"One day Workshop with Hands-on Training on Machine Learning and Its Applications"

Dr. D. Asir Antony Gnana Singh., B.E.,M.E., M.B.A., Ph.D.,
Department of CSE, Anna University,
BIT Campus, Tiruchirappalli-620 0024

at

UCE, Anna University,
BIT Campus, Tiruchirappalli-620 0024

on

28th September 2019

26.1

# TOPICS FOR DISCUSSION

1. Introduction
2. Knowledge discovery from data
3. Human and Machine learning
4. Developing models
5. Performance evaluation
6. Performance improvement


Dr. D. Asir Antony Gnana Singh, CSE,UCE,AU,BIT-CAMPUS

2

# Introduction

- **Learning**
  - The cognitive process of acquiring skill or knowledge

Dr. D. Asir Antony Gnana Singh, CSE,UCE,AU,BIT-CAMPUS

3

# Introduction

- **What is Knowledge?**
  - **The psychological result of perception, learning and reasoning**
  - **Facts, information, and skills acquired through experience or education; the theoretical or practical understanding of a subject.**
  - **Ability to**
    - Recognize
    - Predict
    - Classify
    - Detect
    - Identify the things or objects, etc. using the attributes

Dr. D. Asir Antony Gnana Singh, CSE,UCE,AU,BIT-CAMPUS

4

# Introduction

- Why Knowledge?
  - Used for making decisions to solve the problems (satisfying the needs)
- Decision making
  - Plays a significant role in day-to-day life
  - In every second the human mind makes decision
    - » Do I want to listen this seminar?
    - » What subject should I read?
    - » When should I reach to the college?
    - » What dress should I wear for today?
    - » Whether watching movie is good or bad?



# Introduction

- The good decision maker
  - Can be successful person
  - Can be a good worker
  - Can be a good leader
  - Can be a good manager
  - Can be seated in higher position



# Knowledge Discovery from Data



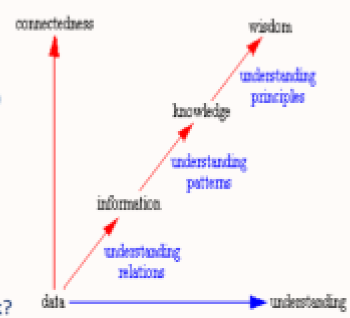

# Knowledge Discovery from Data

**Data:** (symbols, raw facts)
Book, $, is, 1, price

**Information:** (processed data)
Price of 1 book is 50$
(pattern/rule)

**Knowledge**: I have 50 $ therefore I can buy only one book (making decision)

**Wisdom:** Who wrote this book? Whether it is worth full or not? Is it very essential for me?



## Knowledge Discovery from Data

- **Data**
  - Raw facts
- **Types of data (based of forms)**
  - Image (E.g.  )
  - Text (E.g. Apple)
  - Audio (E.g.  )
  - Video
    - (Sequence of images) with or without audio 
  - Multimedia



## Knowledge Discovery from Data

- **Types of data** (Statistical perspective)
  - Numerical data
    - Discrete data
      - E.g. (Height, Width, Length)
    - Continuous data
      - E.g. Temperature
  - Categorical data
    - E.g. ({Yes/no}, { Average, Good, Excellent})
  - Ordinal data
    - Mixes numerical and categorical data



## Knowledge Discovery from Data

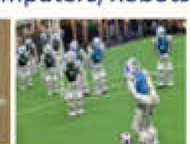
- **Knowledge discovery**
  - Knowledge is discovered through learning the data
- **Learning**
  - The cognitive process of acquiring skill or knowledge
  - Learning is a process of developing/ updating and training the models from data



## Learning

- Human learning (Human)


- Machine learning (Computers/Robots)


## Human Learning

- **Types of learners based on comfort with data**
  - Visual learner (visual) (listen the demonstration/ reading the book)
  - Auditory learner (listening the audio)
  - Kinesthetic learner (Experience)
- **Phases of human learning**
  - Learning/Training phase
  - Testing phase
  - Performance evaluation phase



## Human Learning

- **Learning/ Training phase**
  - Learning the name of the animals



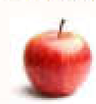

## Human Learning

- **Learning/ Training phase**
  - Learning the name of the fruits




## Human Learning

- **Testing phase**
  - Prediction/detection/classification/recognition



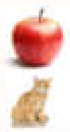
**? Apple**

**? Cat**

**Attributes :** colour, size, shape , texture

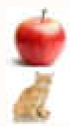
# Human Learning

- **Performance evaluation phase**
  - Student1 recognizes the objects as follows:




Accuracy = 2/2 *100=100%

  - Student2 recognizes the objects as follows:




Accuracy = 1/2 *100=50%



# Machine Learning

- **Machine learning**
  - Machine learning is a field of computer science and a field of artificial intelligence
  - Enables the computers/machines/ robots to learn the data for acquiring knowledge and make data-driven decisions



# Applications of Machine Learning

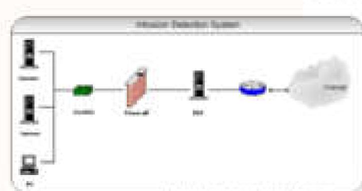
- **Applications of machine learning** (based on terminology)
  - Recognition (E.g. Face or object recognition)
  - Prediction (E.g. Weather prediction)
  - Classification (E.g. Document classification)
  - Detection (E.g. Intrusion detection)



# Applications of Machine Learning

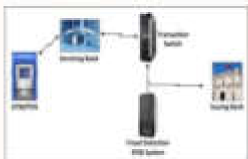
- **Computer networks**
  - Antivirus
  - Intrusion detection

F-PROT AVG McAfee PANDA SOPHOS ESET KASPERSKY



## Applications of Machine Learning

- **Banking**
  - Loan approval prediction
  - Fraudulent detection
  - Currency recognition (Fake currency identification)

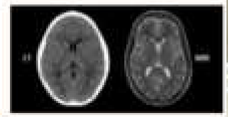




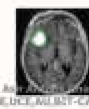
## Applications of Machine Learning

- **Medical**
  - Disease diagnosis

Computed tomography (CT or CAT) scan



Magnetic resonance imaging (MRI) SCAN





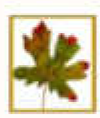
## Applications of Machine Learning

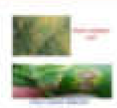
- **Recruitment**
  - Automated job suggestion for recruitment



## Applications of Machine Learning

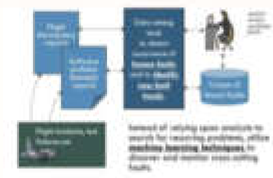
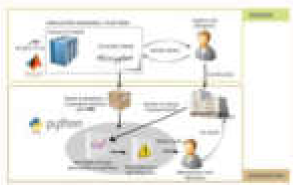
- **Agriculture**
  - Plant disease detection
  - Leaf disease
  - Soil classification
  - Soil fertility prediction



## Applications of Machine Learning

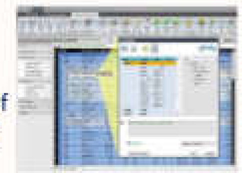
- **Software engineering**
  - Software fault detection



## Applications of Machine Learning

- **Construction industry**
  - Automated cost estimation
  - Analyzing the performance of suppliers and sub-contractors
  - Estimation of materials (cement, sand, water, bricks, etc.)
  - Estimation of resources such as fuels, electricity, etc.





## Applications of Machine Learning

- **Advertisement**
  - Advertisement in blog sites and emails (Google AdSense)



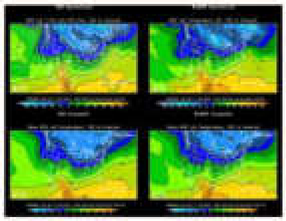

## Applications of Machine Learning

- **Commercial**
  - Buyer behavior analysis
  - Market basket analysis



## Applications of Machine Learning

- **Metrological**
  - Weather prediction
    - Cyclone
    - Storm
    - Rain
  - Weather forecasting





## Applications of Machine Learning

- **Social media**
  - Friends suggestion
  - Categorize subscribers

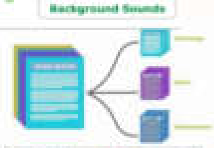



## Other Applications of Machine Learning

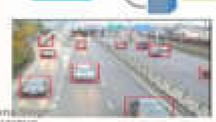
- Face recognition

- Speech recognition


- Document classification



- Vehicle detection





## Phases of Machine Learning

- Learning/Training phase
- Testing phase
- Performance evaluation phase

# Phases of Machine Learning

- **Machine learning algorithm**
  - The machine learning algorithm receives the numerical or categorical data to develop the model
  - The data is in various forms such as image, audio, text and multimedia
  - Hence, the different data acquisition devices are used to acquire the data from the real world
  - Moreover, different techniques are used to convert the data into machine understandable data such as numerical and categorical



# Phases of Machine Learning

- **Data acquisition devices**

| Sensory organ | Sense | Data | Data Acquisition Devices |
|---|---|---|---|
| **Eyes** | Visual | Image | Cameras |
| **Ears** | Hearing | Audio | Voice recorder |
| **Nose** | Small | data | Gas sensor |
| **Tongue** | Taste | Data | Chemical sensor |
| **Skin** | Touch | Data | Temperature/wind/humidity sensors, etc. |



# Phases of Machine Learning

- **Technique for data conversion** (Numerical/categorical)

| Data | Technique for Conversion | Application Example |
|---|---|---|
| Image | Digital image processing | Object recognition |
| Text | Data processing | Text classification |
| Audio | Digital signal processing | Signal classification |



# Phases of Machine Learning

- **Representation of dataset**

| Attributes | | | | | Class/Label |
|---|---|---|---|---|---|
| $X_1$ | $X_2$ | $X_3$ | ... | $X_N$ | C |
| $x_{11}$ | $x_{21}$ | $x_{31}$ | ... | $x_{N1}$ | $c_1$ |
| $x_{12}$ | $x_{22}$ | $x_{32}$ | ... | $x_{N2}$ | $c_2$ |
| $x_{13}$ | $x_{23}$ | $x_{33}$ | ... | $x_{N3}$ | $c_1$ |
| . | . | . | | . | . |
| . | . | . | ... | . | . |
| . | . | . | | . | . |
| $x_{1M}$ | $x_{2M}$ | $x_{3M}$ | ... | $x_{NM}$ | $c_2$ |

## Phases of Machine Learning

- **Models**
  - Model is also called as **Pattern**
  - The models are used to
    - Recognize
    - Predict
    - Classify
    - Detect the label of the unknown data.



## Phases of Machine Learning

- **Example of Model**

Mathematical Model of a Straight Line

**y = mX +b**

Where X – Axis variable
m – Slope
Y – Axis variable
b – Intercept

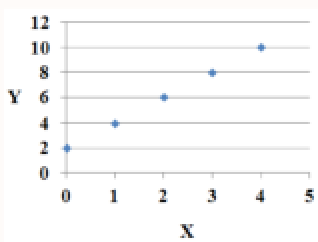


## Phases of Machine Learning

- **Example of Model**

Predict the value of variable Y from model **Y=mX+b**
where **m=2, b=2**

X=0 Y=2*0+2=2
X=1 Y=2*1+2=4
X=2 Y=2*2+2=6
X=3 Y=2*3+2=8
X=4 Y=2*4+2=10
X=5 Y=?






## Overview of Knowledge Discovery from Data

- **Database**
  - Collection of data from various sources
- **Data selection**
  - Select the needed data from the database
  - Transformation process can be carried out
- **Dataset**
  - Structured numerical or categorical data
- **Data preprocessing**
  - Data cleaning (to remove noise and inconsistent data)

# Overview of Knowledge Discovery from Data

- **Preprocessed dataset**
  - The preprocessed dataset is obtained after preprocessing
- **Machine learning algorithm**
  - Learn the data and build the pattern/model
- **Patterns/model**
  - Statistical or mathematical model is used for prediction/ classification /detection /recognition
- **Evaluation of the model**
  - Testing the model to obtain the knowledge (predicted data /unknown data for known data) for decision making



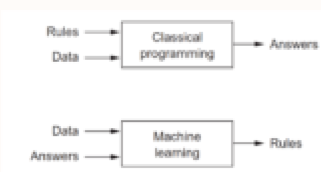
# Classification of Machine Learning Algorithms

- **Unsupervised learning**
  - Learning with unlabeled data Eg. Clustering algorithm
- **Supervised learning**
  - Learning with labeled data Eg. Classification algorithms.
- **Types of Supervised learning**
  - **Decision Tree based**
    - J48 Java implementation decision Tree
  - **Probability based**
    - Naive Bayes (NB)
  - **Instant based**
    - IB1 (instance based learning)



# Classical Programming vs Machine learning

- Ex. Classical
  - Data: a=10; b=10;
  - Rule : c=a+b;




- Ex: Machine learning
  - Dataset:
    - data +answer (label)



# Example of Prediction

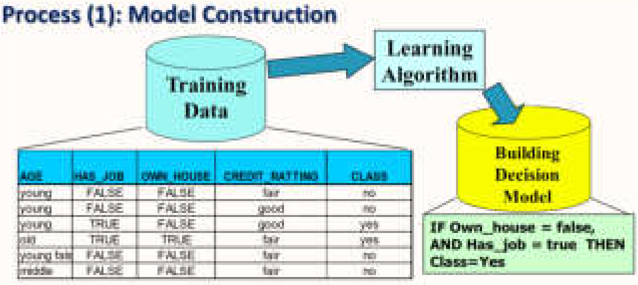
**Dataset :** Historical Loan data of Bank

**Decision :** The Bank Manager has to predict which person shall able to repay the loan amount.

**Historical Loan data form the Bank Enterprise**

| Customer | AGE | HAS_JOB | OWN_HOUSE | CREDIT_RATTING | (Loan repay) CLASS |
|---|---|---|---|---|---|
| 1 | young | FALSE | FALSE | fair | no |
| 2 | young | FALSE | FALSE | good | no |
| 3 | young | TRUE | FALSE | good | yes |
| 4 | old | TRUE | TRUE | fair | yes |
| 5 | young | FALSE | FALSE | fair | no |
| 6 | middle | FALSE | FALSE | fair | no |

## Process (1): Model Construction




| AGE | HAS_JOB | OWN_HOUSE | CREDIT_RATTING | CLASS |
|---|---|---|---|---|
| young | FALSE | FALSE | fair | no |
| young | FALSE | FALSE | good | no |
| young | TRUE | FALSE | good | yes |
| old | TRUE | TRUE | fair | yes |
| young b/w | FALSE | FALSE | fair | no |
| middle | FALSE | FALSE | fair | no |

## Process (2): Prediction of unknown data


| AGE | HAS_JOB | OWN_HOUSE | CREDIT_RATTING |
|---|---|---|---|
| young | FALSE | FALSE | fair |



## Demonstration

With
Decision Tree based Supervised Learner

PRACTICAL SESSION



## Construction and testing the model using decision tree based supervised learner



## Dataset

Class-labeled training tuples from the *AllElectronics* customer database.

| *RID* | *age* | *income* | *student* | *credit_rating* | *Class: buys_computer* |
|---|---|---|---|---|---|
| 1 | youth | high | no | fair | no |
| 2 | youth | high | no | excellent | no |
| 3 | middle_aged | high | no | fair | yes |
| 4 | senior | medium | no | fair | yes |
| 5 | senior | low | yes | fair | yes |
| 6 | senior | low | yes | excellent | no |
| 7 | middle_aged | low | yes | excellent | yes |
| 8 | youth | medium | no | fair | no |
| 9 | youth | low | yes | fair | yes |
| 10 | senior | medium | yes | fair | yes |
| 11 | youth | medium | yes | excellent | yes |
| 12 | middle_aged | medium | no | excellent | yes |
| 13 | middle_aged | high | yes | fair | yes |
| 14 | senior | medium | no | excellent | no |

## Decision tree Construction by Information Gain

- Select the attribute with the highest information gain
- Let $p_i$ be the probability that an arbitrary tuple in D belongs to class $C_i$, estimated by $|C_{i,D}|/|D|$
- **Expected information** (entropy) needed to classify a Tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- **Information needed** (after using A to split D into v partitions) to classify D:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$$

- **Information gained** by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$



Class-labeled training tuples from the *AllElectronics* customer database.

| RID | age | income | student | credit_rating | Class: buys_computer |
|---|---|---|---|---|---|
| 1 | youth | high | no | fair | $\boxed{\text{no}}$ |
| 2 | youth | high | no | excellent | $\boxed{\text{no}}$ |
| 3 | middle_aged | high | no | fair | yes |
| 4 | senior | medium | no | fair | yes |
| 5 | senior | low | yes | fair | yes |
| 6 | senior | low | yes | excellent | no |
| 7 | middle_aged | low | yes | excellent | yes |
| 8 | youth | medium | no | fair | $\boxed{\text{no}}$ |
| 9 | youth | low | yes | fair | $\boxed{\text{yes}}$ |
| 10 | senior | medium | yes | fair | yes |
| 11 | youth | medium | yes | excellent | $\boxed{\text{yes}}$ |
| 12 | middle_aged | medium | no | excellent | yes |
| 13 | middle_aged | high | yes | fair | yes |
| 14 | senior | medium | no | excellent | no |



Yes No

youth

$$Info_{age}(D) = \frac{5}{14} \times \left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right)$$

Middle-aged

$$+\frac{4}{14} \times \left(-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}\right)$$

Senior

$$+\frac{5}{14} \times \left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right)$$

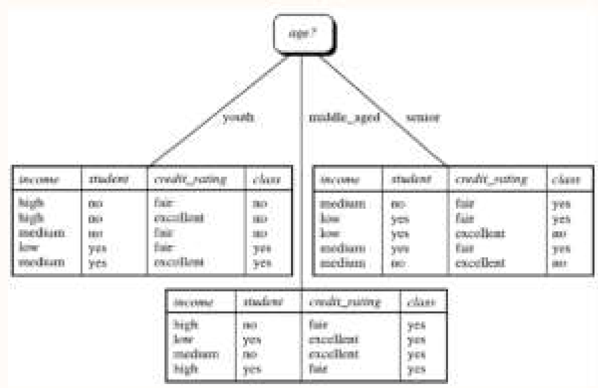
$$= 0.694 \text{ bits.}$$

$$Info(D) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940 \text{ bits.}$$

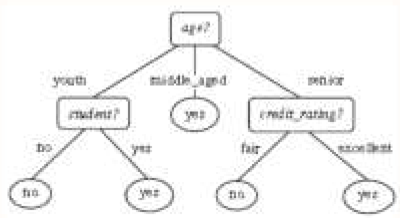
$$Gain(age) = Info(D) - Info_{age}(D) = 0.940 - 0.694 = 0.246 \text{ bits.}$$

Similarly, we can compute $Gain(income) = 0.029$ bits, $Gain(student) = 0.151$ bits, and $Gain(credit\_rating) = 0.048$ bits. Because age has the highest information gain among





## Now you can Make Decision by this Decision Tree




| Age | Income | Student | Credit rating | Class |
|---|---|---|---|---|
| Youth | High | No | Fair | ? |
| Youth | High | No | Excellent | ? |
| Middle aged | High | No | Fair | ? |

**Unlabelled data** **Demo**



## Construction and testing the probabilistic model using naive Bayes based supervised learner



## Naive Bayes Model Construction

**Bayes' Theorem:** is a way of finding a probability when we know certain other probabilities.

The formula is:

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

P(A|B) : how often A happens given that B happens
P(B|A) : how often B happens given that A happens,
P(A) : how likely A is on its own
P(B) : how likely B is on its own

**Remembering**
First think "AB AB AB" then remember to group it like: "AB = A BA / B"



Let us say

**P(Fire)** means how often there is fire
**P(Smoke)** means how often we see smoke

**P(Fire|Smoke)** means how often there is fire when we can see smoke
**P(Smoke|Fire)** means how often we can see smoke when there is fire

So the formula kind of tells us "forwards" P(Fire|Smoke) when we know "backwards" P(Smoke|Fire)

Example: If dangerous fires are rare (1%) but smoke is fairly common (10%) due to barbecues, and 90% of dangerous fires make smoke then:

$$P(Fire|Smoke) = \frac{P(Fire)\,P(Smoke|Fire)}{P(Smoke)} = \frac{1\% \times 90\%}{10\%} = 9\%$$

So the "Probability of dangerous Fire when there is Smoke" is 9%

## Dataset

Class-labeled training tuples from the *AllElectronics* customer database.

| RID | *age* | *income* | *student* | *credit_rating* | *Class: buys_computer* |
|---|---|---|---|---|---|
| 1 | youth | high | no | fair | no |
| 2 | youth | high | no | excellent | no |
| 3 | middle_aged | high | no | fair | yes |
| 4 | senior | medium | no | fair | yes |
| 5 | senior | low | yes | fair | yes |
| 6 | senior | low | yes | excellent | no |
| 7 | middle_aged | low | yes | excellent | yes |
| 8 | youth | medium | no | fair | no |
| 9 | youth | low | yes | fair | yes |
| 10 | senior | medium | yes | fair | yes |
| 11 | youth | medium | yes | excellent | yes |
| 12 | middle_aged | medium | no | excellent | yes |
| 13 | middle_aged | high | yes | fair | yes |
| 14 | senior | medium | no | excellent | no |



P(X|H) is the posterior probability of X conditioned on H.

That is, it is the probability that customers, X, is 35 years old and earns $40,000, given that we know the customer will buy a computer.
P(X) is the prior probability of X. Using our example, it is the probability that a person from our set of customers is 35 years old and earns $40,000.

"How are these probabilities estimated?"

P(H), P(X|H), and P(X) may be estimated from the given data, as we shall see next.

Bayes' theorem is useful in that it provides a way of calculating the posterior probability, P(H|X), from P(H), P(X|H), and P(X).

Bayes' theorem is $P(H|X) = \frac{P(X|H)\,P(H)}{P(X)}$



Class-labeled training tuples from the *AllElectronics* customer database.

| RID | *age* | *income* | *student* | *credit_rating* | *Class: buys_computer* |
|---|---|---|---|---|---|
| 1 | youth | high | no | fair | no |
| 2 | youth | high | no | excellent | no |
| 3 | middle_aged | high | no | fair | yes |
| 4 | senior | medium | no | fair | yes |
| 5 | senior | low | yes | fair | yes |
| 6 | senior | low | yes | excellent | no |
| 7 | middle_aged | low | yes | excellent | yes |
| 8 | youth | medium | no | fair | no |
| 9 | youth | low | yes | fair | yes |
| 10 | senior | medium | yes | fair | yes |
| 11 | youth | medium | yes | excellent | yes |
| 12 | middle_aged | medium | no | excellent | yes |
| 13 | middle_aged | high | yes | fair | yes |
| 14 | senior | medium | no | excellent | no |



Predict the person X will buy the computer or not ?

X = (age = youth, income = medium, student = yes, credit rating = fair)

That is, the naive Bayesian classifier predicts that tuple X belongs to the class $C_i$ if and only if

$$P(C_i|X) > P(C_j|X) \text{ for } 1 \le j \le m,\ j \ne i.$$

Thus, we maximize $P(C_i|X)$.

The class $C_i$ for which $P(C_i|X)$ is maximized is called the maximum posteriori hypothesis.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

$$= P(x_1|C_i) \times P(x_2|C_i) \times \cdots \times P(x_n|C_i).$$

X = (age = youth, income = medium, student = yes, credit rating = fair)

We need to maximize $P(X|C_i)P(C_i)$, for i = 1, 2. $P(C_i)$, the prior probability of each class, can be computed based on the training tuples:

P(buys computer = yes) = 9/14 = 0.643
P(buys computer = no) = 5/14 = 0.357



To compute $P(X|C_i)$, fori = 1, 2, we compute the following conditional probabilities:
P(age = youth | buys computer = yes) = 2/9 = 0.222
P(age = youth | buys computer = no) = 3/5 = 0.600
P(income = medium | buys computer = yes) = 4/9 = 0.444
P(income = medium | buys computer = no) = 2/5 = 0.400
P(student = yes | buys computer = yes) = 6/9 = 0.667
P(student = yes | buys computer = no) = 1/5 = 0.200
P(credit rating = fair | buys computer = yes) = 6/9 = 0.667
P(credit rating = fair | buys computer = no) = 2/5 = 0.400

Using these probabilities, we obtain

P(X|buys computer = yes) =
P(age = youth | buys computer = yes) ×
P(income = medium | buys computer = yes) ×
P(student = yes | buys computer = yes) ×
P(credit rating = fair | buys computer = yes)
= 0.222 × 0.444 × 0.667 × 0.667 = 0.044.
Similarly, P(X|buys computer = no) =
0.600 × 0.400 × 0.200 × 0.400 = 0.019



To find the class, $C_i$ , that maximizes $P(X|C_i)P(C_i)$,

We compute

P(X|buys computer = yes)P(buys computer = yes) = 0.044 × 0.643 = 0.028
P(X|buys computer = no)P(buys computer = no) = 0.019 × 0.357 = 0.007

Therefore,

The person X will buy the computer.



# Performance Evaluation

## Performance Evaluation Techniques

- The performance of the model is evaluated using test dataset
- **Types of dataset**
  - Training dataset
    - For model development
  - Testing dataset
    - For testing performance of the model



## Performance Evaluation Techniques

- **Performance evaluation techniques**
  - Classified based on how the training dataset is supplied for performance evaluation
    1. Use training dataset as test dataset
    2. Supplied test dataset
    3. N-Fold cross - validation
    4. Percentage split
- **Performance evolution metrics**
  - Accuracy
  - Precision
  - Recall
  - ROC (Receiver operating characteristics)



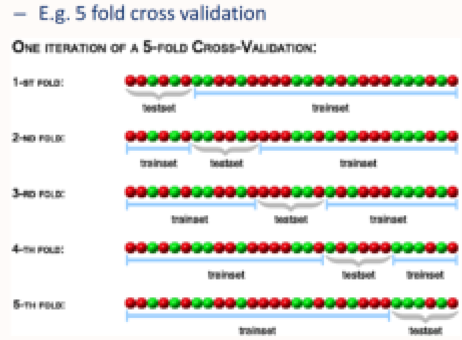
## Performance Evaluation Techniques

- **Use training dataset as test dataset**
  - The entire training dataset is given as the test dataset
- **Supplied test dataset**
  - The separate test dataset is given for testing



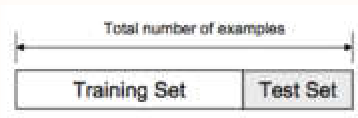
## Performance Evaluation Techniques

- **N-Fold-Cross- Validation**
  - E.g. 5 fold cross validation



## Performance Evaluation Techniques

- **Percentage split**
  - A portion of the whole dataset is splitted (in %) for test dataset
  - E.g.






## Performance Evaluation Metrics

- Accuracy
- Precession
- Recall
- True positive rate (TPR)
- False positive rate (FPR)
- ROC (Receiver operating characteristics)



## Performance Evaluation Metrics

- **Confusion matrix**
  - Confusion matrix (Positive /Negative)

| | | Predicted class | |
|---|---|---|---|
| | | Positive | Negative |
| Actual class | Positive | TP | FN |
| | Negative | FP | TN |

**True positive (TP):** Number of correctly predicted positive instances.

**False negative (FN):** Number of incorrectly predicted positive instances

**False positive (FP):** Number of incorrectly predicted negative instances.

**True negative (TN):** Number of correctly predicted negative instances.



## Performance Evaluation Metrics

- Accuracy

  Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ x 100

- Precision

  Precision = $\frac{TP}{TP+FP}$ x 100

- Recall

  Precision = $\frac{TP}{TP+FP}$ x 100

## Performance Evaluation Metrics

- True positive rate (TPR)

$$TPR = \frac{TP}{TP+FN}$$

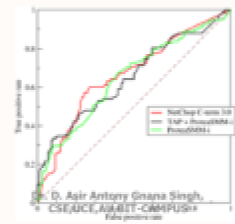
- False positive rate (FPR)

$$FPR = \frac{FP}{FP+TN}$$



## Performance Evaluation Metrics

- **Receiver operating characteristic**
  - The ROC curve is created by plotting the **true positive rate (TPR)** against the **false positive rate (FPR)** at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection in machine learning.





## Performance Evaluation Metrics

- Example of accuracy calculation for Diabetes dataset

=== Confusion Matrix ===

a b <-- classified as

407 93 | a = tested_negative = 500

108 160 | b = tested_positive = 268

Total instances 768

((407+160)/768)X100= 73.8281 %



## Performance Improvement

- **Attribute selection**
  - Also called as:
    - Attribute selection
    - Dimensionality reduction
    - Variable selection
    - Feature selection
  - Removing the redundant attribute
  - Removing the irrelevant attributes that are not participating in prediction

## Performance Improvement

- **Types of attribute selection**
  - Filter
    - E.g. Correlation, information gain, chi-squared
  - Wrapper
  - Embedded