Lecture 9

ML as a random experiment

- Let's make prediction whether a fish is a Salmon or a Mackerel by its size

$X = \mathbb{R}$

$Y = \{ \underset{\text{Mackerel}}{1}, \underset{\text{salmon}}{2} \}$

R.V. R.V.

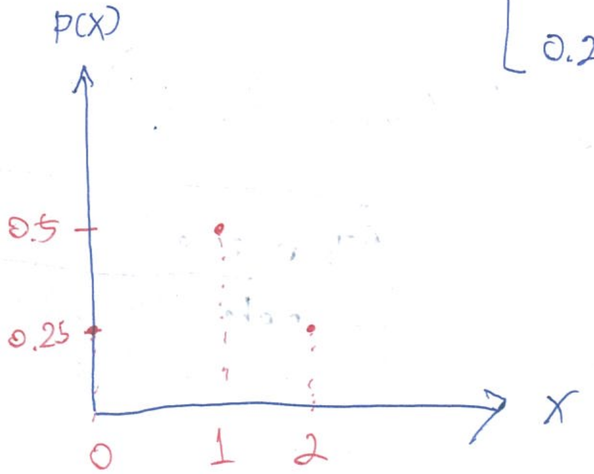
- Each sample $(x_i, y_i) \sim \underbrace{P(\underset{\sim}{X}, \underset{\sim}{Y})}$

Joint probability distribution which we have no access to

- **Probability Distribution**: A function that gives the probability of occurrence of different outcomes

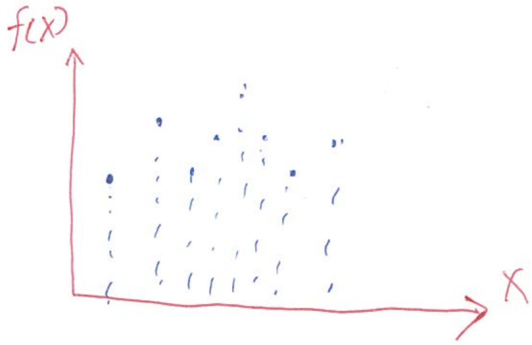
Ex $P(X_i) = \begin{cases} 0.5 & \text{if } x = 1 \\ 0.25 & \text{if } x = 0 \\ 0.25 & \text{if } x = 2 \end{cases}$

Here, X is a discrete R.V. denoting the number of heads observed in a double coin toss



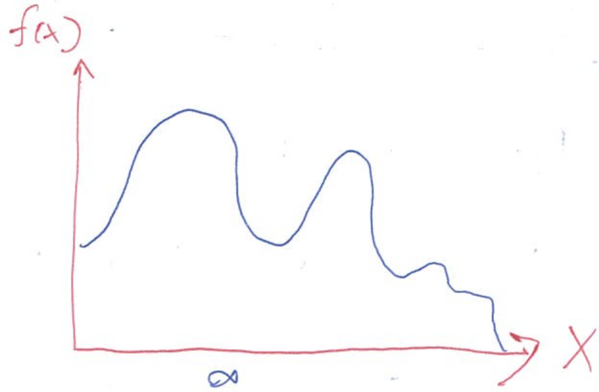
- **Probability mass function**: A function that gives the probability that a discrete R.V. is equal to some value

Probability density function: A function ~~that~~ whose value at any given sample is a likelihood that the value of r.v. equal that sample p.m.f.



$$\sum_i f(x_i) = 1$$

p.d.f.



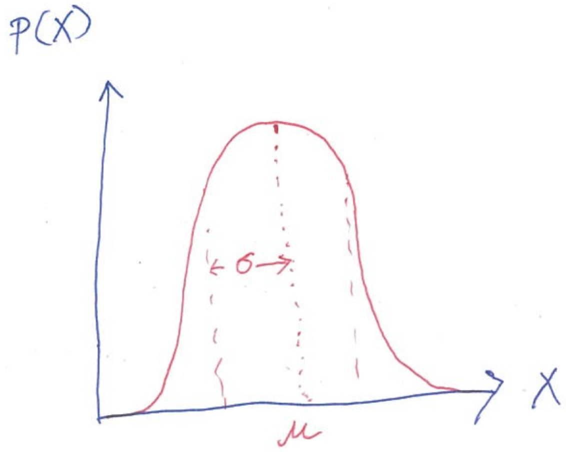
$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

Our distribution $P(X)$:

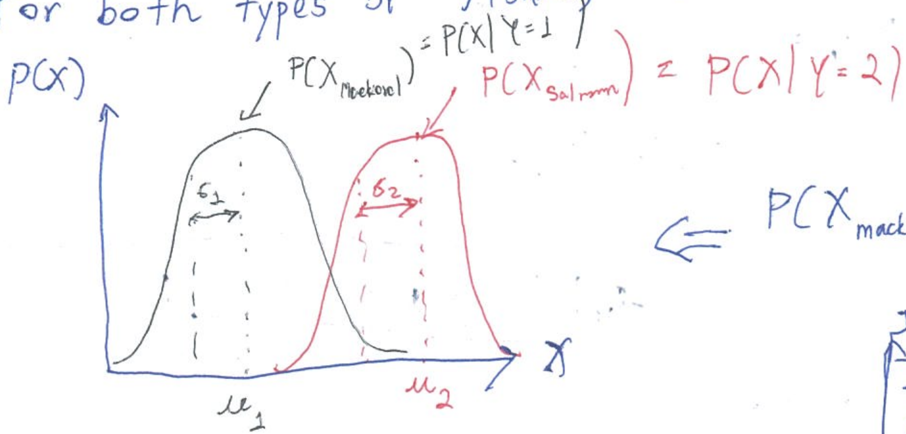
- Condiser one type of fishes, $x_i \sim P(X)$

R.V.

We shall assure a normal/Guassian distribution



- For both types of fishes, we would have

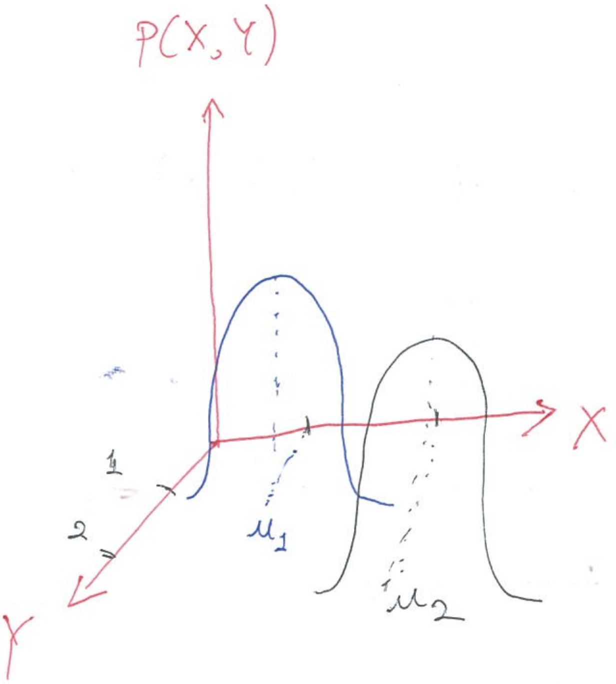


$\Leftarrow P(X_{mackerel}, X_{salmon})$

Joint probability distribution

A probability distribution of at least two jointly r.v.

- How to think of $P(X, Y)$?



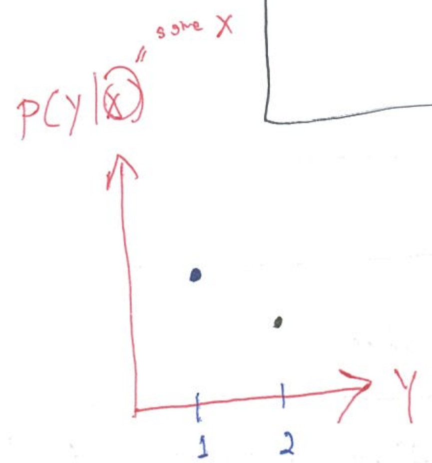
- Let's say if we have ~~no~~ access to $P(X,Y)$, then we could build an optimal classifier that make prediction by

Bayes classifiers

$$\boxed{h(x) = \arg\max_{\forall y} P(Y|x)}$$

strict at x

Ex



$\Rightarrow$ optimal prediction: $h(x) = 1$

- **Conditional probability distribution:** Given two jointly r.v., $P(Y|X)$ is the distribution of Y when X is known to be a particular value

- Huge observation: If accessing $P(X,Y)$ was possible, we then would have an optimal classifier.

Two ML approches:

① Approximate / estimate $P(Y|X)$ directly

Discriminative learning

(E.g. k-NN, Perceptrons)

② Approximate / estimate $P(X, Y)$.

Then, apply Bayes classifier

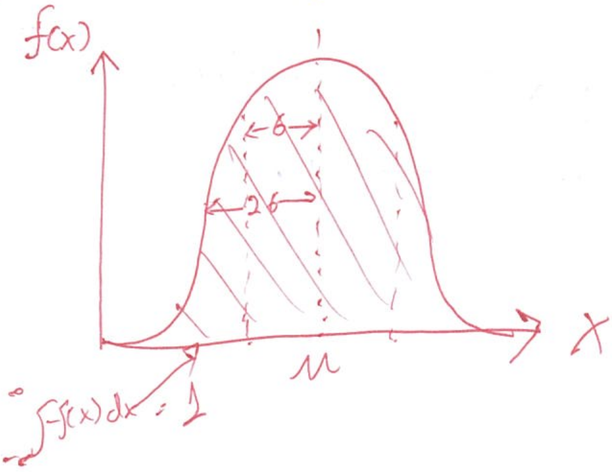
Generatie learning

Big-Q: How to learn the joint distribution from data?

Chain rule for R.V.:
$P(X,Y) = \underline{P(X|Y) \cdot P(Y)}$
$\quad\quad\;\; = \underline{P(Y|X) \cdot P(X)}$

Estimating Probabilities from Data

- Goal: Build a distribution that models (approximate) the real distribution
- Ex. The modelling distribution is a Guassion



p.d.f. $f(x \,|\, \underbrace{(\mu, \sigma)}_{\text{parameter of the modeling distribution)}}) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

- For a Guassian p.d.f. $f(x|(\mu, \sigma))$, ~~each~~ the value of the function at any x is the likelihood (not probability, but can be viewed alike).

Simple scenario I: Coin Toss

- Experiment: Toss a coin n times. How would we estimate $P(H)$?

Ex
$$D = \underbrace{\{H, T, T, H, T, H, H, T, T, T\}}_{n=10}$$

What is $P(H)$?

- Intuitively, $P(H) \approx \frac{n_H}{n_H + n_T} = \frac{4}{10} = 0.4$

- Let's try to formalize what we just did here

Maximum Likelihood Estimation:

1. Make an explicit modelling assumption about what type of distribution your data was sampled from,
2. Set the parameters of this distribution so that the data you obsered is as likely as possible.

Coin Toss (cont.)

- Let θ be a parameter of the modelling distribution

- For our coin toss scenario,

$$D = \{\underset{x_1}{H}, T, T, H, T, H, H, T, T, \underset{x_n}{T}\}.$$

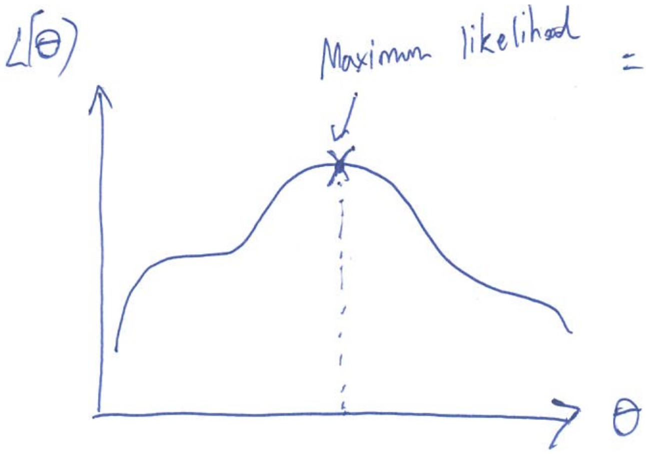
X_1 through X_n are a random $\underset{\text{i.i.d}}{\uparrow}$ sample from $f(x|\theta)$

- To measure the likelihood, we define the likelihood function

$$L(\theta | X_1, \ldots, X_n) = f(x_1, \ldots, x_n | \theta)$$

$$= f(x_1|\theta) \cdots f(x_n|\theta)$$

$$= \prod_i f(x_i|\theta)$$



a binomial distribution

- Let's assume $\boxed{f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}}$.

- Our goal is to find $\theta_{ML} = \underset{\theta}{\arg\max}\, L(\theta, x)$