BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani

Pilani Campus

Session 4: Bayes theorem and Naïve Bayes theorem

(Session 4: 10th/11th June 2023)

IMP Note to Self

RECAPTIULATION:

- Introduction to Probability
- Review of Set Theory
- Counting Principles
- Definition of Probability,
- Axioms of Probability,
- Addition Rule of Probability
- Conditional Probability,
- Independent events and,
- Total Probability

Contact Session 4

Contact Session	List of Topic Title	Reference
CS - 4	Bayes theorem(with proof),Introduction to Naïve Bayes concept.	T1 & T2
HW	Problems on Bayes theorem	T1 & T2
Lab	Bayes theorem & Naïve Bayes Concept	Lab 2

Agenda

- Bayes Theorem
- Introduction to Naïve Bayes concept

Text Books

No	Author(s), Title, Edition, Publishing House
T1	Statistics for Data Scientsists, An introduction to probability, statistics and Data Analysis, Maurits Kaptein et al, Springer 2022
T2	Probability and Statistics for Engineering and Sciences, 8th Edition, Jay L Devore, Cengage Learning

BAYES THEOREM:

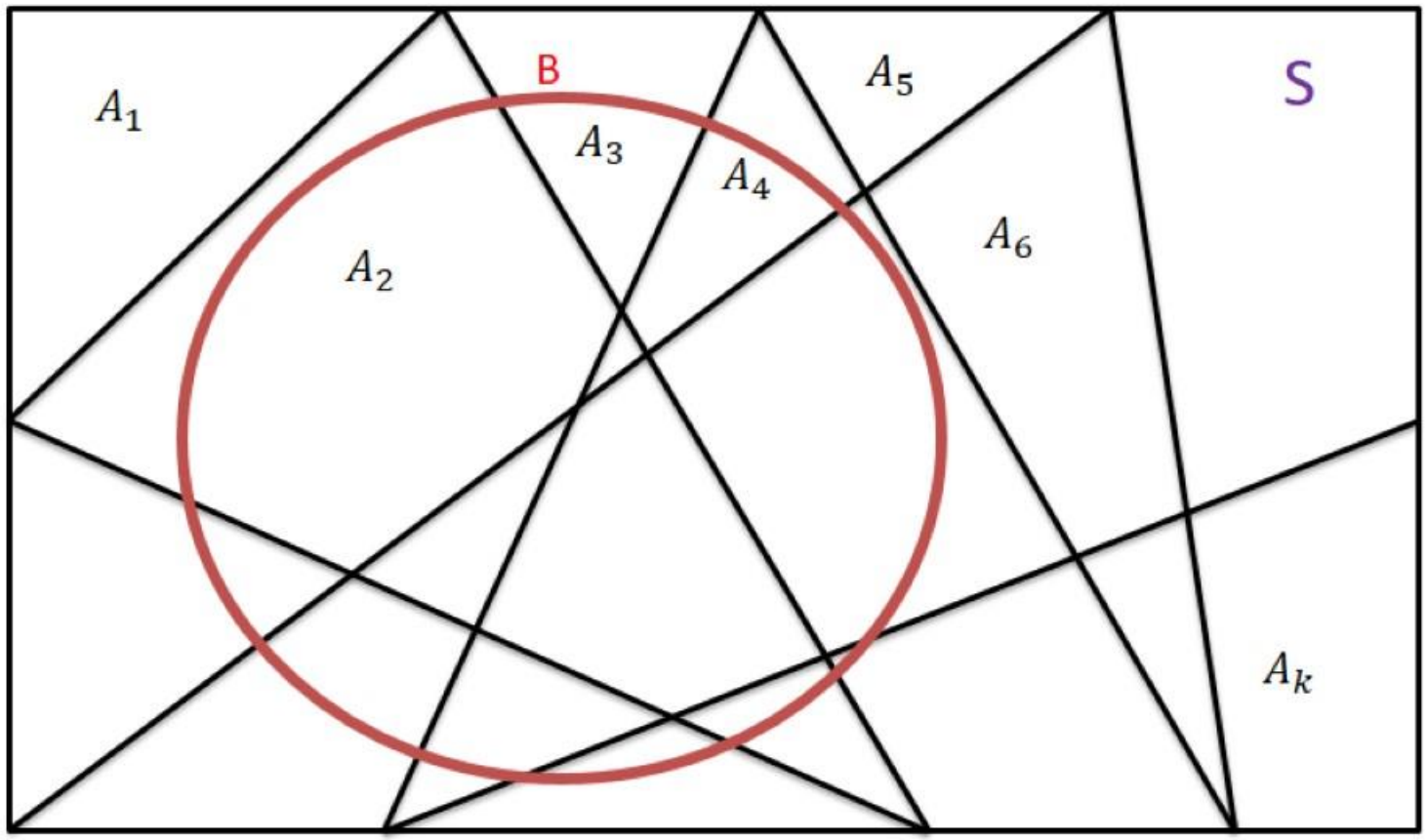
Let $P = \{A_1,\ A_2,\ A_3,\ \ldots\ , A_k\}$ be a set of exhaustive and mutually exclusive events of a sample space S with $P(A_i) \neq 0$ For each i. If B is any other event associated with A_i with $P(B) \neq 0$, then

$$P\left(\frac{A_i}{B}\right) = \frac{P(A_i)P\left(\frac{B}{A_i}\right)}{P(B)}$$

$$= \frac{P(A_i)P\left(\frac{B}{A_i}\right)}{\sum_{i=1}^{n} P(A_i)P\left(\frac{B}{A_i}\right)}$$

Geometrical Representation of Bayes theorem

$S = A_1 \cup A_2 \cup \ldots \cup A_n$ and $A_1 \cap A_2 \cap \ldots \cap A_n = \emptyset$



$$\therefore B = B \cap S = B \cap \{A_1 \cup A_2 \cup A_3, \ldots \cup A_n\}$$

Proof:

The conditional Probability of A_i for any i given B is given as

$$P(A_i \mid B) = \frac{P(A_i \cap B)}{P(B)}$$

But we know that

$$P(A_i \cap B) = P(A_i) \cdot P(B \mid A_i) \;\&\; P(B) = \sum_{i=1}^{i=n} P(A_i) P(B \mid A_i)$$

$$P(A_i \mid B) = \frac{P(A_i) \cdot P(B \mid A_i)}{\sum_{i=1}^{i=n} P(A_i) P(B \mid A_i)}$$

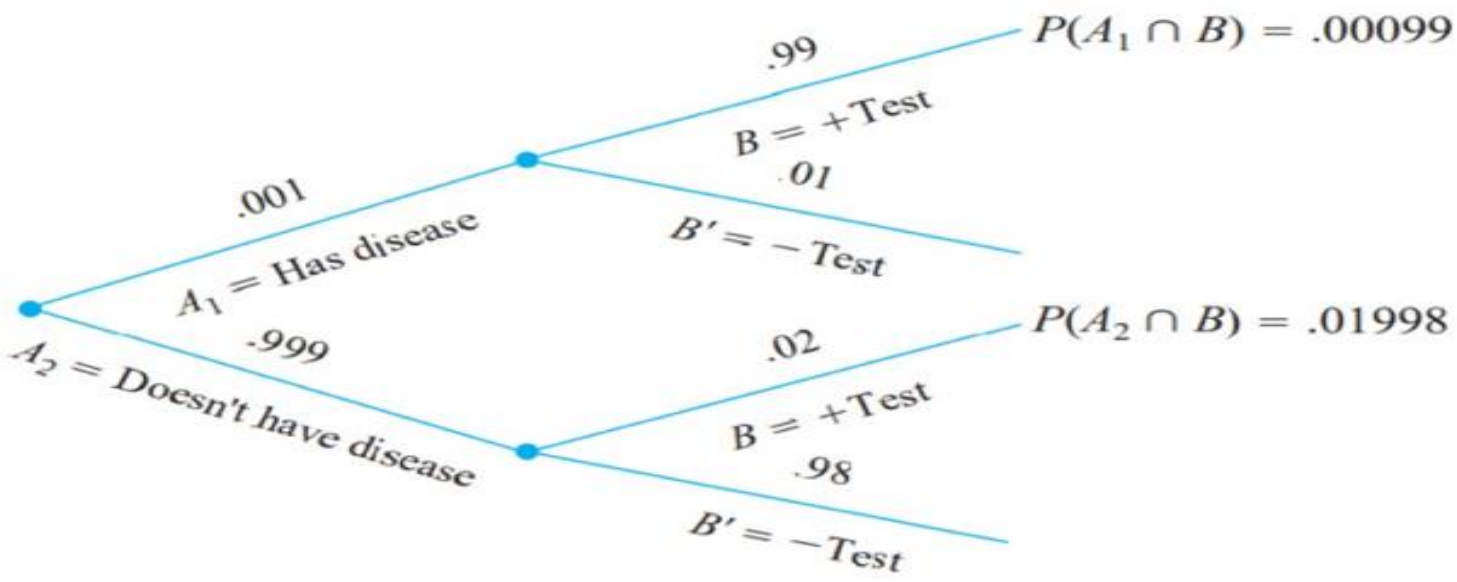
Hence the theorem proved

Example: 1

Only 1 in 1000 adults is afflicted with a rare disease for which a diagnostic test has been developed. The test is such that when an individual actually has the disease, a positive result will occur 99% of the time, whereas an individual without the disease will show a positive test result only 2% of the time. If a randomly selected individual is tested and the result is positive, what is the probability that the individual has the disease?

Solution:

To use Bayes' theorem, let A_1 = individual has the disease, A_2 = individual does not have the disease, and B = positive test result. Then $P(A_1) = .001$, $P(A_2) = .999$, $P(B|A_1) = .99$, and $P(B|A_2) = .02$. The tree diagram for this problem is in Figure



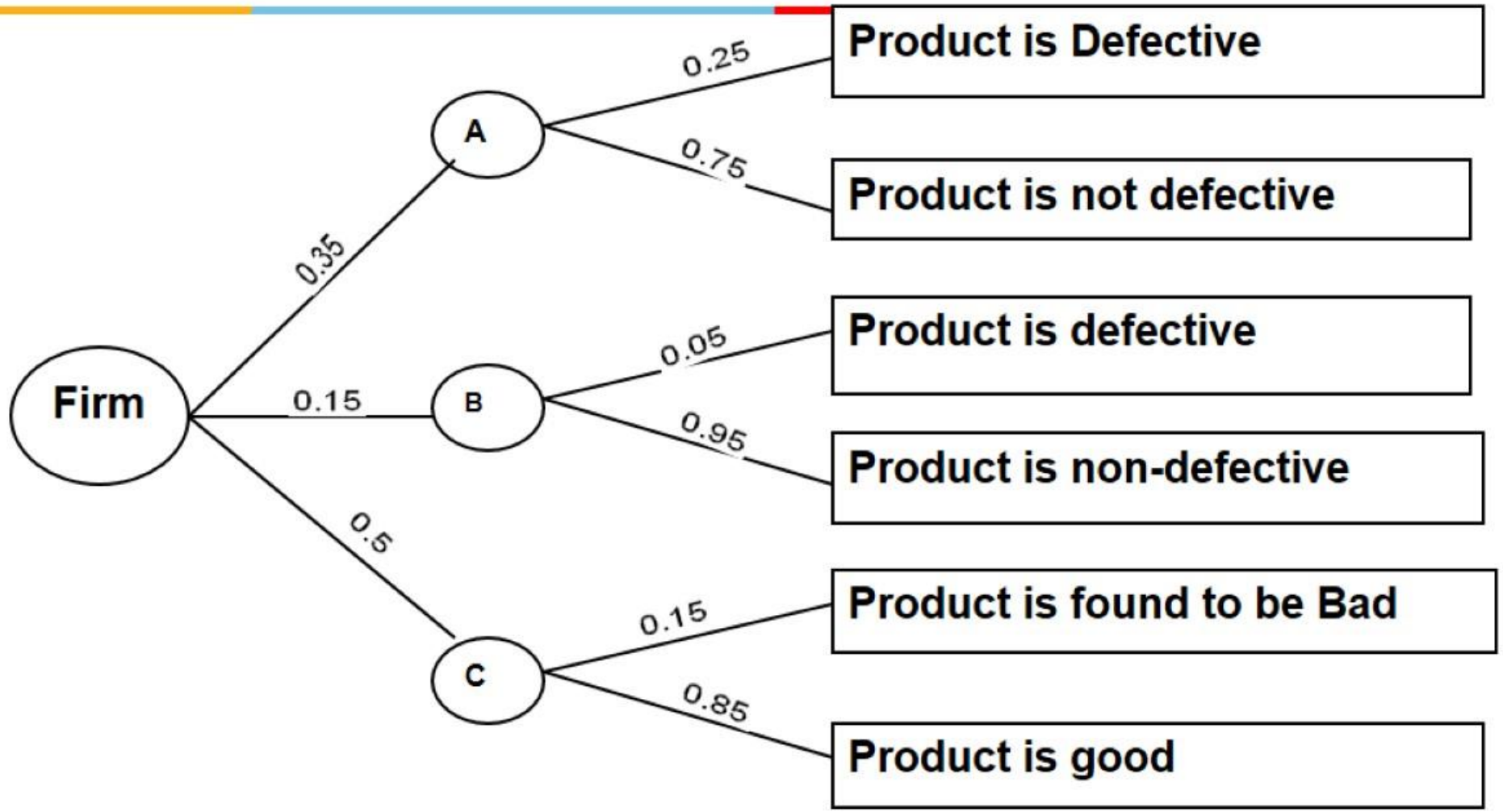
Next to each branch corresponding to a positive test result, the multiplication rule yields the recorded probabilities. Therefore, $P(B) = .00099 + .01998 = .02097$, from which we have

$$P(A_1 | B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{.00099}{.02097} = .047$$

Example: 2

A certain firm has plants A, B, C producing, respectively 35 %, 15% and 50% of the total output. The probabilities of a non – defective product are, respectively, 0.75, 0.95 and 0.85.A Customer receives a bad product, what is the Chance that product came from the plant C?

Tree Diagram



Solution

Let X: “ Customer receives a defective product”.

Clearly, P(X)= $P(A)P\left[\frac{X}{A}\right]+P(B)P\left[\frac{X}{B}\right]+P(C)P\left[\frac{X}{C}\right]$

=0.17

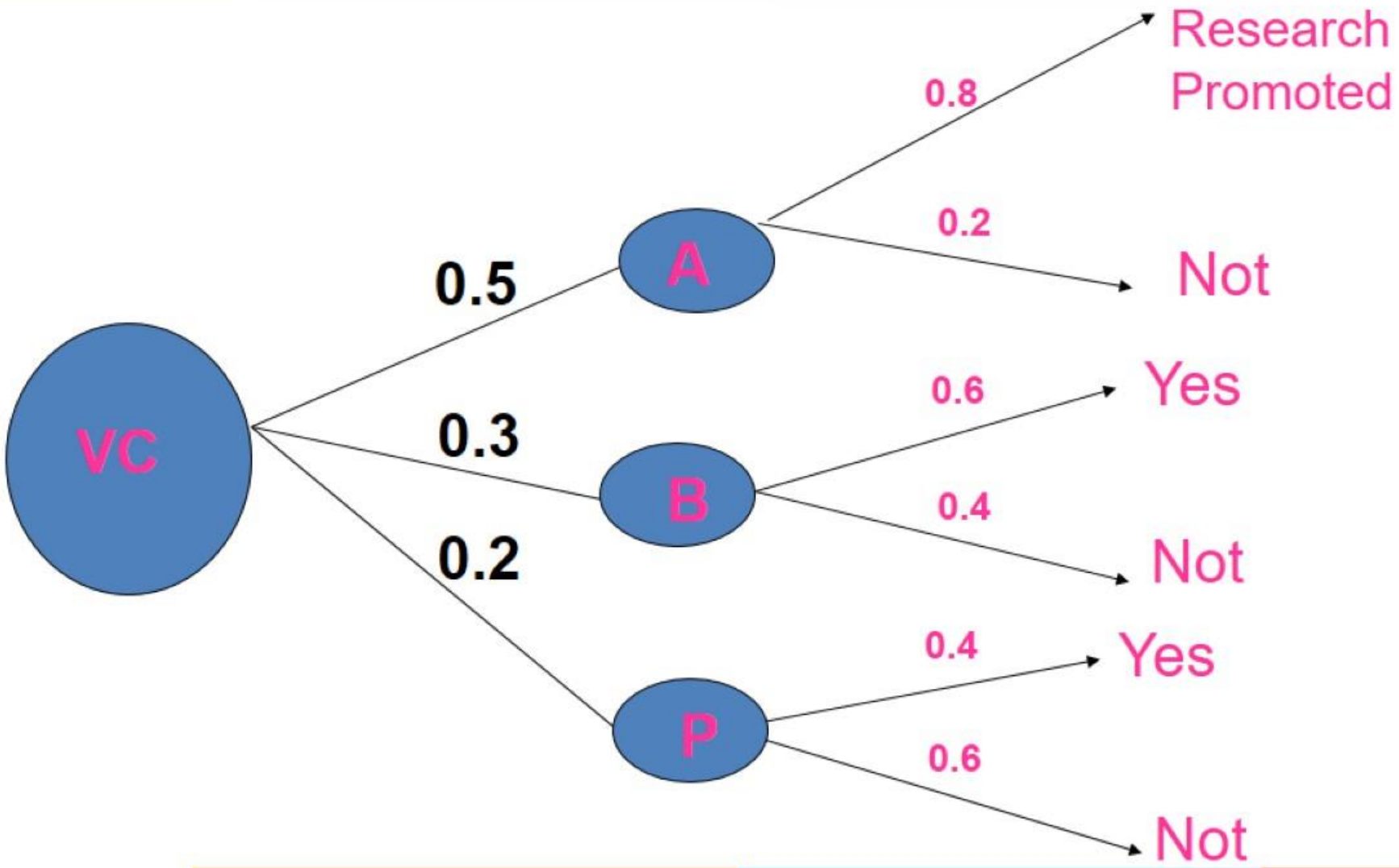
Therefore, the chance that product is manufactured by the plant C is

$$P(C \mid X)=\frac{P(C \cap X)}{P(X)} = \frac{0.5 \cdot 0.15}{0.17} = 0.4412$$

Example: 2

The chances that an academician, a business man and a politician becoming Vice Chancellor of an university are 0.5, 0.3 and 0.2 respectively. The probability that research work will be promoted in the university by these 3 gentlemen are respectively are 0.8, 0.6 and 0.4. It is found Research work has been promoted by the university. What is the chance that an academician has become the VC?

Tree Diagram



Let **X: " Research work is promoted"**

Clearly, $P(X) = 0.5 \times 0.8 + 0.3 \times 0.6 + 0.2 \times 0.4 = 0.66$

Now to find P[" An Academician is VC" / "Research work is promoted i.e. event X"] $= \frac{0.5 \times 0.8}{0.66} = 0.6061$

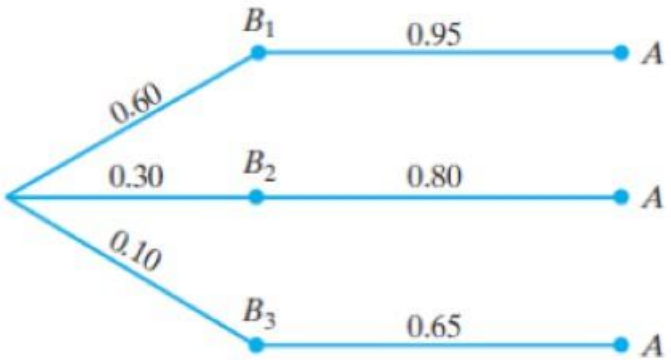
Example: A manufacturer of tablets receives its LED screens from three different suppliers, 60% from supplier $B1$, 30% from supplier $B2$, and 10% from supplier $B3$. In other words, the probabilities that any one LED screens received by the plant comes from these three suppliers are 0.60, 0.30, and 0.10. Also suppose that 95% of the LED screens from $B1$, 80% of those from $B2$, and 65% of those from $B3$ perform according to specifications.

1) What is the probability that any one LED screen received by the plant will perform according to specifications?

2) Determine the probability that a particular LED screen, which is known to perform according to specifications, came from supplier $B3$.

If A denotes the event that a LED screen received by the plant performs according to specifications, and $B1$, $B2$, and $B3$ are the events that it comes from the respective suppliers

$$A = A \cap [B1 \cup B2 \cup B3]$$
$$= (A \cap B1) \cup (A \cap B2) \cup (A \cap B3)$$
$$= P(A \cap B1) + P(A \cap B2) + P(A \cap B3)$$
$$P(A) = P(B1) \cdot P(A \mid B1) + P(B2) \cdot P(A \mid B2) + P(B3) \cdot P(A \mid B3)$$
$$P(A) = (0.60)(0.95) + (0.30)(0.80) + (0.10)(0.65)$$
$$= 0.875$$



$$P(B3 \mid A) = \frac{(0.10)(0.65)}{(0.60)(0.95) + (0.30)(0.80) + (0.10)(0.65)} = 0.074$$

Example: 3

Example: Two firms V and W consider bidding on a road-building job, which may or may not be awarded depending on the amounts of the bids. Firm V submits a bid and the probability is ¾ that it will get the job provided firm W does not bid. The probability is ¾ that W will bid, and if it does, the probability that V will get the job is only1/3. (a) what is the probability that V will get the job? (b) If V gets the job, what is the probability that W did not bid?

Solution:

Answer: Given $P(V/W^1)=3/4$, $P(W) = 1/3$, $P(V/W)=1/3$ $P(W^1) = 2/3$

(a) $V = (V \cap W) \cup (V \cap W^1) \Rightarrow P(V) = P(V \cap W) + P(V \cap W^1)$

$$P(V) = P(V/W)\,P(W) + P(V/W^1)\,P(W^1) = 11/18$$

(b) $P(W^1/V) = P(V/W^1)\,P(W^1) / P(V) = 9/11$

Suggested problems:

Example . An office has 4 secretaries handling respectively 20%, 60%,15% and 5% of the files of all government reports. The probability that they misfile such reports are respectively 0.05, 0.1, 0.1 and 0.05. Find the probability that the misfiled report can be blamed on the first secretary.

Example . In a class 70% are boys and 30% are girls. 5% of boys and 3% of girls are irregular to the classes. What is the probability of a student selected at random is irregular to the classes and what is the probability that the irregular student is a girl?

Example:

Three machines A, B and C produce respectively 60%, 30% and 10% of the total number of items of a factory. The percentage of defective outputs of these machines are 2%, 3% and 4%. An item is selected at random and is found to be defective. (i) Find the probability that the item was produced by machine C? (ii) What is the probability that the item was produced by machine C or B?

Bayesian Learning

- Naive Bayes is a set of simple and efficient machine learning algorithms for solving a variety of classification and regression problems.
- Naive Bayes assumes conditional independence where Bayes theorem does not. This means the relationship between all input features are independent.
- Bayesian learning algorithms that calculate explicit probabilities for hypotheses, such as the naive Bayes classifier, are among the most practical approaches to certain types of learning problems
- For example: Problem of learning to classify text documents such as electronic news articles.
- For such learning tasks, the naive Bayes classifier is among the most effective algorithms known

Features of Bayesian learning

- Prior knowledge can be combined with observed data to determine the final probability of a hypothesis.
- Prior knowledge is provided by asserting
  - prior probability for each candidate hypothesis, and
  - probability distribution over observed data for each possible hypothesis.
- New instances can be classified by combining the predictions of multiple hypotheses, weighted by their probabilities.

Bayes Theorem

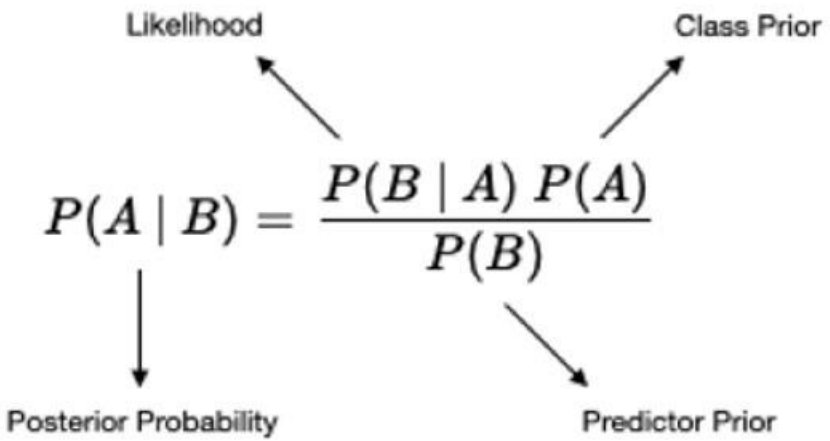
- $P(h)$ = prior probability of hypothesis h, before seeing the training data
- $P(D)$ = prior probability of training data D
- $P(h|D)$ = probability of h given D
- $P(D|h)$ = probability of D given h

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

Machine Learning

- Generative models
  - Build model to estimate the posterior probability P(Y|X) by estimating
  - likelihood of data given target (hypothesis) P(X|Y)
  - Prior probabilities over target P(Y)
  - In general, for a specific class Y=ck,

$$P(Y = c_k|X) = \frac{P(X|Y = c_k)*P(Y=c_k)}{P(X)}$$



Choosing Hypotheses

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- Generally want the most probable hypothesis given the training data
- *Maximum a posteriori* hypothesis h_{MAP}:

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(h|D) \\ &= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D|h)P(h) \end{aligned}$$

- If assume $P(h_i) = P(h_j)$ then can further simplify, and choose the *Maximum likelihood* (ML) hypothesis

$$h_{ML} = \arg\max_{h_i \in H} P(D|h_i)$$

Brute Force MAP Hypothesis

- For each hypothesis *h* in *H*, calculate the posterior probability

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

- Output the hypothesis h_{MAP} with the highest posterior probability

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} P(h|D)$$

MAP Hypothesis

- Using Bayes theorem, we compute the MAP hypothesis for all probable hypothesis (or all unique class labels)
- Identify the best hypothesis describing the data as

$$h_{MAP} = argmax_{h \in H} P(h|D)$$

$$= argmax_{h \in H} \frac{P(D|h)P(h)}{P(D)}$$

$$= argmax_{h \in H} P(D|h)P(h)$$

H: set of all hypothesis

P(D) is independent of h and is same for all hypothesis, therefore dropped

Maximum Likelihood Estimation

- When no prior information is available, all hypothesis are equally likely i.e p(hi) = p(hj)
- This is also true for a balanced class problem where all the classes are equally likely
- This is known as uniform prior
- MAP hypothesis further simplified to
- h_{ML} = argmax P(D/h) (where h belongs to H)

Conditional independence

- **Definition**: X is conditionally independent of Y given Z, if the probability distribution governing X is independent of the value of Y, given the value of Z

$$(\forall i, j, k)\ P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z_k)$$

$$P(X|Y, Z) = P(X|Z)$$

Example:

$$P(\text{Thunder}|\text{Rain}, \text{Lightning}) = P(\text{Thunder}|\text{Lightning})$$

Applying conditional independence

Naïve Bayes assumes X_i are conditionally independent given Y

e.g., $P(X_1|X_2,Y) = P(X_1|Y)$

$$P(X_1,X_2|Y) = P(X_1|X_2,Y)P(X_2|Y)$$
$$= P(X_1|Y)P(X_2|Y)$$

General form: $P(X_1,\cdots,X_n|Y) = \prod_{j=1}^{n} P(X_j|Y)$

How many parameters to describe $P(X_1,\cdots,X_n|Y)$? $P(\mathrm{Y})$?

Without conditional independence assumption?

With conditional independence assumption?

Naïve Bayes Independence assumption

Assumption:

$$P(X_1, \cdots, X_n|Y) = \prod_{j=1}^{n} P(X_j|Y)$$

i.e., X_i and X_j are conditionally independent given Y for $i \neq j$

Naïve Bayes classifier

- **Bayes rule:**

$$P(Y = y_k | X_1, \cdots, X_n) = \frac{P(Y = y_k) P(X_1, \cdots, X_n | Y = y_k)}{\sum_j P(Y = y_j) P(X_1, \cdots, X_n | Y = y_j)}$$

- **Assume conditional independence among X_i's:**

$$P(Y = y_k | X_1, \cdots, X_n) = \frac{P(Y = y_k) \Pi_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \Pi_i P(X_i | Y = y_j)}$$

- **Pick the most probable (MAP) Y**

$$\hat{Y} \leftarrow \underset{y_k}{\operatorname{argmax}} P(Y = y_k) \Pi_i P(X_i | Y = y_k)$$

Prior Probability

MLE

NAÏVE BAYES CLASSIFIER

- Assume independence among attributes X_i when class is given:
  - $P(X_1, X_2, \ldots, X_d \mid Y_j) = P(X_1 \mid Y_j)\, P(X_2 \mid Y_j) \ldots P(X_d \mid Y_j)$
  - Now we can estimate $P(X_i \mid Y_j)$ for all X_i and Y_j combinations from the training data
  - New point is classified to Y_j if $P(Y_j) \prod P(X_i \mid Y_j)$ is maximal.

Example 1:

If the weather is sunny,
then the player will play
or not?

i.e. Play/ Sunny = Yes or No

Note if we know P(Yes/Sunny) and P(No/Sunny) then we can answer the question asked

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Steps to Apply Bayes Theorem

Step 1- View or collect "raw" data.

Step 2 - Convert long data to a frequency table

	Play		
	no	yes	Row Total
weather			
Sunny	2	3	5
Overcast	0	4	4
Rainy	3	2	5
Column Total	5	9	14

Step 3 - Row and column sums to get probabilities

Weather probabilities

sunny = 5/14 , rainy = 5/14

Overcast = 4/14

Play probabilities

no = 5/14

yes = 9/14

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Steps to Apply Bayes Theorem

	Play		
	no	yes	Row Total
weather			
Sunny	2	3	5
Overcast	0	4	4
Rainy	3	2	5
Column Total	5	9	14

P(Sunny)= 5/14

P(Overcast) = 4/14

P(Rainy)=5/14

P(no)=5/14 P(yes)=9/14

Step 4 - Apply probabilities from frequency table to Bayes theorem

$$P(yes \mid sunny) = \frac{P(sunny \mid yes)\, P(yes)}{P(sunny)}$$

Weather	No	Yes
Rainy	3	2
Sunny	2	3
overcast	0	4
Total	5	9

$5/14 = 0.36$

$5/14 = 0.36$

$4/14 = 0.29$

$5/14 = 0.36$ $\quad$ $9/14 = 0.64$

Now $P(\text{Yes}/\text{sunny}) = \dfrac{P(\text{Sunny}/\text{Yes})\, P(\text{Yes})}{P(\text{sunny})}$

$= \dfrac{(3/9)(9/14)}{5/14} = \underline{0.60}$ ✓

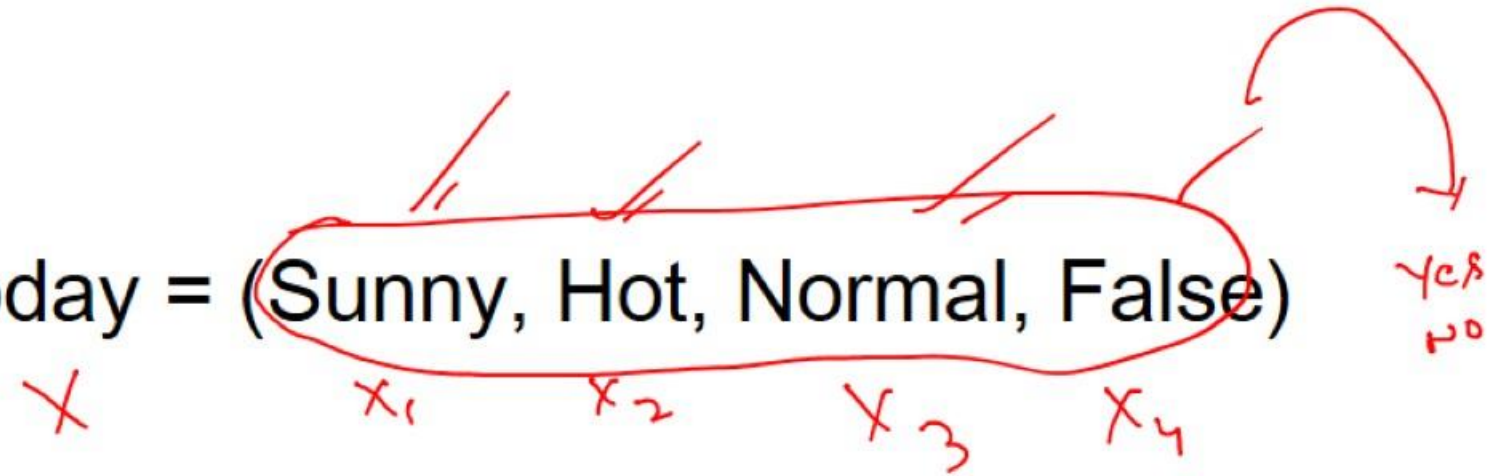
$P(\text{No}/\text{sunny}) = \dfrac{(2/5)(5/14)}{5/14} = \underline{0.40}$

Example 2:

If the features of today = (Outlook is Sunny, Temp is Hot, Humidity is Normal, Windy is False), then the player will play or not?

S. No	Outlook	Temp	Humidity	Windy	Play Tennis
1	Rainy	Hot	High	False	No
2	Rainy	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Sunny	Mild	High	False	Yes
5	Sunny	Cool	Normal	False	Yes
6	Sunny	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Rainy	Mild	High	False	No
9	Rainy	Cool	Normal	False	Yes
10	Sunny	Mild	Normal	False	Yes
11	Rainy	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Sunny	Mild	High	True	No

today = (Sunny, Hot, Normal, False)



$$P(y_1 \mid X) = \frac{P(X \mid yes)\, P(yes)}{P(X)}$$

~~$P(no \mid X)$~~

$$= \frac{P(x_1, x_2, x_3, x_4 \mid yes)\, yes}{P(x_1, x_2, x_3, x_4)} = \frac{P(x_1 \mid y)\, P(x_2 \mid y)\, P(x_3 \mid y)\, P(x_4 \mid y)\, P(y)}{P(x_1)\, P(x_2)\, P(x_3)\, P(x_4)}$$

Outlook

	Yes	No	P(yes)	P(no)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5
Total	9	5	100%	100%

Temperature

	Yes	No	P(yes)	P(no)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5
Total	9	5	100%	100%

Humidity

	Yes	No	P(yes)	P(no)
High	3	4	3/9	4/5
Normal	6	1	6/9	1/5
Total	9	5	100%	100%

Wind

	Yes	No	P(yes)	P(no)
False	6	2	6/9	2/5
True	3	3	3/9	3/5
Total	9	5	100%	100%

Play		P(Yes)/P(No)
Yes	9	9/14
No	5	5/14
Total	14	100%

Example 3:

Magazine Promotion	Watch Promotion	Life Insurance Promotion	Credit Card Insurance	Sex
Yes	No	No	No	Male
Yes	Yes	Yes	Yes	Female
No	No	No	No	Male
Yes	Yes	Yes	Yes	Male
Yes	No	Yes	No	Female
No	No	No	No	Female
Yes	Yes	Yes	Yes	Male
No	No	No	No	Male
Yes	No	No	No	Male
Yes	Yes	Yes	No	Female

New Instance: Magazine Promotion = Yes , Watch Promotion = Yes, Life Insurance Promotion = No, Credit Card Insurance = No then Sex = ?

D = { magazine promotion, watch promotion, Life Insurance promotion, credit card insurance }

h_i = male or Female

P (male // Yes, Yes, NO, NO) = ?

P (Female / Yes, Yes, NO, NO) = ?

	Magazine Promotion		Watch Promotion		L.I Promotion		Credit Card Promotion	
	Male	Female	Male	Female	Male	F	M	F
YES	4	3	2	2	2	3	2	1
NO	2	1	4	2	4	1	4	3
Ratios (YES)	$\frac{4}{6}$	$\frac{3}{4}$	$\frac{2}{6}$	$\frac{2}{4}$	$\frac{2}{6}$	$\frac{3}{4}$	$\frac{2}{6}$	$\frac{1}{4}$
NO	$\frac{2}{6}$	$\frac{1}{4}$	$\frac{4}{6}$	$\frac{2}{4}$	$\frac{4}{6}$	$\frac{1}{4}$	$\frac{4}{6}$	$\frac{3}{4}$

$$P(\text{male}/E)$$

$$= \frac{P(E/\text{male})\,P(\text{male})}{P(E)}$$

$$= \frac{\left(\frac{4}{6}\cdot\frac{2}{6}\cdot\frac{4}{6}\cdot\frac{4}{6}\right)\left(\frac{3}{5}\right)}{P(E)}$$

$$= \frac{0.0593}{P(E)}$$

YES YES NO NO

$\frac{6}{10}\cdot\frac{3}{5}$

$$P(\text{Female}/E)$$

$$= \frac{P(E/\text{Female})\ P(\text{Female})}{P(E)}$$

$$= \frac{\left(\frac{3}{4}\right)\left(\frac{2}{4}\right)\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)\cdot\frac{2}{5}}{P(E)}$$

$$= \frac{\left(\frac{9}{128}\right)\left(\frac{2}{5}\right)}{P(E)} = \frac{0.0281}{P(E)}$$

$0.0593 > 0.0281$

i.e. male ✓

Example 4:

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

A: attributes

M: mammals

N: non-mammals

$$P(A \mid M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A \mid N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A \mid M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A \mid N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

Issues with Naïve Bayes Classifier

Naïve Bayes Classifier:

P(Refund = Yes | No) = 3/7
P(Refund = No | No) = 4/7
P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/7
P(Marital Status = Divorced | No) = 1/7
P(Marital Status = Married | No) = 4/7
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0

For Taxable Income:
If class = No: sample mean = 110
sample variance = 2975
If class = Yes: sample mean = 90
sample variance = 25

- P(Yes) = 3/10
  P(No) = 7/10

- P(Yes | Married) = 0 x 3/10 / P(Married)
  P(No | Married) = 4/7 x 7/10 / P(Married)

Issues with Naïve Bayes Classifier

Consider the table with Tid = 7 deleted

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Naïve Bayes Classifier:

P(Refund = Yes | No) = 2/6
P(Refund = No | No) = 4/6
P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/6
P(Marital Status = Divorced | No) = 0
P(Marital Status = Married | No) = 4/6
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0/3
For Taxable Income:
If class = No: sample mean = 91
sample variance = 685
If class = No: sample mean = 90
sample variance = 25

Given X = (Refund = Yes, Divorced, 120K)

$P(X \mid No) = 2/6 \times 0 \times 0.0083 = 0$

$P(X \mid Yes) = 0 \times 1/3 \times 1.2 \times 10^{-9} = 0$

Naïve Bayes will not be able to classify X as Yes or No!

Naïve Bayes for Text Classification

- Naïve Bayes is commonly used for text classification
- For a document with k terms $d = (t_1, \dots, t_k)$

Fraction of documents in c

$$P(c|d) = P(c)P(d|c) = P(c)\prod_{t_i \in d} P(t_i|c)$$

- $P(t_i|c)$= Fraction of terms from all documents in c that are t_i.

Number of times t_i appears in some document in c

$$P(t_i|c) = \frac{N_{ic} + 1}{N_c + T}$$

Laplace Smoothing

Number of unique words (vocabulary size)

Total number of terms in all documents in c

- Easy to implement and works relatively well
- Limitation: Hard to incorporate additional features (beyond words).
  - E.g., number of adjectives used.

A Simple Example:

Text	Tag
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports

Which tag does the sentence *A very close game* belong to? i.e. P(sports| *A very close game*)

Feature Engineering: Bag of words i.e use word frequencies without considering order

Using Bayes Theorem:

$$P(\text{sports}|\ \textit{A very close game}) = \frac{P(\textit{A very close game}|\ \text{sports})\ P(\text{sports})}{P(\textit{A very close game})}$$

We assume that every word in a sentence is **independent** of the other ones

"close" doesn't appear in sentences of sports tag, So P(close | sports) = 0, which makes product 0

A Simple Example

Draw | Naive Bayes → Text classification

Text	Tag
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports

Which tag does the sentence "*A very close game*" belong to? i.e. P(sports| *A very close game*)

Feature Engineering: Bag of words i.e use word frequencies without considering order

Using Bayes Theorem:

P(sports| *A very close game*)

$$= \frac{\text{P(}\textit{A very close game}\text{| sports) P(sports)}}{\text{P(}\textit{A very close game}\text{)}}$$

We assume that every word in a sentence is **independent** of the other ones

P(A very close game) = P(A) P(very) P(close) P(game)

P(A very close game | sports) = P(a | sports) P(very | sports) P(close | sports) P(game | sports)

"close" doesn't appear in sentences of sports tag, So P(close | sports) = 0, which makes product 0

$$\frac{P(\text{sports} \mid \text{A very close game})}{P(\text{not sports} \mid \text{A very close game})} = \frac{P(\text{A very close game} \mid \text{sports}) \cdot P(\text{sports})}{P(\text{A very close game})}$$

Laplace smoothing

- Laplace smoothing: we add 1 or in general constant k to every count so it's never zero.
- To balance this, we add the number of possible words to the divisor, so the division will never be greater than 1
- In our case, the possible words are ['a', 'great', 'very', 'over', 'it', 'but', 'game', 'election', 'clean', 'close', 'the', 'was', 'forgettable', 'match'].

Apply Laplace Smoothing

Word	P(word | Sports)	P(word | Not Sports)
a	2+1 / 11+14	1+1 / 9+14
very	1+1 / 11+14	0+1 / 9+14
close	0+1 / 11+14	1+1 / 9+14
game	2+1 / 11+14	0+1 / 9+14

$$P(a|Sports) \times P(very|Sports) \times P(close|Sports) \times P(game|Sports) \times P(Sports)$$
$$= 2.76 \times 10^{-5}$$
$$= 0.0000276$$

$$P(a|Not\ Sports) \times P(very|Not\ Sports) \times P(close|Not\ Sports) \times P(game|Not\ Sports) \times P(Not\ Sports)$$
$$= 0.572 \times 10^{-5}$$
$$= 0.00000572$$

Example :

Doc No	Text
1	I LOVED THE MOVIE
2	I HATED THE MOVIE
3	A GREAT MOVIE ,GOOD MOVIE
4	POOR ACTING
5	GREAT ACTING , A GOOD MOVIE
NEW	I HATED THE POOR ACTING

Example :

Doc No	Text	
1	I LOVED THE MOVIE	_POSITIVE
2	I HATED THE MOVIE	_ NEGATIVE
3	A GREAT MOVIE ,GOOD MOVIE	_POSITIVE
4	POOR ACTING	_NEGATIVE
5	GREAT ACTING , A GOOD MOVIE	_POSITIVE
NEW	I HATED THE POOR ACTING	_????

$P(c/x)$

$P(+ / \text{I hated the poor acting}) =$

$P(- / \text{I hated the poor acting}) =$

Based on these probabilities, we can decide the class which the new text belongs

$P(+ \mid \text{I hated the acting})$

ie $P(c_1 \mid x) = \dfrac{P(x \mid c_1)\, P(c_1)}{\cancel{P(x)}}$

$= P(\text{I} \mid +)\, P(\text{hated} \mid +)\, P(\text{the} \mid +)\, P(\text{acting} \mid +)\, P(+)$

$\dfrac{P(\text{I}, +)}{P(+)}$ $\qquad$ $\dfrac{P(\text{hated}, +)}{P(+)}$

words	positive	Negative
I	1	1
loved	1	0
the	1	1
movie	4	1
hated	0	1
a	2	0
great	2	0
Poor	0	1
acting	1	1
good	2	0

$$P(\pm/+)$$

$$= \frac{1+1}{14+10}$$

$$P(I/-)$$

$$= \frac{1+1}{6+10}$$

"I hated the poor acting"

word	positive	negative
I	$\frac{1+1}{14+10} = 0.0833$	$\frac{1+1}{6+10} = 0.125$
hated	$\frac{0+1}{14+10} = 0.0417$	$\frac{1+1}{6+10} = 0.125$
the	$\frac{1+1}{14+10} = 0.0833$	$\frac{1+1}{6+10} = 0.125$
poor	$\frac{0+1}{14+10} = 0.0417$	$\frac{1+1}{6+10} = 0.125$
acting	$\frac{1+1}{14+10} = 0.0833$	$\frac{1+1}{6+10} = 0.125$

x : I hate the Poor acting

$P(+/x)$

$= (\quad)(\quad)(\quad)(\quad)(\quad) \times P(+)$

$P(+) \downarrow 3/5$

$= 6.03 \times 10^{-7}$

$P(-/x)$

$= (\quad)(\quad)(\quad)(\quad)(\quad) P(-1)$

$P(-1) \downarrow 2/5$

$= 1.22 \times 10^{-5}$

$\therefore$ Negative class

Example:

Suppose we got the new message with the words '**Dear Friend**', Decide whether this new message is a normal or spam message?

i.e. Normal/ Dear, Friend = Yes or No

Note if we know P (Normal/ Dear, Friend) and P(Spam/ Dear, Friend) then we can answer the question asked

Email word	Spam	Email word	Spam
Dear	Yes	Friend	No
Friend	No	Friend	Yes
Dear	No	Dear	No
Dear	No	Lunch	No
Dear	No	Friend	No
Friend	No	Dear	No
Lunch	No	Dear	No
Friend	No	Dear	No
Lunch	No	Dear	No
Dear	Yes	Money	Yes
Money	Yes	Money	No
Money	Yes	Money	Yes

Steps to Apply Bayes Theorem

Step 1- View or collect "raw" data.

Step 2 - Convert long data to a frequency table

	Play		
	normal	spam	Row Total
word			
Dear	8	2	10
Friend	5	1	6
Lunch	3	0	3
Money	1	4	5
Column Total	17	7	24

Step 3 - Row and column sums to get probabilities

Email word	Spam	Email word	Spam
Dear	Yes	Friend	No
Friend	No	Friend	Yes
Dear	No	Dear	No
Dear	No	Lunch	No
Dear	No	Friend	No
Friend	No	Dear	No
Lunch	No	Dear	No
Friend	No	Dear	No
Lunch	No	Dear	No
Dear	Yes	Money	Yes
Money	Yes	Money	No
Money	Yes	Money	Yes

Steps to Apply Bayes Theorem

Play	normal	spam	Row Total
word			
Dear	8	2	10
Friend	5	1	6
Lunch	3	0	3
Money	1	4	5
Column Total	17	7	24

As P (N/ D, F) > P (S/ D, F) , we can decide that Dear Friend is Normal message.

Step 4 - Apply probabilities from frequency table to Bayes theorem

$$P(N/D,F) = \frac{P(D,F/N).P(N)}{P(D,F)} = \frac{P(D/N).P(F/N).P(N)}{P(D).P(F)} = \frac{\left(\frac{8}{17}\right).\left(\frac{5}{17}\right).\left(\frac{17}{24}\right)}{\left(\frac{10}{24}\right).\left(\frac{6}{24}\right)} = \frac{0.098}{0.104} = 0.9423$$

$$P(S/D,F) = \frac{P(D,F/S).P(S)}{P(D,F)} = \frac{P(D/S).P(F/S).P(S)}{P(D).P(F)} = \frac{\left(\frac{2}{7}\right).\left(\frac{1}{7}\right).\left(\frac{7}{24}\right)}{\left(\frac{10}{24}\right).\left(\frac{6}{24}\right)} = \frac{0.012}{0.104} = 0.1153$$

Example continued:

Suppose we got the new message contains the word '**Lunch Money Money Money Money**' , Decide whether this new message is a normal or spam message?

Email word	Spam	Email word	Spam
Dear	Yes	Friend	No
Friend	No	Friend	Yes
Dear	No	Dear	No
Dear	No	Lunch	No
Dear	No	Friend	No
Friend	No	Dear	No
Lunch	No	Dear	No
Friend	No	Dear	No
Lunch	No	Dear	No
Dear	Yes	Money	Yes
Money	Yes	Money	No
Money	Yes	Money	Yes

Steps to Apply Bayes Theorem

Play

word	normal	spam	Row Total
Dear	8	2	10
Friend	5	1	6
Lunch	3	0	3
Money	1	4	5
Column Total	17	7	24

We can observe that we have to classify any message with Lunch as Normal message, no matter how many times we see the word Money and that's the problem.

To work around this problem add 1 count to the frequency table to each word(Laplace smoothing)

Step 4 - Apply probabilities from frequency table

$P(N).\ P(L/N).\ P(M/N)^4=\left(\frac{17}{24}\right).\left(\frac{3}{17}\right).\left(\frac{1}{17}\right)^4 =0.0000015$

$P(S).\ P(L/S).\ P(M/S)^4=\left(\frac{7}{24}\right).\left(\frac{0}{7}\right).\left(\frac{4}{7}\right)^4 =0$

Steps to Apply Bayes Theorem

Play

word	normal	spam	Row Total
Dear	8+1	2+1	12
Friend	5+1	1+1	8
Lunch	3+1	0+1	5
Money	1+1	4+1	7
Column Total	21	11	32

Step 4 - Apply probabilities from frequency table

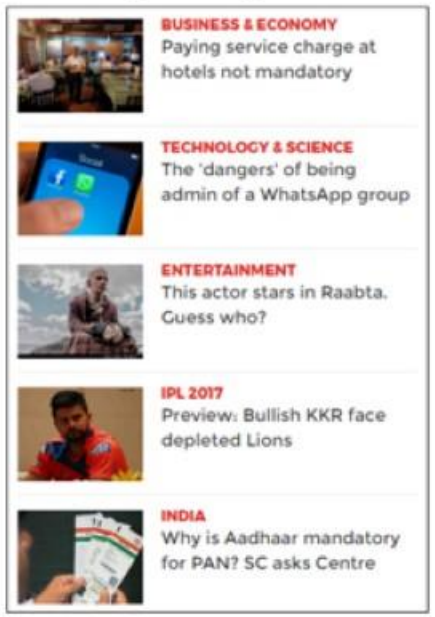
$$P(N).\ P(L/N).\ P(M/N)^4=\left(\frac{21}{32}\right).\left(\frac{4}{21}\right).\left(\frac{2}{21}\right)^4=0.00001$$

$$P(S).\ P(L/S).\ P(M/S)^4=\left(\frac{11}{32}\right).\left(\frac{1}{11}\right).\left(\frac{5}{11}\right)^4=0.00133$$

As $P(S/ L, M^4) > P(N/ L, M^4)$, we can decide that **Lunch Money Money Money Money is Spam** message.

Naïve Bayes Classifier Applications

Categorizing News



Email Spam Detection

Face Recognition

Sentiment Analysis

Naive Bayes Classifier

- ✓ Along with decision trees, neural networks, one of the most practical learning methods.
- ✓ When to use
  - ✓ Moderate or large training set available
  - ✓ Attributes that describe instances are conditionally independent given classification
- ✓ Successful applications:
  - ✓ Diagnosis
  - ✓ Classifying text documents

Learning to Classify Text

- Why?
  - Learn which news articles are of interest
  - Learn to classify web pages by topic

- Naive Bayes is among most effective algorithms
- What attributes shall we use to represent text documents??

Baseline: Bag of Words Approach

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	
...	
oil	
…	
Zaire	

Practical Issues of Bayesian learning

- Require initial knowledge of many probabilities
  - Often estimated based on background knowledge, previously available data, and assumptions about the form of the underlying distributions.
- Significant computational cost required to determine the Bayes optimal hypothesis in the general case (linear in the number of candidate hypotheses)

HW: Exercise

Consider the car theft problem with attributes Color, Type, Origin, and the target, Stolen can be either Yes or No.

we need to classify whether the car is stolen, given the features of the car.

Given the Red color Domestic SUV car Find the probability of whether the car is stolen?

Color	Type	Origin	Stolen?
Red	Sports	Domestic	Yes
Red	Sports	Domestic	No
Red	Sports	Domestic	Yes
Yellow	Sports	Domestic	No
Yellow	Sports	Imported	Yes
Yellow	SUV	Imported	No
Yellow	SUV	Imported	Yes
Yellow	SUV	Domestic	No
Red	SUV	Imported	No
Red	Sports	Imported	Yes

HW: Exercise

If the weather is Snowy, then the player will play or not?

weather	Player play
Sunny	yes
Rainy	no
Cloudy	yes
Sunny	no
Sunny	yes
snowy	no
Rainy	yes
Cloudy	no
Cloudy	yes
Sunny	yes
snowy	no
Cloudy	yes
Rainy	no
snowy	no
snowy	yes

Thanks