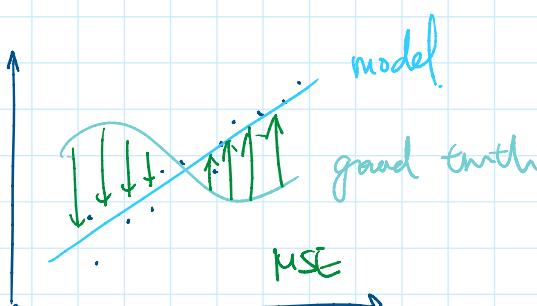
- Naive Bayes.
- Linear model.

Bias - variance trade off

III MSE

How much error that a model makes.

{ continuous target
 discrete target



IV Contingency table (Confusion matrix)

		$\theta=Y$	$\theta=N$	sum
		True positive	False Negative	Yes
		False Positive	True Negative	No
sum	positive	negative	total	

$$\text{Specificity.} = \frac{TP + TN}{\text{Total}}$$

$$\text{Sensitivity} = \frac{TN}{N_0} = \frac{TN}{FP + TN}$$

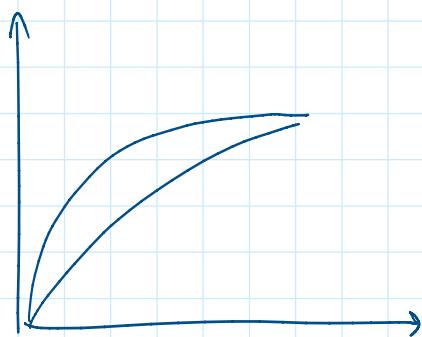
(Recall / Power)

Accuracy. =

False positive rate. =

Precision =
(PPV)

R.O.C (A.U.C)



§ Regression v.s. Classification.

Supervised Learning

D_R η_0 weight

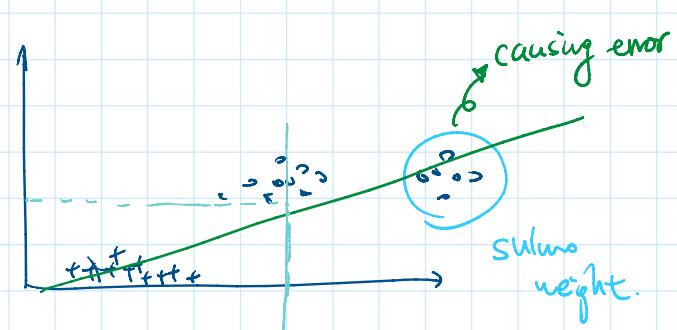
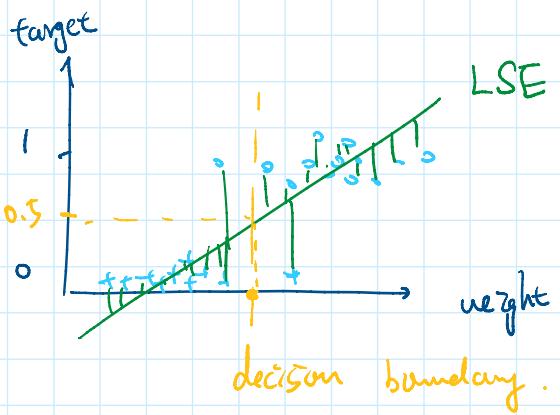
+ : female o : male .

het male : (→

$$\text{female : } 0 \quad f(+1) = 0$$

$$f(+1) = 0$$

indicator function.



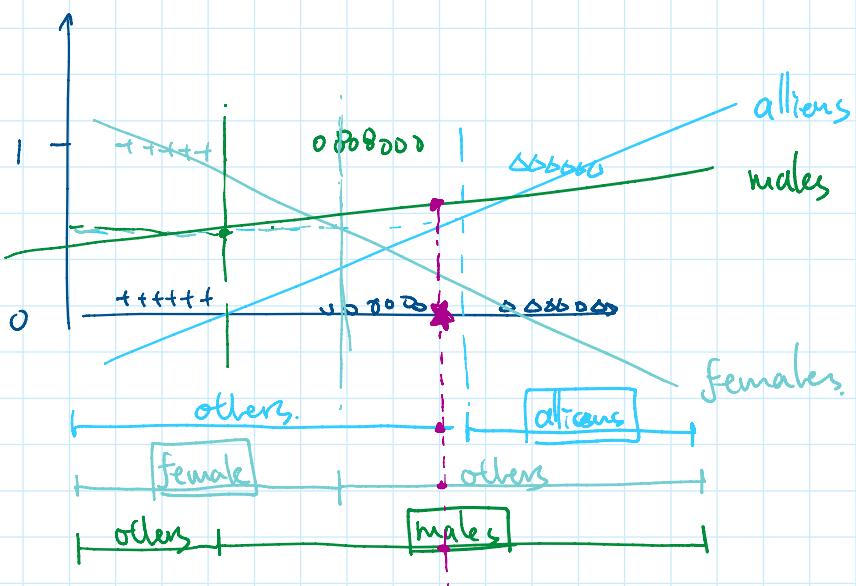
To solve \Rightarrow using other loss function

§ Loss Function.

⦿ multi class

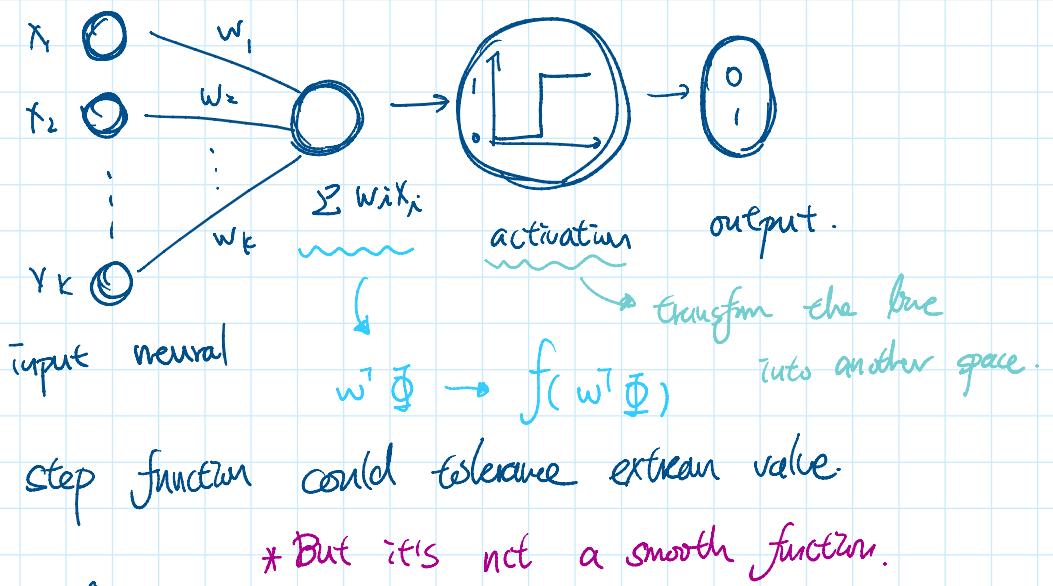
* one of k coding.

o male	1	0	0
+ female	0	1	0
△ alien	0	0	1

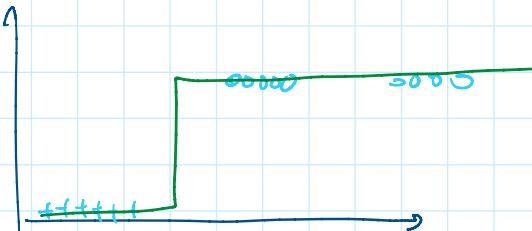


⦿ Fisher's linear discriminant.

- Perception algorithm (simple neural network)



- * Step function could tolerate extreme value.
- * But it's not a smooth function.



§ Gradient descent.

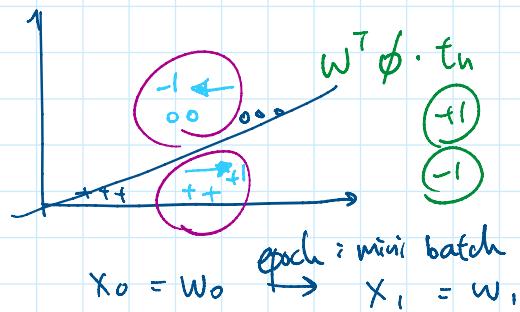
Newton's method 2nd order
 (Taylor's expansion) approximation
 Gradient descent 1st order.

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

γ
learning rate.

- Perception criterion

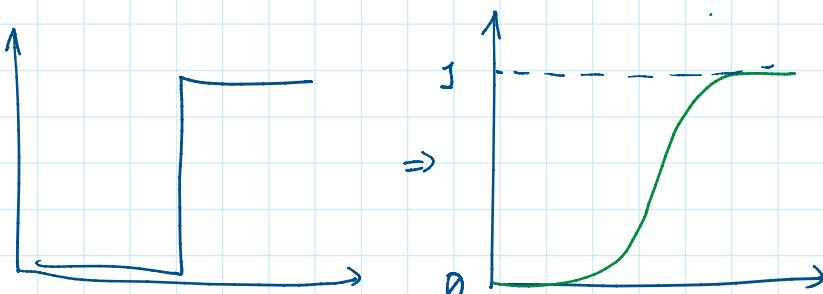
- Correct prediction gives no error
- use misclassified data to update the line



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

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

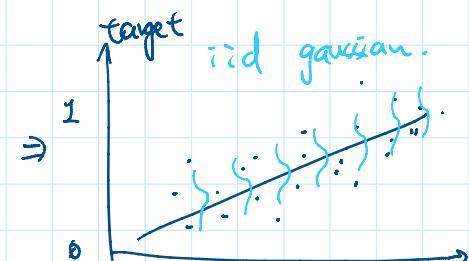
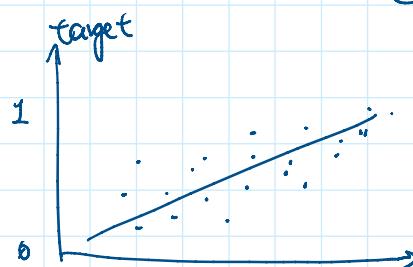
Logistic Regression



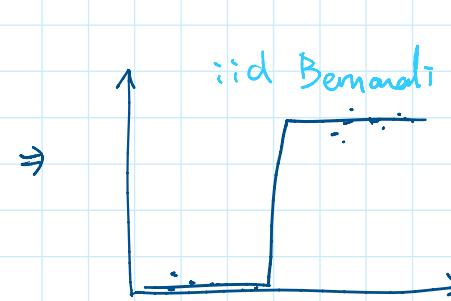
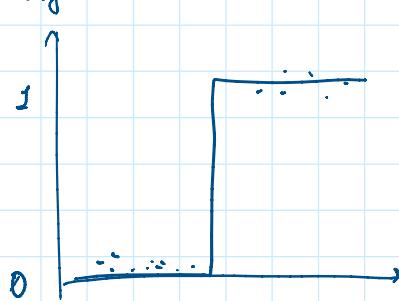
$$f(x) = \frac{1}{1+e^{-x}}$$

probit function
CDF of gaussian.

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



\Rightarrow MLE. MAP...



Given $D = (\underline{x}_i, \underline{y}_i)$

$y_i \sim \text{Bernoulli}(f(w^T \phi))$ $B(x|p) = p^x(1-p)^{1-x}$

[MLE]

$$\text{likelihood} := \prod \text{Bernoulli}(y_i | f(w_i \phi))$$

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

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

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

$$\begin{aligned} \frac{\partial}{\partial w_j} -\log P(D|w) & \rightarrow [w_0 \dots w_k] \begin{bmatrix} x_0 \\ \vdots \\ x_k \end{bmatrix} \\ \frac{\partial}{\partial w_j} \left(-\log \frac{1}{1+e^{-w^T x_i}} \right) & = w_0 x_0 + \dots + w_k x_k \end{aligned}$$

$$\text{let } u = 1+e^v, v = -w^T x_i$$

$$\begin{aligned} \frac{\partial}{\partial w_j} \left(\log \frac{1}{1+e^{-w^T x_i}} \right) &= \frac{\partial (-\log u)}{\partial u} \cdot \frac{\partial u}{\partial v} \cdot \frac{\partial v}{\partial w_j} \\ &= \frac{1}{1+e^{-w^T x_i}} \cdot e^{-w^T x_i} \cdot x_{ij} \quad \text{--- ①} \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial w_j} \left(\log \frac{e^{-w^T x_i}}{1+e^{-w^T x_i}} \right) &= \frac{\partial}{\partial w_j} \left(\log e^{-w^T x_i} - \log 1+e^{-w^T x_i} \right) \\ &= -x_{ij} + x_{ij} \frac{e^{-w^T x_i}}{1+e^{-w^T x_i}} \\ &= -x_{ij} \cdot \left(\frac{1}{1+e^{-w^T x_i}} \right) \quad \text{--- ②} \end{aligned}$$

from ①. ②.

$$\begin{aligned}
 & \frac{\partial}{\partial w_j} - \text{by } P(D|w) \\
 &= - \sum \left[y_i x_{ij} \left(\frac{e^{-w^T x_i}}{1+e^{-w^T x_i}} \right) - ((-y_i) x_{ij}) \left(\frac{1}{1+e^{-w^T x_i}} \right) \right] \\
 &= - \sum \left[y_i x_{ij} \left(1 - \frac{1}{1+e^{-w^T x_i}} \right) - ((-y_i) x_{ij}) \left(\frac{1}{1+e^{-w^T x_i}} \right) \right] \\
 &= - \sum \left[y_i x_{ij} - \frac{y_i x_{ij}}{1+e^{w^T x_i}} - \frac{x_{ij}}{1+e^{w^T x_i}} + \frac{y_i x_{ij}}{1+e^{w^T x_i}} \right] \\
 &= \sum x_{ij} \left(\frac{1}{1+e^{w^T x_i}} - y_i \right) = 0
 \end{aligned}$$

* Recall: Newton's Method

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

$$\begin{aligned}
 \nabla f &= \begin{bmatrix} \frac{\partial J}{\partial w_1} \\ \vdots \\ \frac{\partial J}{\partial w_k} \end{bmatrix} \quad \sum_w x_{ij} \left(\frac{1}{1+e^{w^T x_i}} - y_i \right) \\
 &\Phi = \begin{bmatrix} x_{11} & \dots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nd} \end{bmatrix}_{n \times d} \\
 &\phi_1, \dots, \phi_d
 \end{aligned}$$

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

$$H = \begin{bmatrix} \frac{\partial^2 J}{\partial w_1 \partial w_1} & \frac{\partial^2 J}{\partial w_1 \partial w_2} & \dots & & \\ \frac{\partial^2 J}{\partial w_2 \partial w_1} & \frac{\partial^2 J}{\partial w_2 \partial w_2} & & & \\ \vdots & & \ddots & & \\ & & & \frac{\partial}{\partial w_k} \left[\sum x_{ij} \left(\boxed{\frac{1}{1+e^{-w^T x_i}}} - y_i \right) \right] & \end{bmatrix}$$

$$\begin{aligned}
 & \frac{\partial}{\partial w_k} \left(1 + e^{-w^T x_i} \right)^{-1}, \quad u = 1 + e^{-w^T x_i} \\
 &= \frac{\partial u^{-1}}{\partial u} \cdot \frac{\partial u}{\partial v} \cdot \frac{\partial v}{\partial w_k} \quad v = w_k x_{ik} \\
 &= -u^{-2} \cdot (-e^{-w_k x_{ik}}) \cdot x_{ik} \\
 &= x_{ik} \frac{e^{-w_k x_{ik}}}{(1 + e^{-w_k x_{ik}})^2} \\
 \Rightarrow & \sum_{k=1}^K \left(x_{ik} \cdot \frac{e^{-w_k x_{ik}}}{(1 + e^{-w_k x_{ik}})^2} \right) \\
 & A^T \cdot D \cdot A \\
 & \left[\begin{array}{c} \vdots \\ \vdots \end{array} \right] \left[\begin{array}{ccc} 0 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 0 \end{array} \right] \left[\begin{array}{c} \vdots \\ \vdots \end{array} \right]
 \end{aligned}$$

§ Knn : k - nearest neighbor

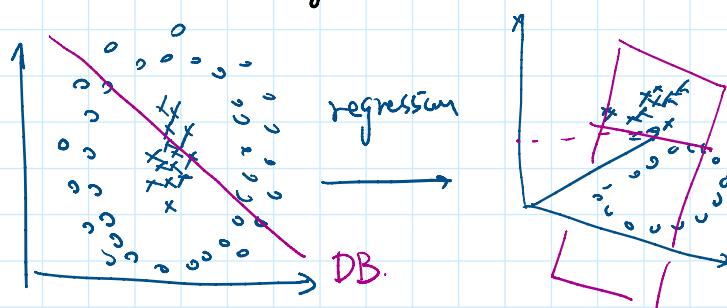
Knn : classification

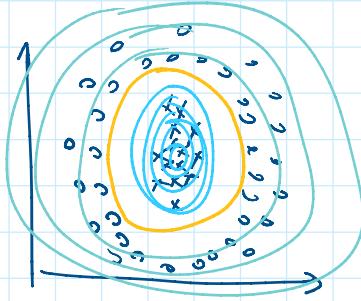
- low bias \Rightarrow overfitting.

Kmeans : clustering.

- high variance.

§ Decision boundary.





Naive Bayes classifier

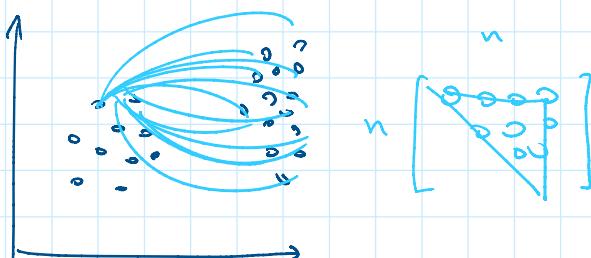
$$N(\mu_i, \Sigma_i) = N(\bar{\mu}, \Sigma)$$

§ Clustering. (unsupervised learning)

Regression
classification
supervised learning.

complete data
 \Rightarrow ~~forget~~ known
~~labels~~

- hierarchical clustering. (k_{nn})



§ E M algorithm.

E

expectation

M

maximization

For complete data :

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

$$\hat{\theta}_{MLE} = \frac{\sum x_i}{n} = \frac{4}{6} = \bar{B}(\theta)$$

— 2 coins flipping.

$$\{C_0, C_1\} \rightarrow \begin{cases} \lambda \cdot P_0 \\ (1-\lambda) \cdot (1-P_1) \end{cases}$$

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

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

(likelihood: $p(X, Z | \theta)$, $\theta = \underline{P_0}, \underline{P_1}, \underline{\lambda}$)

C_0 head 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 likelihood:

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

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

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

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

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

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

$$P_0 = \frac{P_{X_1=0} - P_{X_1=0} z_i}{n - \sum z_i} = \frac{5}{6}$$

- Incomplete data.

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

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

EM algorithm * responsibility -

$$Z = \{ 0, 1 \} \xrightarrow{\text{approximate}}$$

$$\underbrace{w}_{\text{weight}} = \{ 0.5, 0.5 \}$$

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

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

Step 1 initialize all parameters.

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

Step E calculate responsibility.

$$w_0 = \frac{P(z_i=0, x_i | \theta)}{P(z_i=0, x_i | \theta) + P(z_i=1, x_i | \theta)}$$

$$w_1 = \frac{P(z_i=1, x_i | \theta)}{P(z_i=0, x_i | \theta) + P(z_i=1, x_i | \theta)}$$

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

$$w_0 = \{ - \cdot - - - - - \}$$

$$[w_1 = \{ \textcircled{0} - - - - - \}]$$

$$1-w_0$$

\Rightarrow using w_0 instead of Z .

* W is a posterior.

$$W = \frac{P(z_i, x_i | \theta)}{\sum P(z_i, x_i | \theta)} = \frac{P(z_i, x_i | \theta)}{P(x_i | \theta)} = P(z_i | x_i, \theta)$$

Step M update parameters.

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

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

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

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

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

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

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

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

$$\text{E step: } P(Z=0, X_i=1 | \theta) = \frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$$

$$P(Z=0, X_i=0 | \theta) = \frac{1}{3} \times \frac{2}{3} = \frac{2}{9}$$

$$P(Z=1, X_i=1 | \theta) = \frac{2}{3} \times \frac{2}{3} = \frac{4}{9}$$

$$P(Z=1, X_i=0 | \theta) = \frac{2}{3} \times \frac{1}{3} = \frac{2}{9}$$

when $X_i = 1$

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

when $X_i = 0$

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

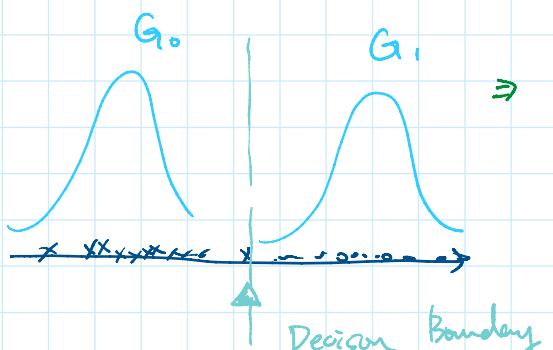
M step:

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

$$P_0^{(1)} = \frac{\sum w_i X_i}{\sum w_i} = \frac{4 \times \frac{1}{5}}{4 \times \frac{1}{5} + 6 \times \frac{1}{2}} = \frac{0.8}{3.8} = \frac{4}{19}$$

$$P_1^{(1)} = \frac{\sum X_i - \sum X_i w_i}{n - \sum w_i} = \frac{4 - 0.8}{10 - 3.8} = \frac{3.2}{6.2} = \frac{16}{31}$$

§ Gaussian Mixture Model (GMM)



$\left\{ \begin{array}{l} \lambda \Rightarrow G_0 \\ \mu_0 (\bar{\mu}_0) \\ \mu_1 (\bar{\mu}_1) \\ \sigma_0^2 (\Sigma_0) \\ \sigma_1^2 (\Sigma_1) \end{array} \right\}$ assume to be 1 (I)

Likelihood

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

Log likelihood

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

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

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

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

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

mean of x_i

Complete \rightarrow Incomplete

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

$$w_0 > w_1, w_1 > w_0$$

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

$$w_0 = \{\dots\}$$

$$w_1 = \{\dots\}$$

Using w_0, w_1 , generate new Z .

and update the parameters with new Z .

\Rightarrow k means clustering

of cluster., here $k=2$

