Unsupervised learning

- Representation learning
  - PCA / kernel PCA
- Clustering
  - LLyod's / k-means.

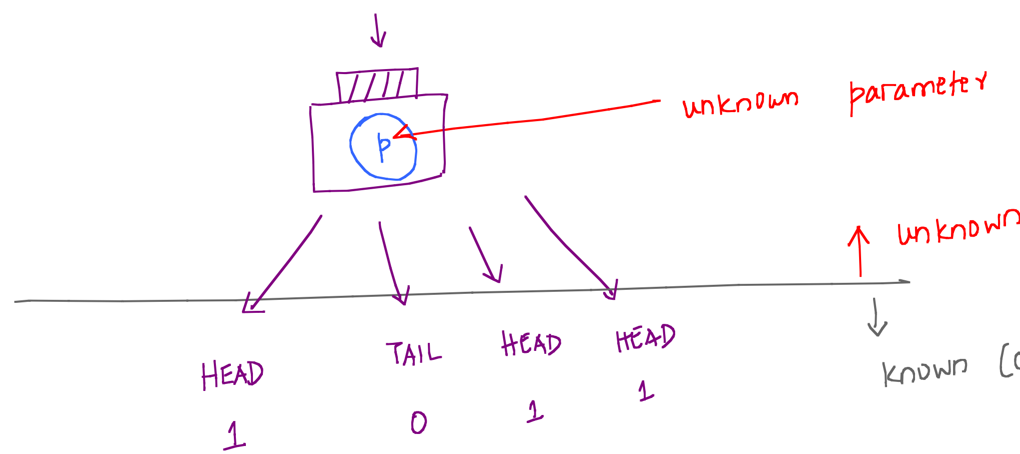
Not Assumed any model for data

Probabilistic assumption

- ESTIMATION

" There is some probabilistic mechanism that generates data about which we don't know "something".. Given data, find/estimate what we don't know"

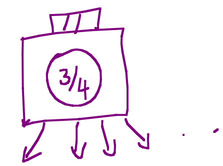
- OBSERVE data
- "ASSUME" a probabilistic model that generates data
- ESTIMATE unknown parameters using data.



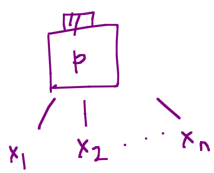
$$\begin{cases} P(X=1) = p \\ P(X=0) = 1-p \end{cases}$$

OBSERVE $\{1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1\}$

ESTIMATE: $9/12 = \boxed{3/4}$



ASSUMPTIONS



OBSERVATIONS ARE

(i) INDEPENDENT

(ii) IDENTICALLY DISTRIBUTED

Guess $= 2/3$? Y

$= 0.0001$? Y

$= 0$? N

$= 1$? N

INDEPENDENCE

$$P(x_i / x_j) = P(x_i) \quad \forall i \neq j$$

IDENTICAL DISTRIBUTION

$$P(x_i = 1) = P(x_j = 1) = p \quad \forall i, j$$

LIKELIHOOD FUNCTION

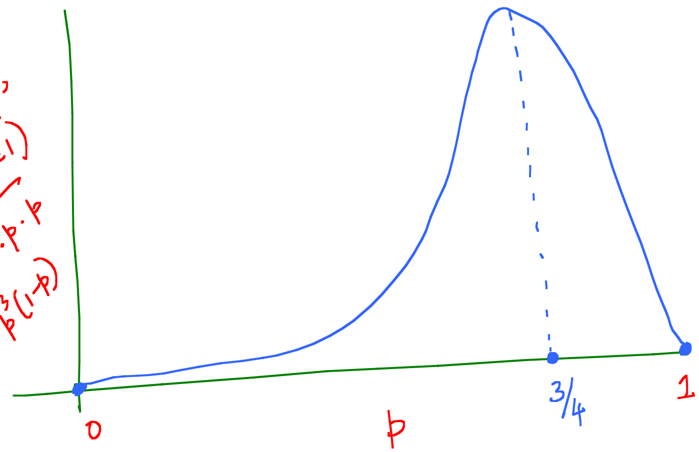
(eg) : $\{1, 0, 1, 1\}$

$3/4$

$P(x_1=1, x_2=0, x_3=1, x_4=1)$

$p \cdot (1-p) \cdot p \cdot p$

$= p^3(1-p)$



FISHER'S PRINCIPLE OF MAXIMUM LIKELIHOOD

$$\mathcal{L}(p ; \{x_1, x_2, \dots, x_n\}) = P(x_1, x_2, \dots, x_n ; p)$$

underlying parameter.

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

[Independence]

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

$$\begin{cases} \text{If } x_i = 1 \Rightarrow p^1 (1-p)^0 = p \\ \text{If } x_i = 0 \Rightarrow p^0 (1-p)^{1-0} = 1-p. \end{cases}$$

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

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

[log is monotonic increasing]

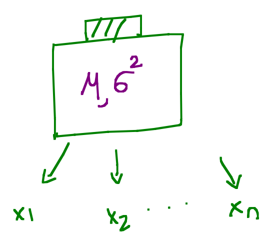
$$= \arg\max_{p} \sum_{i=1}^{n} \left[x_i \log p + (1-x_i) \log (1-p) \right]$$

Take derivative of $\log L(p)$, set it to 0 to get

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

Fraction of 1's

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



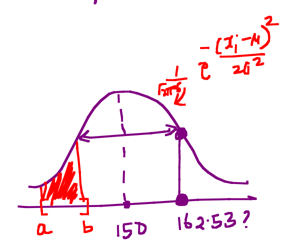
$x_i \sim \text{Gaussian}(\mu, \sigma^2)$ $\forall i$

$\mu \to$ unknown ; $\sigma^2 \to$ known.

$$L(\mu, \sigma^2, \{x_1, \dots, x_n\}) = P(x_1, \dots, x_n ; \mu, \sigma^2)$$

$$= \prod_{i=1}^{n} \underbrace{P(x_i ; \mu, \sigma^2)}_{0} \quad \times$$

$x_i = 162.53$ cm.



$$L(\mu, \sigma^2, \{x_1, \dots, x_n\}) = f_{X_1, \dots, X_n}(x_1, \dots, x_n ; \mu, \sigma^2)$$

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

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

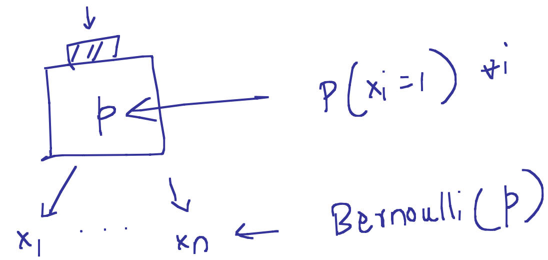
$$\log L\left(\mu, \sigma^2, \{x_1, \dots, x_n\}\right) = \sum_{i=1}^{n} \left[\underbrace{\log\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)}_{\times} - \frac{(x_i-\mu)^2}{2\sigma^2} \right]$$

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

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

BAYESIAN MODELING

Consider the coin examples.

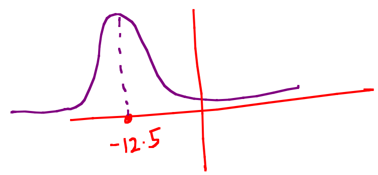
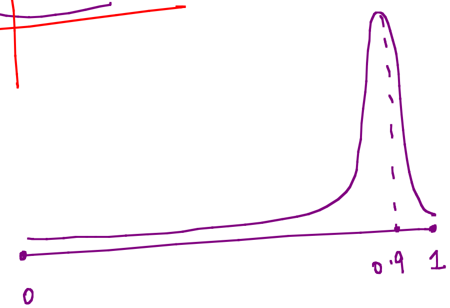
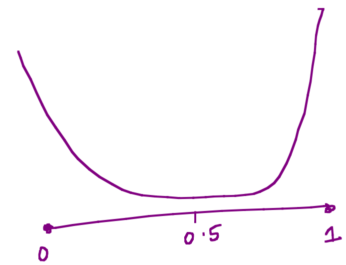


Let's say someone says: "I believe the bias p is somewhere close to 1"

We may have "HUNCH" about parameters

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

<u>APPROACH</u>: Think of the parameter to estimate as a "random" variable.

<u>EARLIER</u>	<u>NOW</u>
μ	
p	 

HUNCH $\rightarrow$ Codified using a probability distribution over θ — $P(\theta)$

PRIOR

$\Downarrow$ DATA.

UPDATED HUNCH — Codified using a prob. distribution — $P(\theta / DATA)$

POSTERIOR

Bayes law

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

A → Parameters θ

B → DATA $\{x_1, \ldots, x_n\}$

LIKELIHOOD

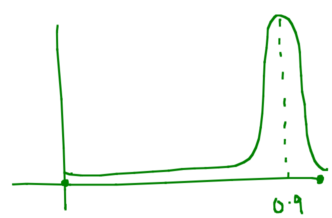
$$\underbrace{P(\theta / \{x_1, \ldots, x_n\})}_{\text{POSTERIOR}} = \left(\frac{P(\{x_1, \ldots, x_n\} / \theta)}{P(\{x_1, \ldots, x_n\})} \right) \cdot \underbrace{P(\theta)}_{\text{PRIOR}}$$

Does not depend on θ.

EVIDENCE

EXAMPLE

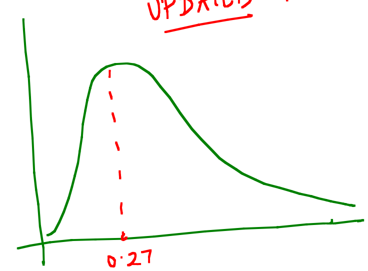
HUNCH



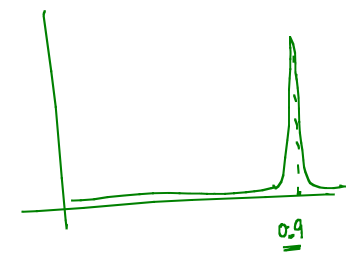
DATA

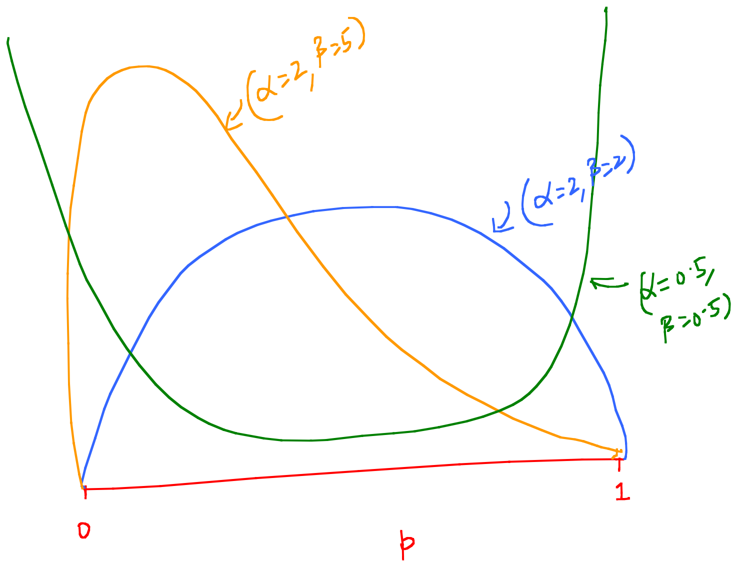
$\{0, 0, 0, 0, 0, 1, 1, 0, 0, 0\}$

UPDATED HUNCH



$\{1,1,1,1,1,1,0,1,1,1,1\}$





DATA - Bernoulli(p)

PRIOR ? $P(\theta)$

BETA PRIOR

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

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

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

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

↳ Same functional form as the PRIOR!!

$$\text{BETA PRIOR } (\alpha, \beta) \xrightarrow[\text{Bernoulli}]{\text{DATA}} \text{BETA POSTERIOR } (\alpha + n_h, \beta + n_t)$$

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

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

MAP Estimator - Maximum Aposteriori Estimator

$\hat{p}_{MAP}$