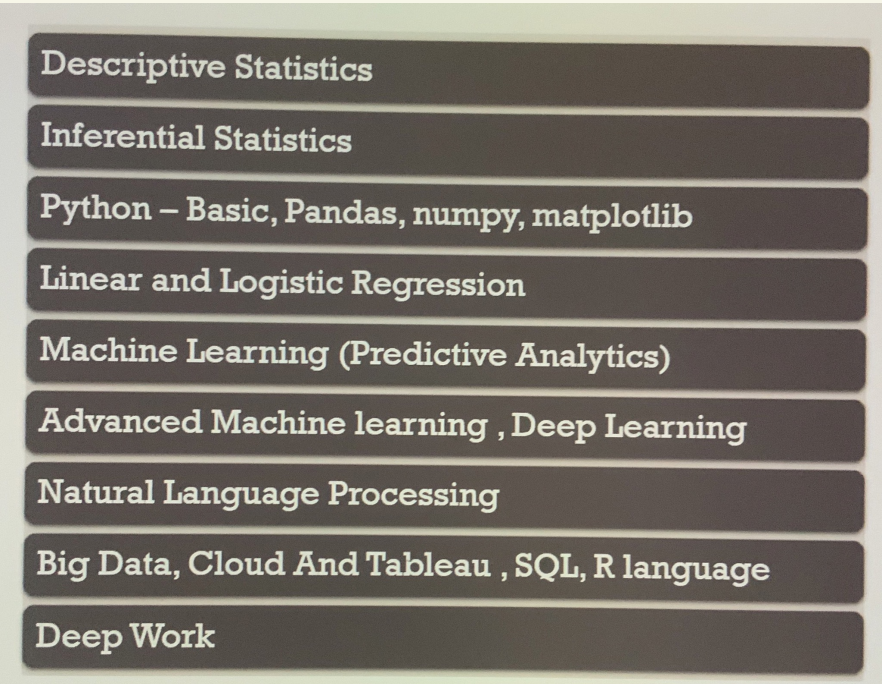
End to End Data Science Notes

DATTATRAY N. NALE

09/Feb/2020

COURSE OVERVIEW



Descriptive Statistics - Data gathered from group inferred to the group.

Inferential Statistics - Data gathered about a group used to refer population.

It is the bridge between descriptive statistics and modeling. This will lay the foundation for prediction.

We would be going over foundations of AI. When we say AI, people talk about machine learning, deep learning, NLP etc. First focus on problems! they are the root of AI.

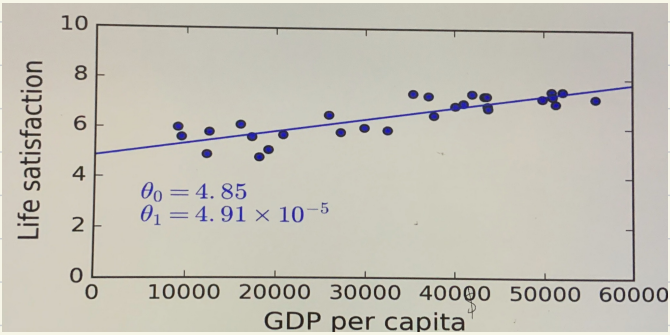
What Problems are we solving?

- Regression
- Classification
- Clustering
- Lowering or increasing dimensions.
- Optimization of any function
- Multi agent Systems

These are the 5 fundamental problems we have that is solved by Data Science. Problems have been consistent for years. The day we understand the problem and the nature of it, it will help you understand any algorithm.

Regression is all about prediction / Forecasting
Eg: Cab price change.
Uber uses battery level to see if you are desperate for a ride.

Anything that has a variable nature, its predictable.



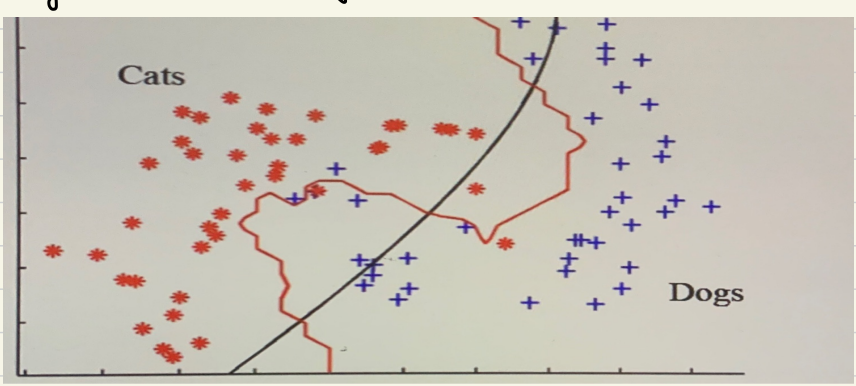
Classification

This helps you classify data. The data can be in any format - text / image

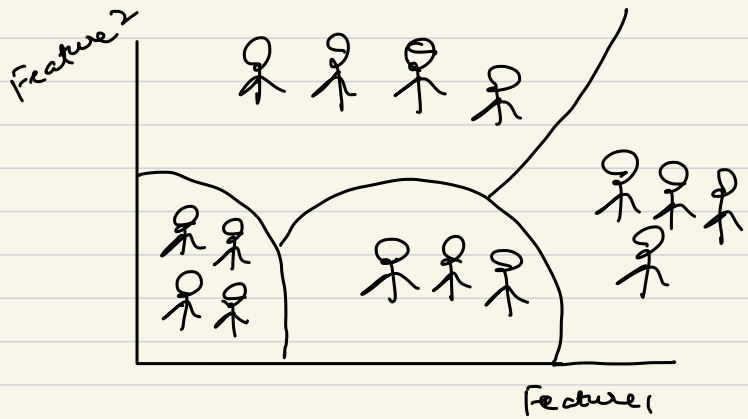
People think that whenever data changes the algorithm changes, or the problem changes. NO!

Data can be image, text, video, speech, number; the problem remains the same.

To classify cats v/s Dog I dont need a numbers. Image / Text anything will work.



CLUSTERING is when you dont have a target variable. What's a target variable? How many times we have gone to a grocery chain and given email id and phone number? You don't give! Grocery chains have details of customer like name, card number etc. But, they don't have details of if customer is young, old, male, female etc.
This happens many times that we do not know accurate details about people. In this case you do clustering.



You cannot tell which customer is male v/s female. But you can group them based on certain attributes.

OPTIMIZATION

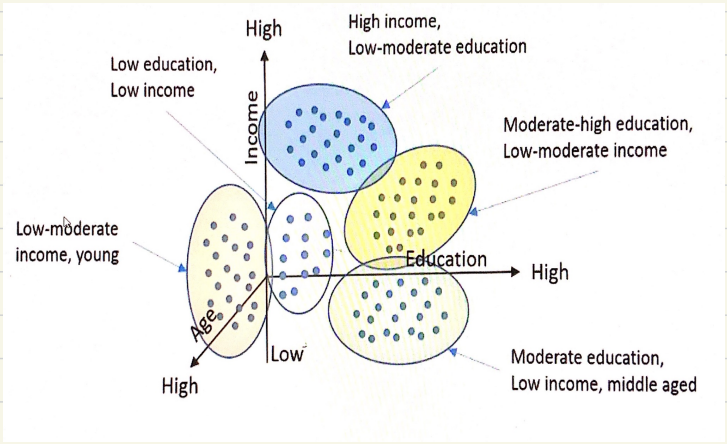
Eg: You would look at a rack at eye level. So if you know height (average) of customers. So companies put kids related items in bottom rack.

How many items to put such that while the shop is on you don't have to replace. It's not about selling more items, you need to sell profitable items and keep customers satisfied as well.

Grocery shelf optimization is classic problem of optimization.

Electioneering Trump Trump carefully engineered his speeches based on crowd. This is a classic example of clustering.

Cluster population and keyword mapping his speeches.



Recommendation Engine

Basically what amazon does when recommending what other customers bought. Or Netflix recommendation.

INTRICACIES OF MODELLING

Lets talk about problem surrounding a problem. The previous problems could be solved by algorithms. Even before these, there are other problems that need to be solved. Thats Exploring Data.

Exploring Data / EDA

- Getting patterns / distributions
- Correlation and Covariance
- Feature Engineering - New features
- Bifurcating users / data and building separate models
- Case study of hot star recommendation engine.

  There are times when a single model may not work. For example, for hotstar, there are 2 models - One for regular users who would benefit from recommendation. Two - for specific users who watch only specific things.

WHY EDA

Exploratory Data Analysis. is mainly to see churning of customers from subscriptim based services.

CHALLENGES FOR MODELLING

1. Lack of training data and Poor quality of datta
2. Too much of Imbalance — churn analysis
3. The problem of Sampling Bias
4. Irrelevant features
5. Poor fitting or over fitting.
6. The biggest one— unexplored waters
   We have explored only 5%. 95% still needs to be explored.
7. Not model always. Model is one of the deliverables of a data scientist, but not always.
   For eg: Data Scientist may recommend an organization to start capturing a specific data so it will impact their profits positively.

So INSIGHT is the solution! What business impact you make?

DIFFERENT COMPONENTS OF AI

- Statistical modeling
- Machine Learning
- Deep Learning
- Reinforcement Learning
- Optimization algorithms.

Statistical Modeling

- Regression Analysis
- Classification
- Time Series

Machine Learning (30 hours)

- Support Vector machines
- Decision Trees
- Clustering
- KNN Classification
- Neural Networks
- Association Rule mining

Advanced Machine Learning

- Bagging Stacking
- Gradient Boosting Machines
- Random Forest
- Data Pipelines

Data Engineering is a small part of Data Science. It deals with where to store data, how to store it, how fast to move it from one server to another, on cloud, on local or hybrid

Data Science is bigger. Includes data engineering and then how to explore, visualize, build models, predict and deploy.

NLP (Natural Language Processing)

- Text Understanding
- Text processing
- Text Modelling
- Case study on description detection.

DEEP LEARNING

- Evolution of deep learning networks
- CNNs
- RNNs
- Keras / Pytorch hands on
- Linear optimization

Deep learning is overrated. Less than 1% have data for deep learning. .001% of them do experiment .001% of those are deployed. And in that .001% solutions make money. Till date there is no real example where companies made money on that.

DEEP WORK

- Why deep work
- Understanding barriers
- Managing Attention
- Rules to attain deep work
- My success with deep work.

PROBABILITY-101

22-Feb-2020

1. What is probability?

Probability is the extent to which an event is likely to occur. measured by below ratio.

$$P = \frac{\text{Num of favorable outcomes for a condition}}{\text{All possible equally likely outcomes}}$$

Eg:

1. probability of a 2 appearing on die of 6 faces.

   1/6

2. What is the probability of heads when you toss a coin?

   1/2

2. What if the outcomes are unequal?

This scenario boils down to a different kind of probability called **Conditional Probability**

When does this occur?

This happens in case of biased coin. A biased coin is a coin that is manipulated to give unfair results.

TWO PARADIGMS OF PROBABILITY

- Bayesian
- Frequentist

Let's understand with examples as below -

Q-1 What is the probability of a 'two' on the die if it was Shakuni (Mahabharata)?

Here key is "If shakuni throws". However if someone else throws, it would be different. Because Shakuni was blessed that in the game of dice, he will get what he wants.

So if Shakuni thinks of 2, he gets 2, else something else.

So probability is only $\frac{1}{2}$ as there are only 2 things that can happen.

Probability is not a fixed value in Bayesian world. It changes according to prior. / prior knowledge of situation.

Q-2 What is the probability that Modiji had tea in the morning?

if you answer 0 or 1, you are Bayesian

if you answer $\frac{1}{2}$, you are frequentist

Frequentists are the people who will compute probability without any prior knowledge,

$$\frac{\text{No tea}}{\text{No tea + Yes tea}} = \frac{1}{2}$$

Frequentists would do thought experiment in their mind. when asked a probability value. However Bayesians think about prior.

Q-3 What is the probability of sun rises in the east?

$$\frac{\text{rise in east}}{\text{rise + not rise}} = \frac{1}{2} \quad (\text{frequentist's way!})$$

SUMMARY

Frequentist – gets sample, long frequency distributions, parameters fixed.

frequency distribution – when you chart graphs based on experiments.

parameter fixed means, they decide the probability and then do experiments to prove the probability.

Bayesian – Calculates prior. They don't fix parameters, they fix data, last 6 yrs etc. But frequentists change data.

Q – Whats probability of a two on die?

Frequentist –

Try throwing die 10K times, count distribution. Say its close to 1/6

Bayesian –

Wait, this dude has cheated previously with die which are always biased. Hence I would like to assess prior for this die. So he asks the guy for past history and creates a distribution for it.

For AI, ML both paradigms are needed!

PROBABILITY OF A SINGLE & MORE EVENTS

29/Feb/2020

SINGLE EVENTS

Below events are single or multiple?

Q-1 Probability of a 6 on die.

This is surely a single event.

Q-2 Probability of sum 'six' for two dies.

Not 2 events as you cannot take probability of 1 die and add it with probability of 2nd die.

To find the probability one has to add favorable outcomes from both dies then divide by total outcomes.

D-1	D-2	two dies together
1	5	→ single event
2	4	
3	3	
4	2	
5	1	

TWO EVENTS

Eg:-

a) Probability of a "4" in first throw and "2" in second throw on a die.

b) Probability of a spade in first pick and Ace in second pick.

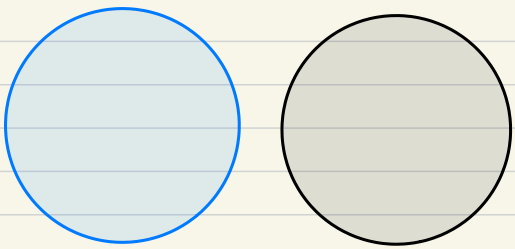
Other Examples of multiple events.

- Probability of 4 6's in an over(6 balls)
- Traffic jam happens and rain happens
- Getting a promotion and having a dinner

EXCLUSIVITY OF EVENTS

Two events may or may not overlap.

(A)



mutually exclusive
(NO overlap)

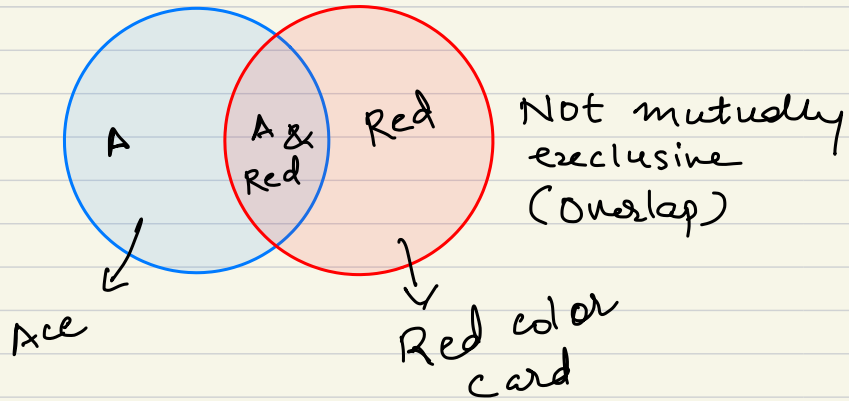
Example -

- I being in Bangalore and Mysore. This cannot happen. I can be in one place at a point of time.

- probability of I having coffee. Probability of I having a fruit punch. in my breakfast. I can't have both!

- probability of 9 AM in USA and 9 AM in India.

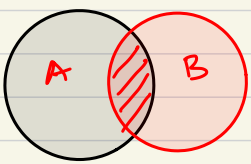
(B)



Exclusivity is mathematical in nature however, dependency is not!

ADDITIVITY OF EXCLUSIVE & NON EXCLUSIVE EVENTS

$$P(A \cup B) = P(A) + P(B)$$



$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

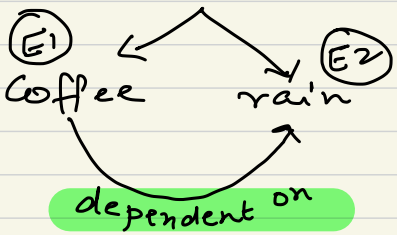
because intersection is double counted!

DEPENDENT AND INDEPENDENT EVENT

Probability of I having coffee → Single event

Probability of I having coffee when it is raining

2 events & also dependent



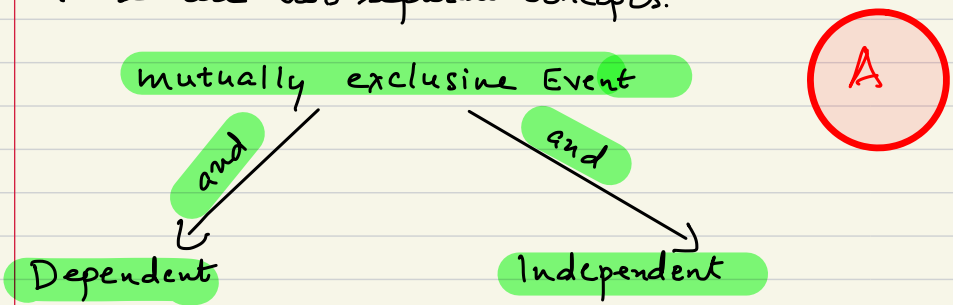
For dependent events we use conditional probability.

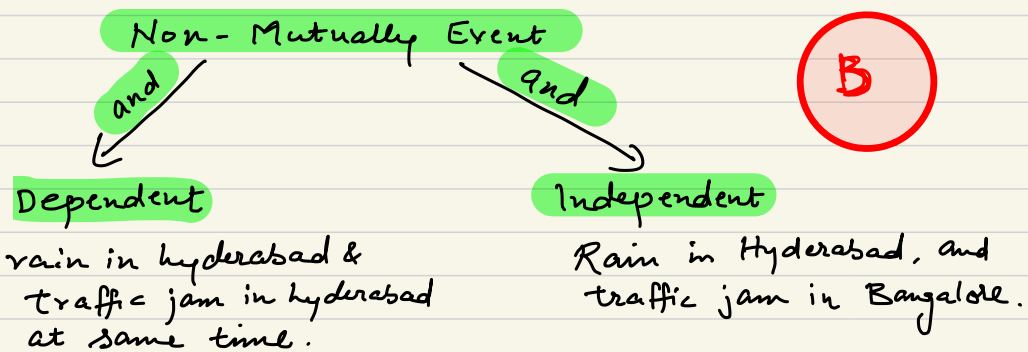
Independent Events

- Probability of rain in Hyderabad and traffic jam in Bangalore. There's no connection!
- Sometimes dependency is a perception. It depends on how customer is evaluating it.
  eg- wife calls you every time you are in a meeting and you leave her a message that you are in a meeting.
  for you - events are independent.
  for your wife - they are dependent. She thinks whenever she calls you, you go to a meeting.

DEPENDENCY & EXCLUSIVENESS

These are two separate concepts.





rain in hyderabad & traffic jam in hyderabad at same time.

Rain in Hyderabad, and traffic jam in Bangalore.

precisely speaking, dependency is logical and exclusiveness is mathematical!

MULTIPLICATIVE RULE

$P(A \cup B)$ — is only for exclusivity

$P(A \cap B)$ — for dependency

$\cup$ = union
$\cap$ = intersection

Independent Events

→ $P(A \cap B) = P(A) \times P(B)$
→ $P(B|A) = P(B)$

How is $P(B|A) = P(B)$

eg: $P(\text{Kid drinking milk} \mid \text{milk is given to the kid})$

is equal to $P(\text{Kid drinking milk})$

Dependent Event

$P(A \cap B) = P(A) \times P(B|A)$

Q-1 What is the probability of not getting a '1' on first and second throw of a fair die of six faces?

Sol this is multiple events and independent events

And means $\cap$

Or means $\cup$

P(NO getting '1' on 1st & 2nd throw of die)

$= P(A) \times P(B)$

$= \frac{5}{6} \times \frac{5}{6} = \frac{25}{36}$

Interview question - There is a drawer with 15 green, 12 blue and 16 red balls. The task is to go on picking balls and place them in your bag till you get a pair of balls of same color in your bag.

How many times should I pick so that I am 100% sure that I have a pair with same color.

Sol

4.

worst case you pick 3 balls different color. picking up 4th will make a pair.

Q-3

A basketball team is down by 2 points with only a few seconds remaining in the game. Given that:

- Chance of making a 2-point shot to tie the game = 50%
- Chance of winning in overtime = 50%
- Chance of making a 3-point shot to win the game = 30%

~~break~~ break

What should the coach do: go for 2- point or 3-point shot?

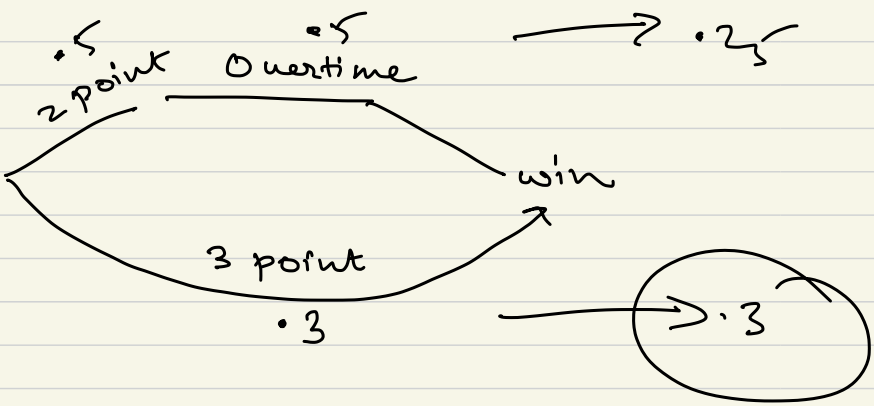
Sol

P(Winning on 2 point shot) = P(2 point shot) × P(winning in overtime)

$= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} = 25\%$

3 point - 30%

They should go with 3 point shot. which has 30% probability V/s 2 point shot with 25% probability.



Python done after this.

LETS BUILD A TREE

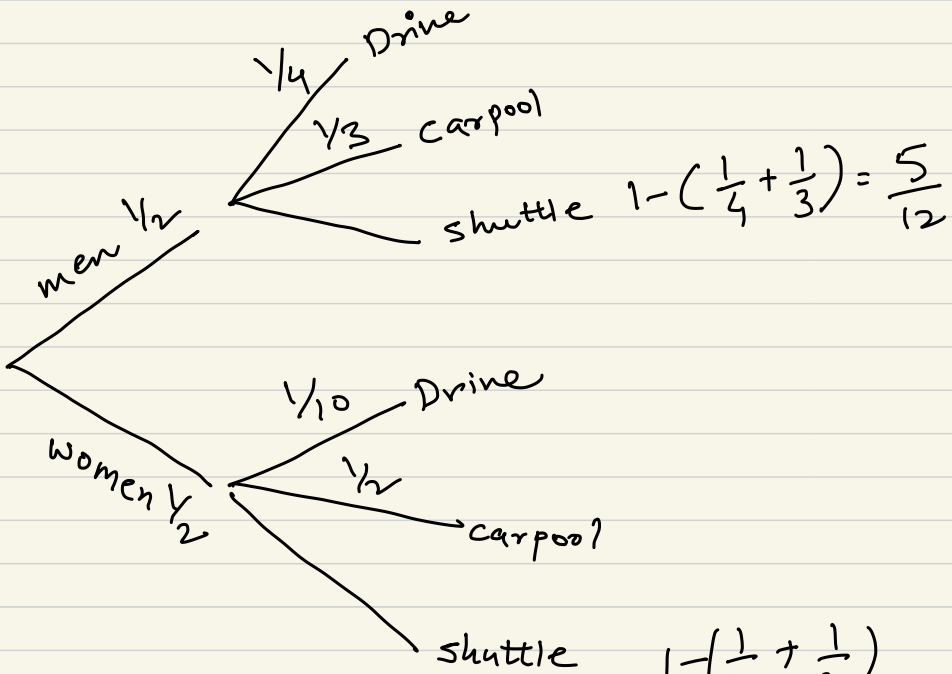
07-Mar-2020

Q-1

- There are equal numbers of men and women in office and you know that 1/4 of the men and 1/10 of the women drive to office every day, 1/3 of the men and 1/2 of the women get a car pool and the rest travel by shuttle, determine

(a) the proportion of the office that are women who go by shuttle;

(b) the proportion of the office that go by shuttle.



men shuttle

$$1-\left(\frac{1}{4}+\frac{1}{3}\right)=1-\left(\frac{3+4}{12}\right)$$

$$=1-\frac{7}{12}=\frac{12-7}{12}=\frac{5}{12}$$

women using shuttle

$= 1 - \left(\frac{1}{10} + \frac{1}{2}\right)$

$= 1 - \left(\frac{2+10}{20}\right) = 1 - \left(\frac{12}{20}\right)$ (×3, +5)

$= 1 - \frac{3}{5} = \frac{5-3}{5} = \frac{2}{5}$

a) Proportion of women by shuttle

$= \frac{1}{2} \times \frac{2}{5} = \frac{1}{5}$

(b) $P(\text{shuttle}) = P(S \cap M) + P(S \cap W)$

$P(S \cap M) = \frac{1}{2} \times \frac{5}{12} = \frac{5}{24}$

$P(S \cap W) = \frac{1}{2} \times \frac{2}{5} = \frac{1}{5}$

$\frac{5}{24} + \frac{1}{5} = \frac{25+24}{24 \times 5} = \frac{49}{120}$

TYPES OF PROBABILITY

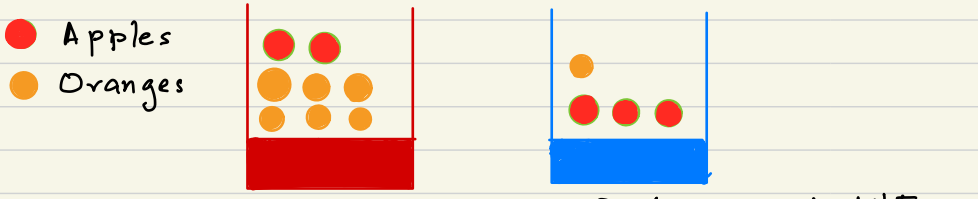
1. Marginal Probability
   Probability of picking an apple.

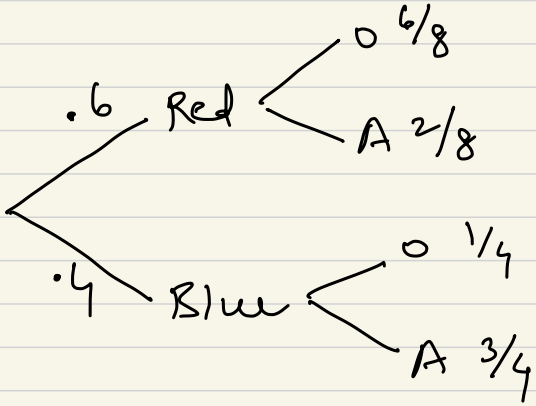
2. Conditional Probability
   Given the bag picked is blue, what's the probability of orange.

3. Joint Probability
   What is the probability of orange and blue box.

1. Marginal Probability / Whole probability

Probability of picking an apple.



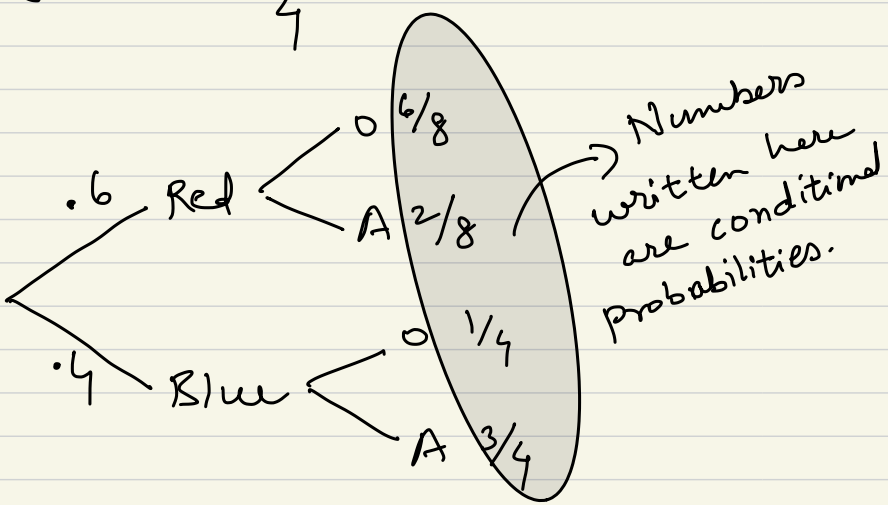


marginal Probability is whole probability. It could be a combination of more than 1 event

CONDITIONAL PROBABILITY

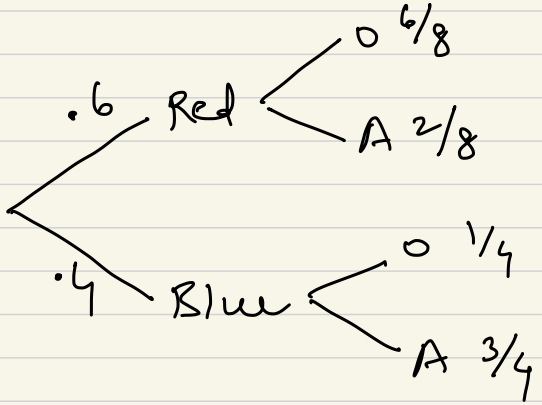
Given the bag picked is blue, what is the probability of orange.

$$P(O|B) = \frac{1}{4}$$



JOINT PROBABILITY

What is the probability of orange and blue box



$$P(O \cap B) = e(O|R) + e(O|B) \cap [e(b)$$

$$[e(O|r) \cap e(b)] + [e(O|b) \cap e(b)]$$

$\downarrow$

0 as no overlap

$$= P(O|b) \times P(b)$$

$$= 1/4 \times 4/10 = 1/10$$

$$\underbrace{P(O \cap b)}_{\text{Joint probability}} = \underbrace{P(O|b)}_{\text{Conditional probability}} \times \underbrace{P(b)}_{\text{Marginal Probability}}$$

Que How are below different? Which one will be more

a) P(Traffic Jam | Rain)

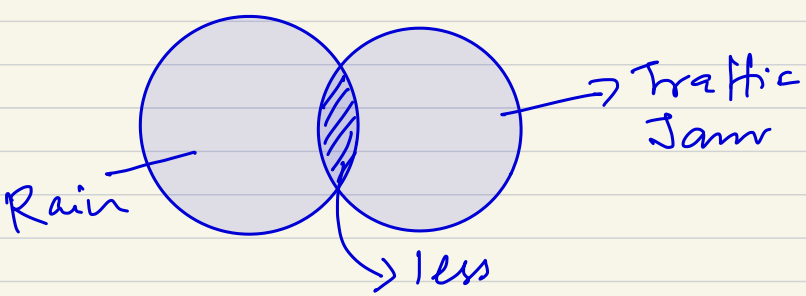
Rain started and then traffic jam occured
There is a causative effect / dependency here.

b) P(Traffic Jam ∩ Rain)

Rain started and Jam happened + Jam already there and rain started

- Intersection does not care about timing / reason / conditionality

P(TJ | Rain) is more



Que You toss a fair coin three times: Given that you have observed at least one heads, what is the probability that you observe at least two heads?

Ans P(2 Heads | 1 head)

	H	T
H	HH	HT
T	TH	TT

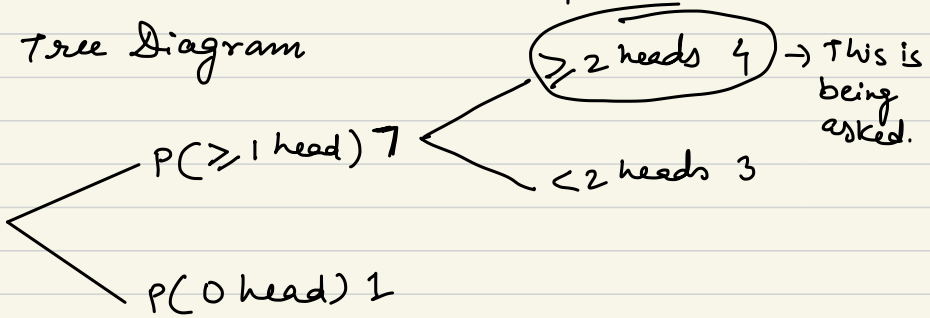
	H	T
HH	HHH	HHT
TH	THH	THT
HT	HTH	HTT
TT	TTH	TTT

favorable events = 4
Total events = 8
$= \frac{4}{8} = \frac{1}{2}$

However you need at list one head so TTT is not to be counted.

So - $\frac{4}{\text{All except TTT}} = \frac{4}{7}$

Tree Diagram



ASSIGNMENT-1

22/Feb/2020

Que

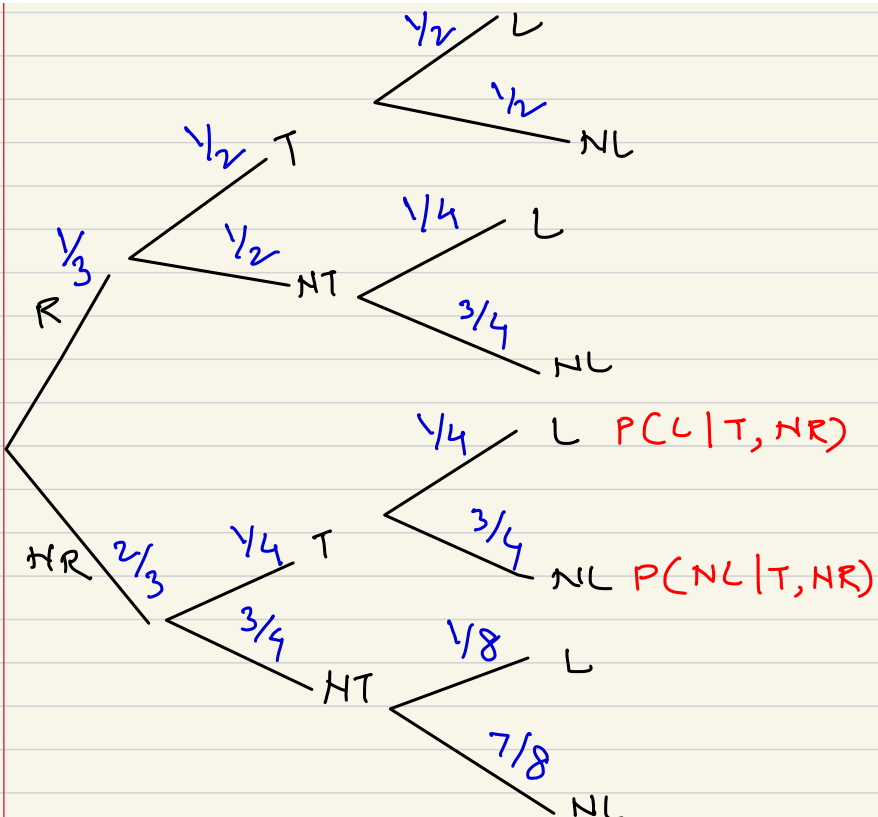
5. In my town, it's rainy for one third of the days. Given that it is rainy, there will be heavy traffic with probability 1/2, and given that it is not rainy, there will be heavy traffic with probability 1/4. If it's rainy and there is heavy traffic, I arrive late for work with probability 1/2. On the other hand, the probability of being late is 1/8 if it is not rainy and there is no heavy traffic. In other situations (rainy and no traffic, not rainy and traffic) the probability of being late is 0.25, 0.25. You pick a random day.

a) What is the probability that it's not raining and there is heavy traffic and I am not late?

b) What is the probability that I am late?

c) Given that I arrived late at work, what is the probability that it rained that day?

Ans



a) $P(NR) \cap P(\text{heavy traffic}) \cap P(\text{Not late}) =$

$P(NR)\, P(\text{heavy traffic} \mid NR)\, P(NL \mid NR, \text{heavy traffic})$

$= \frac{2}{3} \times \frac{1}{4} \times \frac{3}{4_2} = \frac{3^1}{24_8} = \frac{1}{8}$

b) $P(L) =$

$\left(\frac{1}{3} \times \frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{3} \times \frac{1}{2} \times \frac{1}{4}\right) + \left(\frac{2}{3} \times \frac{1}{4} \times \frac{1}{4}\right)$

$+ \left(\frac{2}{3} \times \frac{3}{4} \times \frac{1}{8}\right)$

$= \frac{1}{12} + \frac{1}{24} + \frac{2}{48} + \frac{6}{96}$

$= \frac{1}{12_4} + \frac{1}{24_{(2)}} + \frac{1}{24_{(2)}} + \frac{1}{16_3}$

$= \frac{4+2+2+3}{48}$

$= \frac{11}{48}$

c) $P(R|L) = \dfrac{P(L|R) \times P(R)}{P(L)}$

$= \dfrac{P(L|R,T) + P(L|R,HT)}{P(L)}$

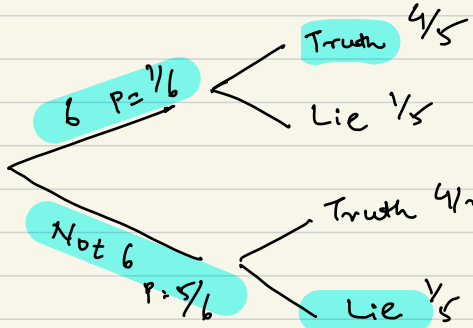
$= \dfrac{\left(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{3}\right) + \left(\frac{1}{4} \times \frac{1}{2} \times \frac{1}{3}\right)}{11/48}$

$= \dfrac{\frac{1}{12} + \frac{1}{24}}{11/48} = \dfrac{3}{24} \times \dfrac{48}{11} = \dfrac{6}{11}$

BAYES THEOREM

14-Mar-2020

Q-1 'A' speaks the truth 4 out of 5 times. A die is tossed. 'A' reports that it is a 6. What are the chances that there actually was a 6?



$$P(\text{Actually 6} \mid \text{A says its a 6}) = \frac{P(6) \times P(\text{Truth} \mid 6)}{P(6) \times P(\text{Truth} \mid 6) + P(\text{Not}) \times P(\text{Lie} \mid \text{Not 6})}$$

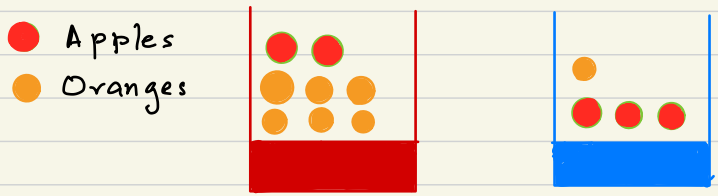
$$= \frac{1/6 \times 4/5}{(1/6 \times 4/5) + (5/6 \times 1/5)}$$

$$= \frac{4/30}{4/30 + 1/6} = \frac{4/30}{\frac{24+30}{30}} = \frac{4}{30} \times \frac{30}{54} = \frac{2}{27}$$

The Reverse Probability

Bayes also ...
Fondly called reverse probability. Not a technical term though!

Q-2



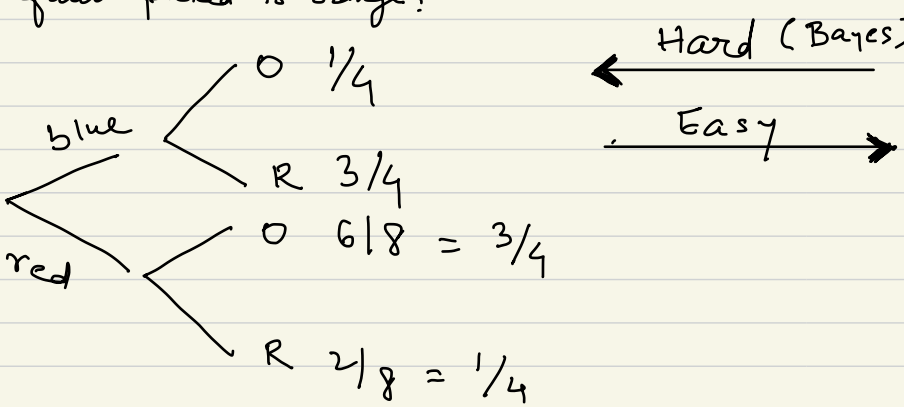
0.6

0.4 → probability of picking from each bag

a) What is the probability of orange given it is a red bag?

$$6/8 = \frac{3}{4}$$

b) What is probability that it is red bag, given that fruit picked is orange?



FORMULATION OF BAYES THEOREM

Event A and B occur and are dependent events

$$P(A \cap B) = P(A|B) \times P(B)$$
$$P(A \cap B) = P(B|A) \times P(A)$$

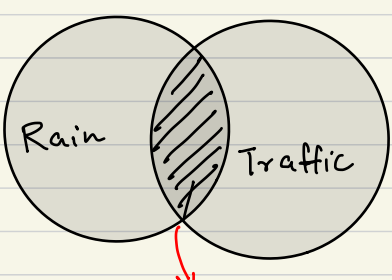
$$P(A|B)\ P(B) = P(B|A) \times P(A)$$

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

$A|B \neq B|A$ why?
A given B is not same as B given A.

For example, Traffic given (because of) rain is not same as, Rain given (because of) traffic.

However probability of both happening at the same time is same because of Intersection, see below



P(Rain ∩ Traffic) will always be same.

Intersection does not talk about what happened first or next. It is not causative in nature, where conditionality is causative in nature. $A \cap B$, we dont care A caused B, or B caused A.

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

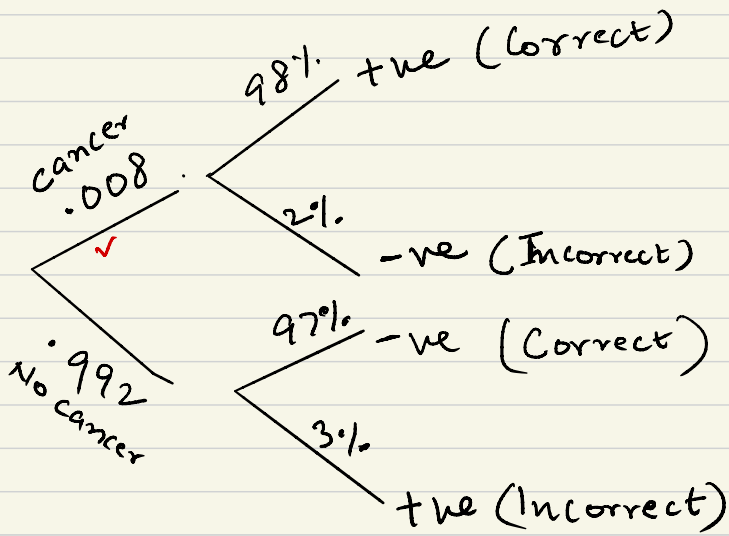
This formula lays the foundation stone for Bayesian Machine Learning.

Here we are talking about prediction. For eg: Yes / No kind of decisions!

Q-3

A patient takes a lab test and the result comes back positive. It is known that the test returns a correct positive result in only 98% of the cases. In the remaining 2% cases, the test returns negative even thought the patient has cancer. The test yields a correct negative result in only 97% of the cases. In the other 3% cases, the test is positive even though the patient does not have cancer. Furthermore, only 0.008 of the entire population has this disease.

- 1. What is the probability that this patient has cancer?
- 2. What is the probability that he does not have cancer?
- 3. What is the diagnosis?



1) $$\frac{P(+ve|C) \times P(C)}{P(+ve)} = \frac{P(+ve|C) * P(C)}{P(+ve|C) * P(C) + P(+ve|NC) * P(NC)}$$

$$= \frac{.98 \times 0.008}{(.98 \times 0.008) + (0.03 * .992)}$$

$$= \frac{.00784}{.00784 + .02976} = \frac{.00784}{0.0376} = 0.208$$

2) $P(NC|Pos) = 1 - P(C|Pos)$

$= 1 - 0.208$

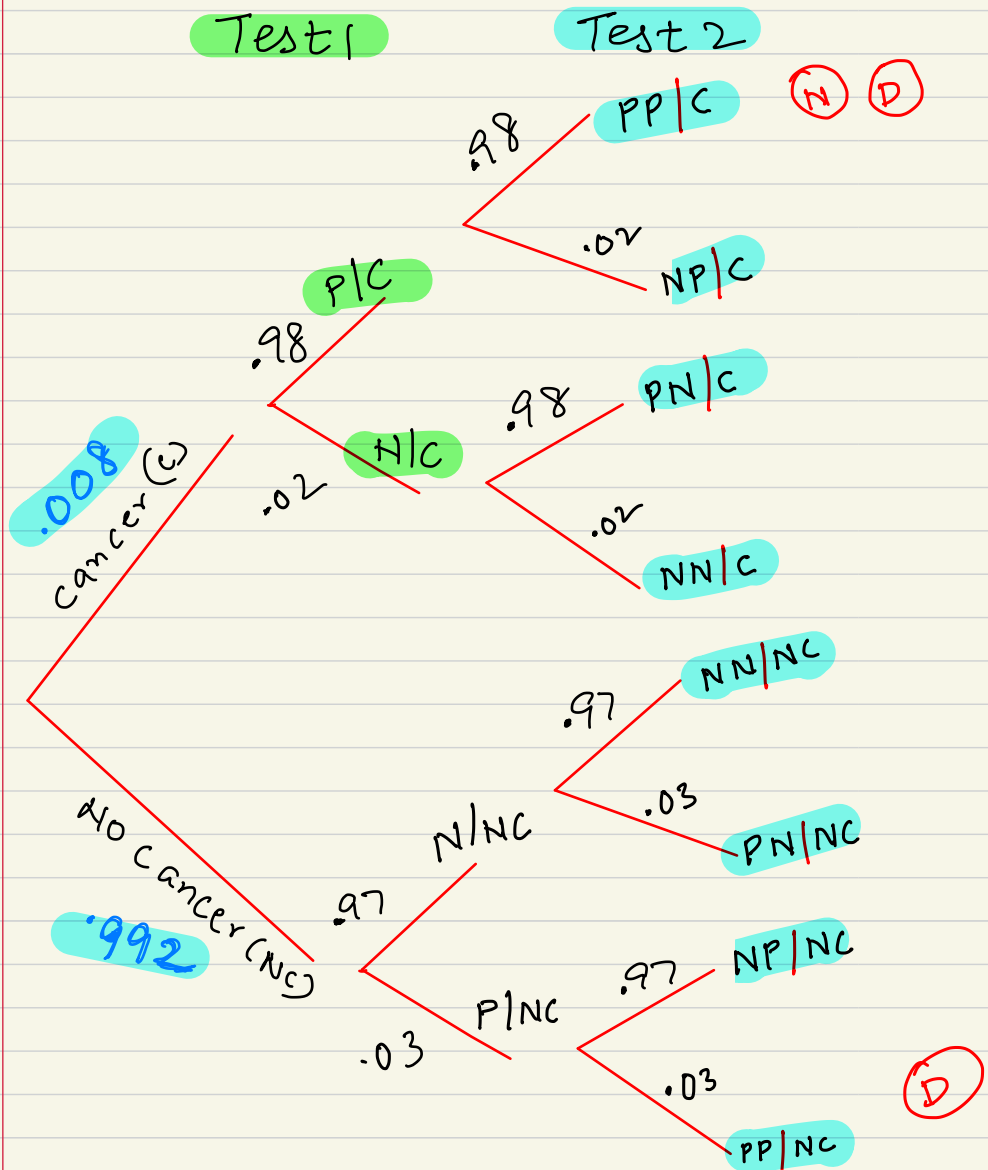
$= 0.792$

3) Problem is in accuracy. In case of +ve test only 20% of the times the test is correct. And 80% of the time it is incorrect.

Reverse labels! and you are 80% accurate

Assignment-2

14/Mar/2020



$P(C \mid PP) = ?$

$P(NC \mid PP)$

a) $P(C|PP) = \dfrac{P(PP|C) \times P(C)}{P(PP)}$

$= \dfrac{.98 \times .98 \times .008}{(.98 \times .98 \times .008) + (.03 \times .03 \times .992)}$

$= \dfrac{.0077}{.0077 + .0009} = \dfrac{.0077}{0.0086} = 89.53\%$

b) $P(HC|PP) = \dfrac{P(PP|NC) \times P(NC)}{P(PP)}$

$= \dfrac{.03 \times .03 \times .992}{(.98 \times .98 \times .008) + (.03 \times .03 \times .992)}$

$= \dfrac{.0009}{.0086} = 0.104 = 10.4\%$


Aayush