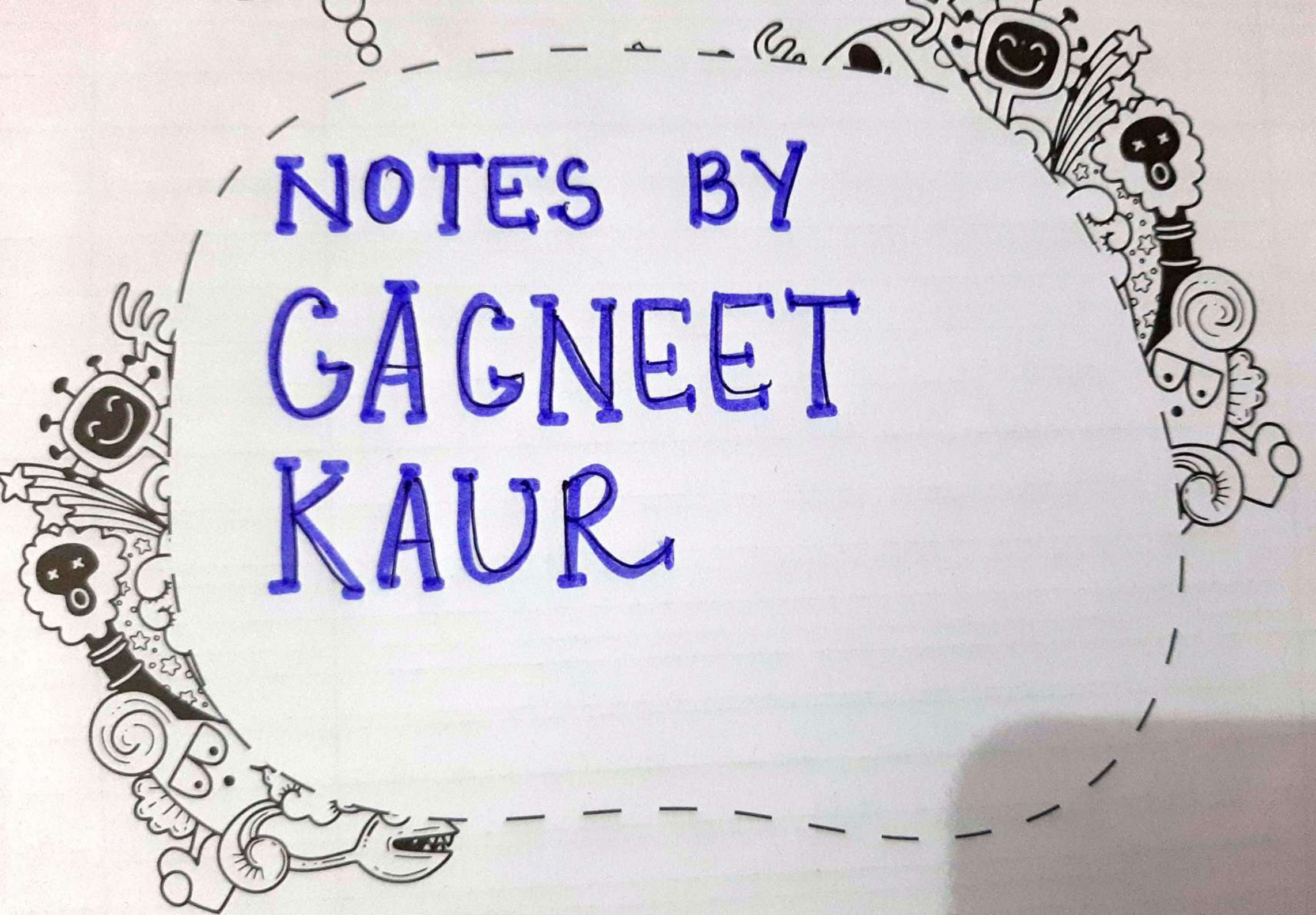

NOTES BY
GAGNEET
KAUR

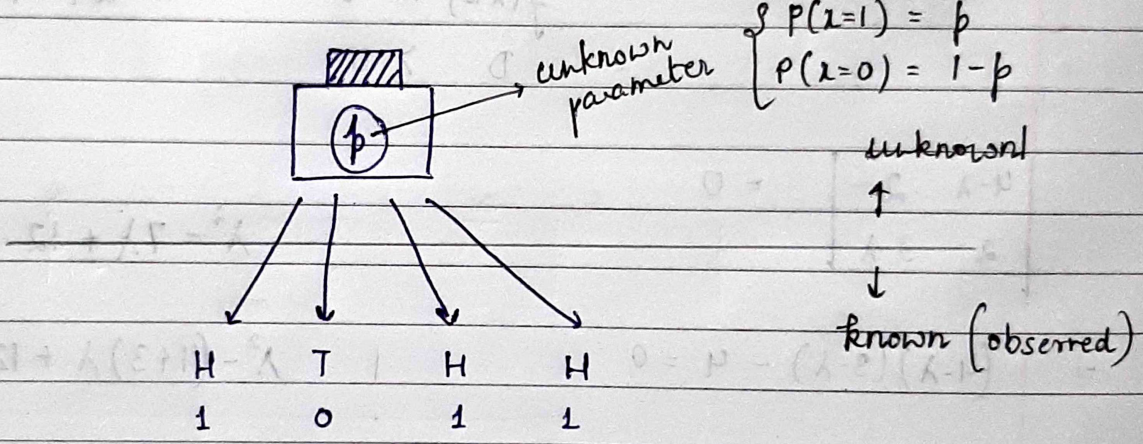
Week 4 MLT

Unsupervised Learning - ESTIMATION

Lecture 1 → Introduction to Estimation

"There is some probabilistic mechanism that generates data"
- about which we don't know something !!

Goal is : Given data find/estimate what we don't know.



Shrinath
PAGE NO.
DATE:

Steps of Estimation Problem : (General Idea Of Estimation)

- Observe data
- 'Assume' a probabilistic model that generates data
- Estimate unknown parameters using data.

ASSUMPTIONS

Data → $\{x_1, x_2, \ldots, x_n\}$

Observations are : (i) independent
(ii) identically distributed

INDEPENDENCE : $P(x_i/x_j) = p(x_i) \quad \forall\, i \neq j$

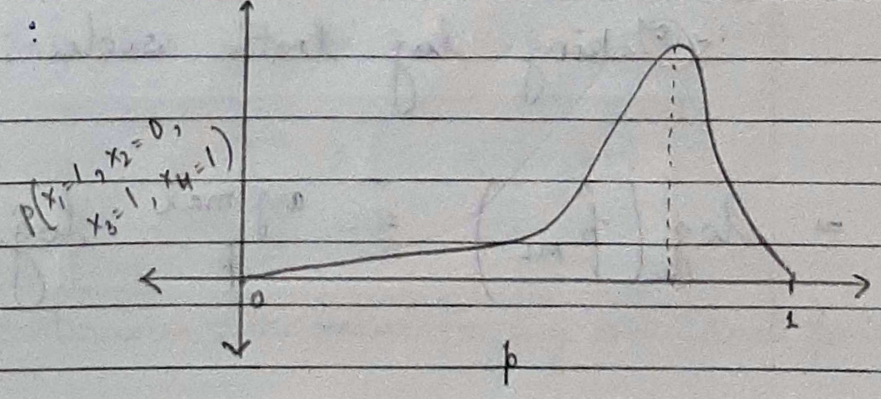
IDENTICAL DIST. : $p(x_i = 1) = p(x_j = 1) = p \quad \forall\, i, j$

Lecture 2 → Maximum Likelihood Estimation

The principled way to get estimators from data - LIKELIHOOD FUNCTION

LIKELIHOOD FUNCTION :

Eg: {1, 0, 1, 1}



For every value of P, we ask the question:
I have seen this data, if the true p was this, then what is the chance that I would have seen the state?

- CURVE PEAK → the value of p for which the chance of seeing the data is highest.

→ the data is most likely for which parameter of P, which choice of P.

Fischer's Principle of Maximum Likelihood

$$L\left(p, \{x_1, x_2, \ldots, x_n\}\right) = P(x_1, x_2, \ldots x_n ; p)$$ (p: underlying parameter)

$$= P(x_1; p) \cdot P(x_2; p) \ldots\ldots P(x_n; p)$$

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

↳ if $\begin{cases} x_i = 1 \Rightarrow p^1 (1-p)^0 = p \\ x_i = 0 \Rightarrow p^0 (1-p)^1 = 1-p \end{cases}$

ESTIMATOR $\hat{p}_{ML} = \underset{p}{\arg\max} \prod_{i=1}^{n} p^{x_i} (1-p)^{(1-x_i)}$

Taking log both sides,

$$\Rightarrow \log\left(\hat{p}_{ML}\right) = \underset{p}{\arg\max} \log\left(\prod_{i=1}^{n} p^{x_i} (1-p)^{(1-x_i)}\right)$$

$$= \arg\max_{p} \sum_{i=1}^{n} \left[x_i \log p + (1-x_i) \log (1-p) \right]$$

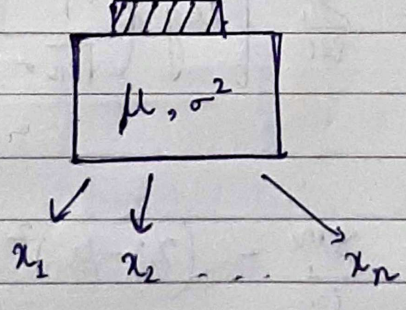
Take derivative of $\log \mathcal{L}(p)$, set it to 0 to get $\hat{p}_{ML}$.

$$\boxed{\hat{p}_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

⇒ Guess for the true 'p' by looking at the data is that value of 'p' which maximizes the chance that we observe this data.

Different form of Data

Data = $\{x_1, x_2, \ldots, x_n\}$ $\quad x_i \in \mathbb{R} \quad \forall i$



$x_i \sim$ Gaussian (μ, σ^2) $\forall i$

$\mu \rightarrow$ unknown ; $\sigma^2 \rightarrow$ known

$$L\left(\mu, \sigma^2, \{x_1, \ldots, x_n\}\right) = P\left(x_1, x_2, \ldots, x_n ; \mu, \sigma^2\right)$$

$$= \prod_{i=1}^{n} P\left(x_i ; \mu, \sigma^2\right)$$

→ because Gaussian is continuous, the value for any individual point is going to be ZERO.

→ Only intervals will have non-zero values for continuous distributions.

Thus the above fn will give us the value ZERO for any mean.

So, Fishers proposal was to not use the probabilities
Instead, replace the probabilities and define the likelihood of the parameter as :

$$\mathcal{L}(\mu, \sigma^2, \{x_1, \ldots, x_n\}) = f_{x_1, \ldots, x_n}(x_1, \ldots, x_n ; \mu, \sigma^2)$$

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

$$= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x_i - \mu)^2 / 2\sigma^2} \right]$$

$$\log \mathcal{L}(\mu, \sigma^2, \{x_1, \ldots x_n\}) = \sum_{i=1}^{n} \left[\log\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right) - \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

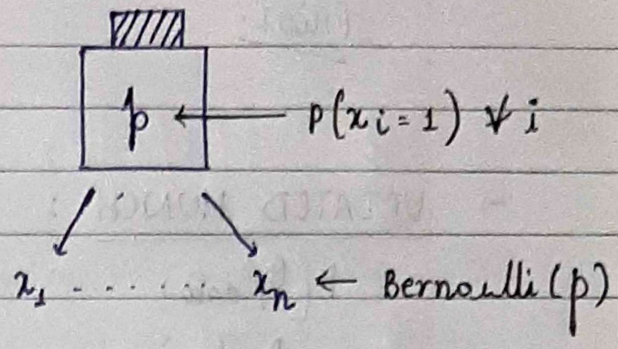
$$\hat{\mu}_{ML} = \underset{\mu}{\arg\max} \sum_{i=1}^{n} -(x_i - \mu)^2$$

Take derivative & set it to zero !!

$$\therefore \quad \boxed{\hat{\mu}_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

Lecture 3 : Bayesian Estimation

Consider the coin example :



Let's say someone says,
'I believe the bias p is somewhere close to 1'

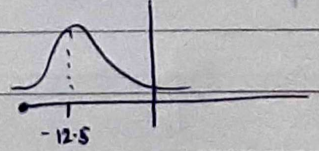
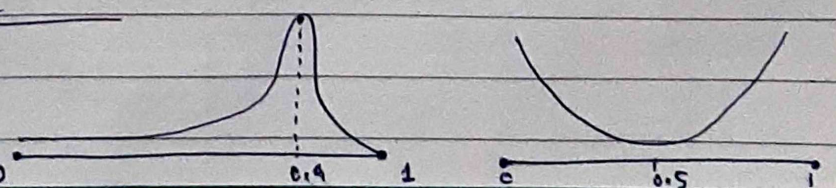
We may have 'HUNCH' about parameters.

⇒ Is there a way we can somehow codify our hunch into mathematically more precise mechanisms that can be incorporated into our estimation procedure?

→ BAYESIAN MODELLING APPROACH

GOAL : Incorporate "hunch / belief" about parameters of interest into the estimation procedure.

APPROACH : Think of the parameter to estimate as a 'random' variable.

Earlier	Now
μ	
p	

→ HUNCH : codified using a probability distribution over θ

$P(\theta)$

Prior.

⇓ DATA

→ UPDATED HUNCH : Codified using prob. distribution

$P(\theta/\text{data})$

Posterior

BAYES LAW : $P(A/B) = \dfrac{P(B/A) \cdot P(A)}{P(B)}$

A → Parameters θ

B → Data $\{x_1, \ldots, x_n\}$

$$\underbrace{P(\theta/\{x_1,\ldots,x_n\})}_{\text{Posterior}} = \left(\frac{\overbrace{P(\{x_1,\ldots,x_n\}/\theta)}^{\text{likelihood}}}{\underbrace{P(\{x_1,\ldots,x_n\})}_{\text{does not depend on } \theta}}\right) \cdot \underbrace{P(\theta)}_{\text{Prior}}$$

EVIDENCE

What is Bayes Theorem Saying ? : You have some initial belief about your prior distribution. To go to a posterior distribution, you have to review your prior dist. You have to make a multiplicative update to this prior.

DATA — Bernoulli (p)

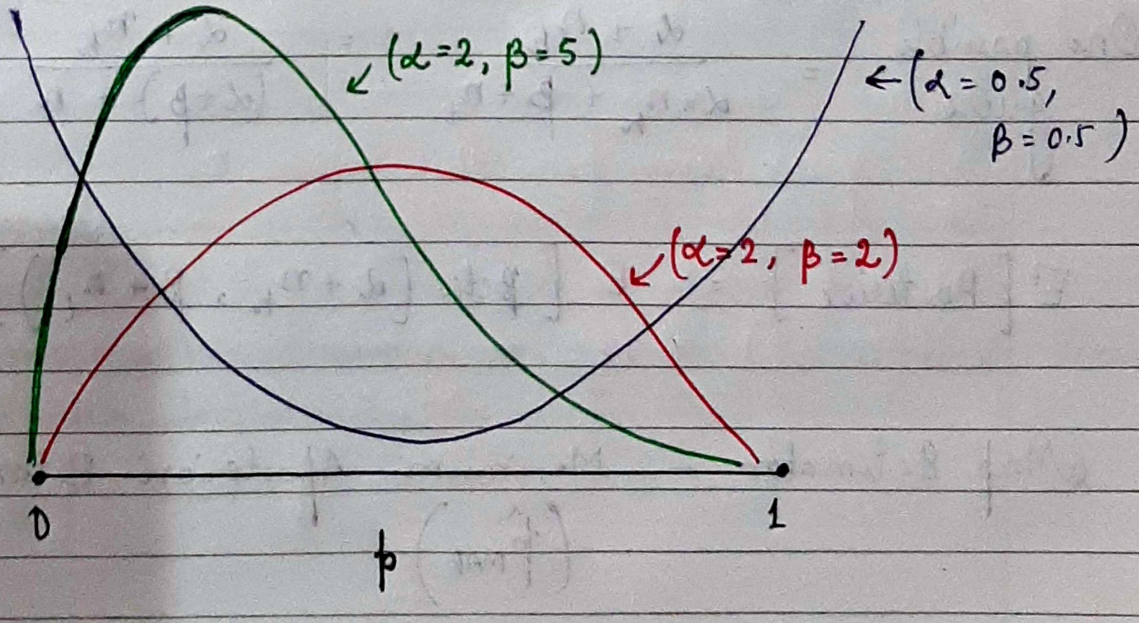
Prior? $P(\theta)$

Beta Prior → $f(p; \alpha, \beta)$

- We want some continuous distribution whose values are between 0 & 1.
- We can't use Gaussian here because it can give me any value between $-\infty$ and ∞, but then the parameter I am trying to estimate (P) takes values b/w 0 & 1.
- So any distribution that we use as a prior to encode our prior knowledge should be supported only in 0 & 1.
- A good choice — BETA PRIOR.

BETA PRIOR

$$f(p; \alpha, \beta) = \begin{cases} \frac{p^{\alpha-1}(1-p)^{\beta-1}}{Z} & \forall\ p \in [0,1] \end{cases}$$



It kind of says that when you have small α & big β, then true p is somewhere closer to 0 than 1.

$$P(\theta/\text{data}) \propto P(\text{data}/\theta) \cdot P(\theta)$$

$$f_{p/\text{data}}(p) \propto \left(\prod_{i=1}^{n} \left[p^{x_i}(1-p)^{(1-x_i)}\right]\right) \cdot \left[p^{\alpha-1}(1-p)^{\beta-1}\right]$$

Likelihood — Prior

$$f_{p/\text{Data}}(p) \propto \underbrace{p^{\sum x_i + \alpha - 1}(1-p)^{\sum (1-x_i) + \beta + 1}}$$

Same functional form as PRIOR !!

Beta Prior $(\alpha, \beta) \xrightarrow[\text{Bernoulli}]{\text{DATA}}$ Beta posterior $(\alpha + n_h, \beta + n_t)$

One possible guess $= \dfrac{\alpha + n_h}{\alpha + n_h + \beta + n_t} = \dfrac{\alpha + n_h}{(\alpha+\beta) + n}$

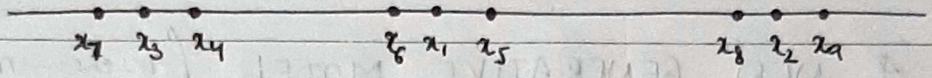
$$E[\text{Posterior}] = E[\text{Beta}(\alpha + n_h, \beta + n_t)] = \frac{\alpha + n_h}{(\alpha+\beta) + n}$$

Map Estimator – Maximum Aposteriori Estimator

$(\hat{p}_{MAP})$

Lecture 4 : Gaussian Mixture Models

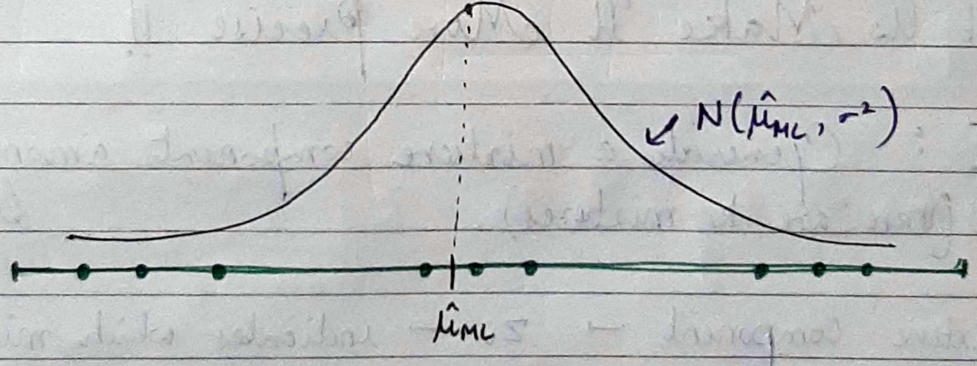
GOAL : Estimation for slightly complicated data



→ Goal is to come up with a model that explains the data.

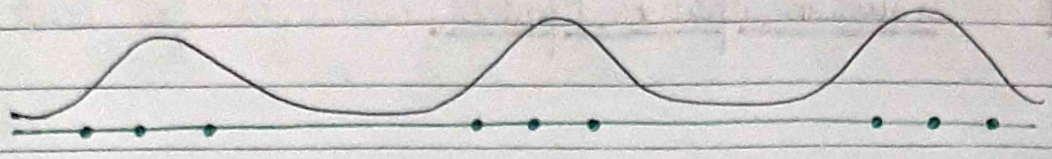
What could be a good generative story?

- Try measuring or modelling this using a Gaussian distribution.
- Let us say we did a principle of maximum likelihood to get the best Gaussian



- The sample mean is the maximum likelihood estimator.

Our explanation via Gaussian is not very satisfactory because there are two low density regions for the Gaussian.



- We want something, a density like above to explain this data. Basically, we want three different models or three places where the density should peak.

A NEW GENERATIVE MODEL (mechanism that generates the data)

The name of this new model that we are going to come up is called as mixtures of Gaussians.

STEP 1: Pick which mixture a data point comes from.
STEP 2: Generate data points from that mixture.

⊙ Let Us Make It More Precise !!

STEP I : Generate a mixture component among $\{1, \ldots, k\}$
(There are k mixtures)

Mixture Component → z_i → indicates which mixture the i^{th} data point comes from that values z_i

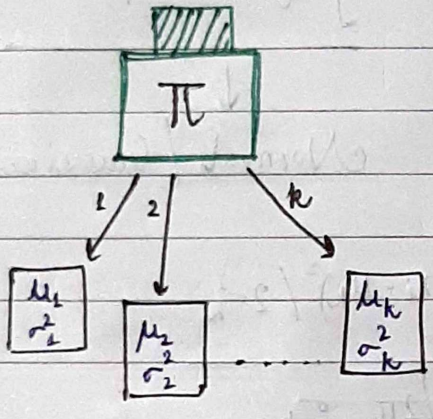
Let us formulate it as Probability Distribution :

$P(z_i = l) = \pi_l$ ← It means that the probability that the i^{th} data point comes from l^{th} mixture is given by π_l. $\left(\sum_{i=1}^{k} \pi_i = 1\right)$

So now we have decided which cluster or mixture the data has to come from !!

STEP II : Generate $x_i \sim N(\mu_{z_i}, \sigma^2_{z_i})$

for each mixture we have a Gaussian with its own mean & variance.



$\{x_1, \ldots, x_n\}$ → OBSERVED

$\{z_1, \ldots, z_k\}$ → UNOBSERVED / LATENT

Latent Variable Models

Our Gaussian Mixture Model is a latent variable model, because the final output that you are observing depends not only on some parameters that you want to estimate, but also on some unobserved latent variables.

HOW MANY PARAMETERS WILL DETERMINE THIS MODEL COMPLETELY ?

$$\pi = [\pi_1 \quad \pi_2 \quad \ldots \quad \pi_k]$$

for each K there are 2 parameters → μ_k & σ^2_k

In Total : $2K + \underset{(\pi)}{k} - 1 = 3K$

We need to estimate $(3K-1)$ parameters from data.

Lecture 5 : Likelihood of GMM

$$\mathcal{L}\left(\begin{matrix}\mu_1,\dots,\mu_K \\ \sigma_1^2,\dots,\sigma_K^2 \\ \pi_1,\dots,\pi_K\end{matrix} \quad x_1,\dots,x_n\right) = \prod_{i=1}^{n} f_{mix}\left(x_i ; \begin{matrix}\mu_1,\dots,\mu_K \\ \sigma_1^2,\dots,\sigma_K^2 \\ \pi_1,\dots,\pi_K\end{matrix}\right)$$

$$= \prod_{i=1}^{n}\left[\sum_{k=1}^{K} \pi_k \cdot \underbrace{f\left(x_i ; \mu_k, \sigma_k^2\right)}_{\downarrow}\right]$$

Normal / Gaussian Density

$$L(\theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \frac{e^{-(x_i-\mu_k)^2/2\sigma_k^2}}{\sqrt{2\pi}\,\sigma_k}$$

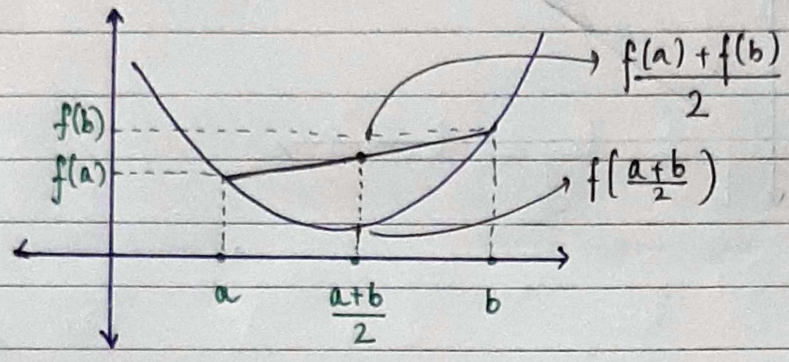
θ ↑ all parameters

$$\log L(\theta) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k \frac{e^{-(x_i-\mu_k)^2/2\sigma_k^2}}{\sqrt{2\pi}\,\sigma_k}\right)$$

- Not possible to solve the above eqn Analytically.
- Need an alternate way to solve this efficiently!

Lecture 6 : Convex Functions & Jensen's Equality

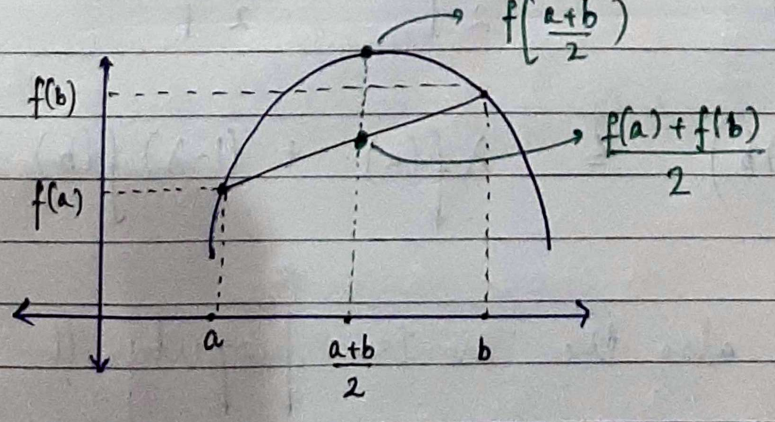
- What is a Convex Function ?



If the linear interpolation at two points has a strictly higher value than the function itself → such a function is called CONVEX FUNCTION.

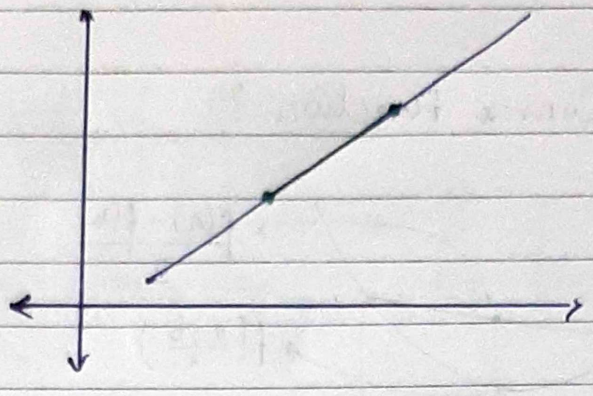
$$f\left(\frac{a+b}{2}\right) \leq \frac{f(a)+f(b)}{2} \qquad \forall\ a, b$$

- What is CONCAVE Function ?



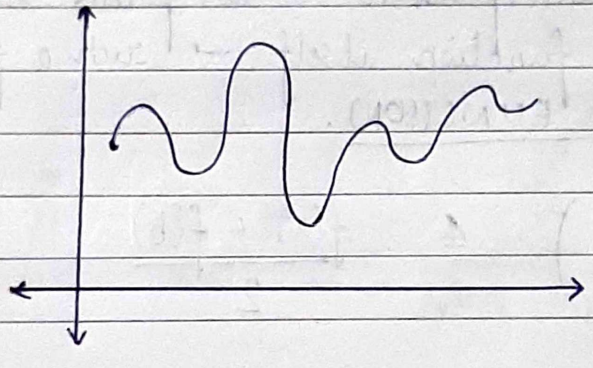
$$\forall\ a, b \qquad f\left(\frac{a+b}{2}\right) \geq \frac{f(a)+f(b)}{2}$$

- If we have a linear fn ⟶ Both Concave & Convex



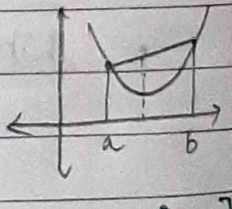
$$f\left(\frac{a+b}{2}\right) = f\frac{(a)+f(b)}{2}$$

- Neither concave nor convex :



Convex :

$$f\left(\frac{1}{2}a + \frac{1}{2}b\right) \leq \frac{1}{2}f(a) + \frac{1}{2}f(b)$$



$$\Rightarrow f\left(\lambda a + (1-\lambda)b\right) \leq \lambda f(a) + (1-\lambda)f(b) \qquad \lambda \in [0,1]$$

Concavity has also the similar property !!

- The line segment joining a and b. You can generally use any lambda — as u vary lambda you are travelling from a to b.

General: Now you can extend it to multiple points :

$$f\left(\lambda_1 a_1 + \lambda_2 a_2 + \ldots\ldots + \lambda_k a_k\right) \geq \lambda_1 f(a_1) + \ldots + \lambda_k f(a_k)$$

$$\boxed{f\left(\sum_{k=1}^{K} \lambda_k a_k\right) \geq \sum_{k=1}^{K} \lambda_k f(a_k)} \quad \text{JENSEN'S INEQUALITY}$$

$$\left(\sum_{k=1}^{K} \lambda_k = 1\right) \qquad \left(0 \leq \lambda_k \leq 1 \;,\; 1 \leq k \leq K\right)$$

- Log is a concave function !
- How can we exploit Jensen's for performing maximum likelihood ?

Lecture 7 : Estimating The Parameters

Recall,

$$\log L(\theta) = \sum_{i=1}^{m} \log\left(\sum_{k=1}^{K}\left(\pi_k \frac{e^{-(x_i-\mu_k)^2/2\sigma_k^2}}{\sqrt{2\pi}\,\sigma_k}\right)\right)$$

- Introduce - for every data point i, the parameters $\{\lambda_1^i, \ldots\ldots, \lambda_k^i\}$ such that $\forall i \sum_{k=1}^{K} \lambda_k^i = 1$, $0 \leq \lambda_k^i \leq 1$ $\forall i, k$

$$\log L(\theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \lambda_k^i \left(\frac{\pi_k \frac{e^{-(x_i-\mu_k)^2/2\sigma_k^2}}{\sqrt{2\pi}\sigma_k}}{\lambda_k^i} \right) \right)$$

By Jensen's Inequality :

$$\log L(\theta) \geq \text{modified\_log } L(\theta, \lambda)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \lambda_k^i \log \left(\frac{\pi_k \, e^{-(x_i-\mu_k)^2/2\sigma_k^2}}{\lambda_k^i \cdot \sqrt{2\pi}\sigma_k} \right)$$

- Note that the above modified log likelihood gives a lower bound for the true log likelihood at θ for any choice of λ

$\lambda \hookrightarrow \begin{Bmatrix} \lambda_1^1, \ldots, \lambda_K^1 \\ \vdots \\ \lambda_1^n, \ldots, \lambda_K^n \end{Bmatrix}$ $\quad$ $\theta \downarrow \begin{Bmatrix} \mu_1, \ldots, \mu_K \\ \sigma_1^2, \ldots, \sigma_K^2 \\ \pi_1, \ldots, \pi_K \end{Bmatrix}$

KEY INSIGHT :

- If we fix λ, it is easy to maximize wrt θ
- If we fix θ, it is easy to maximize wrt λ.

(I) FIX λ and MAXIMIZE OVER θ

$$\max_{\theta} \sum_{i=1}^{n} \sum_{k=1}^{K} \lambda_k^i \left[\log \left(\pi_k \frac{e^{-(x_i-\mu_k)^2/2\sigma_k^2}}{\sqrt{2\pi}\sigma_k} \right) \cdot \frac{1}{\lambda_k^i} \right]$$

λ_k^i → Probability that the i^{th} point goes into the k^{th} cluster !!

Shriram
PAGE NO.
DATE:

$$= \max_{\theta} \sum_{i=1}^{n} \sum_{k=1}^{K} \left[\lambda_k^i \log \pi_k - \lambda_k^i \frac{(x_i - \mu_k)^2}{2\sigma_k^2} - \lambda_k^i \log \sqrt{2\pi}\,\sigma_k \right]$$

Take derivative wrt μ, σ to get :

$$\hat{\mu}_k^{MML} = \frac{\sum_{i=1}^{n} \lambda_k^i x_i}{\sum_{i=1}^{n} \lambda_k^i}$$

means of the data points assigned to a particular cluster

$$\hat{\sigma}_k^{2\,MML} = \frac{\sum_{i=1}^{n} \lambda_k^i \left(x_i - \hat{\mu}_k^{MML}\right)^2}{\sum_{i=1}^{n} \lambda_k^i}$$

sample variance of data points assigned to a particular cluster

Maximizing π :

$$\max_{\pi_1, \ldots, \pi_K} \sum_{i=1}^{n} \left(\sum_{k=1}^{K} \lambda_k^i \log \pi_k \right) \quad \text{such that} \quad \sum_k \pi_k = 1 \; ; \; \pi_k \geq 0$$

can solve using method of Lagrange multipliers,

$$\hat{\pi}_k^{MML} = \frac{\sum_{i=1}^{n} \lambda_k^i}{n}$$

From fixing λ, we get :

$$\hat{\mu}_k^{MML} = \frac{\sum_{i=1}^{n} \lambda_k^i x_i}{\sum_{i=1}^{n} \lambda_k^i} \; ; \; \hat{\sigma}_k^{2\,MML} = \frac{\sum_{i=1}^{n} \lambda_k^i \left(x_i - \hat{\mu}_k^{MML}\right)^2}{\sum_{i=1}^{n} \lambda_k^i}$$

$$\hat{\pi}_k^{MML} = \frac{\sum_{i=1}^{n} \lambda_k^i}{n}$$

(II) Fix θ & Maximize λ wrt

$$= \sum_{i=1}^{n} \left[\sum_{k=1}^{K} \lambda_k^i \log(a_{ik}) - \lambda_k^i \log \lambda_k^i \right]$$

where $a_{ik} = \left(\pi_k \, e^{-(x_i - \mu_k)^2 / 2\sigma_k^2} \cdot \frac{1}{\sqrt{2\pi}\sigma_k} \right)$

For any i,

$$\max_{\lambda_1^i, \ldots \lambda_K^i} \sum_{k=1}^{K} \left[\lambda_k^i \log(a_{ik}) - \lambda_k^i \log \lambda_k^i \right]$$

such that $\sum_{k=1}^{K} \lambda_k^i = 1$, $0 \le \lambda_k^i \le 1$

Can be solved analytically, $\left(P(z_i = k \mid x_i) = \frac{P(x_i / z_i = k) \cdot P(z_i = k)}{P(x_i)} \right)$

$P(x_i / z_i = k)$ ↓

$$\hat{\lambda}_k^{i\ MML} = \frac{\left(\frac{1}{\sqrt{2\pi}\sigma_k} e^{-(x_i - \mu_k)^2 / 2\sigma_k^2} \right) \cdot \pi_k}{\sum_{l=1}^{K} \left(\frac{1}{\sqrt{2\pi}\sigma_l} e^{-(x_i - \mu_l)^2 / 2\sigma_l^2} \cdot \pi_l \right)}$$

← $P(z_i = k)$

↑ $P(z_i = k \mid x_i)$

↑ $P(x_i)$

Lecture 8 : EM Algorithm

→ Initialize $\theta^0 = \{ \mu_1^0, \ldots, \mu_K^0 \; ; \; \sigma_1^{0^2}, \ldots, \sigma_K^{0^2} \; ; \; \pi_1^0, \ldots, \pi_K^0 \}$

→ until convergence ($\| \theta^{t+1} - \theta^t \| \le \epsilon$) ← Tolerance Parameter

Expectation step: $\lambda^{t+1} = \arg\max_{\lambda}$ modified $\log L(\theta^t, \lambda)$

Maximization step: $\theta^{t+1} = \arg\max_{\theta}$ modified $\log L(\theta, \lambda^{t+1})$

→ end

- EM produces "soft clustering"
- EM takes variances into account.
- EM need not be normal regions!