Inspire…Educate…Transform.

Foundations of Statistics and Probability for Data Science

Basic Probability Concepts, Probability Distributions

Prof Anuradha Sharma

8th December,2018

MATERIAL CONTENT FROM Dr. SRIDHAR PAPPU


FLASHBACK
I REMEMBER IT LIKE IT WAS YESTERDAY...
CMCOTT. 04/13/12 #138

MAXIMUM SECURITY

$50.7 BILLION SPENT FOR DEFENCE DEVELOPMENT IN 2016 PLACES INDIA AMONG WORLD'S TOP FIVE DEFENCE SPENDERS

INDIA IS ahead of Saudi Arabia and Russia's expenditure

THE US, China and the UK remain the top three defence spenders ahead of India's fourth place

$46.6 bn

INDIA SPENT $46.6 billion last year, as per a report released on Monday

THE REPORT said that India is set to overtake UK's budget by 2018

$1.6 trillion

The worldwide outlook shows that global defence spending rose by 1 per cent to $1.6 trillion this year, against 0.6 per cent in 2015.

DEFENCE EXPENDITURE

US	$622 bn
China	$191.7 bn
UK	$5.8 bn
Saudi	$48.68bn
Russia	$48.44 bn

"Over the next three years, India will re-emerge as a key growth market for defence suppliers
– Craig Caffrey, principal analyst for Asia-Pacific at 'IHS Janes'

38/35 in math, physics: In Bihar, some students score more than total

Faryal Rumi | TNN | Updated: Jun 9, 2018, 16:33 IST

A- A+

Students checking their results on mobile phones after BSEB release of Intermediate results on website in Patn... Read More

HIGHLIGHTS

- The Bihar School Examination Board was again in the limelight when some class XII students claimed that they scored higher marks than the total.
- Some others complained that they received marks in papers they never appeared for.

PATNA: Two years after the infamous topper scam, the Bihar School Examination

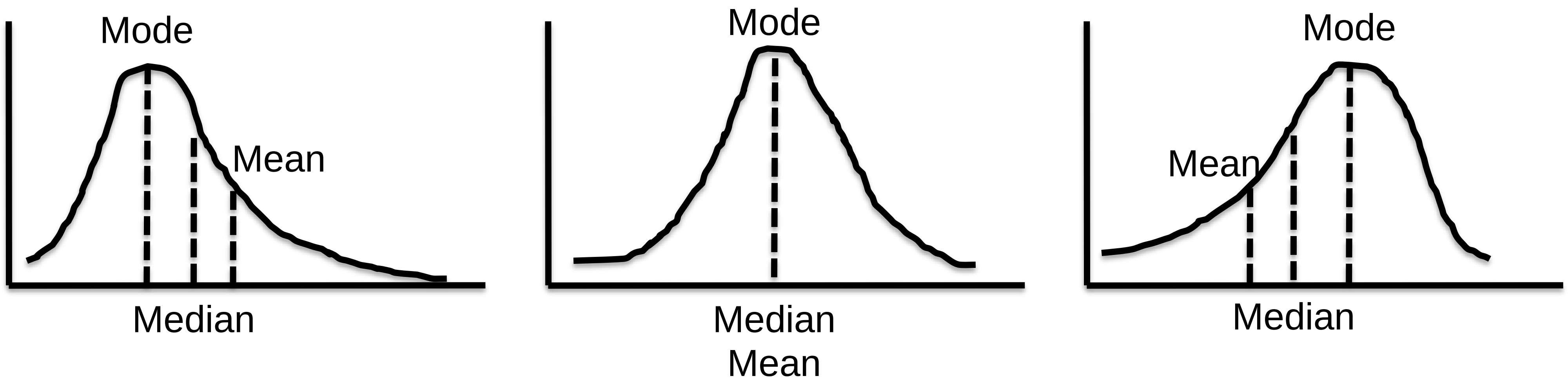
Data Types – Recent Interview Question

A sample of 400 Bangalore households is selected and several variables are recorded. Which of the following statements is correct?

- Socioeconomic status (recorded as “low income”, “middle income”, or “high income”) is nominal level data
- The number of people living in a household is a discrete variable
- The primary language spoken in the household is ordinal level data (recorded as “Kannada”, “Tamil”, etc)

The Central Tendencies

Identify where the MODE, MEDIAN and MEAN lie in the below distributions.



Measures of Spread – Recent Interview Question

The spread of the data in a dataset could be studied using ________

- Interquartile range
- Variance
- Standard Deviation
- Range (max-min)
- All of the above

Measures of Spread – Recent Interview Question

Given the numbers are 68, 83, 58, 84, 100, 64, the second quartile is:

- 74.5
- $\boxed{75.5}$
- 75
- 74

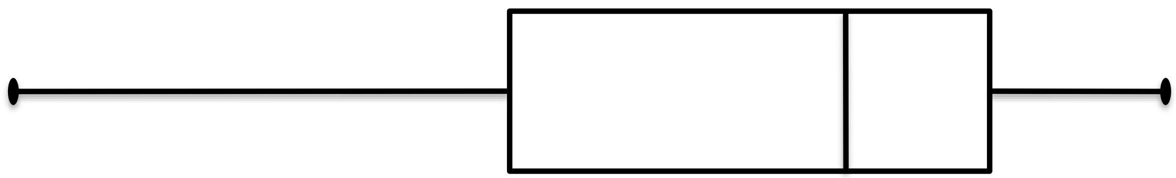
Measures of Spread – Recent Interview Question

Which of the following plot is used to analyze interquartile range

- Scatterplot
- Histogram
- Lineplot
- Boxplot
- All of the above

Measures of Spread – Recent Interview Question

What term would best describe the shape of the given boxplot?



- Symmetric
- Skewed with right tail
- Skewed with left tail
- All the above

Measures of Spread (Dispersion)

Just as Quartiles divide data into 4 equal parts, Deciles divide it into 10 equal parts and Percentiles into 100 equal parts.

Given the above, find the 25th, 50th, 75th and the 90th percentiles for the top 16 global marketing sectors for advertising spending for a recent year according to *Advertising Age*. Also, find Q2, 5th decile and IQR. Data in next slide.

Sector	Ad spending (in $ million)
Automotive	22195
Personal Care	19526
Entertainment and Media	9538
Food	7793
Drugs	7707
Electronics	4023
Soft Drinks	3916
Retail	3576
Restaurants	3553
Cleaners	3571
Computers	3247
Telephone	2448
Financial	2433
Beer, Wine and Liquor	2050
Candy	1137
Toys	699

Measures of Spread (Dispersion)

Sector	Ad spending (in $ million)
Automotive	22195
Personal Care	19526
Entertainment and Media	9538
Food	7793
Drugs	7707
Electronics	4023
Soft Drinks	3916
Retail	3576
Cleaners	3571
Restaurants	3553
Computers	3247
Telephone	2448
Financial	2433
Beer, Wine and Liquor	2050
Candy	1137

25^{th}Percentile = 25*(n+1)/100
25^{th}Percentile = 25*(16+1)/100
25^{th}Percentile = 4.25 (Between Financial & Telephone)
25^{th}Percentile = (2433+2448)/2
25^{th}Percentile (Q1) = 2440.5

50^{th} Percentile = 50*(n+1)/100
50^{th} Percentile = 50*(16+1)/100
50^{th} Percentile = 8.5 (Cleaners & Retail)
50^{th} Percentile = (3571+3576)/2
50^{th} Percentile (MEDIAN/Q2) = 3573.5

75^{th} Percentile (Q3) = 7750

90^{th} Percentile = 20860.5

5^{th}Decile(Median/Q2)=5*(n+1)/10=3573.5

IQR = Q3-Q1 = 7750 – 2440.5 = 5309.5

PROBABILITY BASICS

Probability - Types

BREAK

- Joint Probability
  - P(A and B) = P(A)*P(B)
- Union Probability
  - P (A or B) = P(A) + P(B) – P(A and B)
- Marginal Probability - Probability of a Single Attribute
  - Only one P(A), P(B)
- Conditional Probability

Probability - Types

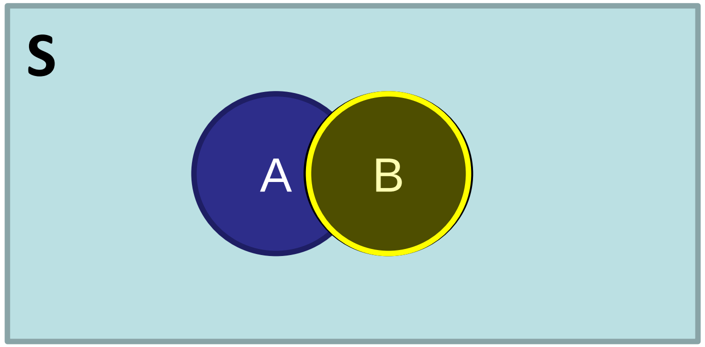
Conditional Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

Probability of *A* occurring **given that** *B* has occurred.

The sample space is restricted to a single row or column.

This makes rest of the sample space irrelevant.



Probability - Types

		Age			
		Young	Middle-aged	Old	Total
Loan Default					
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

Conditional Probability

What is the probability that a person will not default on the loan payment **given** she is middle-aged?

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	10,503	27,368	259	38,130
	Yes	3,586	4,851	120	8,557
	Total	14,089	32,219	379	46,687

P(No | Middle-Aged) = ?

Probability - Types

		Age			
		Young	Middle-aged	Old	Total
Loan Default					
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

$$\text{Conditional Probability} = P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	10,503	27,368	259	38,130
	Yes	3,586	4,851	120	8,557
	Total	14,089	32,219	379	46,687

Note that this is the ratio of **Joint Probability** to **Marginal Probability**

$$\text{P(No | Middle-Aged)} = \frac{P(\text{Middle aged and } NO)}{P(\text{Middle})} = \frac{0.586}{0.690} = 0.85$$

$$\text{P(No | Middle Aged)} = \frac{P(\text{Middle aged and } NO)}{P(\text{Middle})} = \frac{27368/46687}{32219/46687} = 0.85$$

Conditional Probability – Order Matters

		Age			Total
		Young	Middle-aged	Old	Total
Loan Default	No				
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

P(No | Middle-Aged) = 0.586/0.690 = 0.85

What is the probability that a person is middle-aged **given** she has not defaulted on the loan payment?

P(Middle-Aged | No) = 0.586/0.816 = 0.72 (<u>Order Matters</u>)

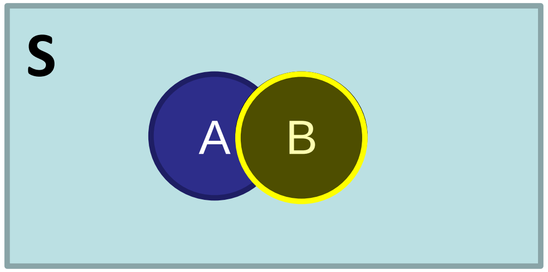
P(Middle-Aged | No) = 27368/38130 = 0.72 (<u>Order Matters</u>)

Probability - Types

Conditional Probability – Visualizing using Probability Tables and Venn Diagrams

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	10,503	27,368	259	38,130
	Yes	3,586	4,851	120	8,557
	Total	14,089	32,219	379	46,687

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

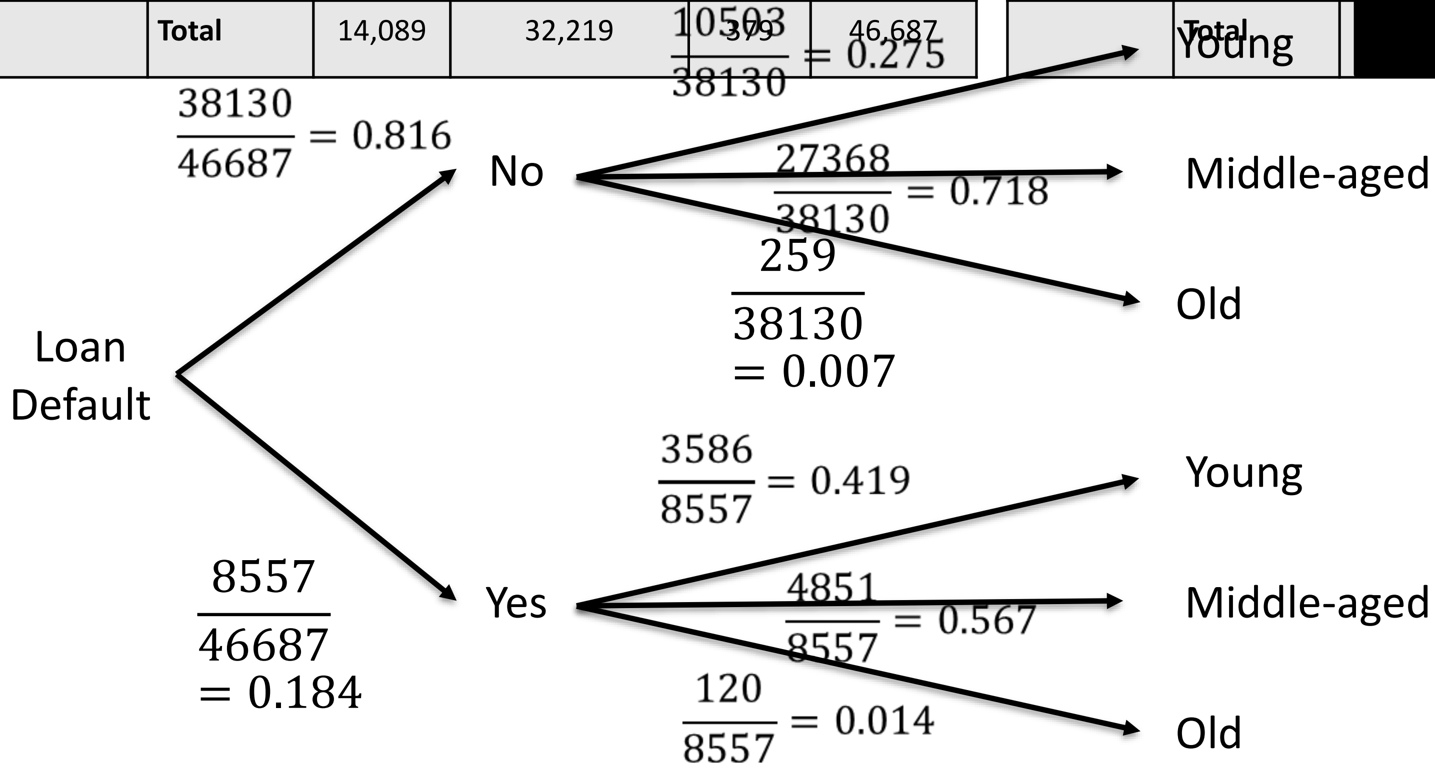


Probability - Types

Conditional Probability – Visualizing using Probability Trees

		Age (Numbers)			
		Young	Middle-aged	Old	Total
Loan Default	No	10,503	27,368	259	38,130
	Yes	3,586	4,851	120	8,557
	Total	14,089	32,219	379	46,687

		Age (Probabilities)			
		Young	Middle-aged	Old	Total
Loan Default	No	[redacted]	0.586	0.005	0.816
	Yes	[redacted]	0.104	0.003	0.184
	Total	[redacted]	0.690	0.008	1.000



Find
P(Young and No)

We know that

P(Young/No) = $\frac{P(Young\ and\ NO)}{P(No)}$

P(Young and No) = P(Young/No) * P(No)

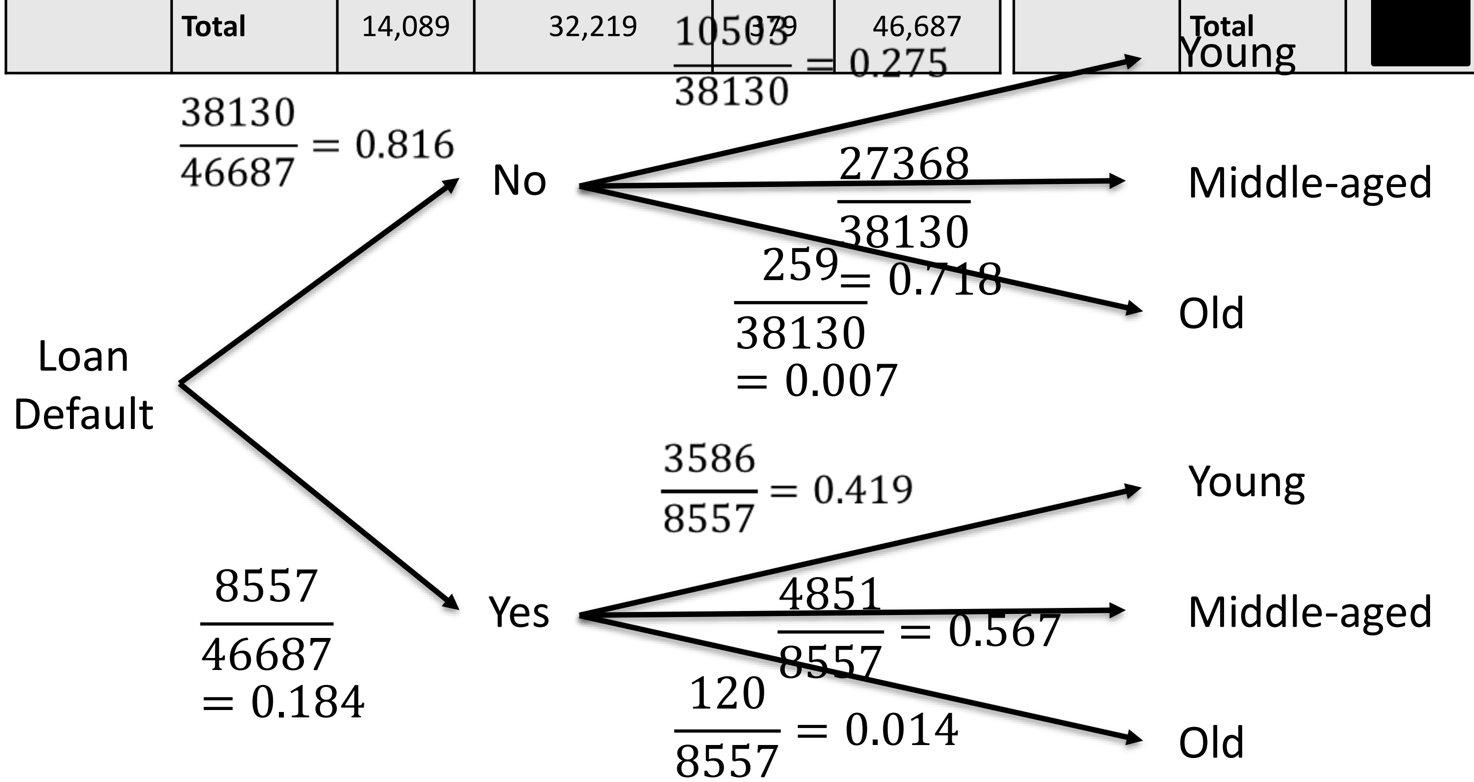
P(Young and No) = 0.275*0.816 = **0.225**

Probability - Types

Conditional Probability – Visualizing using Probability Trees

		Age (Numbers)			
		Young	Middle-aged	Old	Total
Loan Default	No	10,503	27,368	259	38,130
	Yes	3,586	4,851	120	8,557
	Total	14,089	32,219	379	46,687

		Age (Probabilities)			
		Young	Middle-aged	Old	Total
Loan Default	No	[redacted]	0.586	0.005	0.816
	Yes		0.104	0.003	0.184
	Total		0.690	0.008	1.000



Find

- P(Young and No)
- P(No and Young)
- P(Young) = P(No) P(Young/No) + P(Yes) *P(Young/Yes)
- P(No)
- P(Young | No)
- P(No | Young)

Probability - Types

Attention Check

Identify the type of probability in each of the below cases:

1. P(Old and Yes)
2. P(Yes and Old)
3. P(Old)
4. P(Yes)
5. P(Old | Yes)
6. P(Yes | Old)
7. P(Young | No)
8. P(Middle-aged or No)
9. P(Old or Young)

		Age (Probabilities)			
		Young	**Middle-aged**	**Old**	**Total**
Loan Default	**No**	0.225	0.586	0.005	**0.816**
	Yes	0.077	0.104	0.003	**0.184**
	Total	**0.302**	**0.690**	**0.008**	**1.000**

1 and 2: **Joint**; 3 and 4: **Marginal**; 5, 6 and 7: **Conditional**; 8 and 9: **Union**

Probability - Types

Conditional Probability

$$P(A|B) = \frac{P(A \; and \; B)}{P(B)} \Rightarrow P(A \; and \; B) = P(B) * P(A|B)$$

Similarly

What happens when A and B are INDEPENDENT?

$$P(B|A) = \frac{P(A \; and \; B)}{P(A)} \Rightarrow P(A \; and \; B) = P(A) * P(B|A)$$

Equating, we get

$$P(A|B) * P(B) = P(A) * P(B|A)$$

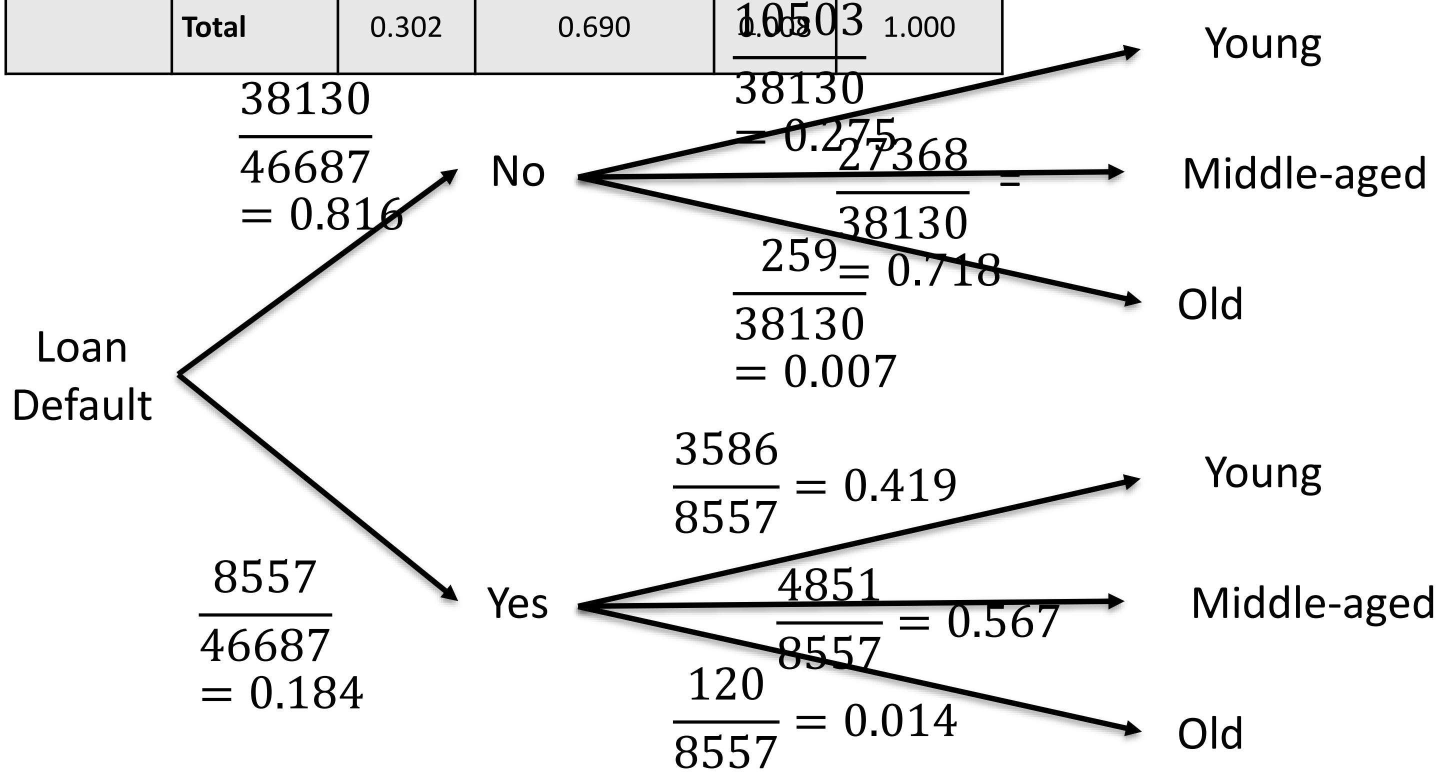
$$\boxed{\therefore P(A|B) = \frac{P(A) * P(B|A)}{P(B)}}$$

Probability - Types

Conditional Probability – Visualizing using Probability Trees

		Age (Probabilities)			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

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



Now find

P(No | Young)

=

$$\frac{\mathrm{P(No)} * \mathrm{P(Young|No)}}{\mathrm{P(Young)}}$$

=

$$\frac{0.816 * 0.275}{(0.275*0.816)+(0.419*0.184)}$$

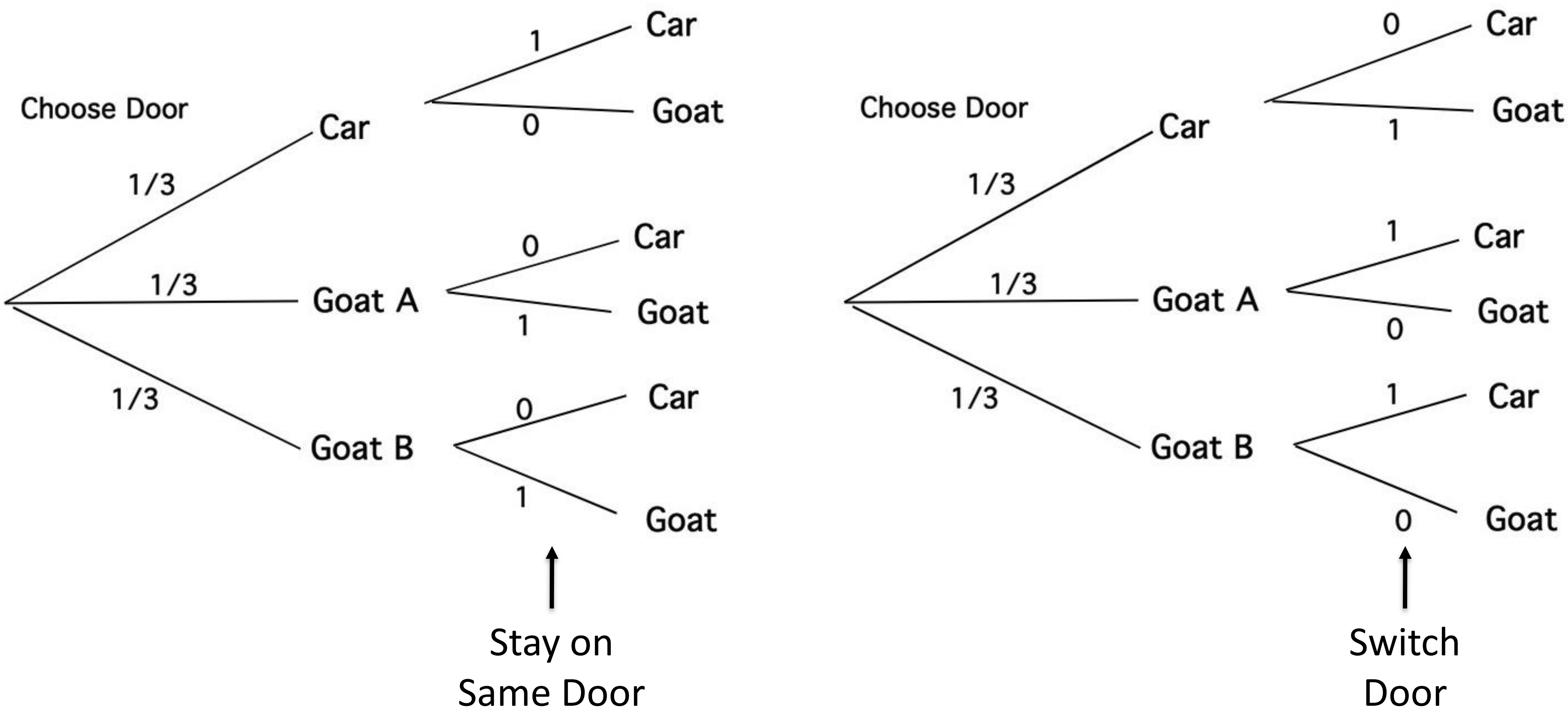
$=0.744$

Probability - Types

Monty Hall Problem - Intuitive

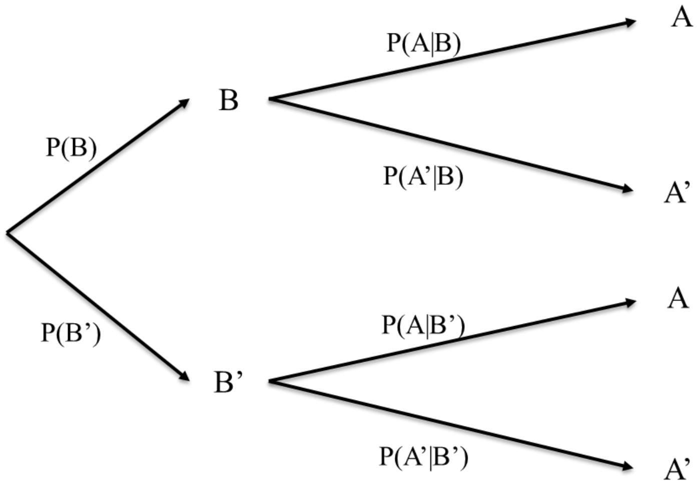
Probability - Types

Monty Hall Problem – Probability Tree



Probability - Types

Conditional Probability -> Bayes' Theorem



$P(B \text{ and } A) = P(B) * P(A|B)$

$P(B \text{ and } A') = P(B) * P(A'|B)$

$P(B' \text{ and } A) = P(B') * P(A|B')$

$P(B' \text{ and } A') = P(B') * P(A'|B')$

Note B' means "not B"

$$P(B|A) = \frac{P(B) * P(A|B)}{P(A)} = \frac{P(A|B) * P(B)}{P(A|B) * P(B) + P(A|not\ B) * P(not\ B)}$$

Bayes' Theorem

Bayes' Theorem allows you to find reverse probabilities, and to allow **revision of original probabilities** with new information.

Case – Clinical trials

Epidemiologists claim that probability of breast cancer among Caucasian women in their mid-50s is 0.005. An established test identified people who had breast cancer and those that were healthy. A new mammography test in clinical trials has a probability of 0.85 for detecting cancer correctly. In women without breast cancer, it has a chance of 0.925 for a negative result. If a 55-year-old Caucasian woman tests positive for breast cancer, what is the probability that she in fact has breast cancer?

P(Cancer) = 0.005
P(Test positive | Cancer) = 0.85
P(Test negative | No cancer) = 0.925
P(Cancer | Test positive) = ?

Bayes' Theorem

Case – Clinical trials

P(Cancer) = 0.005 (*aka* Prior Probability)

P(Test positive | Cancer) = 0.85 (*aka* Likelihood)

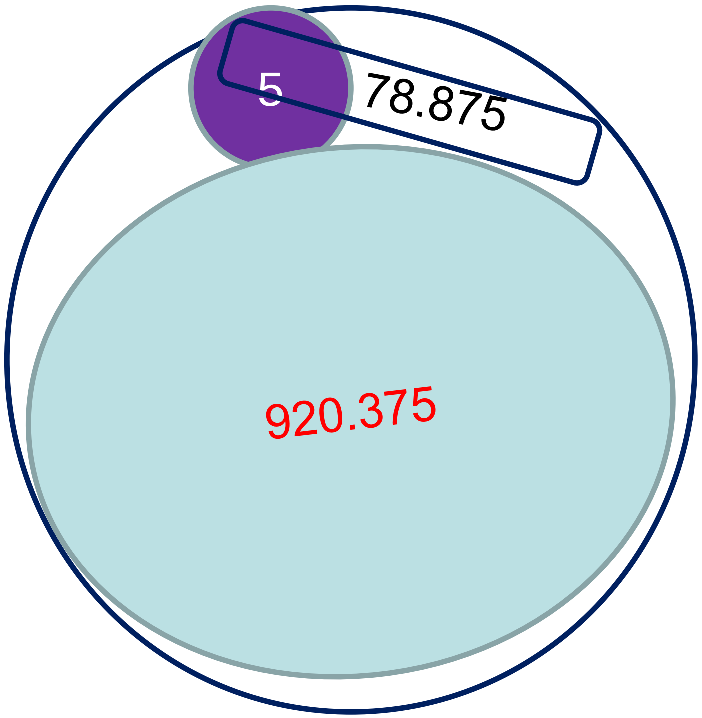
P(Test negative | No cancer) = 0.925

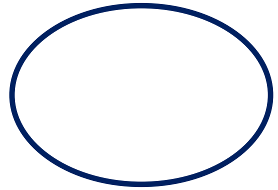
P(Cancer | Test positive) = ? (*aka* Posterior or Revised Probability)

P(Test Positive) *aka* Evidence

$$Posterior\ Probability = \frac{Prior\ Probability * Likelihood}{Evidence}$$

Cancer Detection – Bayes Theorem

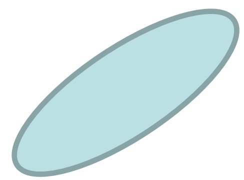




Let us assume 1000 women in the mid 50

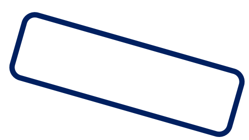
P(Cancer) = 0.005 or 5/1000. So for every 1000 women in their mid 50s, 5 get breast cancer

P(No Cancer) = 1- 0.005 = 0.995 = 995 women



P(test –ve / No cancer) = 0.925 = 92.5% of 995 = 920.375

P(test +ve / No cancer) = 74.625 (995 – 920.375)



P(Test +ve) = P(test +ve /Cancer) + P(test +ve/No Cancer)
0.85*5 + 74.625 = 78.875

Bayes' Theorem

Case – Clinical trials

P(Cancer) = 0.005 (*aka* Prior Probability)

P(Test positive | Cancer) = 0.85 (*aka* Likelihood)

P(Test negative | No cancer) = 0.925

P(Cancer | Test positive) = ? (*aka* Posterior or Revised Probability)

P(Test Positive) *aka* Evidence

$$P(Cancer|Test +) = \frac{P(Cancer) * P(Test +| Cancer)}{P(Test+)}$$

$$P(Cancer|Test +) = \frac{P(Cancer) * P(Test + |Cancer)}{P(Test + |Cancer) * P(Cancer) + P(Test + |No\ cancer) * P(No\ cancer)}$$

$$= \frac{0.005 * 0.85}{0.85 * 0.005 + 0.075 * 0.995} = \frac{0.00425}{0.078875} = 0.054$$

Homework

Draw a Probability Table and a Probability Tree for the above case.

Bayes' Theorem

Case – Spam filtering

Apache SpamAssassin™

Latest News

2015-04-30: SpamAssassin 3.4.1 has been released! Highlights include:

- improved automation to help combat spammers that are abusing new top level dc
- tweaks to the SPF support to block more spoofed emails;
- increased character set normalization to make rules easier to develop and stop sp
- continued refinement to the native IPv6 support; and
- improved Bayesian classification with better debugging and attachment hashing.

SpamAssassin works by having users train the system. It looks for patterns in the words in emails marked as spam by the user. For example, it may have learned that the word "free" appears in 20% of the mails marked as spam, i.e., P(Free | Spam) = 0.20. Assuming 0.1% of non-spam mail includes the word "free" and 50% of all mails received by the user are spam, find the probability that a mail is spam if the word "free" appears in it.

Bayes' Theorem

BREAK

Case – Spam filtering

P(Spam) = 0.50

P(Free | Spam) = 0.20

P(Free | No spam) = 0.001

P(Spam | Free) = ?

$$P(Spam|Free) = \frac{P(Spam) * P(Free|Spam)}{P(Free)}$$

$$P(Spam|Free) = \frac{P(Spam) * P(Free|Spam)}{P(Free|Spam) * P(Spam) + P(Free|No\ spam) * P(No\ spam)}$$

$$= \frac{0.5 * 0.2}{0.2 * 0.5 + 0.001 * 0.5} = \frac{0.1}{0.1005} = 0.995$$

This helps the spam filter automatically classify the messages as spam.

A slight detour

HOW GOOD IS YOUR CLASSIFICATION?

Confusion Matrix

Spam filtering		Predicted		Total
		Positive	Negative	
Actual	Positive	952	526	1478
	Negative	167	3025	3192
Total		1119	3551	4670

		Predicted		METRICS
		Positive	Negative	
Actual	Positive	True +ve	False –ve	Recall/Sensitivity/True Positive Rate (Minimize False –ve)
	Negative	False +ve	True –ve	Specificity/True Negative Rate (Minimize False +ve)
		Precision		Accuracy, F_1 score

Confusion Matrix - Metrics

		Predicted		
		Positive	Negative	
Actual	Positive	True +ve	False –ve	Recall/Sensitivity/True Positive Rate (Minimize False –ve)
	Negative	False +ve	True –ve	Specificity/True Negative Rate (Minimize False +ve)
		Precision		Accuracy, F_1 score

$$\text{Recall (Sensitivity)} = \frac{\text{True+ve}}{\text{Actual+ve}}$$

$$\text{Recall (Sensitivity)} = \frac{\text{True+ve}}{\text{True+ve + False−ve}}$$

$$\text{Specificity} = \frac{\text{True −ve}}{\text{Actual−ve}}$$

$$\text{Specificity} = \frac{\text{True −ve}}{\text{False+ve +True −ve}}$$

$$\text{Precision} = \frac{\text{True +ve}}{\text{Predicted +ve}}$$

$$\text{Precision} = \frac{\text{True+ve}}{\text{True+ve + False+ve}}$$

$$\text{Accuracy} = \frac{\text{True +ve + True−ve}}{\text{Total}}$$

$$\text{Accuracy} = \frac{\text{True+ve + True −ve}}{\text{True+ve + False−ve + False+ve + True −ve}}$$

$$F_1\ \text{Score} = 2^* \frac{\text{Precision} * \text{Recall}}{\text{Precision+Recall}}$$

Confusion Matrix

Spam filtering		Predicted		Total	
		Positive	Negative		
Actual	Positive	952	526	1478	Recall(Sensitivity)
	Negative	167	3025	3192	Specificity
Total		1119	3551	4670	
		Precision			Accuracy, F1 Score

$$Recall(Sensitivity) = \frac{952}{1478} = 0.644$$

$$Specificity = \frac{3025}{3025 + 167} = \frac{3025}{3192} = 0.948$$

$$Precision = \frac{952}{1119} = 0.851$$

$$Accuracy = \frac{952 + 3025}{952 + 3025 + 526 + 167} = \frac{3977}{4670} = 0.852$$

Which measure(s) is/are more important?

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} = \frac{2 * 0.851 * 0.644}{0.851 + 0.644} = \frac{1.096}{1.495} = 0.733$$

Confusion Matrix

Which measure(s) is/are more important?

Court System – Death Sentence		**Verdict**		
		Guilty	Not Guilty	
Actual	Guilty	True +ve	False –ve	Recall/Sensitivity/True Positive Rate (Minimize False –ve)
	Not Guilty	False +ve	True –ve	Specificity/True Negative Rate (Minimize False +ve)
		Precision		Accuracy, F_1 score

Confusion Matrix

Breast cancer detection		Predicted		Total
		Positive	Negative	
Actual	Positive	852	126	978
	Negative	67	1025	1092
Total		919	1151	2070

$$Recall\ (Sensitivity) = \frac{852}{978} = 0.871$$

$$Precision = \frac{852}{919} = 0.927$$

$$Accuracy = \frac{852 + 1025}{852 + 1025 + 126 + 67} = \frac{1877}{2070} = 0.907$$

$$Specificity = \frac{1025}{1025 + 67} = \frac{1025}{1092} = 0.939$$

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} = \frac{2 * 0.871 * 0.927}{0.871 + 0.927} = \frac{1.615}{1.798} = 0.898$$

Which measure(s) is/are more important?

Confusion Matrix

Which measure(s) is/are more important?

Anti Virus Detection		Detection		
		Virus	No Virus	
Actual	Virus	True +ve	False –ve	Recall/Sensitivity/True Positive Rate (Minimize False –ve)
	No Virus	False +ve	True –ve	Specificity/True Negative Rate (Minimize False +ve)
		Precision		Accuracy, F_1 score

Confusion Matrix

Which measure(s) is/are more important?

Organ Matching from Donors		Predicted		
		Match	No Match	
Actual	Match	True +ve	False –ve	Recall/Sensitivity/True Positive Rate (Minimize False –ve)
	No Match	False +ve	True –ve	Specificity/True Negative Rate (Minimize False +ve)
		Precision		Accuracy, F_1 score

Confusion Matrix

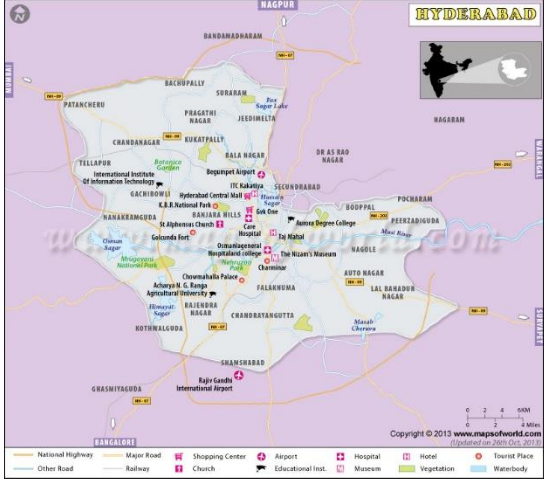
Which measure(s) is/are more important?

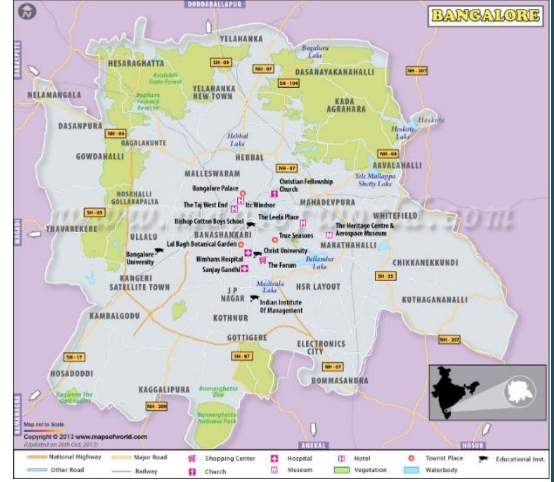
Credit Card Fraud Detection		Detection		
		Fraud	No Fraud	
Actual	Fraud	True +ve	False –ve	Recall/Sensitivity/True Positive Rate (Minimize False –ve)
	No Fraud	False +ve	True –ve	Specificity/True Negative Rate (Minimize False +ve)
		Precision		Accuracy, F_1 score

Confusion Matrix

Which measure(s) is/are more important?

Image Text Classification		**Predicted Word**		
		CAT	DOG	
Actual Word	CAT	True +ve	False –ve	Recall/Sensitivity/True Positive Rate (Minimize False –ve)
	DOG	False +ve	True –ve	Specificity/True Negative Rate (Minimize False +ve)
		Precision		Accuracy, F_1 score





Inspire...Educate...Transform.

HYDERABAD
2nd Floor, Jyothi Imperial, Vamsiram Builders, Old Mumbai Highway, Gachibowli, Hyderabad - 500 032
+91-9701685511 (Individuals)
+91-9618483483 (Corporates)

BENGALURU
Floors 1-3, L77, 15th Cross Road, 3A Main Road, Sector 6, HSR Layout, Bengaluru – 560 102
+91-9502334561 (Individuals)
+91-9502799088 (Corporates)

Social Media

Web:	http://www.insofe.edu.in
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	http://www.youtube.com/InsofeVideos
SlideShare:	http://www.slideshare.net/INSOFE
LinkedIn:	http://www.linkedin.com/company/international-school-of-engineering

This presentation may contain references to findings of various reports available in the public domain. INSOFE makes no representation as to their accuracy or that the organization subscribes to those findings.