Supervised learning
- Classification
- Regression

} Generative
} Discriminative.
└ you need large amount of data for this

Features → $X^{(1)}, X^{(2)}, \ldots \longrightarrow \overline{X}$

Labels → {0, 1}
└ classification problem

Sample = $\{\overline{x_i}, y_i\}$.

Labels ⇒ $\in \mathbb{R}$
└ Regression problem.

Generative model
- probabilistic model.

underlying assumption (x, y) is from some distribution $\mathbb{P}$.
$(X, Y) \sim \mathbb{P}$

Simplest model
Classification problem.
Ex: Credit default problem.

$Y = \{0, 1\}$

We will assign some prior probabilities
$P_0 \triangleq P(Y=0) = 0.8$
$P_1 \triangleq P(Y=1) = 0.2$
} estimated from historical data.

$h_0 \rightarrow$ loan to everybody

$h_1 \rightarrow$ loan to noone.

$h_2(z) \rightarrow \begin{cases} 0 & \text{if } z=0 \\ 1 & z=1 \end{cases}$ $\quad P(z=1) = 0.2$

How to compare h_0, h_1, h_2?

Loss $\Rightarrow L(h, Y)$

0-1 Loss,

$L(1,1) = L(0,0) = 0$

$L(1,0) = L(0,1) = 1$

$$E_y(L(h_0, Y)) = \sum_y L(0,y) P_y$$

$$= L(0,0) P_0 + L(0,1) P_1$$

$$= \underline{0.2}$$

$$E_y(L(h_1, Y)) = \sum L(1,y) P_y$$

$$= \underline{0.8}$$

$E_{z,y}(L(h(z), Y)) =$

$L(0,0) P(z=0, Y=0) + L(1,0) P(z=1, Y=0) +$
$L(0,1) P(z=0, Y=1) + L(1,1) P(z=1, Y=1)$

$z \perp y$ (independant)

$= 1(P(z=1)) P(y=0) + P(z=0) P(y=1)$

$= (1-\delta) 0.2 + \delta (0.8)$

Best value here $\delta = 0$

$E(h(z), Y) = 0.2$

"Simply creating a new source of randomness might not help you" :(

$$X = \begin{cases} 1 \rightarrow \text{late tax payer} \\ 0 \rightarrow \text{not a late tax payer} \end{cases}$$

$p_0 = P(Y=0) = 0.8$

$p_1 = P(Y=1) = 0.2$

Class conditional density (likelihood).

$f_1(x) = P(X=x \mid Y=1)$

$f_0(x) = P(X=x \mid Y=0)$

$f_1(1) = P(X=1 \mid Y=1) = 0.95$

$f_0(1) = P(X=1 \mid Y=0) = 0.1$

Setup 2

$X = 1 \text{ or } 0$

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

$h(0) = ?$ $\quad$ $h(1) = ?$

X	h_1	h_2	h_3	h_4
0	0	0	1	1
1	1	0	1	0

find the loss for everything ? find minimum loss that would be h.

$$E\left(L\left(h_f(\underline{x}), \underline{y}\right)\right)$$

$$= \sum_x \sum_y L(h_f(x), y))\, P(x,y)$$

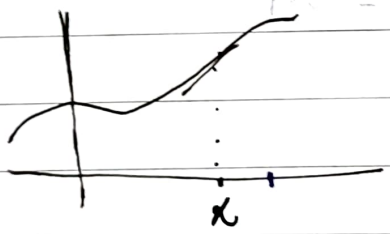
$$= L(1,0)\, P(0,0) + L(1,1)\, P(0,1)$$
$$L(0,0)\, P(1,0) + L(1,0)\, P(1,1)$$

$$= P(0,0) + P(1,1)$$
$$= P(Y=0)\, P(X=0 \mid Y=0) + P(Y=1)\, P(X=1 \mid Y=1)$$
$$= (0.8)\,(0.9) + 0.2\,(0.95)$$
$$= 0.72 + 0.190$$
$$= 0.91$$

Huge Loss!! → Bad classifier

Lecture 2 — 10|1|23

Derivative



The derivative is the slope.

$$f'(x) = \lim_{h \to 0} \left[\frac{f(x+h) - f(x)}{h}\right]$$

$$f(x+h) \cong f(x) + f'(x)\,h \quad \text{(approximate as a line)}$$

Generalisation

$$\vec{f}(\vec{x} + \vec{h}) = f(\vec{x}) + T_f(h)$$

$\vec{f}$ → is a vector valued fn

$\mathbb{R}^n \longrightarrow \mathbb{R}^m$

$\vec{f} : (f_1, f_2, \cdots, f_m) \quad f_i :: \mathbb{R}^n \to \mathbb{R}$

$T_f(h) \to$ linear transformation

special kind of transformation satisfying the following conditions:

1) $T(x+y) = T(x) + T(y)$
2) $T(\alpha x) = \alpha T(x)$

consider a vector in $\mathbb{R}^n$

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \alpha_1 \overbrace{\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}}^{e_1} + \alpha_2 \overbrace{\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}}^{e_2} + \alpha_3 \overbrace{\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}}^{e_3}$$

$$T\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \alpha_1 T\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \alpha_2 T\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + \alpha_3 T\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

$T(e_1) = e_2 \quad T(e_2) = e_3 \quad T(e_3) = e_1$

Here T is a simple rotation.

$$T\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = 1\,T(e_1) + 2\,T(e_2) + 3\,T(e_3)$$

$$= 1\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + 2\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + 3\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

$$= \begin{pmatrix} 3 \\ 1 \\ 2 \end{pmatrix}$$

This is same as.

$$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \\ 2 \end{pmatrix}$$

$$\begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & & \frac{\partial f_m}{\partial x_n} \end{pmatrix}_{m \times n} \qquad h \to n \times 1$$

T_f

$$f(x+h) = f(x) + T_f h$$

Rules of differentiation.

1. $h = f + g$

$f: \mathbb{R}^n \to \mathbb{R}^m \qquad g: \mathbb{R}^n \to \mathbb{R}^m$

$$T(h(x)) = T(f(x)) + T(g(x))$$

2. Chain Rule:

$f :: \mathbb{R}^n \to \mathbb{R}^k$

$g :: \mathbb{R}^k \to \mathbb{R}^m$

$$h = g(f(x)) \qquad h :: \mathbb{R}^n \to \mathbb{R}^m$$

3. $T_h = T_g(f(x)) \cdot T_f(x)$

composition upon is nothing but matrix multiplication.

$T_1 = \mathbb{R}^n \to \mathbb{R}^k \longrightarrow (k \times n)$ matrix

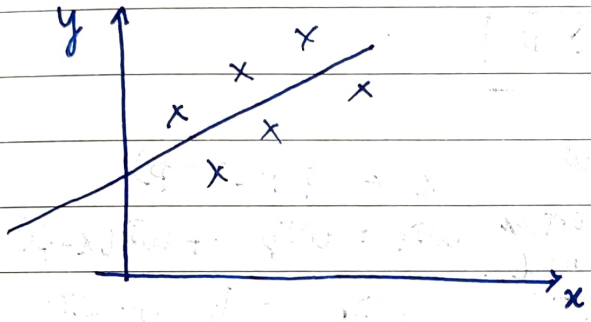
$T_2 = \mathbb{R}^k \to \mathbb{R}^m \longrightarrow (m \times k)$ matrix

Ex: $f :: \mathbb{R}^2 \longrightarrow \mathbb{R}^3$

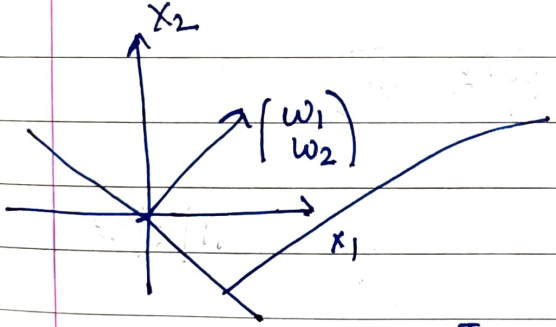
$$f(x, y) = (\underbrace{x^2+y}_{f_1}, \underbrace{xy}_{f_2}, \underbrace{x+y}_{f_3})$$

"Chain Rule is matrix multiplication"

A Simple Neural Network - Linear Regression



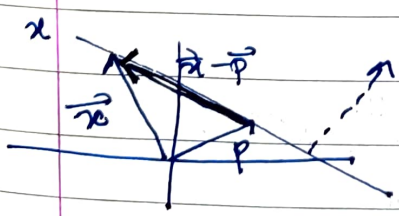
$$y = w^T x + b$$



this line is the collection of vectors st.

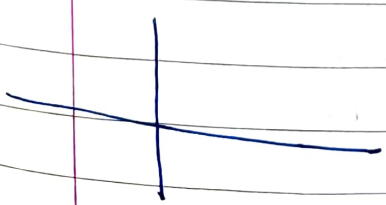
$$\boxed{w^T x = 0}$$

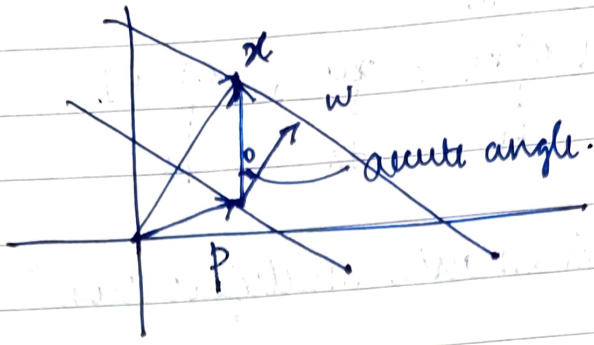
$$w^T x \Rightarrow \|w\| \|x\| \cos\theta$$



$$w^T x = w^T (p + x - p)$$
$$= b + w^T (x - p)$$

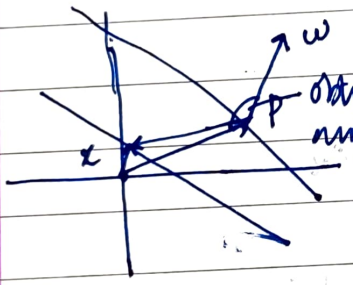
$$\boxed{w^T x = b}$$





$$x = p + (x - p)$$
$$w^T x = w^T p + w^T (x - p)$$
$$w^T x \not= b + \downarrow > 0$$

$$\boxed{w^T x > b}$$



$$x = p + (x - p)$$
$$w^T x = w^T p + w^T (x - p)$$
$$w^T x = b + < 0$$

$$\boxed{w^T x < b}$$

Lecture 3 — 11/1/23

Regression problem

Given the feature $\bar{X} = X^{(1)}, X^{(2)}, \ldots$
we need to predic $Y \in \mathbb{R}$

Ex: $X \rightarrow$ features of a person
$Y \rightarrow$ family income.

Through linear regression, we aim to find a line, plane, & hyperplane. that classifies the problem.
(1D, 2D, 3D)

$w^T x = b \longrightarrow$ eqn of the line

$Y = w^T x + w_0 \longrightarrow$ (Rearrange)

$$\begin{cases} \overline{w_1} x + \overline{w_2} Y + b = 0 \\ Y = \dfrac{-\overline{w_1}}{w_2} X - \dfrac{b}{w_2} \end{cases}$$

Goal:

Input $\{x_i, Y_i\}$

$x_i^{(1)}, \ldots, x_i^{(d)}, \quad Y \in \mathbb{R}$

Find $w = w_1, \ldots, w_d, w_0$

$w^T x_i + w_0 \approx Y_i$ (

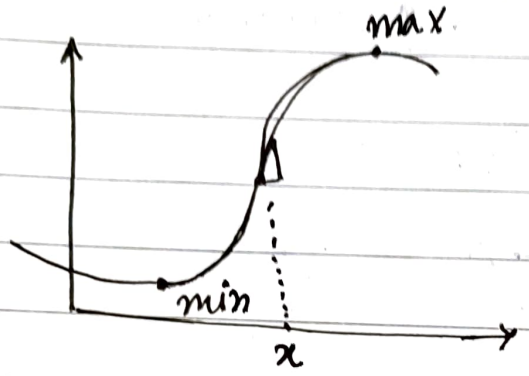
↳ How do we measure this?
↳ Using the loss fn.

- the standard used to approximate the error is the root mean square error.
- $L(w_0, \ldots, w_d) = \frac{1}{n} \sum_{i=1}^{n} (w^T x_i + w_0 - Y_i)^2$

↳ mean squared error (MSE).

$$\min_{w} \frac{1}{n} \sum^{} (w^T x_i + w_0 - y_i)^2$$

↳ unconstrained optimization problem

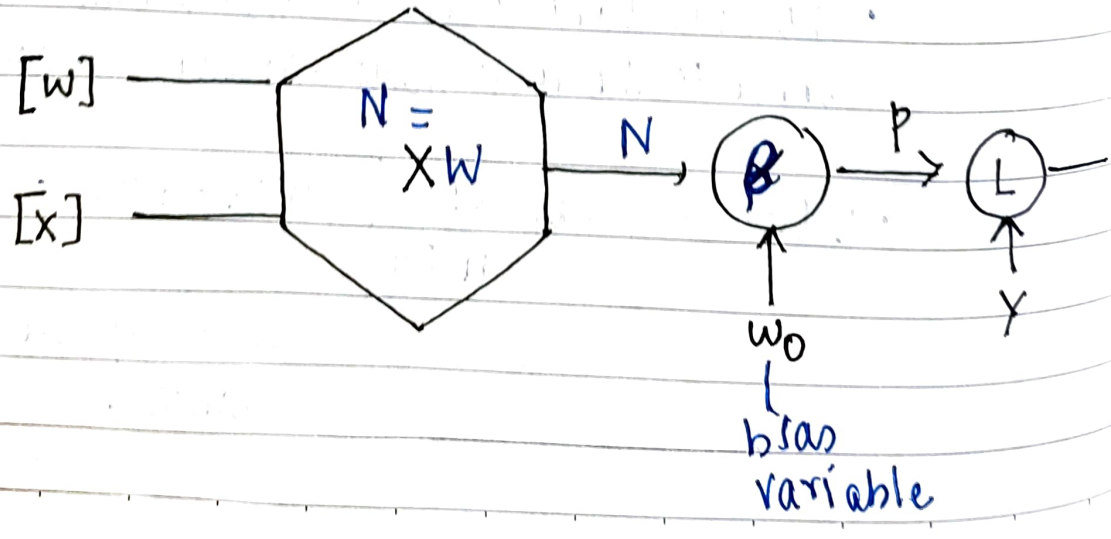


$x + \eta f'(x) \rightarrow$ increase
$x - \eta f'(x) \rightarrow$ decrease
} in 2D space.

In General, going in the direction of increasing gradient will increase y.

So to minimise the loss, we will move in the direction of decreasing gradient.

A simple Neural Network model (linear regression)



$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{n3} \end{bmatrix} \qquad W = \begin{bmatrix} w_{11} \\ w_{21} \\ w_{31} \end{bmatrix}$$

$$N = XW \qquad P = XW + w_0$$

$$L = \sum (P_i - Y_i)^2$$

$$L : \mathbb{R}^n \times \mathbb{R}^{\#} \longrightarrow Y$$

$$L(P, Y) = \frac{\partial L}{\partial P_1} \frac{\partial L}{\partial P_2} \cdots \frac{\partial L}{\partial P_n}$$

$$\frac{\partial L}{\partial P_i} = 2 (P_i - Y_i)$$

$$P : \mathbb{R}^n \times \mathbb{R} \longrightarrow \mathbb{R}^n$$

$$\downarrow \qquad \downarrow$$

$$(N_1, \cdots, N_n) \quad w_0$$

$$P = (N_1 \cdots N_n\}, w_0) + (P_1 \cdots P_N)$$

$$P_i = N_i + w_0$$

$$\frac{\partial L}{\partial N_i} = \frac{\partial L}{\partial P_i} \cdot \underbrace{\frac{\partial P_i}{\partial N_i}}_{1} = \frac{\partial L}{\partial P_i}$$

$$\frac{\partial L}{\partial w_0} = \frac{\partial L}{\partial P_1} \frac{\partial P_1}{\partial w_0} + \frac{\partial L}{\partial P_2} \frac{\partial P_2}{\partial w_0} + \cdots + \frac{\partial L}{\partial P_n} \frac{\partial P_n}{\partial w_0}$$

$$\frac{\partial L}{\partial w_0} = \sum_{i=1}^{n} \frac{\partial L}{\partial P_i}$$

$$X : \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ \vdots & & \\ x_{n1} & x_{n2} & x_{n3} \end{bmatrix}$$

$$W : \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \end{bmatrix} \rightarrow \text{parameters}$$

Linear regression is a simple neural network, here W will be an $n \times 1$ matrix, but generally it will be an $n \times l$ matrix (l is the number of layers)

$$N : \begin{bmatrix} N_{11} & \cdot \\ N_{21} & \cdot \\ N_{n1} & N_{n2} \end{bmatrix}$$

$$\begin{bmatrix} \frac{\partial L}{\partial w_{11}} & \frac{\partial L}{\partial w_{12}} \\ \vdots & \\ \frac{\partial L}{\partial w_{31}} & \end{bmatrix} \quad \begin{bmatrix} \frac{\partial L}{\partial N_{11}} & \cdots & \frac{\partial L}{\partial N_{12}} \end{bmatrix}$$

$$\frac{\partial L}{\partial x_{11}} = \frac{\partial L}{\partial N_{11}} \cdot \frac{\partial N_{11}}{\partial x_{11}} + \frac{\partial L}{\partial N_{12}} \frac{\partial N_{12}}{\partial x_{11}}$$

$$\frac{\partial L}{\partial N_{11}} w_{11} + \frac{\partial L}{\partial N_{12}} w_{12}$$

Input to the backward pass :

$$\begin{bmatrix} \frac{\partial L}{\partial N_{11}} & \frac{\partial L}{\partial N_{12}} \\ \vdots & \\ \frac{\partial L}{\partial N_{n1}} & \end{bmatrix}_{n \times 2} \quad \begin{bmatrix} w_{11} & w_{21} & w_{31} \\ w_{12} & w_{22} & w_{32} \end{bmatrix}_{2 \times 3} \quad (W^T)$$

$$= \begin{bmatrix} \frac{\partial L}{\partial x_{11}} & \frac{\partial L}{\partial x_{12}} & \frac{\partial L}{\partial x_{13}} \\ \vdots & & \end{bmatrix}_{n \times 3}$$

Lecture 4

16/1/23

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{1d} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$ } first sample. (d features)

why $\frac{dL}{dx}$? Because we can have many layers later.

$w_1 x_1 + w_2$

$\frac{dL}{dw}$

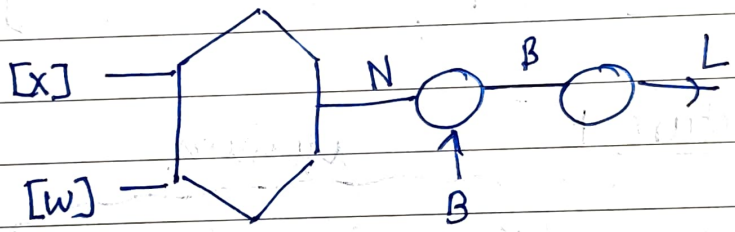
$$N_{11} = x_{11} w_1 + x_{12} w_{21} + \cdots$$

$$N_{21} = x_{21} w_{11} + x_{22} w_{21} + \cdots$$

$$W = \begin{bmatrix} w_{11} & w_{22} \\ w_{21} & \vdots \\ \vdots & \\ w_{d1} & w_{d2} \end{bmatrix} \qquad N = \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \\ \cdot & \cdot \\ \vdots & \vdots \end{bmatrix}$$

$$\frac{dL}{\partial w_{11}} = \frac{dL}{\partial N_{11}} \times \frac{dN_{11}}{dw_{11}} + \frac{dL}{dN_{21}} \frac{dN_{21}}{dw_{11}} + \cdots$$

$$= \frac{dL}{dN_{11}} x_{11} + \frac{dL}{dN_{21}} x_{21} + \cdots \frac{dL}{dN_{m1}} \frac{dN_{m1}}{dw_{11}}$$



$$\begin{bmatrix} \frac{dL}{dN_{11}} & \frac{dL}{dN_{12}} \\ \frac{dL}{dN_{21}} & \frac{dL}{dN_{22}} \\ \vdots & \vdots \\ \frac{dL}{dN_{m1}} & \frac{dL}{dN_{m2}} \end{bmatrix}$$

$$\begin{bmatrix} \frac{dL}{dw_{11}} & \frac{dL}{dw_{12}} \\ \frac{dL}{dw_{12}} & \\ \vdots & \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} \cdots x_{n1} \\ \vdots & \\ x_{13} & x_{n3} \end{bmatrix}_{d \times n} \begin{bmatrix} \frac{dL}{dN_{11}} & \frac{dL}{dN_{12}} \\ \vdots & \\ \frac{dL}{dN_{n1}} & \frac{dL}{dN_{n2}} \end{bmatrix}_{n \times 2}$$

"It is a very modular approach, no matter how many layers we have."

$$S_n = \{ (x_i, y_i) \}_{i=1}^{n}$$

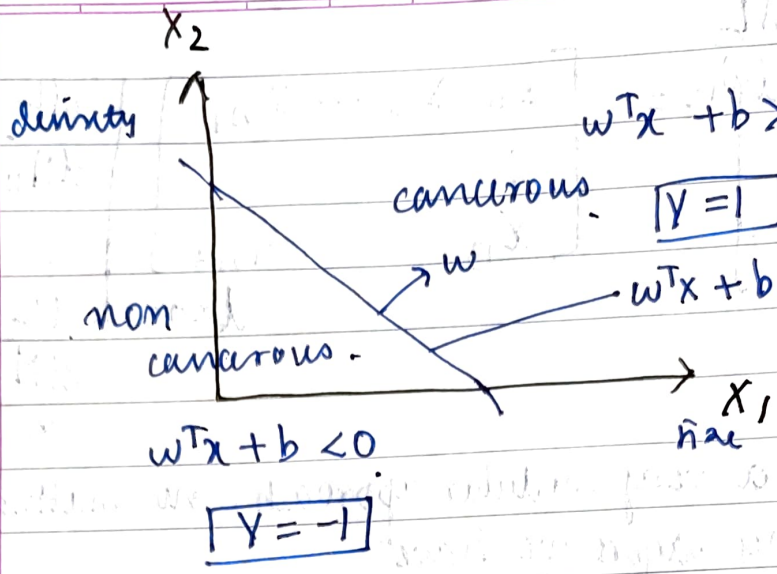
$FP(S_n)$ $BP(S_n)$ $\quad w^{new} = w^{old} - \alpha \frac{dL}{dw}$

→ If you use the entire training sample, it is called standard gradient descent.

→ If you take random batches of training sample, it is called stochastic gradient descent.

→ Practically, its been shown that stochastic GD works better.

Classification:
We can reuse the same model. ~~to~~ used in regression model to solve classification problem., you will only have to tweak the loss fn.



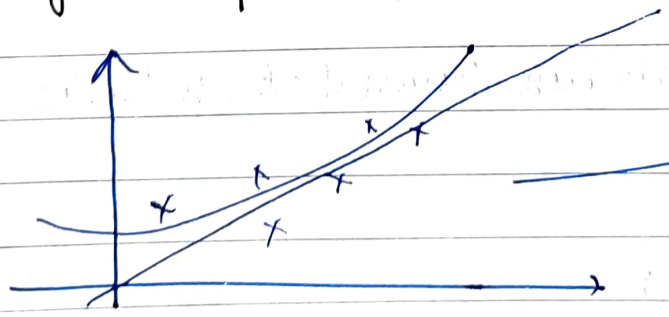
We want to find $(Y_i)(w^T x + b) \geq 0$

Linear model for classification is called as perceptron.

Loss is called as cross-entropy loss.

lecture - 5 17/1/23.

Logistic Regression



$w^T x + b \rightarrow$ range : $(-\infty, \infty)$

$w^T x + b \xrightarrow{f} [0,1]$

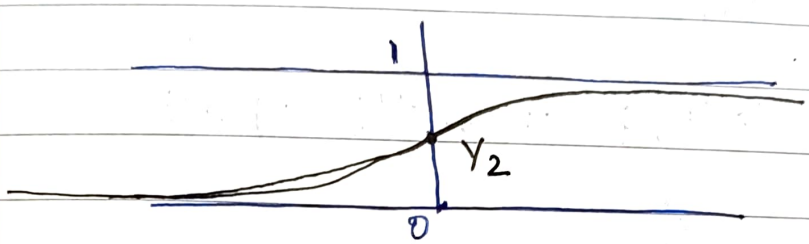
↳ we can treat the num as a probabilities

sigmoid function :

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

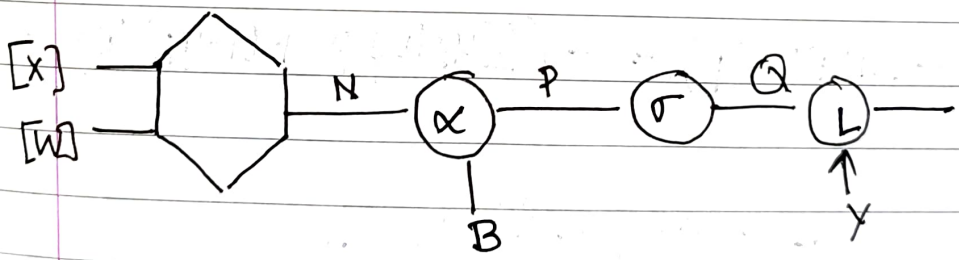
$z \to \infty \quad \sigma(z) \to 1$

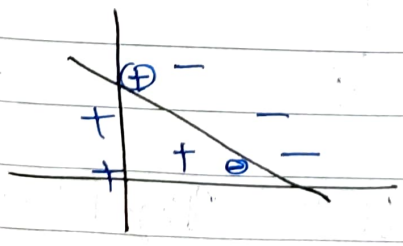
$z \to -\infty \quad \sigma(z) \to 0$



Idea $P(Y=1 \mid X) = \sigma(w^T x + b)$

↳ Basically if the value is > than some probability (threshhold), then we predict as 1





$q_i = \sigma(P_i)$

D = H H H T T H H

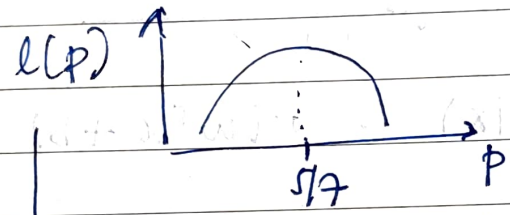
$P(D \mid p) = p^5 (1-p)^2$

$D = 1\,1\,1\,0\,0\,1\,1$
$y_1\ y_2\ y_3$

the i^{th} value is $y_i \Rightarrow \{0,1\}$

$$P(D|p) = \prod_{i=1}^{n} (p)^{y_i} (1-p)^{(1-y_i)}$$

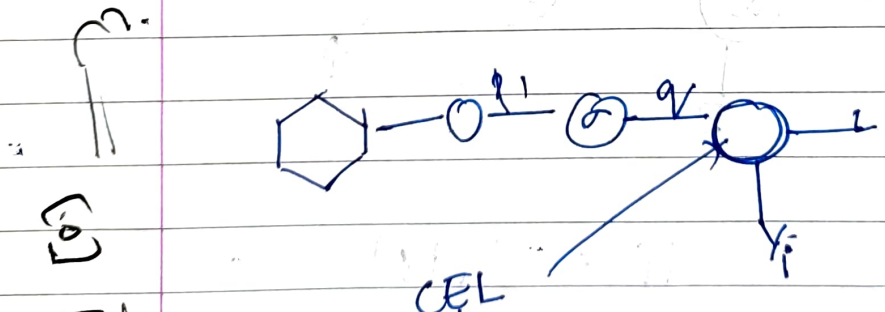
$$\ln P(D|p) = \sum y_i \log p + (1-y_i) \log(1-p)$$



→ maximum likelihood estimator.

cross entropy loss:

$$-\ln P(D|p) = -\left(\sum y_i \log p + (1-y_i) \log(1-p)\right)$$



$$= -\sum y_i \log q_i + (1-y_i)(\log(1-q_i))$$

$y_i \to q$. In the optimized ver.

$\log(q_i) \to q_i \to 0$

→ minimize the loss.

$$\frac{dL}{dq_i} = -\left(\frac{y_i}{q_i} + \frac{(1-y_i)}{(1-q_i)}(-1)\right)$$

$$= \frac{1-y_i}{1-q_i} - \frac{y_i}{q_i}$$

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

$$\sigma'(z) = \frac{e^{-z}}{(1)}$$

$$\frac{dL}{dp_i} = \frac{dL}{dq_i}\frac{dq_i}{dp_i}$$

$$= \frac{dL}{dq_i}(q_i)(1-q_i)$$

whole training sample — Batch gradient descent.

small portions — minibatch SGD.

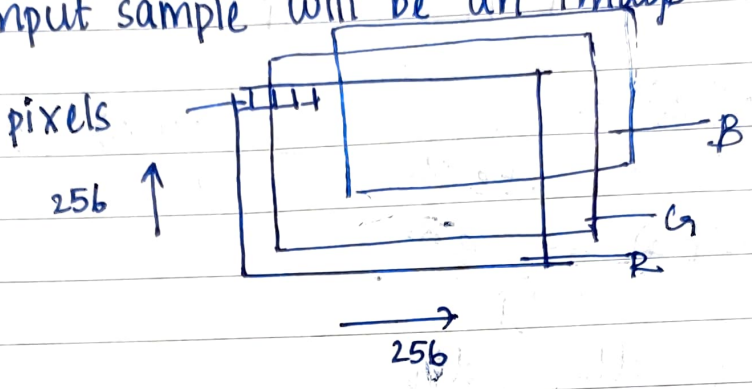
One training — Stochastic gradient descent.

$$\sum_{k} \frac{dL}{d}$$

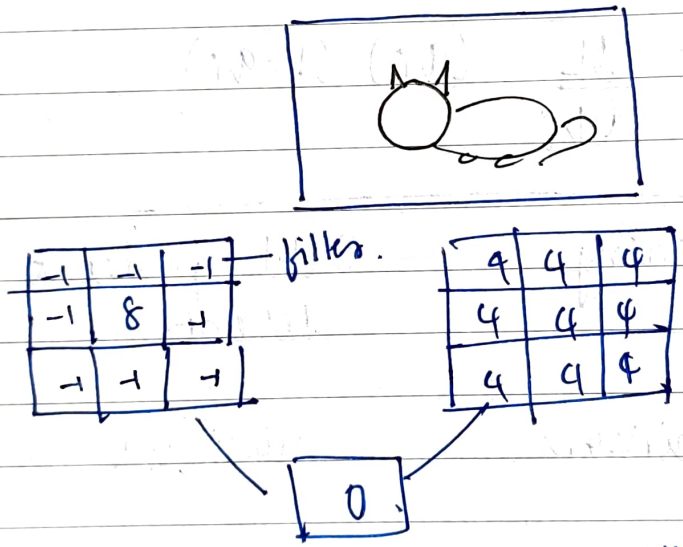
Convolutional Neural Network (CNN)

- used for computer vision problem.

- input sample will be an image



let us assum that is only a single channel.



using this filter we will be able to detect the edges.

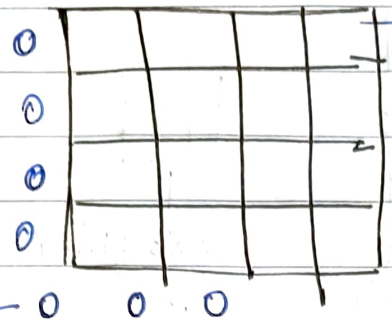
convolution will give 0 if underlying pixels are all the same. ie.

filter of dimension

$$\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{bmatrix}$$

we are trying to learn these values

design choices? filter size, number of filters to use



→ Output image will of dimension 2×2. with a 3×3 filter.

padding — 1 padding for 3×3, 2 padding for 5×5. etc...

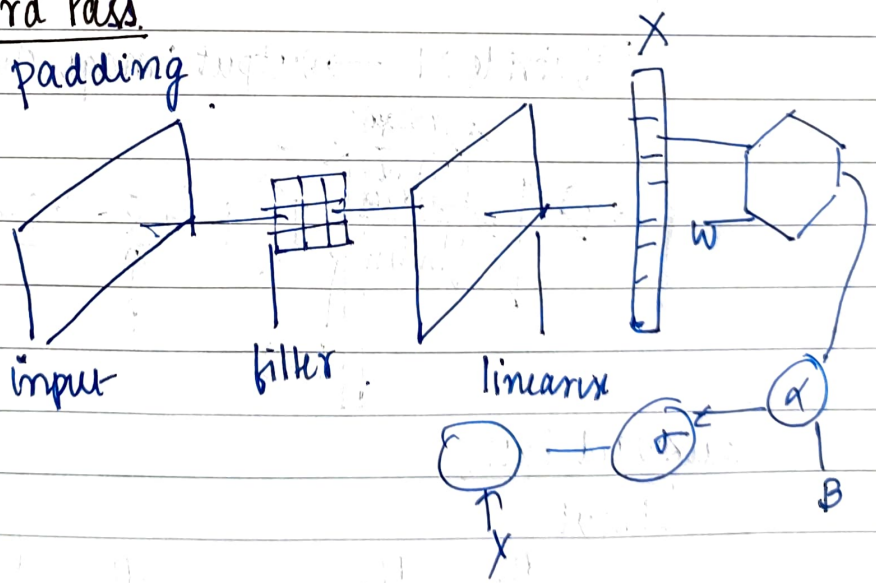
→ now output will be of dimension 4×4.

stride → by how many units do we move the filter.

→ If stride is higher, you will get a smaller image.

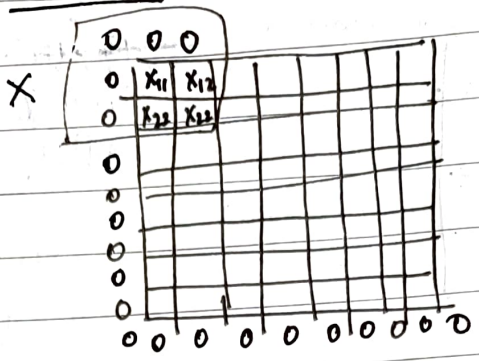
Forward Pass.

→ padding.



17/1/23

Lecture 5



Input image

3×3 filter

↓

layer 0 padding

↓

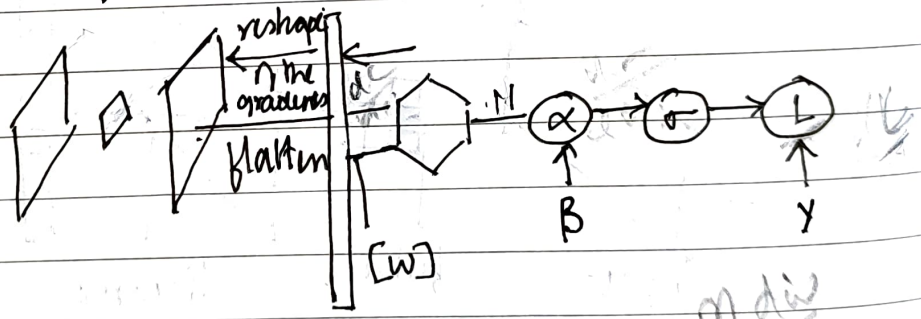
to maintain the dimension of input image.

$$\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{bmatrix}$$

$$out_{11} = x_{11}w_{22} + x_{12}w_{23} + x_{21}w_{32} + x_{22}w_{33}$$

stride = 1

↳ If stride > 1 → Output image shrinks

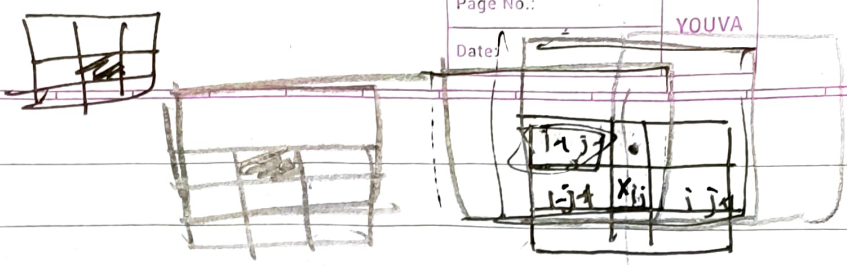


Backward Pass.

Input:

$\frac{dL}{dO_{11}}$	$\frac{dL}{dO_{12}}$	...	$\frac{dL}{dO_{1m}}$
$\frac{dL}{dO_{21}}$	$\frac{dL}{dO_{22}}$		

$\frac{dL}{dx_{ij}}$



If we change x_{ij}, what and all outputs change?

↳ $O_{(i-1)(j-1)}$, $O_{(i-1)(j)}$, $O_{(i-1)(j+1)}$ ....

$\frac{dL}{dx_{ij}}$

$\frac{dL}{dO_{(i-1)(j-1)}} w_{33}$	$\frac{dL}{dO_{(i-1)j}} w_{32}$	$\frac{dL}{dO_{(i-1)(j+1)}} w_{31}$
$\frac{dL}{dO_{i(j-1)}} w_{23}$	$\frac{dL}{dO_{ij}} w_{22}$	$\frac{dL}{dO_{i(j+1)}} w_{21}$
w_{13}	w_{12}	w_{11}

$$\frac{dL}{dx_{ij}} = \frac{dL}{dO_{(i-1)(j-1)}} w_{33} + \cdots + \frac{dL}{dO_{(i+1)(j+1)}} w_{11}$$

$$\frac{dL}{dw_{11}} = ?$$

Does w_{11} affect O_{11}? No

~~What is that~~

what does w_{11} affect?

$$O_{22} = x_{11} w_{11} + x_{12} w_{12} + x_{13} w_{13} + \cdots$$

$$\frac{dL}{dw_{11}} = \frac{dL}{dO_{22}} x_{11} + \frac{dL}{dO_{23}} x_{12} + \frac{dL}{O_{24}} x_{13}$$

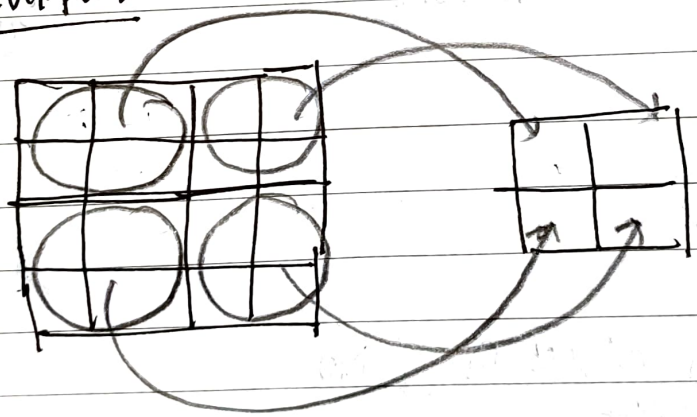
Claim

0	0	0	0	0 0
0	x_{11}	x_{12}	x_{13}	x_{14} 0
0	x_{21}	x_{22}	x_{23}	x_{24} 0
0	x_{31}	x_{32}	x_{33}	x_{34} 0
0	0	0	0	0 0

Input image - 3X4
filter - 3X3
Output image - 3X4

convolve → Output Image with Input Imag.

Maxpool -



Avg pool:

Softmax → Read about it!!

Lecture -6

18/1/23

Generative Models: → we are trying to model (X, Y).
↳ joint distribn.

Ex: Credit Default Problem.

$Y \in \{0, 1\}$
↳ No default, ↳ Default

priors:

$P_1 = P(Y=1) = 0.2$ $\quad P_0 = P(Y=0) = 0.8$

I $h_1 = P(Y=1) = 0.2$
$P(Y=0) = 0.8$
(Always predict 1)

II $h_2 \equiv 0$ (Always predict 0)
$Perr(h_2) = P(h_2 \neq Y) = P(Y=1) = 0.2$

III $h(z) = \begin{cases} 1 & \text{if } z=1 \\ 0 & \text{if } z=0 \end{cases}$

$Perr(h_2) = P(h_2 \neq Y)$
$= P(h(z)$

X ← feature space

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

- 1 → late tax payer.
- 0 → not a late tax payer

priors:

$p_1 = P(Y=1) = 0.2$

$p_0 = P(Y=0) = 0.8$

$f_y(x) = P(X=x \mid Y=y)$

$f_1(1) = P(X=1 \mid Y=1) = 0.95$

$f_0(1) = P(X=1 \mid Y=0) = 0.1$

Goal: We need to find a fn h

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

Bayes Rule

$$q_y(x) = P(Y=y \mid X=x)$$

$$= \frac{}{}$$

$$\therefore q_1(x) = \frac{p_1 f_1(x)}{p_0 f_0(x) + p_1 f_1(x)}$$

$$= \frac{P(Y=1)\ P(X=x \mid Y=1)}{P(Y=0)\ P(X=x \mid Y=0) +}$$

$h_0(0) = 0 \quad h_1(1) = 1$

M T W T F S S | Page No.: | Date: | YOUVA

$$q_1(1) = \frac{0.2 \times 0.95}{0.8 \times 0.1 + 0.2 \times 0.95} = 0.7$$

$$q_0(1) = \frac{0.8 \times 0.1}{0.8 \times 0.1 + 0.2 \times 0.95} = 1 - q_1(0) = 0.3$$

$$q_1(0) = \frac{0.2 \times 0.05}{0.8 \times 0.9 + 0.2 \times 0.05} = 0.013$$

$$q_0(0) = 1 - q_1(0) = 0.987$$

Type I error : False Alarm

There is no alarm, but my classifier is predicting 1.

TI error $= P(h(x) = 1 \mid y = 0)$

$= P(h(x) = 1 \mid x = 1, y = 0)\, P(x = 1 \mid y = 0) + P(h(x) = \cdot$

$P(h(x) = 1 \mid y = 0)$

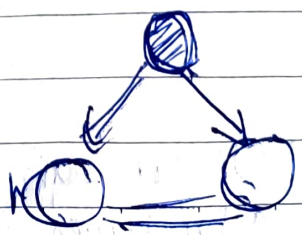
$= P(h(x) = 1 \mid x = 1, y = 0)\, P(x = 1 \mid y = 0) + P(h(x) = 1 \mid x = 0, y = 0)\, P(x = 0 \mid y = 0)$

($P(h(x) = 1 \mid x = 0, y = 0) \to 0$)

$= P(h(x) = 1 \mid x = 1, y = 0)\, P(x = 1 \mid y = 0)$

$= P(h(x) = 1 \mid x = 1)\, P(x = 1 \mid y = 0)$ ($h(x) = 1$ if $x = 1$)

$= 0.1$



HW

2) Type II error (missed detection)

$P_{err}(h) = P_1 \text{ Type1} + P_0 \text{ Type2}$

Bayes classifier

$h_B(x) = 1\{q_1(x) > q_0(x)\}$

$P_{err}(h_B) = P(h_B(X) \neq Y)$

$= \sum P(h_B(X) \neq Y \mid X = x)\, P(X = x)$

(We are trying to prove Bayes classifier, is the most optimal classifier : proof of optimality).

$= \sum_x \left[\sum_y P(h_B(x) \neq Y \mid X = x, Y = y)\, P(Y = y \mid X = x) \right] P(X=$

$= \sum_x \left[\sum_y L(h_B(x), y)\, q_y(x) \right] P(X = x)$

$q_{y^*}(x) = \max_y q_y(x)$

$= \sum_x \left[(1 - q_{y^*}(x)) \right] P(X = x)$

all the terms will be there except $\max_y q_y(x)$

we are trying to minimise this loss, and in Bayes Class. We maximis $q_{y^*}(x)$.

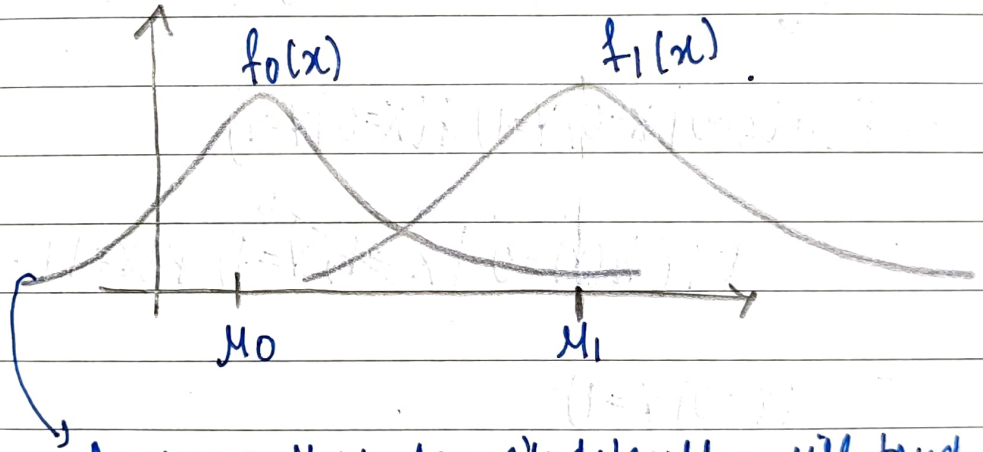
Hence the classifier is optimal.

Continuum of Features:

$X \in \mathbb{R}$. (Distribution of time when taxes are paid)

$p_0 = 0.8$, $p_1 = 0.2$

$f_1(x) = N(\mu_1, \sigma_1^2)$
$f_0(x) = N(\mu_0, \sigma_0^2)$



A person that doesn't default will tend to pay taxes early.

$q_1(x) = P(Y=1 \mid X=x)$ — X cuz x is continous.

$q_1(x) = P(Y=1 \mid x \in (x - \delta/2, x + \delta/2))$

$$q_1(x) = \frac{p_1 f_1(x)}{p_0 f_0(x) + p_1 f_1(x)}$$

$$h_B(x) = 1\{q_1(x) \geq q_0(x)\}$$

h: $h(0) = 0$
$h(1) = 1$

HW Qn

Type 2 error (missed detection).

$h(x) = 0$ $\quad Y = 1$

Type 2 error

$= P(h(x) = 0 \mid Y = 1)$

$= \sum_x P(h(x) = 0, x \mid Y = 1)$

$= P(h(x) = 0, x = 0 \mid Y = 1) + P(h(x) = 0, x = 1 \mid Y =$

$= P(h(x) = 0 \mid x = 0, Y = 1)\, P(x = 0 \mid Y = 1)$

$+ P(h(x) = 0 \mid x = 1, Y = 1)\, P(x = 1 \mid Y = 1)$ (0)

$= P(Y = 0 \mid Y = 1)$

$= 0.05$

$P_{err}(h) = P_1$ type 1 err $+ P_2$ type 2 err.

$P_{err}(h) = P(h(x) \neq Y)$

$= P(h(x) = 0, Y = 1) + P(h(x) = 1, Y = 0)$

$= P(h(x) = 0 \mid Y = 1)\, P(Y = 1) + P(h(x) = 1 \mid Y = 0)\, P(Y =$

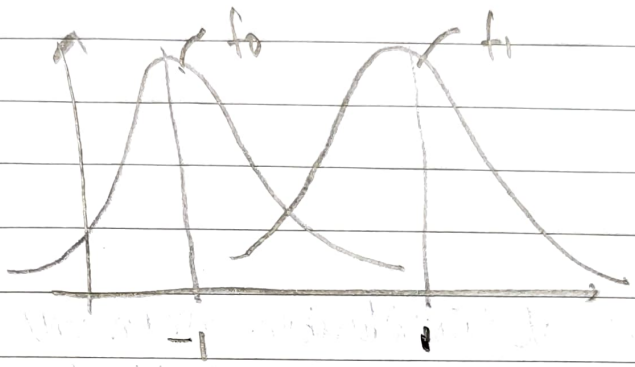
$=$ (type 1 error) P_1 + (type 2 error) P_2

Lecture 7

23/1/23

$X = \mathbb{R}$ ~ times when taxes are priced.

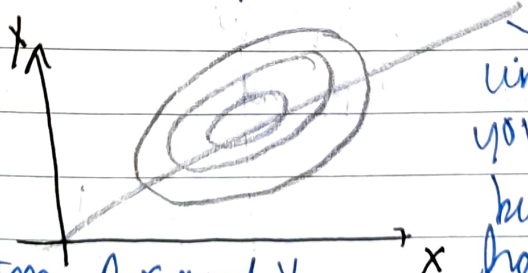
$f_0(x) \sim N(-1, \sigma^2)$ $\quad p_0 = 1/2$
$f_1(x) \sim N(1, \sigma^2)$ $\quad p_1 = 1/2$



Conditional densities (recap...)

$X \sim$ height $\quad Y \sim$ weight
$\sim N(5'5, \sigma_1^2) \quad \sim N(65, \sigma_2^2)$
└ Marginals.

$$X, Y \sim N\left(\begin{pmatrix} 5'5 \\ 65 \end{pmatrix}, \Sigma\right)$$



~ If it's a straight line, if you know x you know y perfectly, but for x and y will have some variance

joint distribution of x and y

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

~~$f(x|y)$~~

$$f(x \mid Y=70) = \frac{P(X=x, Y=70)}{P(Y=70)}$$

$$\approx P\left[X \in [x-\delta/2, x+\delta/2], Y \in [70-\delta/2, 70+\delta/2]\right]$$

for normal distribution when we fix y, it becomes a normal distribution

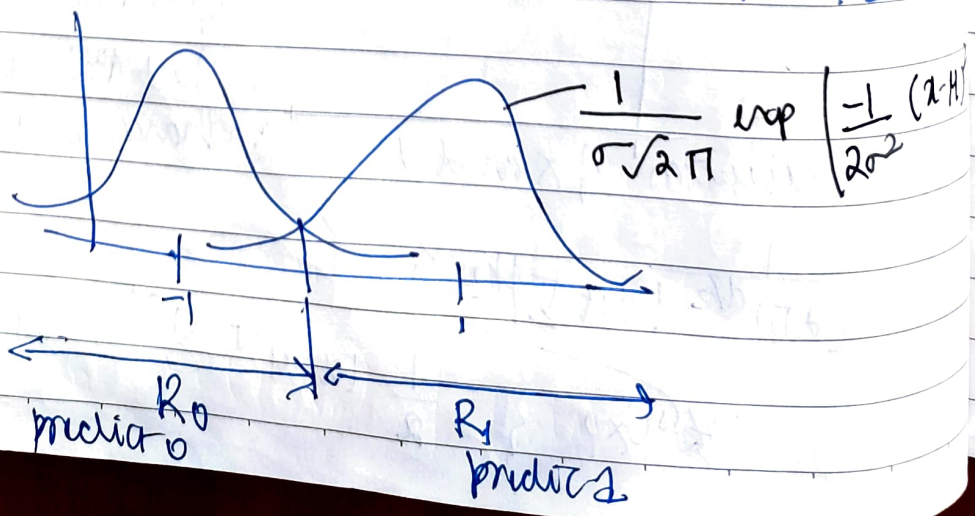
$$P(Y=1 \mid X=x) \triangleq q_1(x) = \frac{p_1 f_1(x)}{Z}$$

$$Z = p_0 f_0(x) + p_1 f_1(x)$$

$$q_0(x) > q_1(x)$$

$$p_0 f_0(x) > p_1 f_1(x)$$

$$f_0(x) > f_1(x) \qquad p_0 = p_1 = 1/2$$



$$-(x+1)^2 > -(x-1)^2.$$

$$\boxed{x < 0}$$

$R_0 \rightarrow$ region in my feature space where we report 0.

$R_1 \rightarrow$ report 1.

$$R_0 = \{ x \in X \mid h(x) = 0 \}$$

Plot:

Type 1 error: $h(x) = 1$ but $Y = 0$.

$$R_I = P(h(x) = 1 \mid Y = 0)$$

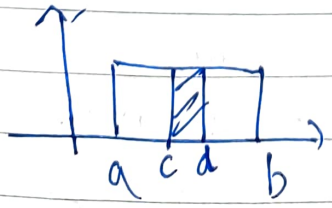
source of randomness $= X$

If X is fixed then $h(x)$.

[Digression

$X \sim$ Unif $[a, b]$

$$P(X \in [c, d]) = \int_{x \in [c,d]} f_X(x)\, dx$$



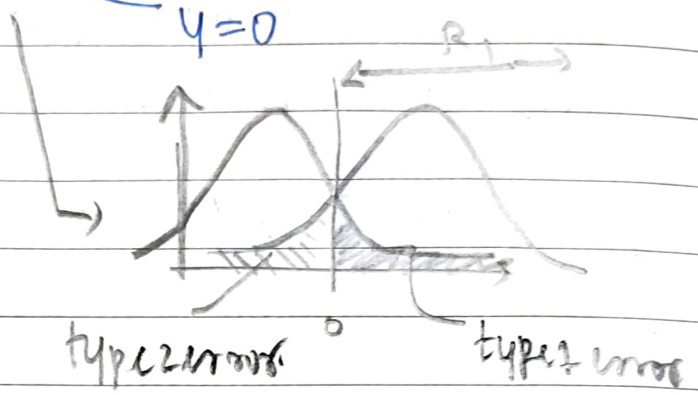
]

$$P_I = P(h(x)=1 \mid X=0)$$

$$= \int_{R_1} f_0(x)\, dx.$$

R_1: $h(x)=1$

f_0: $Y=0$



2) Neyman Pirson.

$$P_{error} = \underbrace{p_0}_{\text{fix}} \underbrace{\text{type I}}_{\text{fix}} + \underbrace{p_1}_{\text{automatic fixed.}} \underbrace{\text{type 2}}_{\text{automatic fixed.}}$$

ROC curve, AUC

Bayesian Decision Theory

$$Y = \{C_1, C_2, \ldots, C_m\}$$

└ there are m classes.

$$p_1 = P(Y=C_1) \qquad f_1(x) = P(X=x \mid Y=C_1)$$

$$\vdots \qquad\qquad \vdots$$

$$p_m = P(Y=C_m) \qquad f_m(x) = P(X=x \mid Y=C_m)$$

Instead of trying to report one of the m classes. we will take an action.

$$A = \{\alpha_1, \alpha_2, \cdots, \alpha_K\}$$

$$L : A \times Y \longrightarrow \mathbb{R}$$

loss function

classifier:

$$h : X \longrightarrow A .$$

what I want to minimize:

Risk of a classifier.

$$R(h) = E\left[L(h(X), Y)\right]$$

$$= E_X\left[E_{Y|X}\left[L(h(X), Y) \mid X\right]\right]$$

$$\left[E[X] = E_Y\left[E\left[X|Y\right]\right]\right.$$

$$E[X|Y] = g(Y)\left.\right]$$

$$= \int E\left(L(h(x), Y) \mid X\right) f(x)\, dx$$

$$R\left(h(x) \mid x\right) = E\left(L(h(x), Y) \mid X = x\right)$$

$$= \sum_{j=1}^{m} L(h(x), C_j)\, q_j(x)$$

Bayes classifier

At choose.

$$\alpha^*_k = \underset{r}{\operatorname{argmin}} \sum_K L(\alpha_r, c_k)\, q_k(x)$$

Lecture 8 — 24|1|23

$X \rightarrow$ feature space

$Y \rightarrow \{0, 1\}$

p_0
p_1
$\vdots$
p_m
} priors

$f_0(x) \sim N(-1, \sigma^2)$

$f_1(x) \sim N(+1, \sigma^2)$

$$q_i(x) = P(Y = i \mid X = x)$$

$L : A \times Y \rightarrow \mathbb{R}$

$R(h(x) \mid X = x)$

$$\underset{\alpha_j}{\operatorname{argmin}} \sum_K L(\alpha_j, c_k)\, q_k(x)$$

↗

Bayes classifier.

Ex – (0-1) loss

P_{err} is minimum.

$$h_B(x) = 1\{q_1(x) \geq q_0(x)\}$$

Data $\xrightarrow{\text{procedure}}$ Estimate parameters.

given : $S_n = \{(x_i, y_i)\}_{i=1}^{n}$

$$p_0 = \frac{\sum 1\{y_i = 0\}}{n} \longrightarrow \text{Estimator for } p_0.$$

Estimators are some function of your data.

$$\hat{\theta}(x_1, x_2, \ldots, x_n) \longrightarrow \theta$$

Ex: Suppose there is a coin whose bias is p.

$$S_n = \{(x_i)\}_{i=1}^{n} \longrightarrow \text{draw } n \text{ iid samples.}$$

$$\hat{p} = \frac{\sum x_i}{n} \longrightarrow \text{used to estimate } p.$$

- We want to find a good estimator and understand the characteristics of a good estimator.

Desirable properties of an estimator:

1) Unbiasedness

$$E[\hat{\theta}] = \theta$$

p (fixed but unknown)

$$S_n = \{X_i\}_{i=1}^{n}$$

$$M_n = \frac{\sum_{i=1}^{n} X_i}{n} \qquad M_{odd} = \frac{X_1 + X_3 + X_5 + \cdots}{n_{odd}}$$

[HW - Expected value of M_n

$E[M_n] = p$

$E[M_{odd}] = p$]

$$E[M_n] = E\left[\frac{\sum X_i}{n}\right]$$

$$= \frac{\sum E(X_i)}{n} = \frac{np}{n} = p$$

X_is are iid,

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

$$= p$$

$$\boxed{E[M_{odd}] = p}$$

Both have the same unbiasedness, but what is the diff?

$$\text{Var}(M_n) = \text{Var}\left(\frac{\sum X_i}{n}\right)$$

$$= \frac{1}{n^2} \text{Var} \sum X_i \quad [X_i \text{ are independent}]$$

$$= \frac{1}{n^2} \sum \text{Var} X_i$$

$$= \frac{(p)(1-p)}{n}$$

$$\text{Var}(M_{odd}) = \frac{p(1-p)}{n_{odd}}.$$

$\text{Var}(M_{odd}) > \text{Var}(M_n)$

variance is big means the errors will be more.

2) The estimator should have low variance.

3) Consistency.

$\hat{\theta}_n \xrightarrow{IP} \theta$ as $n \to \infty$

↳ must happen

convergence in probability

Date: YOUVA

$$\lim_{n \to \infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0 \qquad \forall \varepsilon > 0$$

↳ weak law of large numbers.

↳ Probability of taking 200 samples and $\hat{p}$ of taking 1000 samples must be closer to p.

1) unbiasedness
2) low variance
3) consistency ✓ → we are only looking at consistency

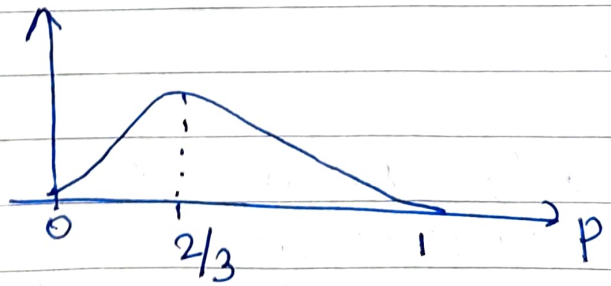
Maximum likelihood Estimation

$$\mathcal{D} = \{H, H, T\}$$

$$\ell(p \mid \mathcal{D}) = P(\mathcal{D} \mid p)$$

If $p = 0$, $\ell(0 \mid \mathcal{D}) = 0$

$\ell(1 \mid \mathcal{D}) = 0$



$$\hat{p}_{MLE} = \underset{p}{\operatorname{argmax}} \; (p \mid \mathcal{D})$$

Lecture 9

25/1/23

Estimator

$\theta(X_1, X_2, \cdots, X_n) \rightarrow \theta$ (fixed, unknown)

$\hat{\theta}_n \Rightarrow$ random variable.

Desirable properties:

1) unbiasedness $E[\hat{\theta}_n]$
2) low variance of $\hat{\theta}_n$
3) consistency $\hat{\theta}_n \rightarrow \theta$

$$\forall \varepsilon > 0 \quad \lim_{n \to \infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0$$

$$\hat{M}_n = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \{X_i\}_{i=1}^{n} \text{ iid bern}(p)$$

fix $\varepsilon > 0$

$$P(|\hat{M}_n - p| > \varepsilon)$$

Markov Inequality

$X > 0$ (non negative rv)

$$E(X) = \sum_{0 \le x < a} x\, p_X(x) + \sum_{x \ge a} x\, p_X(x)$$

(positive num.)

$$E(X) \ge \sum_{x \ge a} x\, p_X(x)$$

$$\ge \sum a\, p_X(x \ge a)$$

$$P[X \geq a] \leq \frac{E(X)}{a}$$ — markov inequality

Union bound:

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum P(A_i)$$

chebyshev inequality (no restriction $X \in \mathbb{R}$)

$$P(|X| \geq a) = P(x^2 \geq a)$$

$$\Rightarrow P(|X-\mu| \geq a)$$

$$\Rightarrow P(|X-\mu|^2 \geq a^2)$$

$$\Rightarrow \leq \frac{E(x-\mu)^2}{a^2} \quad \text{(markov inequality)}$$

$$\Rightarrow \leq \frac{Var(x)}{a^2}$$

$$\Rightarrow \hat{\mu}_n = \frac{1}{n} \sum X_i$$

fix $\epsilon > 0$

$$P(|\hat{\mu}_n - p| > \epsilon)$$

$$= P(|\mu_n - p|^2 > \epsilon^2)$$

$$\leq \frac{E(\hat{\mu}_n - p)^2}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2}$$

$$\therefore \lim_{n\to\infty} P(|M_n - p| \geq \varepsilon) = 0$$

$$\tilde{M}_n = \frac{1}{n-1} \sum_{i=1}^{n} X_i$$

unbiased? $E[\tilde{M}_n] = \frac{n}{n-1} \mu \neq \mu$

$\mu = p$

consistency? — Yes.

$$P(|\tilde{M}_n - p| > \varepsilon) \leq \frac{Var(\tilde{M}_n)}{\varepsilon^2}$$

$$\leq \frac{n\sigma^2}{(n-1)^2 \varepsilon^2}. \qquad \sigma^2 = (p)(1-p)$$

As $n \to \infty$ $\quad P(|M_n - p| > \varepsilon) \to 0.$

Maximum Likelihood Estimator.

$$L(\theta|\mathcal{D}) \triangleq p(\mathcal{D}|\theta) \qquad \left(\mathcal{D} = \{x_i\}_{i=1}^{n}\right)$$
$$= \prod p(x_i|\theta) \qquad \text{iid}$$

MLE

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}}\ L(\theta|\mathcal{D})$$

MLE for Bernoulli

Assume p ($0 < p < 1$ fixed but unknown)

└ $\{X_i\}_{i=1}^{n}$ iid $P(X_i = 1) = p$

$$\ell(p|\mathcal{D}) = \prod_{i=1}^{n} p^{x_i} (1-p)^{(1-x_i)}$$

Ex: ~~{H, H, T, T}~~ $\{1, 1, 0, 0\}$

log likelihood

$$\hat{p}_{MLE} = \underset{p}{\text{argmax}} \sum x_i \log p + (1-x_i) \log (1-p)$$

$$\frac{d}{dp} L(p) = \frac{\sum x_i}{p} -$$

$$= \sum \frac{x_i}{p} - \frac{(1-x_i)}{(1-p)} = 0$$

$$= \sum \frac{x_i (1-p) - (1-x_i) p}{p (1-p)}$$

$$= \frac{\sum x_i}{n}$$

$$p_{MLE} = \frac{\sum x_i}{n}$$

MLE for multinomial.

$X_i \in \{1, 2, \dots, K\}$

$p(X_i = 1) = p_1$

$p(X_j = 2) = p_2$

$$\sum_{i=1}^{K} p_i = 1$$

$X_i \rightarrow$ one hot encoding

$X_i = 1, 1, 2, 2, 3, 1, \dots$

$$\begin{pmatrix}1\\0\\0\\0\end{pmatrix}, \begin{pmatrix}1\\0\\0\\\vdots\end{pmatrix}, \begin{pmatrix}0\\1\\0\\\vdots\end{pmatrix} \begin{pmatrix}0\\0\\\vdots\end{pmatrix} \begin{pmatrix}0\\0\\1\\\vdots\end{pmatrix}$$

$$X_i \rightarrow \tilde{X}_i = \begin{pmatrix} x_i^{(1)} \\ \vdots \\ x_i^{(K)} \end{pmatrix}$$

MLE

$$l(p_1 p_2 \cdots p_K \mid \mathcal{D}) = \prod_{i=1}^{n} \prod_{j=1}^{K} p_j^{x_i^{(j)}}$$

$$L(l(p_1, p_2 \cdots \mid \mathcal{D})) = \sum_i \sum_j x_i^{(j)} \log p_j$$

$$p_1, p_2, \cdots = \underset{p_1, p_2 \cdots}{\arg\max} = \sum_i \sum_j x_i^{(j)} \log p_j$$

$$\text{s.t. } \sum p_j = 1 \quad p_j > 0$$

(Bishop - Appendix)
Optimization of constrained

$$\max \sum_{i=1}^{n} \sum_{j=1}^{k} x_i^{(j)} \log_e (p_j)$$

subject to $\sum p_i = 1$

$$L(x, \lambda) = f(x) + \lambda g(x)$$

→ we're making a constrained problem unconstrained

→ constrained $g(x) = 0$

$$\max_{x, \lambda} L(x, \lambda)$$

$$L(p_1, p_2, \cdots, p_k, \lambda)$$

$$= \sum_{i=1} \sum_{j=1} x_i^{(j)} \ln (p_j) + \lambda \left(\sum p_i - 1 \right)$$

$$\frac{\partial L}{\partial p_j} = \sum_{i=n}^{n} \frac{x_i^j}{p_j} + \lambda = 0$$

$$\frac{\partial L}{\partial \lambda} = 0$$

$$= \sum_{i=1}^{n} x_i^{(j)} + \lambda p_j = 0$$

Summing over $j = \{1, 2, \ldots, k\}$

$$= \sum_{j=1}^{k} \sum_{i=1}^{n} x_i^{(j)} + \lambda \sum_{j=1}^{k} p_j = 0$$

$$= n + \lambda = 0$$

$$\lambda = -n$$

$$\frac{\partial L}{\partial p_j} = \sum x_i^{(j)} + \lambda p_j = 0$$

$$= \boxed{p_{j,MLE} = \frac{\sum x_i^{(j)}}{n}}$$

30/1/23

MLE for Normal distribution.

$$\mathcal{D} = \{x_i\}_{i=1}^{n} \qquad X_i \sim N(\mu, \sigma^2) \text{ iid.}$$

$$L(\theta|\mathcal{D}) = P(\mathcal{D}|\theta)$$

$$= \prod_{i=1}^{n} f(x_i|\mu,\sigma^2)$$

$$f(x_i|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right\}$$

$$\ell(\mu,\sigma^2|\mathcal{D})$$

~~$\ell(\theta|\mathcal{D})$~~ $= \sum_{i=1}^{n} \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) \frac{-1}{2\sigma^2}(x_i-\mu)^2$

$$\frac{\partial L}{\partial \mu} = 0$$

$$\boxed{\mu_{MLE} = \frac{\sum x_i}{n}}$$

$$\frac{\partial L}{\partial \sigma} = 0$$

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

Bias variance tradeoff

MSE → mean squared error

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

$$\hat{\theta}(x_1, x_2, \ldots, x_n)$$

$$= E\left[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta\right]^2$$

$$= E\left[(\hat{\theta} - E(\hat{\theta}))^2 + (E\hat{\theta} - \theta)^2 + 2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta)\right]$$

$$= E(\hat{\theta} - E(\hat{\theta}))^2 + (E(\hat{\theta}) - \theta)^2 + 2(E(\hat{\theta}) - \theta)\, E[\hat{\theta} - E(\hat{\theta})]$$

$$= E\left(\hat{\theta} - E[\hat{\theta}]\right)^2 + \left(E[\hat{\theta}] - \theta\right)^2.$$

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

$$\text{MSE}(\hat{\theta}) = \underbrace{\text{Var}(\hat{\theta})}_{\text{How big the bag is?}} + \underbrace{\text{Bias}^2}_{\text{to cope with the variance you have to reduce the amount of bias in the "bag..", this will give us a bias.}}$$

Bayesian Estimators:

Ex: $P(\mathcal{D}) = 0.1$

↳ probability of disease.

$$P(+|\mathcal{D}) = 0.7 \qquad P(+|\mathcal{D}^c) = 0.01$$

$$P(\mathcal{D}|+) = \frac{P(+|\mathcal{D})\,P(\mathcal{D})}{P(+|\mathcal{D})\,P(\mathcal{D}) + P(+|\mathcal{D}^c)\,P(\mathcal{D}^c)}$$

$$= \frac{(0.01)(0.7)}{(0.7)(0.01) + (0.9)(0.01)}$$

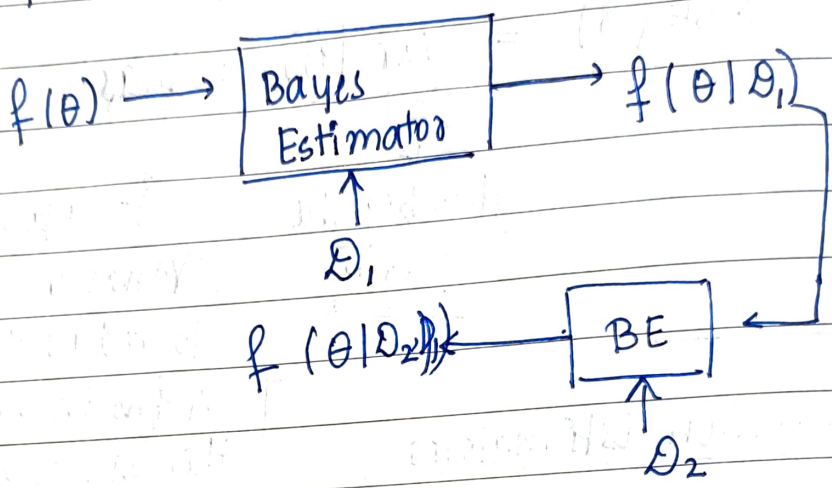
$$= \frac{70}{79} = 0.875$$

$P(\mathcal{D}) = 0.1$ → Bayes rule → $P(\mathcal{D}|E_1) = 0.877$ → $P(\mathcal{D}|E_1, E_2)$ ← (BR)

In MLE - we assumed θ was a fixed unknown

In Bayes Estimator → θ (Random Distribution)
└ you have a belief, but it improves more and more as we see more values

$f(\theta)$ → prior for θ.



$$f(\theta|D_1) = \frac{f(\theta)\, f(D_1|\theta)}{\int f(\tilde{\theta})\, f(D_1|\tilde{\theta})\, d\tilde{\theta}}$$

$$= \frac{f(\theta)\, f(D_1|\theta)}{f(D_1)}$$

[Here MLE cannot be used, because we're getting data piece by piece].

- When we have some prior knowledge, its better to use Bayesian estimator.

conjugate pairs

$$f(\theta|\Theta) = \frac{f(\theta)\, f(\Theta|\theta)}{c}$$

form

coin toss prior:

$$f(p) = \text{Beta}(\alpha, \beta) \quad \begin{pmatrix} \alpha > 0 \\ \beta > 0 \end{pmatrix}$$

$$= \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\int_0^1 p^{\alpha-1}(1-p)^{\beta-1}\, dp}$$

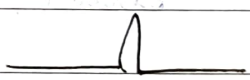
$$= p^{\alpha-1}(1-p)^{\beta-1}\,(K)$$

$$f(\Theta|p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

$$f(p|\Theta) = \text{Beta}(\alpha + \sum x_i,\ \beta + n - \sum x_i)$$

Beta (1,1) ⇝ uniform (0,1)

25 H, 5T Beta (26,6).



31/1/23

Bayes Estimator

Θ ~ Real Random variable

$f(\theta)$ ~ Prior (comes from knowledge of the problem) (we choose it)

$f(\Theta|\theta)$ ~ likelihood

$[L(\theta|\Theta)]$.

→ $f(\theta)$ and $f(\mathcal{D}|\theta)$ will form a conjugate pair, if $f(\theta|\mathcal{D})$ has the same form as $f(\theta)$

$$f(\theta|\mathcal{D}) = \frac{f(\theta)\, f(\mathcal{D}|\theta)}{\int f(\theta')\, f(\mathcal{D}|\theta')\, d\theta'}$$

↳ posterior distribution

(Goal is to estimate θ)

→ **Example:**

Beta distribution

$$\text{Beta}(p|\alpha,\beta)$$

α, β are fixed parameters that we don't know.

$$\text{Beta}(p|\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\int_0^1 \tilde{p}^{\alpha-1}(1-\tilde{p})^{\beta-1}\, d\tilde{p}}$$

↳ density function for p.

$$f(p) \sim \text{Beta}(p|\alpha,\beta)$$

$f(\mathcal{D}|p)$ → bias of coin

↳ sequence of coin tosses

$\sum x_i$ → number of heads

$$f(\theta|p) = p^{\sum x_i} (1-p)^{n-\sum x_i}$$

$$f(p|\theta) = \frac{f(\theta|p)\, f(p)}{\int_0^1 f(\tilde{p})\, f(\theta|\tilde{p})\, d\tilde{p}}$$

↳ this is a constant

$$f(p|\theta) = c \times p^{\sum x_i} (1-p)^{n-\sum x_i} p^{(\alpha-1)} (1-p)^{\beta-1}$$

$$= c \times p^{\sum x_i + \alpha - 1} \times (1-p)^{n-\sum x_i + \beta - 1}$$

$$= \text{Beta}(\alpha + \sum x_i,\ n - \sum x_i + \beta)$$

Beta $\int_0^1 \tilde{p}^{\alpha-1} (1-\tilde{p})^{\beta-1} d\tilde{p} = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

$\Gamma(a) = \int_0^\infty e^{-x} x^{a-1} dx$ → solve the integration by parts

$\Gamma(a+1) = a\,\Gamma(a)$ — recurrence relation

$\Gamma(n) = (n-1)!$

↳ positive integers

"Γ will come up randomly in various problems, its weird way in which nature arranges itself."

Date: YOUVA

$$E[p] = \int_0^1 p\, f(p)$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 p\,(p)^{\alpha-1}(1-p)^{\beta-1}\,dp$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int p^{\alpha} \cdot p^{\alpha+\beta-1}\,dp$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \left[\frac{p^{\alpha+1}}{\alpha+1} - \frac{p^{\alpha+\beta}}{\alpha+\beta} \right]_0^1$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 (p)^{\alpha}(1-p)^{\beta-1}\,dp$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int p^{(\alpha+1)-1}(1-p)^{\beta-1}\,dp$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+1)\,\Gamma(\beta)}{\Gamma(\alpha+\beta+1)}$$

$$= \frac{\alpha}{\alpha+\beta}$$

prior = Beta (α, β)

$\downarrow$ Σx_i , $n - \Sigma x_i$

$f(p|\theta) = \text{Beta}(\alpha + \Sigma x_i, \beta + n - \Sigma x_i)$

What do we predict in the end?

option ① $E(f(p|\theta)) = \dfrac{\alpha + \Sigma x_i}{\alpha + \beta + n}$ → mean of posterior

option ② MAP

└ maximum A priori

$\max_p f(p|\theta)$

$f'(p|\theta) = 0$

$p^{\alpha + \Sigma x_i - 1} (1-p)^{\beta + n - \Sigma x_i - 1}$

$\Rightarrow$

$f = k\, p^{\alpha' - 1} (1-p)^{\beta' - 1}$

$= k$

$\ln(f(p)) = \ln k + (\alpha - 1) \ln p + (\beta - 1) \ln(1-p)$

$\dfrac{d \ln(fp)}{dp} = \dfrac{\alpha - 1}{p} + \dfrac{\beta - 1}{p} = 0$

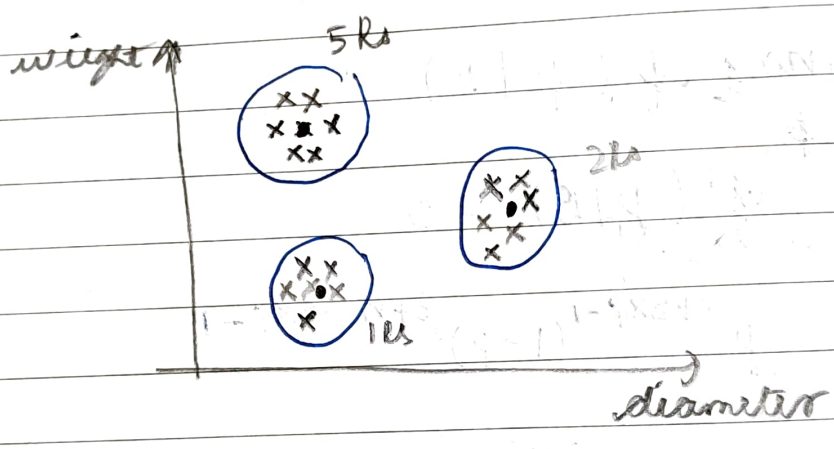
$p(\alpha - 1) + p(\beta - 1) = 0$

$p\alpha - p + p\beta - p = 0.$

$$P_{map} = \frac{\alpha - 1}{\alpha + \beta - 2}$$

Unsupervised learning (no labels are given)
K Means Algorithm

problem: you go to an ATM and it has to give out coins



Given this data can we identify which is 1Re, 2Rs, 5Rs coin.

given features $\{x_i\}_{i=1}^{n}$, K (fixed)

labels are not given we have to cluster the features and give them labels.

Required output:

Goal: For all i $Z_{ij} = 1$ if point i belongs to the j^{th} cluster.

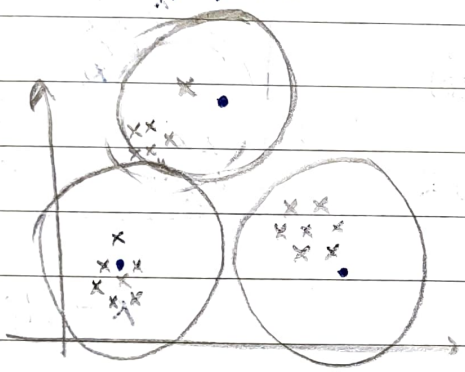
what do we wanna ~~estimate~~ optimise?

↳ we want close clusters

μ_1^*, μ_2^*, μ_3^* are centroids

$$\mu_j^* = \frac{\sum_i z_{ij} x_i}{\sum_i z_{ij}}$$

goal: $\min \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \, ||x_i - \mu_j^*||$



1/2/23

what we have as input? $\{x_i\}_{i=1}^{n}$, $K > 0$.

output? for every point x_i to some cluster

$z_{ij} = 1$ (if i^{th} point belongs to the j^{th} cluster)

$i \in \{1, \dots, n\}$ $j \in \{1, \dots, k\}$

Objective: $\min \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \, ||x_i - \mu_j^*||$

↳ the average distance from centroids is minimum.

k-means algorithm:

1) Random initialisation of

$$\mu_j^{*(0)} \{ j = 1, \dots, k \}$$

2) Repeat till convergence (t loop variable)

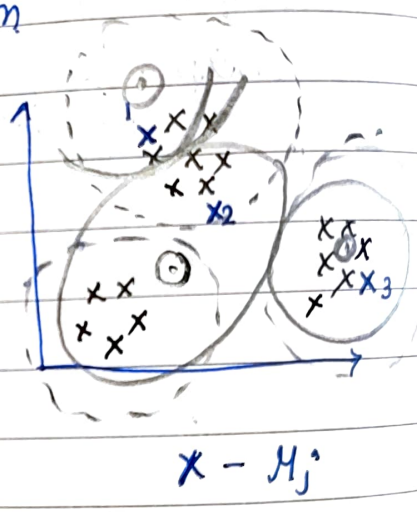
$$\rightarrow z_{ij}^{(t)} = 1 \quad \text{if } j \in \underset{m}{\operatorname{argmin}} \| x_i - \mu_m^{(t-1)} \|^2$$

(break ties arbitrarily)

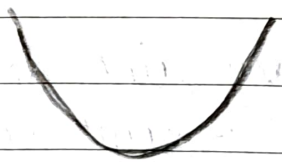
One run of this algorithm

$$\rightarrow \mu_j^{*(t)} = \frac{\sum z_{ij}^{(t)} x_i}{\sum z_{ij}^{(t)}}$$

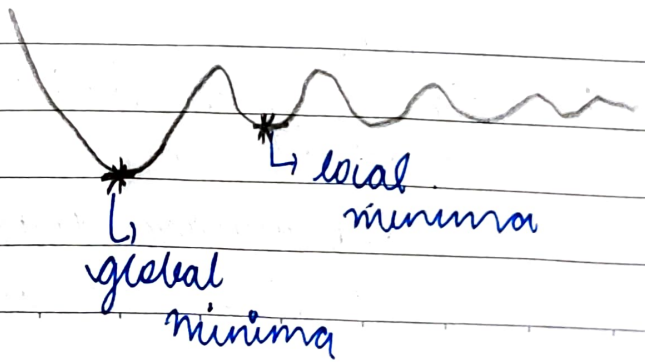
$$\forall j \in 1, 2, \dots, k$$



Convex functions (shaped like a bowl)

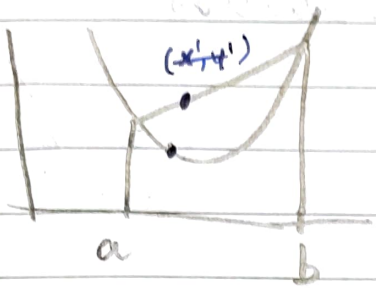


General function



→ Saving k means optimally is an NP hard.

→ K means gives local minima.



$x = \lambda a + (1-\lambda) b$

$\lambda \in (0,1)$

$xy' = \lambda f(a) + (1-\lambda) f(b)$

$f(\lambda a + (1-\lambda) b) \leq \lambda f(a) + (1-\lambda) f(b)$

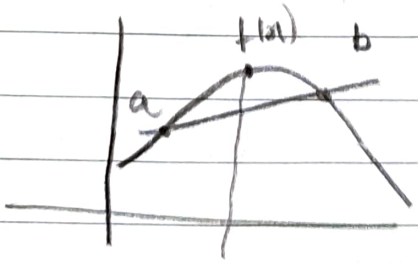
↳ one possible condition for convexity

easier condition :

↳ $f'(x)$ is increasing

↳ $f''(x) > 0$

concave function



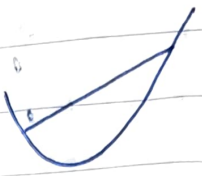
$f(\lambda a + (1-\lambda) b) \geq \lambda f(a) + (1-\lambda) f(b)$

$f'(x)$ is decreasing

$f''(x) < 0.$

convex functions have a global minima (concave - maxima).

lets take a convex function



Jensen's Inequality

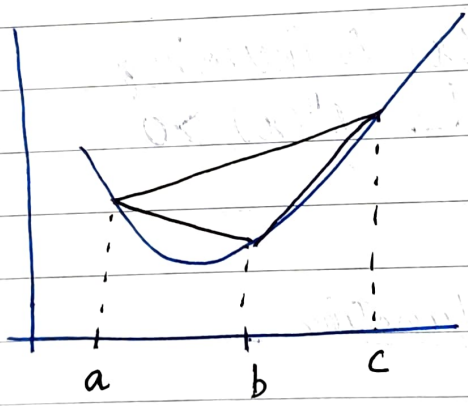
$E[f(x)] \geq$

$f(\lambda a + (1-\lambda) b) \leq \lambda f(a) + (1-\lambda) f(b)$

probable prob

$E[f(x)] \geq f[E[x]]$

How did we generalize?



$p = \lambda_1 a + \lambda_2 b + \lambda_3 c \quad (\lambda_1 + \lambda_2 + \lambda_3 = 1)$

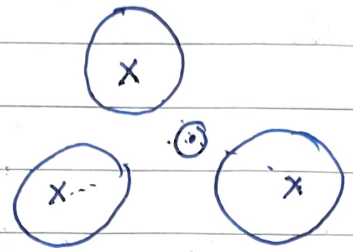
$a \leq \lambda_1 a + \lambda_2 b + \lambda_3 c \leq c$

$E[X] = p.$

$f(E(x)) \leq E[f(x)]$

K means has a limitations:

1. Assume there are three points



1) circular
2) same size.
   ↳ ideally.
   ↳ in general encourages same size

K means algorithm:

Input: $\{x_i\}_{i=1}^{n}$, $K>0$

Output: cluster assignment

$z_{ij} = 1$ if i^{th} pt belongs to j^{th} cluster

$\hat{\mu}_j^*$ → centroid

Random init: choose $\hat{\mu}_j^{(0)}$ (or) $z_{ij}^{(0)}$ at random

Date: 2/2/23 YOUVA

Limitations of K-means algorithm

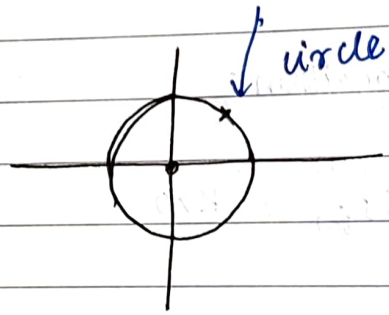
← K means doesn't work

$$||x_i - \hat{M}_j^*|| = \sqrt{(x - M_j^*)^T (x_i - M_j^*)}$$

$||x_i - \hat{M}_j^*||$ — vector

norm tells the distance from the origin.

$$(x_i - M_j^*)^T \; I \; (x - M_j)$$



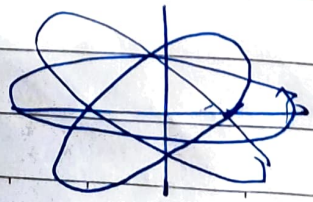
positive (semi) Definite

$$x^T A x \geq 0 \quad (> \text{for +ve definite})$$

We can define new norms

$$(x_i - M_j^*)^T \; A \; (x_i - M_j)$$

this can have all kinds of shape.



multivariate gaussian. $\frac{1}{(2\pi)^{n/2}|\det \Sigma|^{1/2}}$

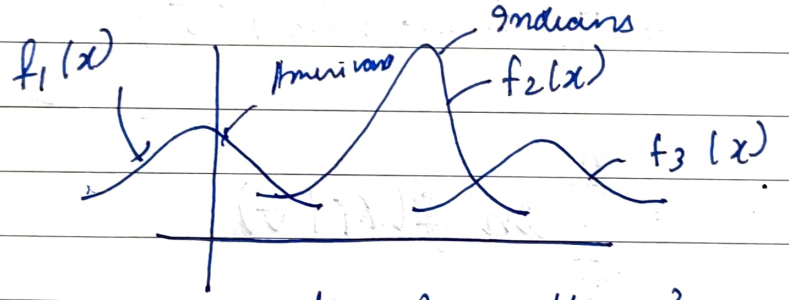
$$\exp\left(\frac{-1}{2}(x-\mu_j)^T \Sigma^{-1}(x-\mu_j)\right)$$

$\Sigma \rightarrow$ covariance matrix $\rightarrow$ positive definite

$\Sigma^{-1} \rightarrow$ is also positive definite.

We're gonna generalise K-means to Gaussian ~~Victor~~ models : Mixture

Single variate.



μ_1, σ_1^2 μ_2, σ_2^2 μ_3, σ_3^2

$$f(x) = \pi_1 f_1(x) + \pi_2 f_2(x) + \pi_3 f_3(x)$$

π_1 ↓ probability you are in the first population

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

$$\theta = \begin{bmatrix} \pi_1 & \pi_2 & \pi_3 \\ \mu_1 & \mu_2 & \mu_3 \\ \sigma_1^2 & \sigma_2^2 & \sigma_3^2 \end{bmatrix}$$

If we had multiple variable.

$\mu_1 = [\ ]$ → vectors

$\Sigma_1 = [\ v\ ]$ → matrices.

$\log(L(\theta/\mathcal{D})) = \Pi f(x_i|\theta)$

z_i → latent variance.
here the category is unknown.

$f(x_1, z_i = c \mid \theta)$
$= \Pi_c f(x_i \mid \mu_c, \sigma_c^2)$

$$l(\theta) = \sum_{i=1}^{n} \ln f(x_i|\theta)$$

$$= \sum_{i=1}^{n} \ln \sum_{c} f(x_i, z_i = c|\theta)$$

$$= \sum_{i=1}^{n} \ln \sum_{c} \Pi_c f(x_i \mid \mu_c, \sigma_c^2)$$

$$\theta_{MLE} = \underset{\tilde{\theta}}{\operatorname{argmax}}\ l(\tilde{\theta}) \quad \Big| \quad \sum \pi_c = 1$$

$$\sum \pi_c = 1$$

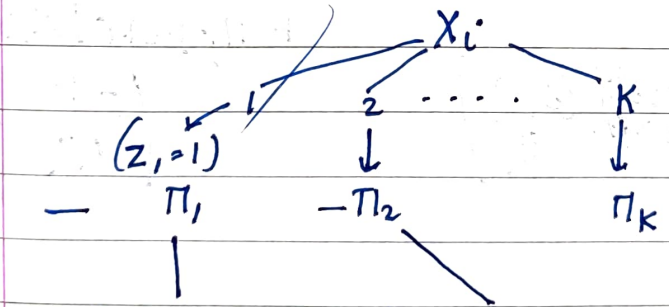
$$\alpha(\theta, \lambda) = \underset{\tilde{\theta}}{\operatorname{argmax}}\ l(\tilde{\theta}) + \lambda (1 - \sum \pi_c)$$

$$\frac{\partial \alpha(\theta, \lambda)}{\partial \mu_c} = \sum_{i=1}^{n} \frac{1}{\sum \pi_c f(x_i | \mu_c, \sigma_c^2)} \pi_c f(x_i | \mu_c \sigma_c^2)$$

$$\times \frac{1}{2\sigma_c^2} 2(x_i - \mu_c)$$

06/02/23

every point X_i can come from one of the k clusters.



gaussian mixture model

compute for MAP

MLE:

$$L(\theta) = \prod_{i=1}^{n} f(x_i | \theta)$$

$$l(\theta) = \sum_i^n \ln f(x_i|\theta)$$

$$= \sum \ln \sum_c f(x_i, z_i=c|\theta)$$

$$= \sum \ln \sum_c \pi_c f(x_i|\mu_c, \sigma_c^2)$$

Lagrangian

$$L(\theta, \lambda) = \sum \ln \sum \pi_c f(x_i|\mu_c, \sigma_c^2) + \lambda(1 - \sum \pi_c)$$

$$\frac{\partial L}{\partial \mu_j} = \sum_{i=1}^n \frac{\pi_j f(x_i|\mu_j, \sigma_j^2)}{\sum_c \pi_c f(x_i|\mu_c, \sigma_c^2)} \times \frac{(x_i - \mu_j)}{\sigma_j^2} = 0$$

$$\frac{\partial L}{\partial \sigma_j} = \sum_{i=1}^n \frac{\pi_j f(x_i|\mu_j, \sigma_j^2)}{\sum_c \pi_c f(x_i|\mu_c, \sigma_c^2)} \times \left(\pi_j f(x_i|\mu_j, \sigma_j^2) \times \left(\frac{1}{\sigma_j^3}(x_i - \mu_j)^2 + \pi_j f(x_i|\mu_j, \sigma_j^2)\right)\left(\frac{-1}{\sigma_j}\right)\right) = 0$$

$$\boxed{uv' + u'v = \frac{(u)}{(v)}(\ell)}$$

$$X_i = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(\frac{-1}{2\sigma_j}(x_i - \mu_j)^2\right)$$

$$\frac{\partial L}{\partial \pi_c} = \sum_{i=1}^n \frac{f(x_i|\mu_c, \sigma_c^2)}{\sum_c \pi_c f(x_i|\mu_c, \sigma_c^2)} - \lambda = 0$$

Put $\Gamma_{ij} = \frac{\pi_j f(x_i|\mu_j, \sigma_j^2)}{\sum \pi_c f(x_i|\mu_c, \sigma_c^2)}$

$$\sum_i \frac{\Gamma_{ij}(x_i - \mu_j)}{\sigma_j^2} = 0 \quad \text{---} (1)$$

$$\mu_j = \frac{\sum \Gamma_{ij} x_i}{\sum \Gamma_{ij}}$$

$$\sum_i^n \Gamma_{ij} \left(\frac{1}{\sigma_j^3} (x_i - \mu_j)^2 - \frac{1}{\sigma_j} \right) = 0.$$

$$\sigma_j^2 = \frac{\sum \Gamma_{ij} (x_i - \mu_j)^2}{\sum \Gamma_{ij}}$$

$$\sum_j \left(\left(\sum_i \Gamma_{ij} \right) - \lambda \Pi_j \right) = 0$$

$$\lambda = \sum_{i=1}^{n} \sum_j \Gamma_{ij} = n.$$

$$\Pi_j = \frac{1}{n} \sum_{i=1}^{n} \Gamma_{ij}$$

Gaussian Mixture model.

choose $\theta^{(0)}$ ↳ at random ↳ to initialize do k means, $\pi_k = \frac{1}{k}$.

Loop $t = 1, \ldots,$ convergence.

$$\gamma_{ij}^{(t+1)} = \frac{\pi_j^{(t)} f(x_i \mid \mu_j, \sigma_j^2)}{\sum \pi_c f(x_i \mid \mu_c, \sigma_c^2)}$$

→ E step

→ M step.

$$\mu_j^{(t+1)} = \frac{\sum_i \gamma_{ij}^{t+1} x_{ij}}{\sum_i \gamma_{ij}}$$

$$\sigma_j^{(t+1)} = \frac{\sum \gamma_{ij} (x_i - \mu_j^{t+1})^2}{\sum \gamma_{ij}}$$

$$\pi_j = \frac{1}{n} \sum_{i=1}^{n} \gamma_{ij}$$

EM algorithm. γ_{ij} = probabilistic assignment

We never saw z, so now let's talk about.

Full info model.

$$\sum\sum z_{ij} \ln\left[\pi_j f(x_i \mid \mu_j, \sigma_j^2)\right]$$

Full Information Model.

$$\mathcal{D}_i = \{(x_i, z_i)\}_{i=1}^{n}$$

z_i : . . . k

will replace z_i with one hot encoding

$$z_{ij} = 1 \text{ if } z_i = j$$

$$l(\theta) = \prod_{i=1}^{n} P(x_i, z_i \mid \theta)$$

$$= \prod_{j=1}^{k} \prod_{i=1}^{n} \left\{\pi_j f(x_i \mid \mu_j, \sigma_j^2)\right\}^{z_{ij}}$$

$$\mu_j = \frac{\sum z_{ij} x_i}{\sum z_{ij}} \qquad \sigma_j^2 = \frac{\sum z_{ij} (x_i - \mu_j)^2}{\sum z_{ij}}$$

$$\pi_j = \frac{\sum z_{ij}}{n}$$

Input: $\{(x_i, z_i)\}_{i=1}^{n}$

replace z_i with z_{ij} one hot.

$$\mu_j = \frac{\sum z_{ij} x_i}{\sum z_{ij}} \qquad \sigma_j^2 = \frac{\sum z_{ij} (x_i - \mu_j)^2}{\sum z_{ij}}$$

$$\pi_j = \frac{\sum z_{ij}}{n}$$

z_{ij} = hard assignment.

$$l(\theta) = \sum\sum z_{ij} \ln[\pi_j \cdot f(x_i, \mu_j, \sigma_j^2)]$$

why does GMM work?

$l(\theta)$

$$\sum_i \sum_j \Gamma_{ij} \ln[P(x_i, z_i = c|\theta)$$

$$= \underset{\Gamma_{ij}}{E} \ln[P(x_i, z_i = c|\theta)$$

$$\Gamma_{ij} = P(z_i = j | x_i, \theta^{(old)})$$

$$l(\theta) = \ln\left(\prod_i \prod_j \{f(x_i, z_i = j|\theta)\}^{z_{ij}}\right)$$ 7/2/23

$$l(\theta) = \sum\sum z_{ij} \ln f(x_i, z_i = j|\theta)$$

$$+ \lambda(\sum \pi_j - 1)$$

$$\pi_j f(x_i | \mu_j, \sigma_j^2)$$

$$\Gamma_{ij}^{(t+1)} = \frac{\pi_j^{(t)} f(x_i | \mu_j^{(t)}, \sigma_j^{2(t)})}{\sum_c \pi_c f(x_i | \mu_c, \sigma_c)}$$

$$= \frac{f(x_i, z_i = j | \theta^{(t)})}{f(x_i | \theta^{(t)})}$$

$$= P(z_i = j | x_i, \theta^{(t)})$$

$$Q(\theta, \theta^t) = \sum_{i=1}^{n} \sum_{j=1}^{K} \gamma_{ij}^{(t+1)} \ln f(x_i, z_i = j \mid \theta)$$

(E step)

$$= E_{z|x,\theta^t} \left[\ln f(x_i, z_i \mid \theta) \right]$$

$$\theta^{t+1} = \text{argmax} \left(Q(\theta, \theta^t) \right)$$

→ Exam there'll be a question with Z as a continous function.

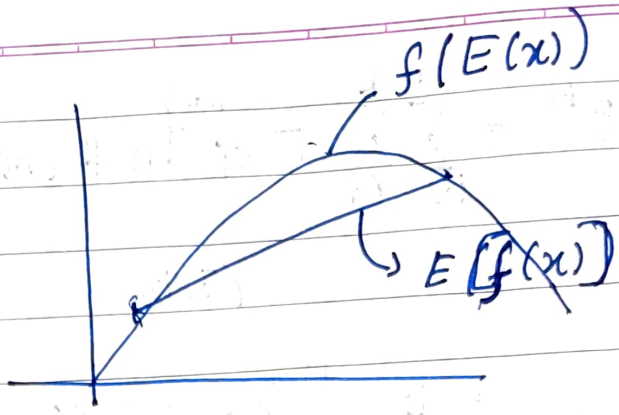
→ We have $P(z_i = j \mid x_i, \theta^{t+1})$
↳ get to $f(x_i, z_i = j \mid \theta)$
↳ argmax.

A PROOF:

claim: EM converges to local maxima

Alternatively: Show that every step

$$l(\theta^{(t)}) \leq l(\theta^{(t+1)})$$



$$\ell(\theta^t) = \sum_{i=1}^{n} \ln \sum_{j=1}^{k} \pi_j \, f(x_i \mid \mu_j, \sigma_j^2)$$

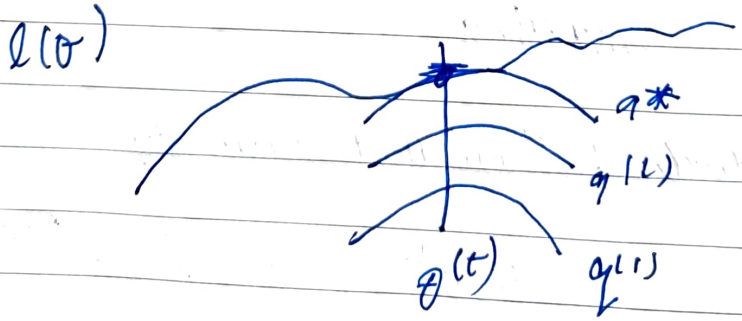
$$= \sum_{i=1}^{n} \ln \sum q(z_i = j) \, \pi_j \, \frac{f(x_i \mid \mu_j, \sigma_j^2)}{q(z_i = j)}$$

$E(x)$

q is any given dist^n

$$\geq \sum_{i=1}^{n} \sum_{j=1}^{n} q(z_i = j) \ln \frac{p(x_i, z_i = j \mid \theta^t)}{q(z_i = j)}$$

(Jensen's inequality)



$$\ell(\theta^t) - \sum\sum \frac{q(z_i=j)\ln P(x_i, z_i=j \mid \theta^{(t)})}{q(z_i=j)}$$

$$= \sum \ln P(x_i \mid \theta^{(t)}) - \sum\sum q(z_i=j)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} q(z_i=j) \ln P(x_i \mid \theta^{(t)}) - \sum_i \sum_j q(z_i=j) \ln\left(\frac{P(x_i, z_i=j \mid \theta^{(t)})}{q(z_i=j)}\right)$$

$$= \sum_i \sum_j q(z_i=j)\left(\ln P(x_i \mid \theta^{(t)}) - \frac{\ln P(x_i, z_i=j \mid \theta^t)}{q(z_i=j)}\right)$$

$$= \sum_i \sum_j q(z_i=j) \ln \frac{P(x_i \mid \theta^{(t)}) \times q(z_i=j)}{P(x_i, z_i=j \mid \theta^t)}$$

$$= \sum\sum q(z_i=j) \ln \frac{q(z_i=j)}{P(z_i=j \mid x_i, \theta_t)}$$

(KL divergence
↳ distance b/w
$q(z_i=j)$ and $P(z_i=j \mid x_i, \theta_t)$)

$= 0$

$q(z_i = j) = P(z_i = j \mid x_1, \theta^{(t)})$

$$\ell(\theta^{(t+1)}) = \sum E_q \ln \frac{P(x_i, z \mid \theta^{(t)})}{q}$$

$$+ \sum_i KL(q \,\|\, P(z \mid x_i, \theta^{(t)}))$$

given: $KL(q \| P) \geq 0$, $KL(q \| q) = 0$

$$= E_q[\ln(x_i, z \mid \theta^{(t)})] - E_q[\ln q] \geq 0$$

Entropy

$$+ \sum KL(q \| P(\_))$$

E step:

$$= Q(\theta^t, \theta^{(t)}) + \text{const} \quad \sum 0 \quad + 0$$

M step:

$$\theta^{t+1} = \arg\max_{\theta} Q(\theta, \theta^t)$$

$$Q(\theta^{t+1}, \theta^t) > Q(\theta^{(t)}, \theta^{(t)})$$

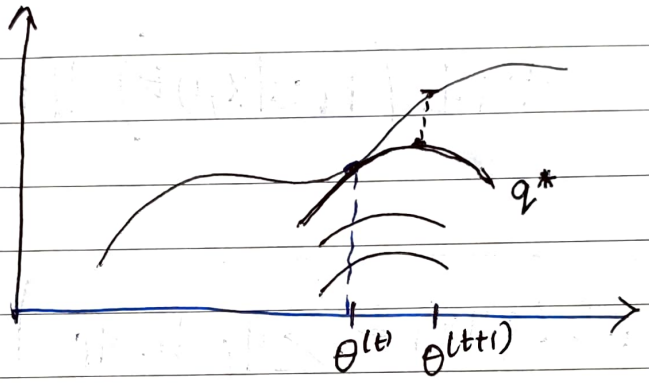
$l(\theta^{t+1})$

9|2|23

$$l(\theta|\mathcal{D}) = E_{q|z}\frac{\ln P(X,Z|\theta)}{q(z)} + KL(q \,||\, P(Z|X,\theta))$$

→ expected complete likelihood

→ entropy term.



at $\theta^{(t)}$ we want KL divergence to be 0

$$l(\theta^{(t)}|\mathcal{D}) = E_{z|x,\theta^{(t)}}\frac{\ln(P(x,z|\theta^{(t)}))}{P(z|x,\theta^{(t)})}$$

$q^*(z) = P(z|x,\theta^{(t)})$ + KL → 0

↳ the distance term goes to zero.

$$\theta^{t+1} \underbrace{E_{z|x,\theta^{(t)}} \ln P(x,z|\theta) - E_{z|x,\theta^{(t)}} \ln(z|x,\theta^{(t)})}_{\text{argmax}_\theta\, Q(\theta,\theta^t)}$$

↓ intermediate steps for explanation.

$$\underbrace{E_{z|x,\theta(t)} \ln P(x, z|\theta^{t+1})}_{\text{maximise}} - \underbrace{E_{z|x,\theta(t)} \ln (z|x, \theta^{(t)})}_{\text{constant}}$$

$$l(\theta^{(t+1)}) =$$

$$E_{z|x,\theta(t)} \ln P(x, z|\theta^{t+1}) - E_{z|x,\theta(t)} \ln P(z|x, \theta^{(t)})$$

$$+ KL\left(P(z|x, \theta^t) \,||\, P(z|x, \theta^{t+1})\right)$$

because there's a -ve the inequality flips

$$\boxed{KL(q||p) = \sum q_i \ln \frac{q_i}{p_i}}$$

HW : MLE map, qn. Bayesian map ← midsem.
↳ read this from Bishop

- 25% of midsem, will be from theory assignments.

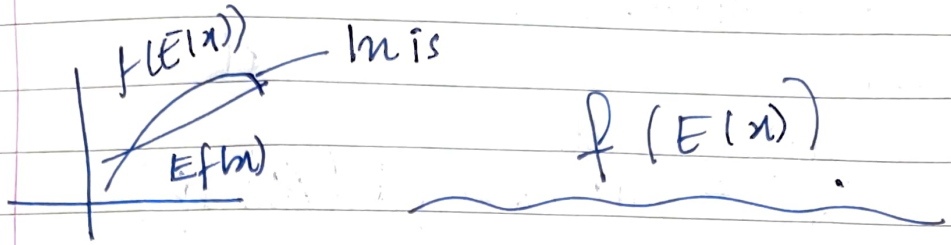
HMM

proof: $KL(q||p) \geq 0 \Rightarrow$ is positive

$$-\underbrace{\sum q(z_i = j)}_{E} \ln \frac{P(z_i = j \mid x, \theta^t)}{q(z_i = j)}$$

expectation of $f(x)$

We'll prove this using Jensen's Inequality.



$$\geq -\ln \sum q(z_i = j) \times \frac{P(z_i = j \mid x, \theta^t)}{q(z_i = j)}$$

because there's a -ve the inequality flips ≥ 0

$\therefore$ $\boxed{KL(q \| p) \geq 0}$ hence proved.

Mixture of Bernoullis.

1) coin 1 → P_1 = head

2) coin 2 → P_2 = head.

3) prob of choosing coin 1 = π_1 $\quad \pi_1 + \pi_2 = 1$

" coin 2 = π_2

Toss the coin for d rounds.

↳ $n_i \Rightarrow$ number of heads recieved.

repeat.

$$\mathcal{D} = \{n_i\}_{i=1}^{n}$$

$\theta = \pi_1, \pi_2, p_1, p_2$.

Estep

$$\gamma_{ij}^{(t+1)} = \frac{\pi_j^{(t)} \times P(x_i | p_j)}{\sum_j \pi_c \; P(x_i | p_j)}$$

$$P(x_i | p_j) = \binom{d}{n} (p_j)^n (1-p_j)^{(d-n)}$$

$P(X_i = n | p_j)$ — Binomial distribution

$$p_j^{t+1} = \sum_i \gamma_{ij}^{t+1} \; x$$

$$\gamma_{ij}^{(t+1)} = \frac{P(z_i = j, n_i | \theta^t)}{\sum_{k=1}^{2} P(z_i = k, n_i | \theta^t)}$$

$$= \frac{\pi_j \; P(n_i | z_i = j, \theta^t)}{\sum_{k=1}^{2} \pi_k \; P(n_i | z_i = j, \theta^t)}$$

$$= \pi_i \binom{d}{n} p_j^{n_i} (1-p_j)^{d-n_i}$$

$$\rho_j^{t+1} = \frac{\sum_i \Gamma_{ij}^{t+1} n_i}{\sum_i \left(\Gamma_{ij}^{t+1}\right) d}$$

$$\pi_j^{(t+1)} = \frac{\sum_i \Gamma_{ij}^{(t+1)}}{d\, n}$$