Lecture Scribe for Machine Learning (CS60050)

Spring 2020-2021

Date: 20-Jan-2021

Instructor: Prof. Aritra Hazra

Till last class what we have learned is that there is an unknown function f:X->Y and a set of training examples $<x_i,y_i>$, then learning goal is to predict a function g from hypothesis set which approximates f (g≈f). For that we have learned concept learning, decision tree learning algorithms (to incorporate the disjunctive property of the attributes).

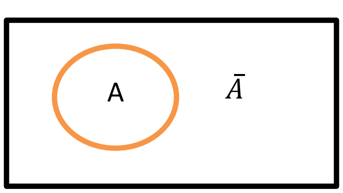
For classification problem, the function could be represented as f:X->Y{1,0}. This problem could also be defined as Prob(y=1) given x which is represented as P(y=1|<x>) and Prob(y=0) given x which is represented as P(y=0|<x>. In next few classes we are going to learn how a classification probability could be estimated for an unknown set of attributes if a set of attributes with their classification probabilities are given as training examples.

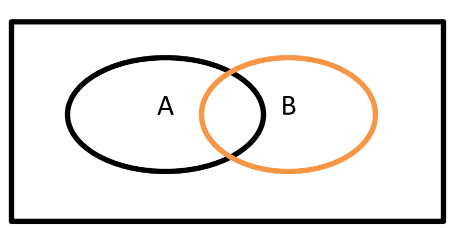
Note, if Prob(y=1|<x>) = p then off course Prob(y=0|<x>) = 1-p .

Probability Basics:

An event or a random variable is outcome of a random experiment.
For example say we draw a random student from this class. What is the probability that the student is a female student? Here the event is [A-> female student], A is the random variable.
Prob(A=f) = $\frac{|Sf|}{|S|}$. Sf is the number of female students in the class and S is the total number of students in the class.

Venn diagram representation of the same is as below.





Axioms:

1. P(A) + P($\bar{A}$) = 1
2. 0≤P(A) ≤1

3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
4. $P(A) = P(A \cap B) + P(A \cap \bar{B})$

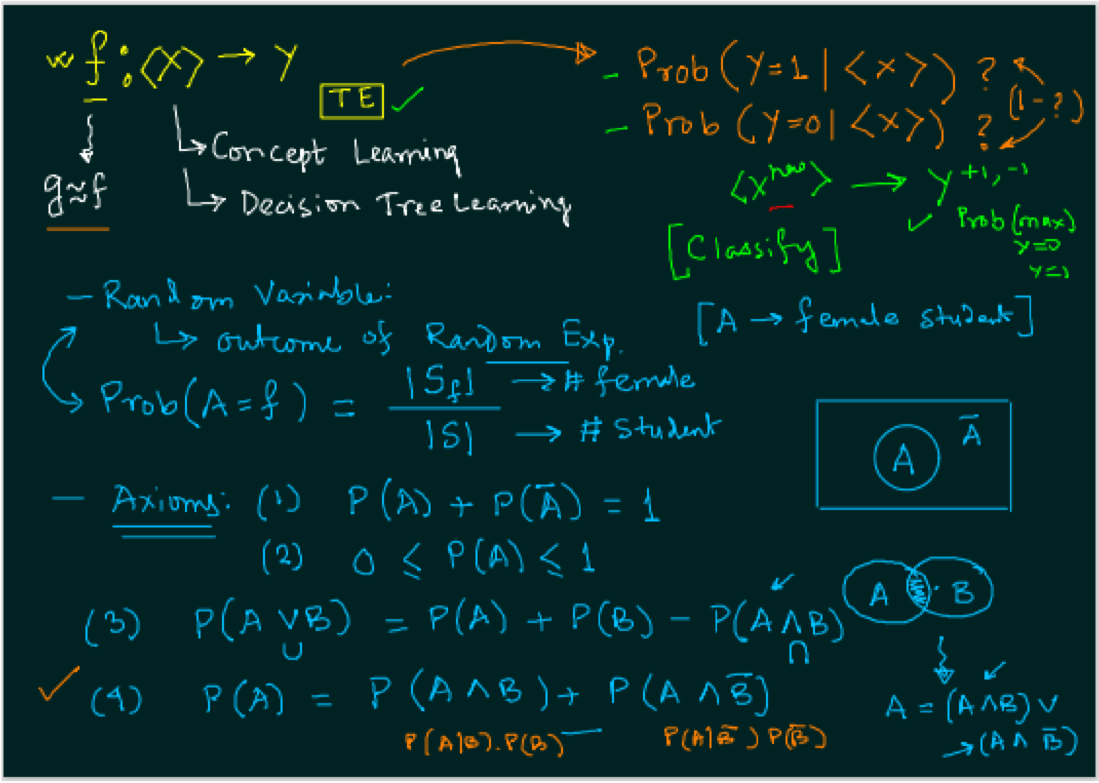


Figure 1 (Handout-04a): Probability Basics and Learning in Probabilistic Way.

Conditional Probability:

What is the probability of A given the probability of B. The same is written as P(A|B).

Mathematically,

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

$\Rightarrow P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$ {Premise of Bayes Rule}

Hence Bayes Rule is $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

Chain rule:

$P(A \cap B \cap C) = P(A|B \cap C)P(B \cap C) = P(A|BC)P(B|C)P(C)$

Chain rule generalization:

$P(A_1A_2......A_n) = P(A_1|A_2A_3...A_n)P(A_2|A_3A_4.....A_n)......P(A_{n-1}|A_n)P(A_n)$

So generalized Bayes theory -

1. $P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A')P(A')}$

2. $P(A|B \cap X) = \frac{P(B|A \cap X)P(A \cap X)}{P(B \cap X)}$

Remember,
P(A=1|B) = 1-P(A=0|B)
but
P(A|B=1) ≠ 1-P(A|B=0)

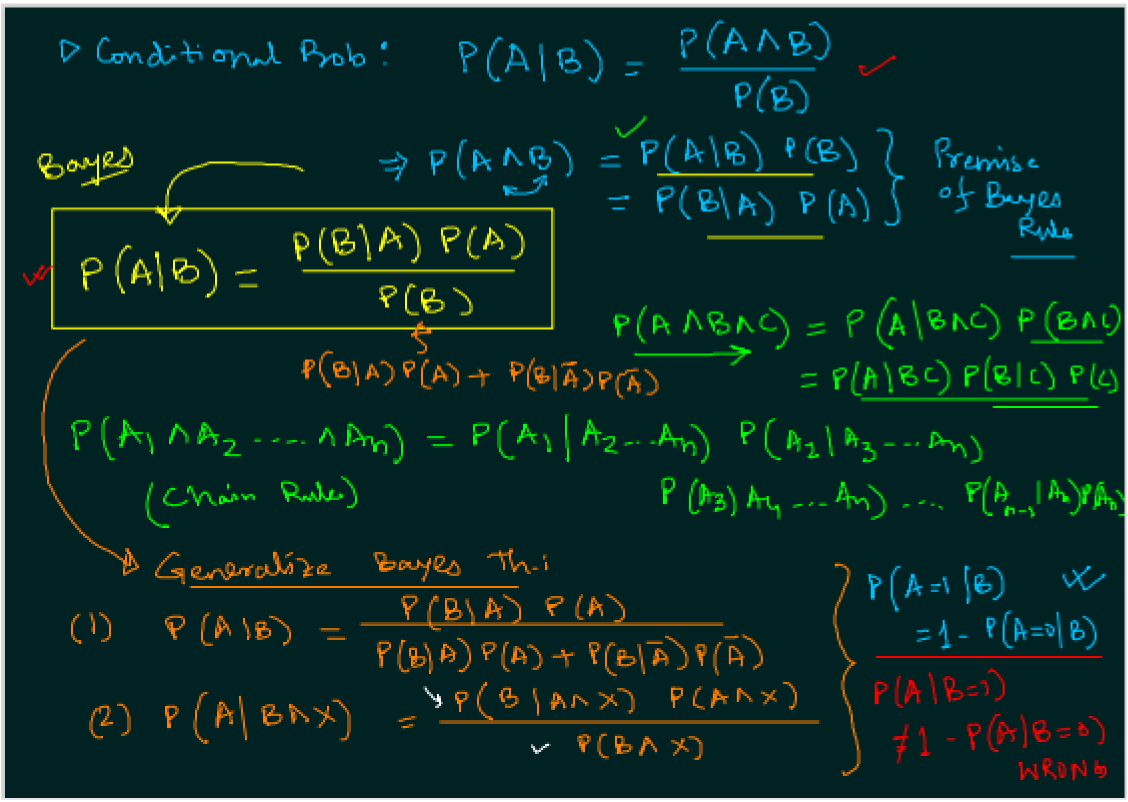


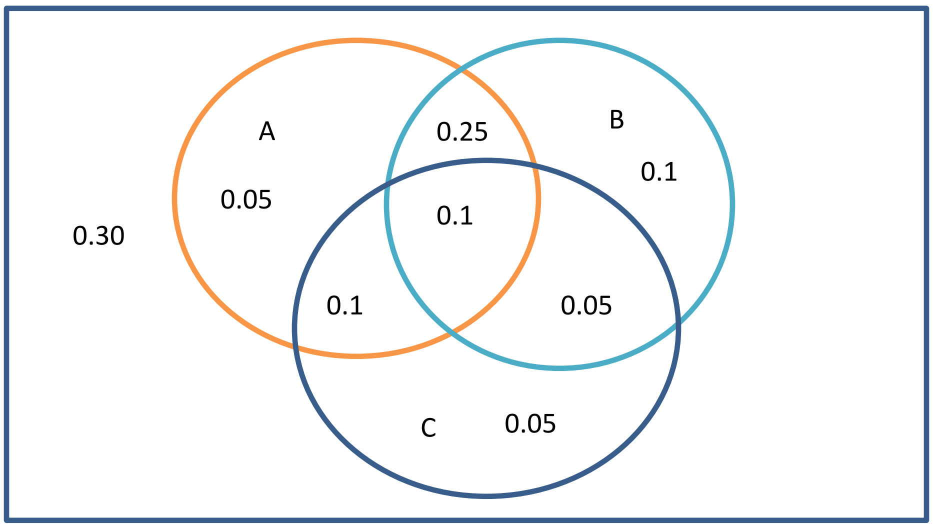
Figure 2 (Handout-04a): Conditional Probability and Bayes Rule

Learning with Joint Distribution:

Suppose following probability distribution is given.

A	B	C	Probability
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

Here A, B, C are binary attributes. In this table probabilities for all possible attribute values have been given though some rows may not be available in reality.
Let's draw a Venn diagram from the table data.



Some of all the probabilities is 1.
P(A=1) = $\sum_{rows\ where\ A=1} P(rows\ in\ JDT)$ = 0.05 + 0.10 + 0.25 + 0.10 = 0.50

Similarly,
P(A∩$\bar{B}$) $=$ Sum of all the rows where A = 1 and B = 0
= 0.05 + 0.10 = 0.15
So basically if full JDT is given, anything could be estimated.

But there is one challenge. For n attributes, we need to have 2^n-1 probabilities or 2^n-1 training examples beforehand.

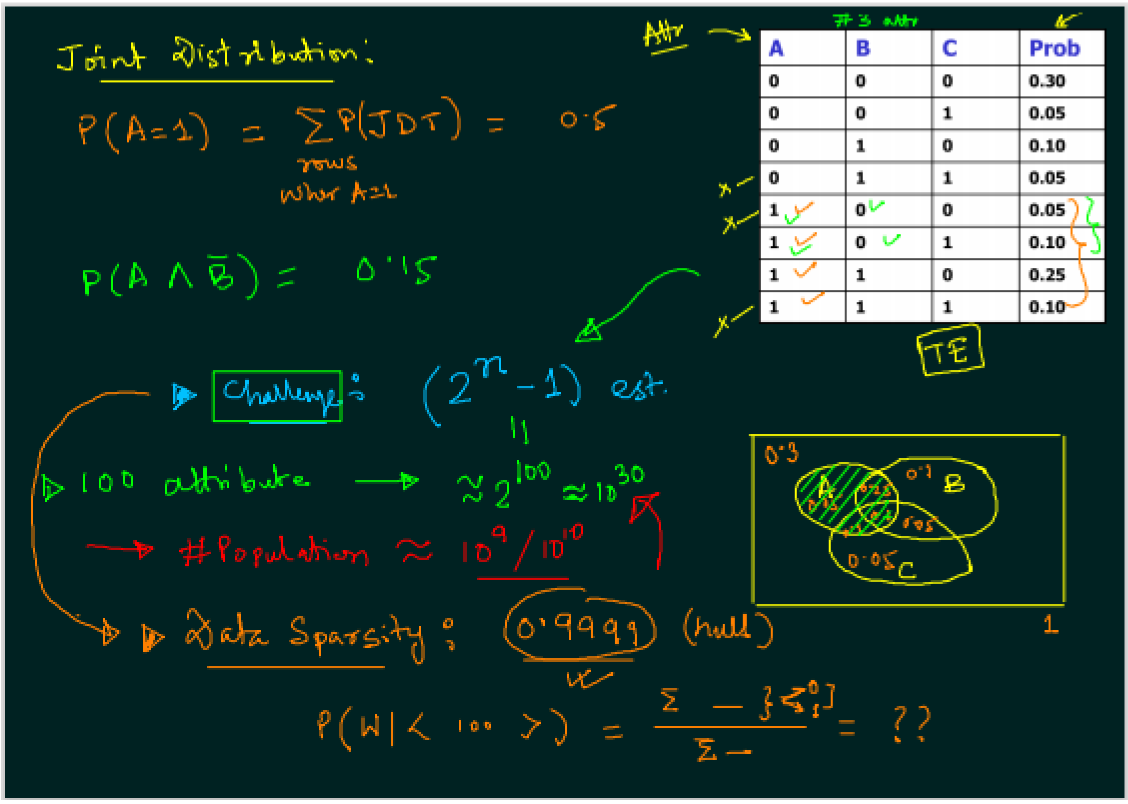


Figure 3 (Handout-04a): Joint Distribution and Its Challenge

Let's take another practical example.

gender	Hrs_worked	wealth	probability
F	V0:40.5-	poor	0.253122
		rich	0.0245895
	V1:40.5+	poor	0.0421768
		rich	0.0116293
M	V0:40.5-	poor	0.331313
		rich	0.0971295
	V0:40.5+	poor	0.134106
		rich	0.105933

The problem could be defined as approximating function f:<HW,G> -> W {poor, rich} from the given data set. Same problem could also be defined as Prob(w=rich|<HW,G>).

Prob(w=poor) = $\sum_{rows=poor} prob$

Prob(male|poor) = $\frac{Prob(male\ and\ poor)}{Prob(poor)} = \frac{0.331313+0.134106}{0.253122+0.0421768+0.331313+0.134106}$ = 0.465/0.65

Similarly, Prob(rich|HW,G) = $\frac{Prob(rich\ and\ hw\ and\ g)}{Prob(hw\ and\ g)}$

If HW = <40.5- and G = female then

$$\text{Prob(rich|HW,G)} = \frac{Prob(rich\ and{<}40.5-\ and\ female)}{Prob({<}40.5-\ and\ female)} = \frac{0.0245895}{0.253122+0.0245895}$$

So if JDT is given then we can calculate all other conditional probabilities.

Are we done?

If Joint Probability Distribution Table is available then we are indeed done. But for n attributes we need (2^n-1) estimations. So if there are 100 attributes in a problem, we need $2^{100} \approx 10^{30}$ estimations.
World has a population of 10^9 or 10^{10}. Assume prediction of tuberculosis requires 100 symptoms. To predict whether someone has tuberculosis we need 10^{30} data but we have only 10^{10} population. So there will be much less data than what we require. This lack of data is called data sparsity.
In this tuberculosis case there will 0.9999 null entries in JPDT.

When we are calculating probabilities, we are summing up some rows from JPDT.

$$\text{Prob(rich|<..100 attr..>)} = \frac{\sum some\ rows}{\sum some\ rows}$$

If some rows are missing then what value we should consider for those missing rows, 0 or something else? If those are wrong then our estimation would also be wrong. We have to do something to estimate probability even though there are not ample data.

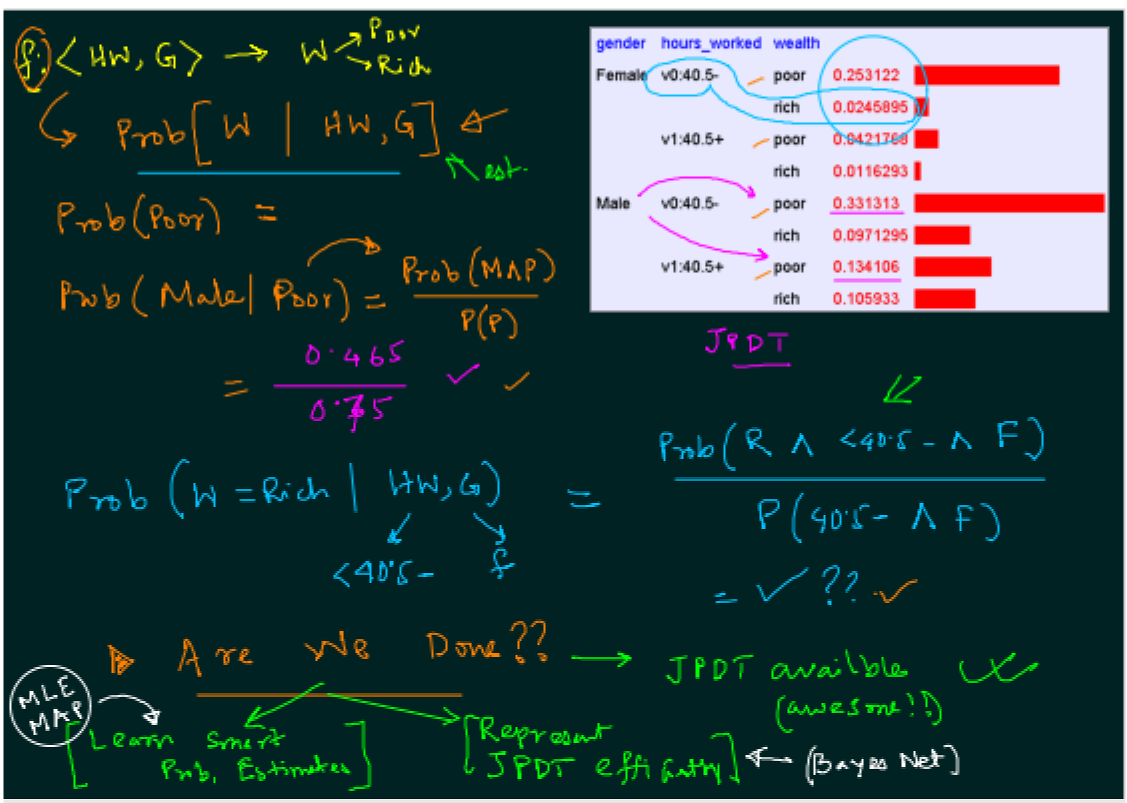


Figure 4 (Handout-04a): JDT for Practical Example and Challenge

What can we do?

To handle the problem due to data sparsity and due to exponential data volume with respect to number of attributes, we'll learn below two approaches.

1> Learn smart probability estimation – something that helps to calculate nearly correct probability even though data missing.
Here we'll learn Maximum Likelihood Estimation (MLE) and Maximum A Posterieori (MAP) to learn smart estimation in case of data sparsity.

2> Representation of JPDT – Even though data available, the volume of data is so high that searching on JPDT table could be very inefficient if JPDT table is not represented in a way to make the search efficient.
Here we'll learn Baye's Net to represent JPDT table efficiently.

How to estimate probability smartly (Here probability is not deterministic. This is learning):

Let's talk about coin flipping problem.

(H) X=1 (T) X=0

In a random flipping if we get α_1 times head and α_0 times tail then $\hat{\theta}$ = P(X=1) = $\frac{\alpha 1}{\alpha 1+\alpha 0}$

Suppose one day someone flips the coin 100 times and got 49 H and 51 T. Then as per the above we can say Prob(X=1) = 49/(49 + 51) = 0.49.
Now another day someone flips the coin 3 times and got 2 H and 1 T. So Prob(X=1) = 2/(2+1) = 0.67.

In the first case where we had 100 flips we got much fair result. There is some issue with second case having only 3 flips. So if training data is abundant, we are getting approximately correct estimation of probability but if there are less training data then estimation is not approximating correctly.

So we have to find another learning algorithm that can cater to both the scenarios correctly.

Here comes something called "A Priori" knowledge.
For coin flip case, Prob(θ) should be ~0.5 for unbiased coin.
So basically my prior knowledge is that out of 20 flips, 10 should be H and 10 should be T.

So for our learning algorithm, if training data is less, we are biased to our prior knowledge and if we have sufficient training data, we'll follow likelihood of the data.

So we can write $\hat{\theta}$ = P(X=1) = $\frac{\alpha 1+10}{(\alpha 1+10)+(\alpha 0+10)}$

So if I have less data, my prior knowledge will dominate and the probability would be closed to 0.5. If we have ample data, then prior knowledge would be almost ignored and the result will still be closed to 0.5.

So if we understand up to this then intuitively,
MAP is nothing but $\hat{\theta}_{map}$= P(X=1) = $\frac{\alpha 1+10}{(\alpha 1+10)+(\alpha 0+10)}$
And MLE is nothing but $\hat{\theta}_{mle}$= P(X=1) = $\frac{\alpha 1}{\alpha 1+\alpha 0}$

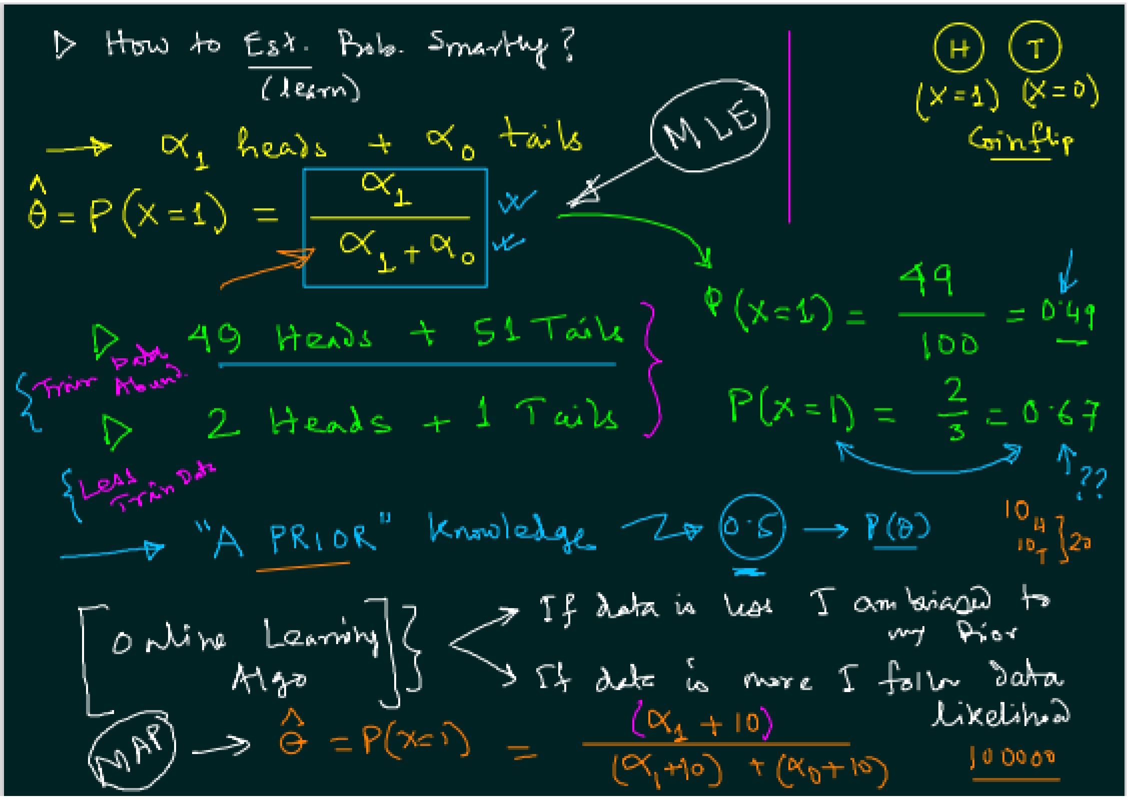


Figure 5 (Handout-04a): Learning Smart Estimation

Mathematically,

MLE: $\hat{\theta}_{mle}$= argmax $_{\theta}$(Prob(Data|θ))

MAP: $\hat{\theta}_{map}$= argmax $_{\theta}$(Prob(θ|Data))

Prob(θ|Data) = $\frac{Prob(Data|\theta)Prob(\theta)}{Prob(Data)}$.

So maximizing probability means, maximizing the numerator as denominator is fixed (called margine). Here, Prob(Data|θ) is likelihood of the data and Prob(θ) is the priori.

Let's assume the coin flip problem again.

P(X=1) = θ

P(X=0) = 1- θ

Say there were five consecutive flips with results 10010. So the likelihood of the data given θ is θ(1- θ)(1- θ) θ(1- θ) = $\theta^2(1-\theta)^3$.

So in general if α1 H and α0 T, then likelihood of data with θ is $\theta^{\alpha1}(1-\theta)^{\alpha0}$

Now we need to maximize θ over $\theta^{\alpha1}(1-\theta)^{\alpha0}$ and that gives us the MLE.

Applying maximization rule $\frac{d}{d\theta}$[..] = 0

$\frac{d}{d\theta}$[α1 ln θ + α0 ln (1- θ)] = 0

- ⇨ $\alpha1.\frac{1}{\theta} + \alpha0\frac{d}{d\theta}\ln(1-\theta).\frac{(1-\theta)}{(1-\theta)} = 0$
- ⇨ $\alpha1.\frac{1}{\theta} + \alpha0.\frac{1}{1-\theta}.(-1)$
- ⇨ $\frac{\alpha1}{\alpha1+\alpha0}$

So we have now mathematically derived our intuitive knowledge $\hat{\theta}_{mle} = \frac{\alpha 1}{\alpha 1+\alpha 0}$

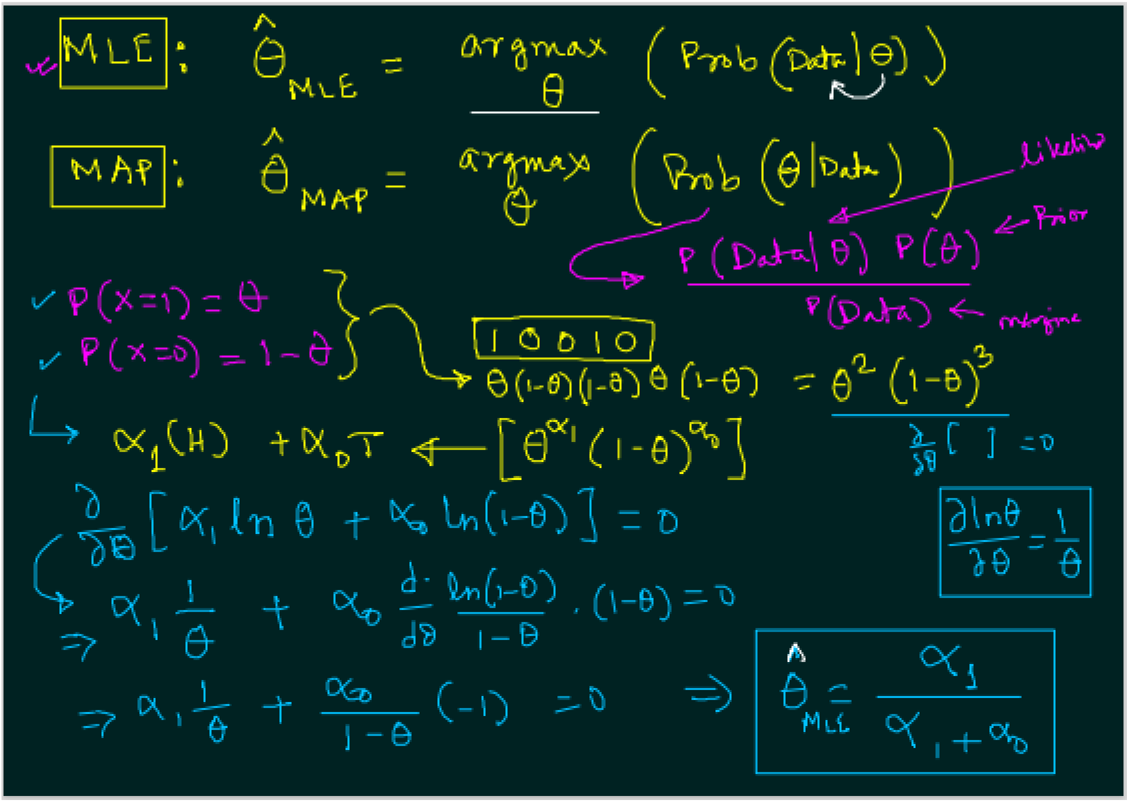


Figure 6 (Handout-04a): Derivation of MLE

In next class we'll learn about MAP.

* Scribe made by Parimal Santra (Roll No: 20CS72E01)*