Assignment

Course Title: Data Mining & Knowledge Discovery

Course Code: IT-5110

Institute of Information Technology,
Jahangirnagar University,
Savar, Dhaka.

Submitted To:
Md. Fazlul Karim Patwary
Professor
IIT, JU.

Submitted by:
Mahmuda Khan Moon
Roll:1068
Batch: 5th

1.Define the term “KNN” classification. Write two limitations of this classification.
Ans:

<u>“KNN” classification:</u>

KNN algorithm is one of the simplest classification algorithm and it is one of the most used learning algorithms. **KNN** is a **non-parametric, lazy** learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point.

In *k-NN classification*, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.

KNN classifier-

- Does not build models explicitly
- Unlike eager learners such as decision tree induction and rule-based systems
- Classifying unknown records are relatively expensive

<u>Two limitations of KNN classification:</u>

1. The main disadvantage of the KNN algorithm is that it is a *lazy learner*, i.e. it does not learn anything from the training data and simply uses the training data itself for classification, which can result in the algorithm not generalizing well and also not being robust to noisy data.
2. To predict the label of a new instance the KNN algorithm will find the *K* closest neighbors to the new instance from the training data, the predicted class label will then be set as the most common label among the *K* closest neighboring points. The main disadvantage of this approach is that the algorithm must compute the distance and sort all the training data at each prediction, which can be slow if there are a large number of training examples. Further, changing *K* can change the resulting predicted class label.

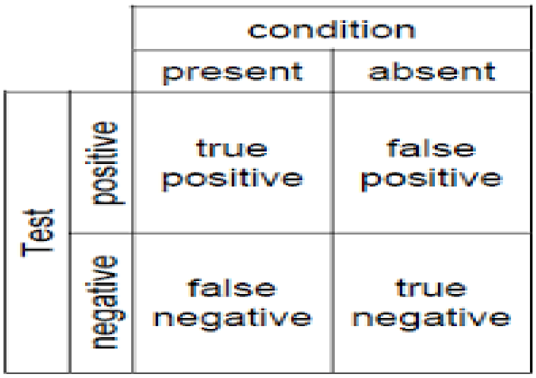
2.Define the term True Positive and False Negative.
Ans:

<u>True Positive:</u>

A true positive test result is one that detects the condition when the condition is present.

<u>False Negative:</u>

A false negative test result is one that does not detect the condition when the condition is present.



		condition	
		present	absent
Test	positive	true positive	false positive
	negative	false negative	true negative

3.Define the terms: accuracy, recall, F-measures and precision.

Ans:

Accuracy:

Accuracy refers to the closeness of a measured value to a standard or known value. For example, if in lab you obtain a weight measurement of 3.2 kg for a given substance, but the actual or known weight is 10 kg, then your measurement is not accurate. In this case, your measurement is not close to the known value.

Precision:

Precision refers to the closeness of two or more measurements to each other. Using the example above, if you weigh a given substance five times, and get 3.2 kg each time, then your measurement is very precise. Precision is independent of accuracy. You can be very precise but inaccurate, as described above. You can also be accurate but imprecise. For example, if on average, your measurements for a given substance are close to the known value, but the measurements are far from each other, then you have accuracy without precision.

In pattern recognition, information retrieval and binary classification, **precision** (also called positive predictive value) is the fraction of relevant instances among the retrieved instances. For example, suppose a computer program for recognizing dogs in photographs identifies 8 dogs in a picture containing 12 dogs and some cats. Of the 8 identified as dogs, 5 actually are dogs (true positives), while the rest are cats (false positives). The program's precision is 5/8

Recall:

Recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. Suppose a computer program for recognizing dogs in photographs identifies 8 dogs in a picture containing 12 dogs and some cats. Of the 8 identified as dogs, 5 actually are dogs (true positives), while the rest are cats (false positives). The program's recall is 5/12.

F-measures:

The F measure (F1 score or F score) is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test, where an F_1 score reaches its best value at 1 (perfect precision and recall) and worst at 0. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results returned by the classifier, and r is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).

4.Define tree based and rule based classification.

Ans:

Tree based classification:

Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modeling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.

Rule based classification:

The term rule-based classification can be used to refer to any classification scheme that make use of IF-THEN rules for class prediction. We can express a rule in the following from −

IF condition THEN conclusion

Let us consider a rule R1,

R1: IF age = youth AND student = yes
THEN buy_computer = yes

- The IF part of the rule is called **rule antecedent** or **precondition**.
- The THEN part of the rule is called **rule consequent**.
- The antecedent part the condition consist of one or more attribute tests and these tests are logically ANDed.
- The consequent part consists of class prediction.

5.Write the process of classification of any one.

Ans: The process of decision tree based classification is given below:

There are several steps involved in the building of a decision tree.

Splitting: The process of partitioning the data set into subsets. Splits are formed on a particular variable

Pruning: The shortening of branches of the tree. Pruning is the process of reducing the size of the tree by turning some branch nodes into leaf nodes, and removing the leaf nodes under the original branch. Pruning is useful because classification trees may fit the training data well, but may do a poor job of classifying new values. A simpler tree often avoids over-fitting.A pruned tree has less nodes and has less sparsity than a unpruned decision tree.

Tree Selection: The process of finding the smallest tree that fits the data. Usually this is the tree that yields the lowest cross-validated error.

Key Factors:

1. **Entropy**

   A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogeneous). ID 3 algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one.

2. **Information Gain**

   The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain (i.e., the most homogeneous branches).

Steps Involved

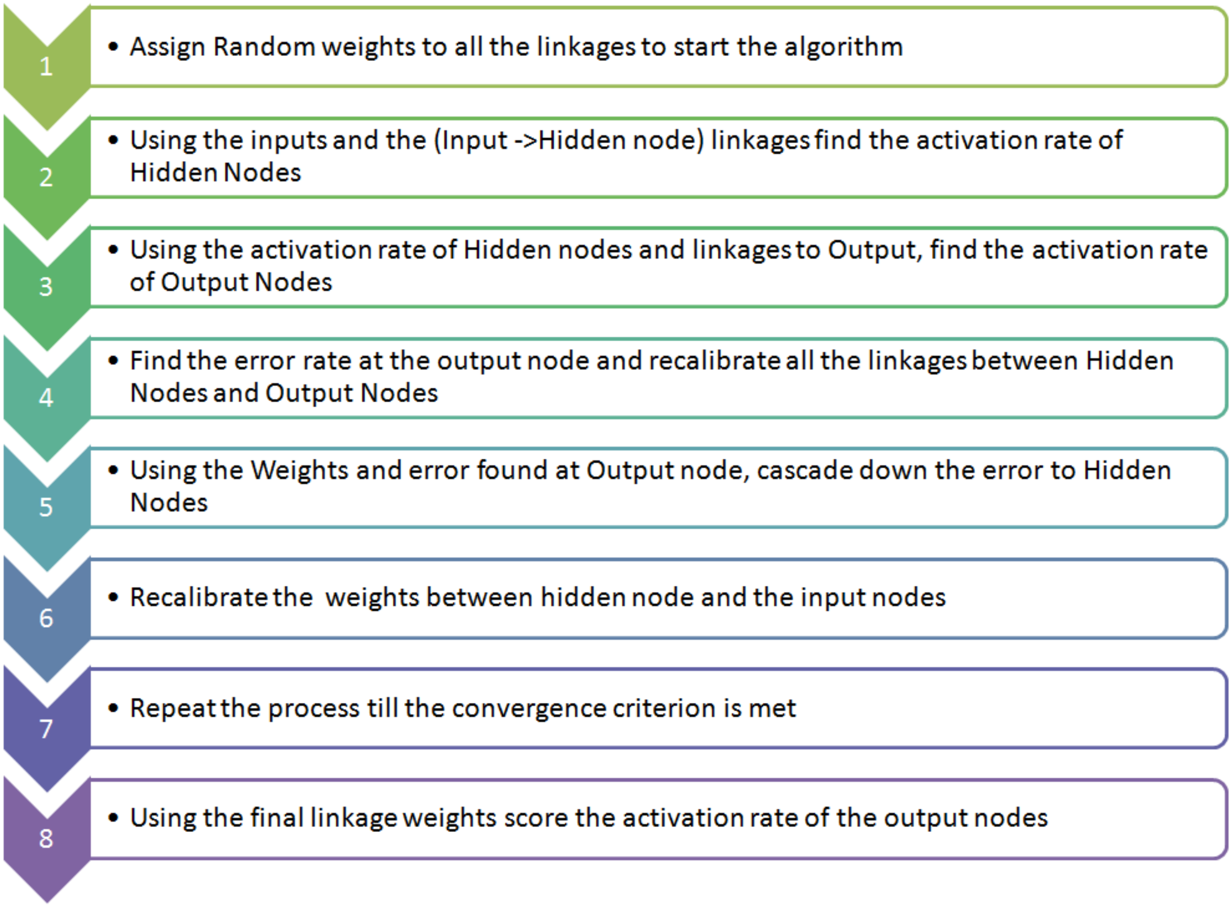
Step 1: Calculate entropy of the target.

Step 2: The dataset is then split on the different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain, or decrease in entropy.

Step 3: Choose attribute with the largest information gain as the decision node, divide the dataset by its branches and repeat the same process on every branch.

6. How ANN classifier works?

Ans: Following is the framework in which artificial neural networks (ANN) classifier works:



1

- Assign Random weights to all the linkages to start the algorithm

2

- Using the inputs and the (Input ->Hidden node) linkages find the activation rate of Hidden Nodes

3

- Using the activation rate of Hidden nodes and linkages to Output, find the activation rate of Output Nodes

4

- Find the error rate at the output node and recalibrate all the linkages between Hidden Nodes and Output Nodes

5

- Using the Weights and error found at Output node, cascade down the error to Hidden Nodes

6

- Recalibrate the weights between hidden node and the input nodes

7

- Repeat the process till the convergence criterion is met

8

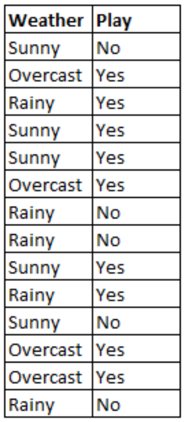
- Using the final linkage weights score the activation rate of the output nodes

7. How Bayes classifier works?

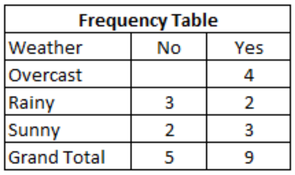
Ans: Following steps with a set of weather data to explain how Bayes classifiers works:

Step 1: Convert the data set into a frequency table

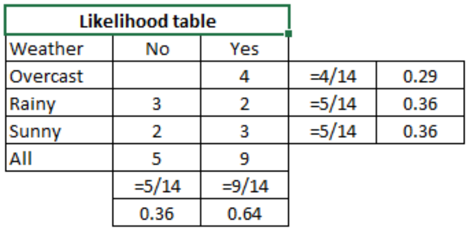
Step 2: Create Likelihood table by finding the probabilities like Overcast probability = 0.29 and probability of playing is 0.64.



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



Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9



Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

<u>Step 3:</u> Now, use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

Problem: Players will play if weather is sunny. Is this statement is correct?

We can solve it using above discussed method of posterior probability.

P(Yes | Sunny) = P(Sunny | Yes) * P(Yes) / P (Sunny)

Here we have P (Sunny |Yes) = 3/9 = 0.33, P(Sunny) = 5/14 = 0.36, P(Yes)= 9/14 = 0.64

Now, P (Yes | Sunny) = 0.33 * 0.64 / 0.36 = 0.60, which has higher probability.

Naive Bayes uses a similar method to predict the probability of different class based on various attributes. This algorithm is mostly used in text classification and with problems having multiple classes.

8. How do you perform KNN? Write the limitations of KNN.

Ans:

<u>To perform KNN:</u>

Requires three things

- The set of stored records
- Distance Metric to compute distance between records
- The value of *k*, the number of nearest neighbors to retrieve

<u>To classify an unknown record:</u>

- Compute distance to other training records
- Identify *k* nearest neighbors
- Use class labels of nearest neighbors to determine the class label of unknown record (e.g., by taking majority vote)

1. Compute distance between two points:

Euclidean distance $$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

2. Determine the class from nearest neighbor list

take the majority vote of class labels among the k-nearest neighbors

Weigh the vote according to distance

weight factor, w = 1/d2

3. Choosing the value of k:

If k is too small, sensitive to noise points

If k is too large, neighborhood may include points from other classes

Limitations of KNN:

- k-NN classifiers are lazy learners
- It does not build models explicitly
- Unlike eager learners such as decision tree induction and rule-based systems
- Classifying unknown records are relatively expensive
- It does not learn anything from the training data and simply uses the training data itself for classification.
- To predict the label of a new instance the KNN algorithm will find the *K* closest neighbors to the new instance from the training data, the predicted class label will then be set as the most common label among the *K* closest neighboring points.
- The main disadvantage of this approach is that the algorithm must compute the distance and sort all the training data at each prediction, which can be slow if there are a large number of training examples.
- Another disadvantage of this approach is that the algorithm does not learn anything from the training data, which can result in the algorithm not generalizing well and also not being robust to noisy data. Further, changing *K* can change the resulting predicted class label.

9. How do you validate a classification model?

Ans: Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

The general procedure is as follows:

1. Shuffle the dataset randomly.
2. Split the dataset into k groups
3. For each unique group:
    1. Take the group as a hold out or test data set
    2. Take the remaining groups as a training data set
    3. Fit a model on the training set and evaluate it on the test set
    4. Retain the evaluation score and discard the model
4. Summarize the skill of the model using the sample of model evaluation scores

Importantly, each observation in the data sample is assigned to an individual group and stays in that group for the duration of the procedure. This means that each sample is given the opportunity to be used in the hold out set 1 time and used to train the model k-1 times.

10. What are the functions of ROC curve in validation?

Ans:

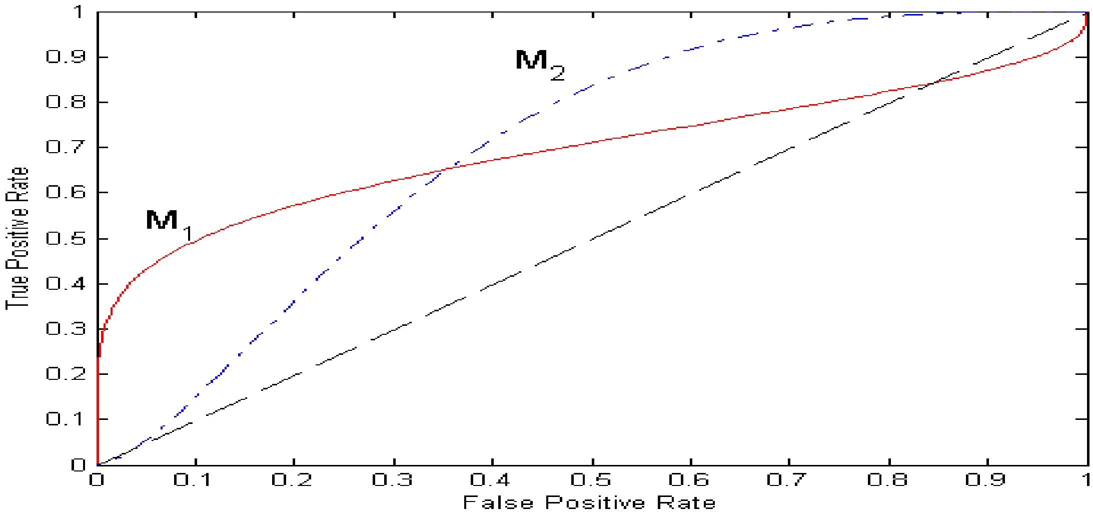
<u>The functions of ROC curve in validation:</u>

Characterize the trade-off between positive hits and false alarms

ROC curve plots TP (on the y-axis) against FP (on the x-axis)

Performance of each classifier represented as a point on the ROC curve

Changing the threshold of algorithm, sample distribution or cost matrix changes the location of the point



No model consistently outperform the other

M_1 is better for small FPR

M_2 is better for large FP

Area Under the ROC curve

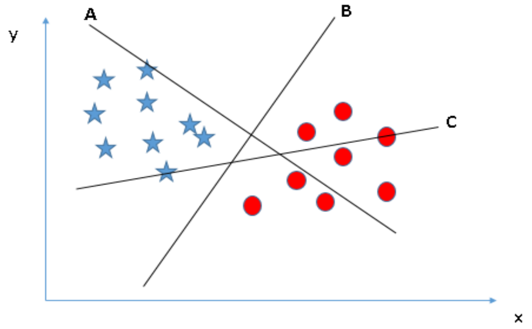
Ideal: Area = 1

Random guess: Area = 0.5

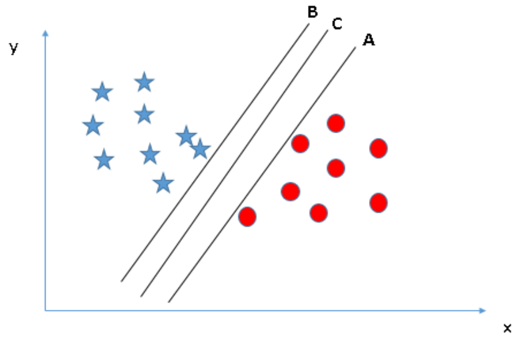
11. How SVM classifier works?

Ans:

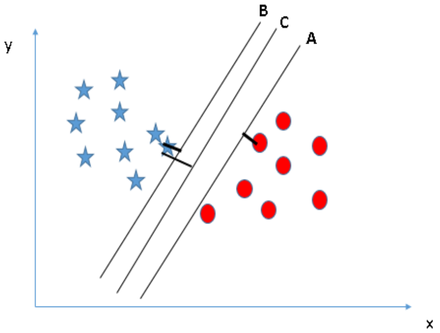
- **Identify the right hyper-plane (Scenario-1):** Here, we have three hyper-planes (A, B and C). Now, identify the right hyper-plane to classify star and circle.



- You need to remember a thumb rule to identify the right hyper-plane: “Select the hyper-plane which segregates the two classes better”. In this scenario, hyper-plane “B” has excellently performed this job.
- **Identify the right hyper-plane (Scenario-2):** Here, we have three hyper-planes (A, B and C) and all are segregating the classes well.

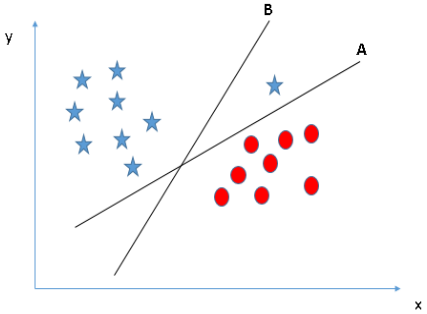


Here, maximizing the distances between nearest data point (either class) and hyper-plane will help us to decide the right hyper-plane. This distance is called as **Margin**. Let’s look



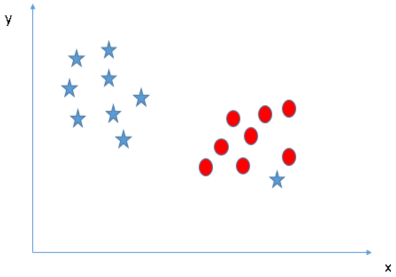
Above, you can see that the margin for hyper-plane C is high as compared to both A and B. Hence, we name the right hyper-plane as C. Another lightning reason for selecting the hyper-plane with higher margin is robustness. If we select a hyper-plane having low margin then there is high chance of miss-classification.

- **Identify the right hyper-plane (Scenario-3):** Hint: Use the rules as discussed in previous section to identify the right hyper-plane

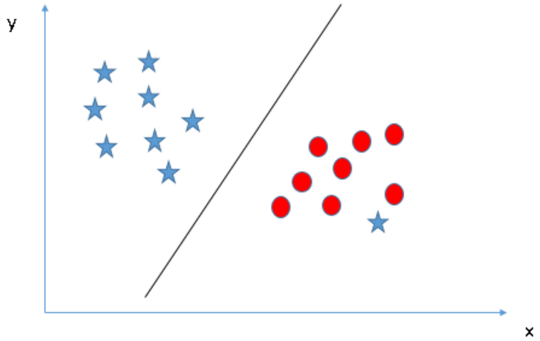


Some of you may have selected the hyper-plane **B** as it has higher margin compared to **A.** But, here is the catch, SVM selects the hyper-plane which classifies the classes accurately prior to maximizing margin. Here, hyper-plane B has a classification error and A has classified all correctly. Therefore, the right hyper-plane is **A.**

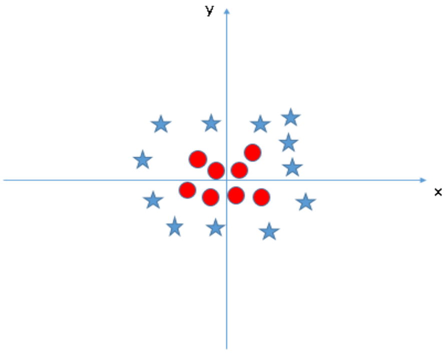
- **Can we classify two classes (Scenario-4)?:** Below, I am unable to segregate the two classes using a straight line, as one of star lies in the territory of other(circle) class as an outlier.



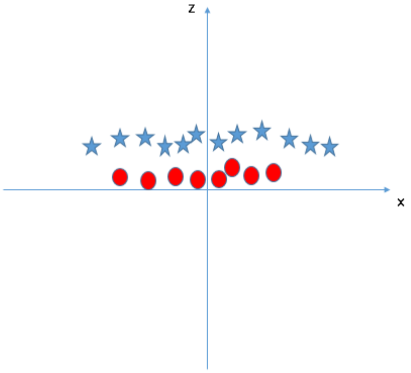
- As I have already mentioned, one star at other end is like an outlier for star class. SVM has a feature to ignore outliers and find the hyper-plane that has maximum margin. Hence, we can say, SVM is robust to outliers.



- **Find the hyper-plane to segregate to classes (Scenario-5):** In the scenario below, we can't have linear hyper-plane between the two classes, so how does SVM classify these two classes? Till now, we have only looked at the linear hyper-plane.



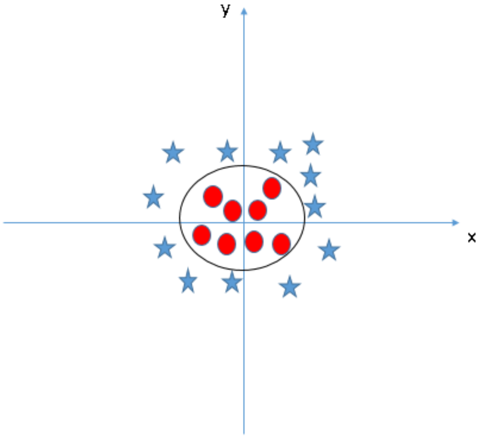
- SVM can solve this problem. Easily! It solves this problem by introducing additional feature. Here, we will add a new feature z=x^2+y^2. Now, let's plot the data points on axis x and z:



In above plot, points to consider are:

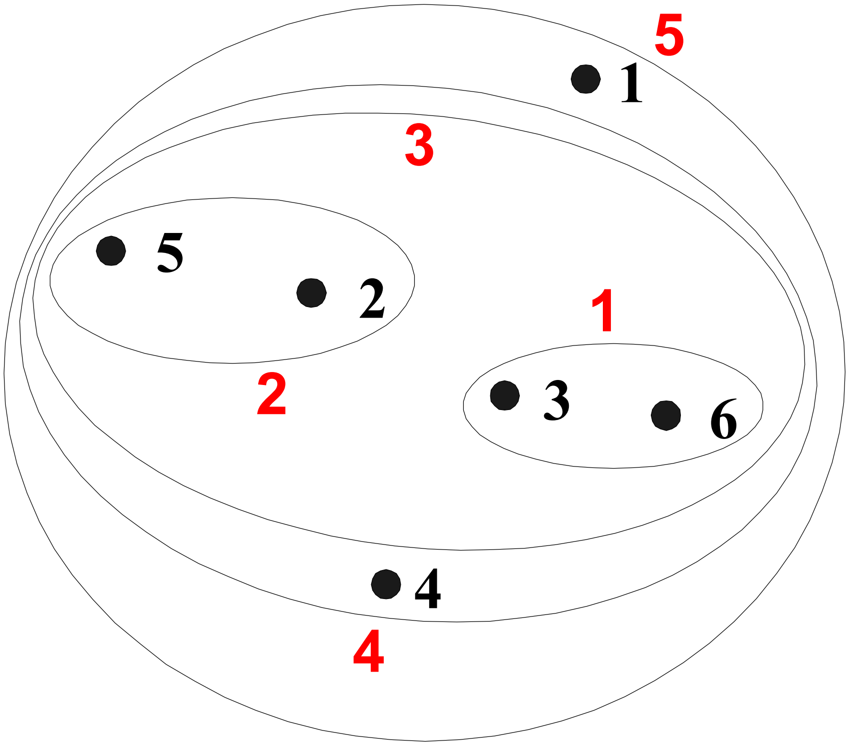
- All values for z would be positive always because z is the squared sum of both x and y
- In the original plot, red circles appear close to the origin of x and y axes, leading to lower value of z and star relatively away from the origin result to higher value of z.

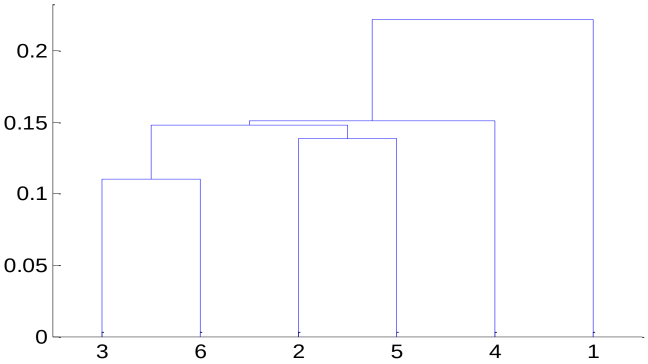
When we look at the hyper-plane in original input space it looks like a circle:



12. Math on Draw dendrogram for hierarchical clustering.

Ans: Dendrogram for hierarchical clustering





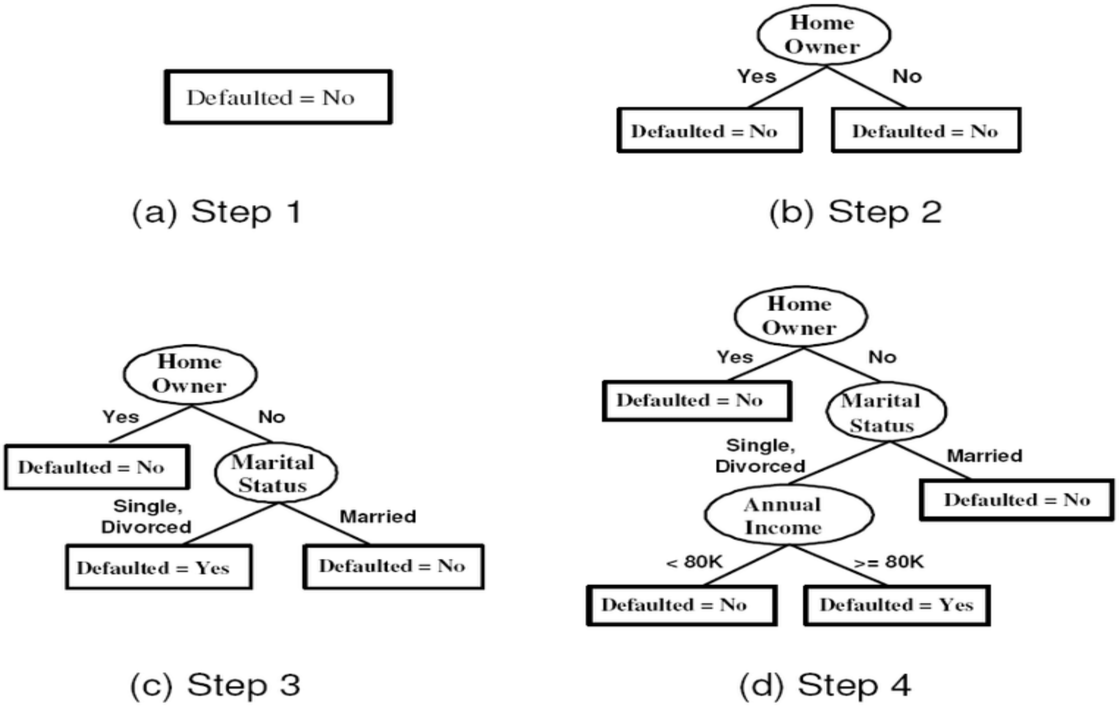
13. If you have a data set with class attribute and a new data without class attribute. You want to predict the value of class attribute of new data. Write the process of classifying this new data using decision tree classification (Hunt's Algorithm).

Ans:

Let D_t be the set of training records that reach a node t

General Procedure:

- If D_t contains records that belong the same class y_t, then t is a leaf node labeled as y_t
- If D_t is an empty set, then t is a leaf node labeled by the default class, y_d
- If D_t contains records that belong to more than one class, use an attribute test to split the data into smaller subsets. Recursively apply the procedure to each subset.



(a) Step 1

(b) Step 2

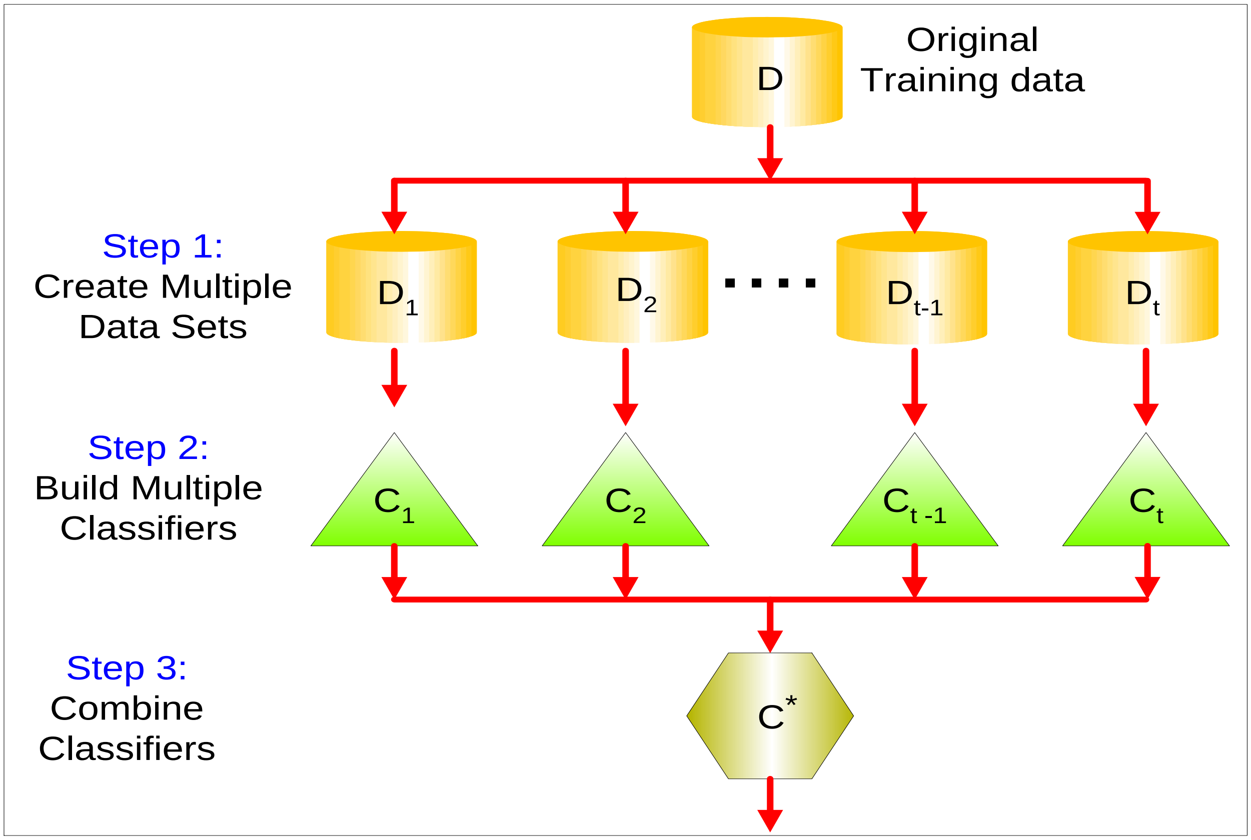
(c) Step 3

(d) Step 4

14. If you have a data set with class attribute and a new data without class attribute. Write the process of classifying this new data using ensemble method.

Ans:

- □ Construct a set of classifiers from the training data
- □ Predict class label of previously unseen records by aggregating predictions made by multiple classifiers



15. If you have a data set with class attribute and a new data without class attribute. Write the process of classifying this new data using KNN.

Ans:

In the classification setting, the K-nearest neighbor algorithm essentially boils down to forming a majority vote between the K most similar instances to a given "unseen" observation. Similarity is defined according to a distance metric between two data points. A popular choice is the Euclidean distance given by

$$d(x,x')=\sqrt{(\ (x1-x'1)^2+(x2-x'2)^2+\ldots+(xn-x'n)^2\ )}$$

but other measures can be more suitable for a given setting and include the Manhattan, Chebyshev and Hamming distance.

More formally, given a positive integer K, an unseen observation *x* and a similarity metric *d*, KNN classifier performs the following two steps:

- It runs through the whole dataset computing *d* between *x* and each training observation. We'll call the K points in the training data that are closest to *x* the set A. Note that K is usually odd to prevent tie situations.

It then estimates the conditional probability for each class, that is, the fraction of points in A with that given class label. (Note *I*(*x*) is the indicator function which evaluates to 1 when the argument *x* is true and 0 otherwise)

$$P(y=j|X=x)=1/K \sum_{i\in A} I(y^{(i)}=j)$$

Finally, our input *x* gets assigned to the class with the largest probability.

An alternate way of understanding KNN is by thinking about it as calculating a decision boundary (i.e. boundaries for more than 2 classes) which is then used to classify new points.

16. **Math on validation?**
    Ans:
    <u>Validation:</u> In this approach, instead of using the training set to estimate the generalization error, the original training data is divided into two smaller subsets. One of the subsets is used for training, while the other, known as the validation set, is used for estimating the generalization error.
    Validation set: Pick algorithm + knob settings
    - Pick best-performing algorithm (NB vs. DT vs…)
    - Fine-tune knobs (tree depth, k in KNN, c in SVM)
    -
17. **In constructing a decision tree, how do we select an attribute and when do we stop the further expansion of the tree?**
    Ans:
    Attribute selection measure is a heuristic for selecting the splitting criterion that partition data into the best possible manner. It is also known as splitting rules because it helps us to determine breakpoints for tuples on a given node. Most popular selection measures are Information Gain, Gain Ratio, and Gini Index.
    If dataset consists of "n" attributes then deciding which attribute to place at the root or at different levels of the tree as internal nodes is a complicated step. By just randomly selecting any node to be the root can't solve the issue. If we follow a random approach, it may give us bad results with low accuracy.

    For solving this attribute selection problem, researchers worked and devised some solutions. They suggested using some criterion like Information gain, Gini index, etc. These criterions will calculate values for every attribute. The values are sorted, and attributes are placed in the tree by following the order i.e. the attribute with a high value (in case of information gain) is placed at the root.

There are two strategies for avoiding model overfitting or further expansion in the context of decision tree induction.

- Pre-pruning (Early Stopping Rule)
- Post-pruning

Pre-pruning: In this approach, the tree-growing algorithm is halted before generating a fully grown tree that perfectly fits the entire training data. To do this, a more restrictive stopping condition must be used; e.g. stop expanding a leaf node when the observed gain in impurity measure (or improvement in the estimated generalization error) falls below a certain threshold.

Post-pruning: In this approach, the decision tree is initially grown to its maximum size. This is followed by a tree-pruning step, which proceeds to trim the fully grown tree in a bottom-up fashion. Trimming can be done by replacing a subtree with (1) a new leaf node whose class label is determined from the majority class of records affiliated with the subtree, or (2) the most frequently used branch of the subtree. The tree-pruning step terminates when no further improvement is observed.

Post-pruning tends to give better results than pre-pruning because it makes pruning decisions based on a fully grown tree.

18. On what principle Bayesian classifier has been built?

Ans:

Bayesian classification is based on Bayes' Theorem. Bayesian classifiers are the statistical classifiers. Bayesian classifiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.

Bayes' Theorem is named after Thomas Bayes. There are two types of probabilities −

- Posterior Probability [P(A/B)]
- Prior Probability [P(A)]

Where B is data tuple and A is some hypothesis.

According to Bayes' Theorem, P (A/B) = P(B/A).P(A) / P(B)

Using Bayes theorem, we can find the probability of A happening, given that B has occurred. Here, B is the evidence and A is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other.

19. Math on Gini Index?

Ans:

Gini Index: Gini index is the most commonly used measure of inequality. Also referred as Gini ratio or Gini coefficient.

Gini Index for a given node t:

$$\text{Gini (t)} = 1\text{-}\sum_j [p(j|t)]\ \text{^}2$$

$p(j \mid t)$ is the relative frequency of class j at node t.

- Maximum $(1 - 1/n_c)$ when records are equally distributed among all classes, implying least interesting information.
- Minimum (0.0) when all records belong to one class, implying most interesting information.

Example of computing Gini: Gini (t) = $1-\sum_{j}[p(j|t)]$ ^2

C1	0
C2	6

P(C1)=0/6 = 0 P(C2) = 6/6 = 1

Gini = $1 - P(C1)^2 - P(C2)^2$ = 1-0-1 =0

C1	1
C2	5

P(C1)=1/6 P(C2) = 5/6

Gini = $1 - P(C1)^2 - P(C2)^2 = 1 - (1/6)^2 – (5/6)^2$ = .278

C1	2
C2	4

P(C1)=2/6 P(C2) = 4/6

Gini = $1 - (2/6)^2 - (4/6)^2$ = .444

20. What are the advantage of tree based classification?

Ans:

Advantage of tree based classification:

- Decision trees are easy to interpret and visualize.
- It can easily capture Non-linear patterns.
- It requires fewer data preprocessing from the user, for example, there is no need to normalize columns.
- It can be used for feature engineering such as predicting missing values, suitable for variable selection.
- The decision tree has no assumptions about distribution because of the non-parametric nature of the algorithm.
- It does not require any domain knowledge.
- It is easy to comprehend.
- The learning and classification steps of a decision tree are simple and fast.

21. What are the main principal of Bayesian Classification?

Ans:

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often perform better in many complex real world situations.

Requires a small amount of training data to estimate the parameters.

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

P(A) : Prior probability of hypothesis A

P(B) : Prior probability of training data B

P(A/B) : Probability of A given B

P(B/A) : Probability of B given A

- Naive assumption: attribute independence $P(x_1,\ldots,x_k|C) = P(x_1|C)\cdot\ldots\cdot P(x_k|C)$
- If ith attribute is categorical: $P(x_i|C)$ is estimated as the relative frequency of samples having value xi as ith attribute in class C.
- If ith attribute is continuous: $P(x_i|C)$ is estimated through a Gaussian density function.

22. What are the sequential steps in doing classification of a set of data?

Ans:

Classification (also known as classification trees or decision trees) is a data mining algorithm that creates a step-by-step guide for how to determine the output of a new data instance. The tree it creates is exactly that: a tree whereby each node in the tree represents a spot where a decision must be made based on the input, and one moves to the next node and the next until one reaches a leaf that tells him the predicted output.

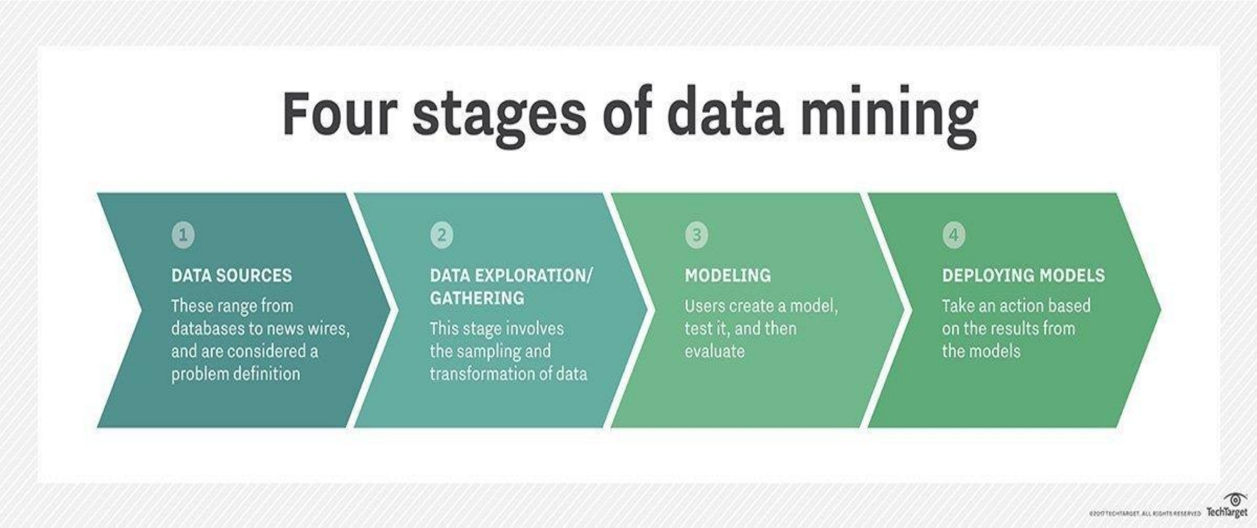
Classification has a two-step process:

Model construction: describing a set of predetermined classes

- Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute.
- The set of tuples used for model construction is training set.
- The model is represented as classification rules, decision trees, or mathematical formulae.

Model usage: for classifying future or unknown objects

- Estimate accuracy of the model
    - The known label of test sample is compared with the classified result from the model.
    - Accuracy rate is the percentage of test set samples that are correctly classified by the model.
    - Test set is independent of training set, otherwise over-fitting will occur.
- If the accuracy is acceptable, use the model to classify data tuples whose class labels are not known.



Four stages of data mining

1 DATA SOURCES
These range from databases to news wires, and are considered a problem definition

2 DATA EXPLORATION/ GATHERING
This stage involves the sampling and transformation of data

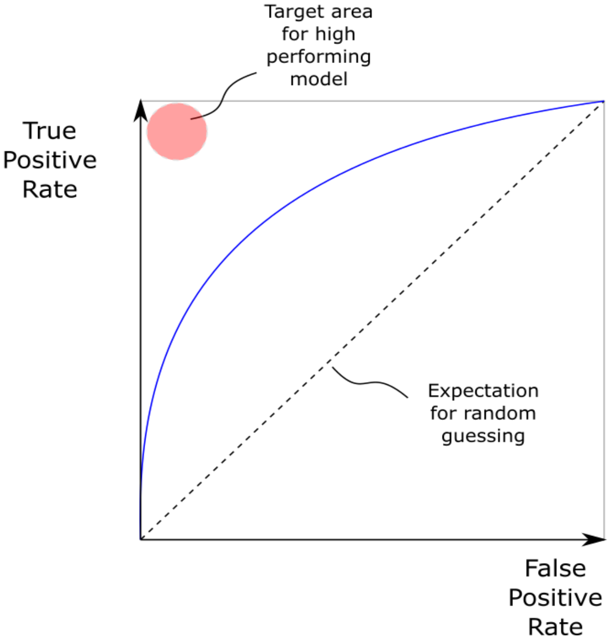
3 MODELING
Users create a model, test it, and then evaluate

4 DEPLOYING MODELS
Take an action based on the results from the models

23. What are the use of ROC curve?

Ans:

<u>ROC curve:</u> A receiver operating characteristic (ROC) curve is a graphical approach for displaying the tradeoff between true positive rate and false positive rate of a classifier. It is a plot of True positive (TP) and false positive (FP) rates (fractions).



<u>Use of ROC curve:</u>

- It is used for evaluating data mining schemes, and comparing the relative performance among different classifiers.

- ROC curves are frequently used to show in a graphical way the connection/trade-off between clinical sensitivity and specificity for every possible cut-off for a test or a combination of tests.
- ROC curves are used in clinical biochemistry to choose the most appropriate cut-off for a test. The best cut-off has the highest true positive rate together with the lowest false positive rate.
- ROC curves are appropriate when the observations are balanced between each class, whereas precision-recall curves are appropriate for imbalanced datasets.
- ROC curves should be used when there are roughly equal numbers of observations for each class.

24. What are the uses of support vector machine (SVM)?

Ans:

<u>Support Vector Machine (SVM):</u> “Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for either classification or regression challenges. However, it is mostly used in classification problems. It uses a technique called the kernel trick to transform user data.

SVM is capable of doing both classification and regression. SVM also works very well with high-dimensional data and avoids the curse of dimensionality problem.

<u>Uses of SVM:</u>

- It works really well with clear margin of separation
- It is effective in high dimensional spaces.
- It is effective in cases where number of dimensions is greater than the number of samples.
- It uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
-

25. What do you mean by the term precision and recall? When do we use these?

Ans:

<u>Precision:</u> Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances.

In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the query:

$$\text{Precision} = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \quad \text{or,}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

Precision is a good measure to determine, when the costs of False Positive is high. For instance, email spam detection. In email spam detection, a false positive means that an email that is non-spam (actual negative) has been identified as spam (predicted spam). The email user might lose important emails if the precision is not high for the spam detection model.

<u>Recall:</u> Recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. In information retrieval, recall is the fraction of the relevant documents that are successfully retrieved.

$$\text{Recall} = \frac{|\{relevant\ documents\} \cap \{\ retrieved\ documents\}|}{|\{relevant\ documents\}|} \quad \text{or,}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

Recall actually calculates how many of the Actual Positives our model capture through labeling it as Positive (True Positive). Applying the same understanding, we know that Recall shall be the model metric we use to select our best model when there is a high cost associated with False Negative.
For instance, in fraud detection or sick patient detection. If a fraudulent transaction (Actual Positive) is predicted as non-fraudulent (Predicted Negative), the consequence can be very bad for the bank.

26. What do you mean by rule base classification?

Ans:

<u>Rule Base Classification:</u> Rule-based classifier makes use of a set of IF-THEN rules for classification. We can express a rule in the following from

IF condition THEN conclusion

Rule: (Condition) ➡ y

Where

Condition is a conjunctions of attributes

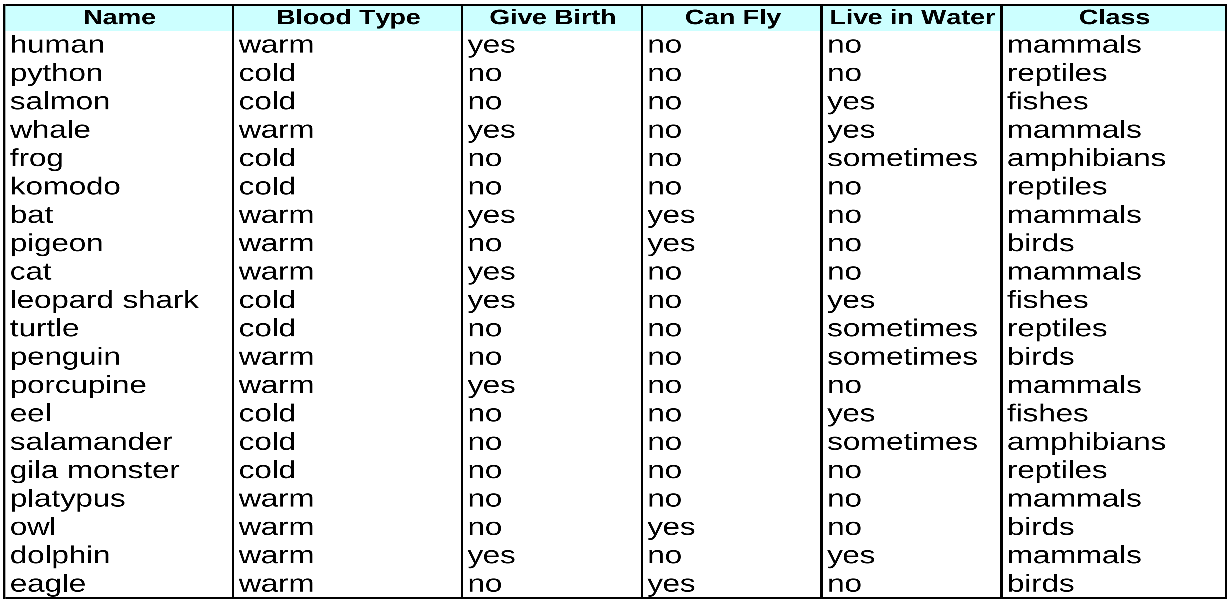
y is the class label

Examples of classification rules:

(Blood Type=Warm) ^ (Lay Eggs=Yes) → Birds

(Taxable Income < 50K) ^ (Refund=Yes) → Evade=No

Let’s go through more examples of classification rules:



Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
human	warm	yes	no	no	mammals
python	cold	no	no	no	reptiles
salmon	cold	no	no	yes	fishes
whale	warm	yes	no	yes	mammals
frog	cold	no	no	sometimes	amphibians
komodo	cold	no	no	no	reptiles
bat	warm	yes	yes	no	mammals
pigeon	warm	no	yes	no	birds
cat	warm	yes	no	no	mammals
leopard shark	cold	yes	no	yes	fishes
turtle	cold	no	no	sometimes	reptiles
penguin	warm	no	no	sometimes	birds
porcupine	warm	yes	no	no	mammals
eel	cold	no	no	yes	fishes
salamander	cold	no	no	sometimes	amphibians
gila monster	cold	no	no	no	reptiles
platypus	warm	no	no	no	mammals
owl	warm	no	yes	no	birds
dolphin	warm	yes	no	yes	mammals
eagle	warm	no	yes	no	birds

R1: (Give Birth = no) $\wedge$ (Can Fly = yes) $\rightarrow$ Birds

R2: (Give Birth = no) $\wedge$ (Live in Water = yes) $\rightarrow$ Fishes

R3: (Give Birth = yes) $\wedge$ (Blood Type = warm) $\rightarrow$ Mammals

R4: (Give Birth = no) $\wedge$ (Can Fly = no) $\rightarrow$ Reptiles

R5: (Live in Water = sometimes) $\rightarrow$ Amphibians

27. What is decision tree classification?

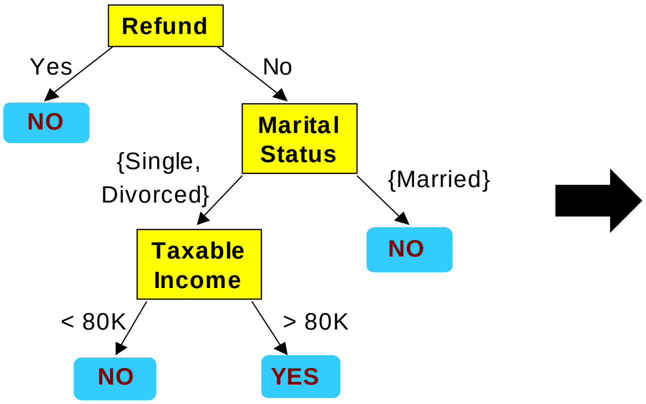
Ans:

<u>Decision Tree Classification</u>: Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy). Leaf node (e.g., Play) represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

<u>Decision Tree Classification Rules:</u>

Rules are mutually exclusive and exhaustive

Rule set contains as much information as the tree



Classification Rules

(Refund=Yes) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income<80K) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income>80K) ==> Yes

(Refund=No, Marital Status={Married}) ==> No

28. What is ensemble method of classification? Explain with pictorial example.

Ans:

<u>Ensemble Method of Classification:</u>

Ensemble of classifiers is a set of classifiers whose individual decisions combined in some way to classify new examples

Simplest approach:

1. Generate multiple classifiers

2. Each votes on test instance

3. Take majority as classification

Classifiers different due to different sampling of training data, or randomized parameters within the classification algorithm

Differ in training strategy, and combination method:

1. Parallel training with different training sets: bagging

2. Sequential training, iteratively re-weighting training examples so current classifier focuses on hard examples: boosting

3. Parallel training with objective encouraging division of labor mixture of experts.

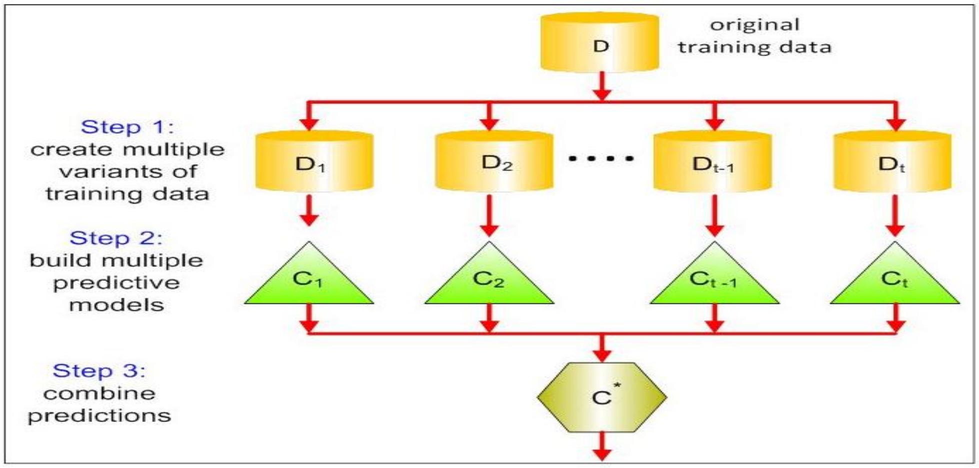


Figure: A Logical view of the Ensemble Learning Method

29. What is the main principal of Gini index? – explain. When we have to use Gini index in splitting?

Ans:

<u>The main Principal of Gini index:</u> The Gini index (i.e. Gini coefficient) is a statistical measure of distribution. It is commonly used as an index of economic inequality that measures income or wealth distribution among the population. The Income Gini Index is more popular than the Wealth Gini index, because it is much easier to measure income as opposed to wealth. The index ranges from 0 to 1 (or 0 – 100 in percent). An index of 0 represents perfect equality, whereas an index of 1 (or 100) describes perfect inequality.

Gini Index for a given node t :

$$GINI(t) = 1 - \sum_{j} [p(j \mid t)]^2$$

- Maximum (1 - 1/n_c) when records are equally distributed among all classes, implying least interesting information
- Minimum (0.0) when all records belong to one class, implying most interesting information

C1	**0**
C2	**6**
Gini=0.000	

C1	**1**
C2	**5**
Gini=0.278	

C1	**2**
C2	**4**
Gini=0.444	

C1	**3**
C2	**3**
Gini=0.500	

Examples:

C1	**0**
C2	**6**

P (C1) = 0/6 = 0   P(C2) = 6/6 = 1

Gini = 1 – P(C1)^2 – P(C2)^2 = 1 – 0 – 1 = 0

C1	**1**
C2	**5**

P(C1) = 1/6   P(C2) = 5/6

Gini = 1 – (1/6)^2 – (5/6)^2 = 0.278

C1	**2**
C2	**4**

P(C1) = 2/6   P(C2) = 4/6

Gini = 1 – (2/6)^2 – (4/6)^2 = 0.444

<u>Splitting Based on GINI:</u>

Used in CART, SLIQ, SPRINT.

When a node p is split into k partitions (children), the quality of split is computed as,

$$\boxed{GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)}$$

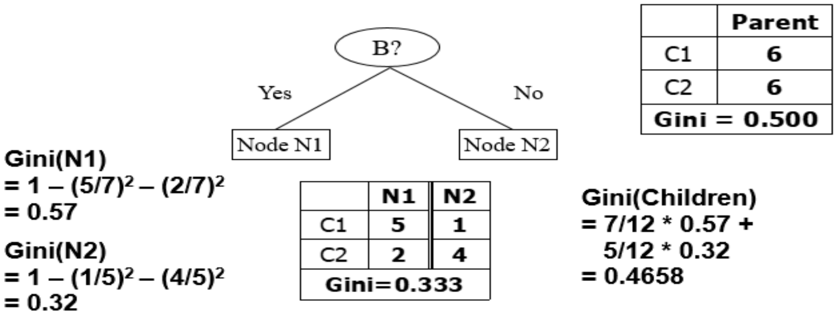
where, ni = number of records at child i,

n = number of records at node p.

Splits into two partitions

Effect of Weighing partitions:

Larger and Purer Partitions are sought for.



	Parent
C1	6
C2	6
Gini = 0.500	

Gini(N1)
$= 1 – (5/7)^2 – (2/7)^2$
= 0.57

Gini(N2)
$= 1 – (1/5)^2 – (4/5)^2$
= 0.32

	N1	N2
C1	5	1
C2	2	4
Gini=0.333		

Gini(Children)
= 7/12 * 0.57 +
5/12 * 0.32
= 0.4658

30. When we have to use Gini index in splitting?

Ans:

Uses of Gini index in splitting:

Splitting Based on GINI:

Used in CART, SLIQ, SPRINT.

When a node p is split into k partitions (children), the quality of split is computed as,

$$\boxed{GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)}$$

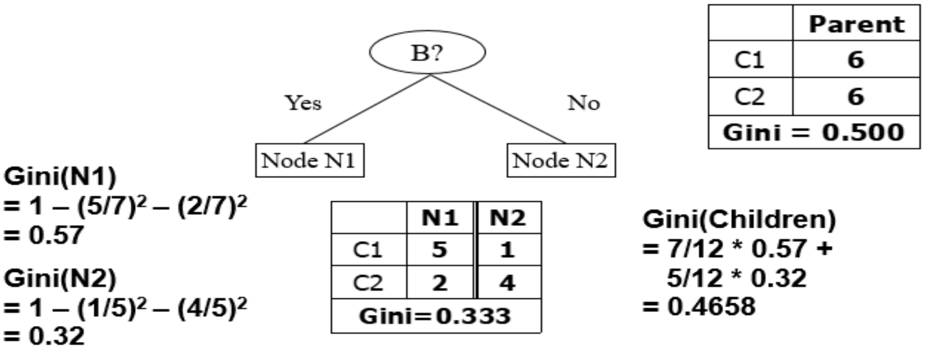
where, ni = number of records at child i,

n = number of records at node p.

Splits into two partitions

Effect of Weighing partitions:

Larger and Purer Partitions are sought for.


B?
Yes
No
Node N1
Node N2

	Parent
C1	6
C2	6
Gini = 0.500	

Gini(N1)
$= 1 – (5/7)^2 – (2/7)^2$
= 0.57

Gini(N2)
$= 1 – (1/5)^2 – (4/5)^2$
= 0.32

	N1	N2
C1	5	1
C2	2	4
Gini=0.333		

Gini(Children)
= 7/12 * 0.57 +
5/12 * 0.32
= 0.4658

31. Which classification technique will you use?

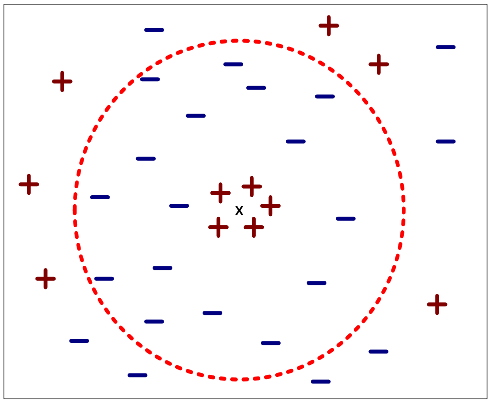
Ans:

I will use the nearest neighbor classification.

- □ Compute distance between two points:
    - ◘ Euclidean distance

$$d(p,q)=\sqrt{\sum_i (p_i - q_i)^2}$$

    - ◘ Determine the class from nearest neighbor list
    - ◘ take the majority vote of class labels among the k-nearest neighbors
    - ◘ Weigh the vote according to distance
        - ■ weight factor, w = 1/d^2
- □ Choosing the value of k:
    - ◘ If k is too small, sensitive to noise points
    - ◘ If k is too large, neighborhood may include points from other classes



- □ k-NN classifiers are lazy learners
    - ◘ It does not build models explicitly
    - ◘ Unlike eager learners such as decision tree induction and rule-based systems
    - ◘ Classifying unknown records are relatively expensive

32. Why and when do researchers like to use SVM classifier?

Ans:

SVM: SVM is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs. Simply put, it does some extremely complex data transformations, then figures out how to separate your data based on the labels or outputs you've defined.

Why use: Well SVM it capable of doing both classification and regression. In this post I'll focus on using SVM for classification. In particular I'll be focusing on non-linear SVM, or SVM using a non-linear kernel. Non-linear SVM means that the boundary that the algorithm calculates doesn't have to be a straight line. The benefit is that you can capture much more complex relationships between your data points without having to perform difficult transformations on your own. The downside is that the training time is much longer as it's much more computationally intensive.

When use: SVM is one of the best classifier but not the best. In fact, no one could be the best. It depends upon the problem which classifier would be suitable.

As for as, SVM is concerned, it is a suitable classifier in following cases:

1) When number of features (variables) and number of training data is very large (say millions of features and millions of instances (data)).

2) When sparsity in the problem is very high, i.e., most of the features have zero value.

3) It is the best for document classification problems where sparsity is high and features/instances are also very high.

4) It also performs very well for problems like image classification, genes classsification, drug disambiguation etc. where number of features are high.

It is one of the best, because of following things:

1) It uses Kernel trick

2) It is Optimal margin based classification technique in Machine Learning.

3) Good number of algorithms are proposed which utilizes problem structures and other smaller-smaller things like problem shrinking during optimization etc.

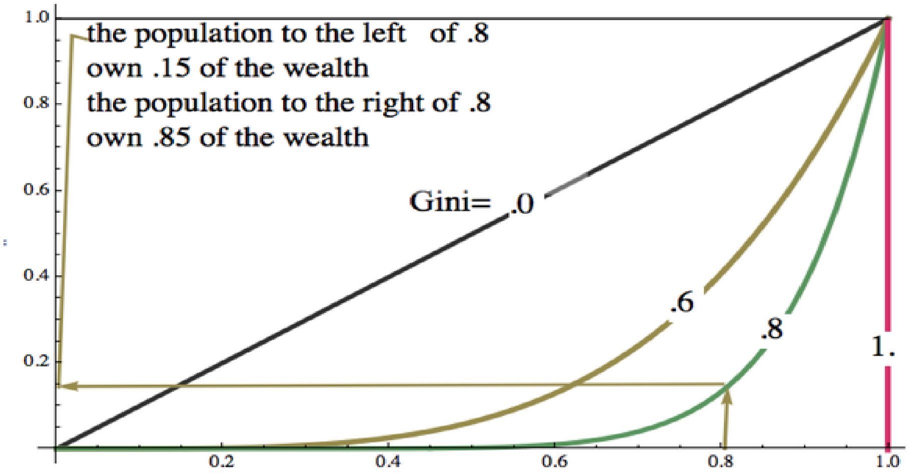
33. Why and when do we use Gini coefficient or entropy?

Ans:

Why: The Gini index measures the **distribution** of wealth, income, or anything else for that matter.

Since wealth distribution is vital to the stability of a society and the well being of its people, lets use it for an explanation of its use.

First, for those who aren't familiar with Dr. Gini's index, in the plot below,

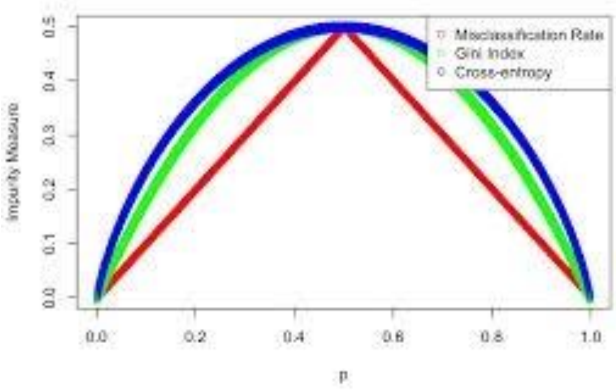


the y axis is the fraction of total wealth owned by the fraction on the x axis.The x axis is the fraction of the population who own that wealth. For the US, the wealth Gini index can be calculated for the populations net wealth with home equity included, around .6 fraction of households, for those who have substantial other assets, those between .9 and .99. And those who have extreme wealth, between .99 and 1.0.

34. Why entropy is used instead of Gini index?

Ans:

From what I can tell, both are used largely for the same purpose. If we visualize these two metrics (and throw in the miss-classification error) for a binary classification, we'd see something like this:



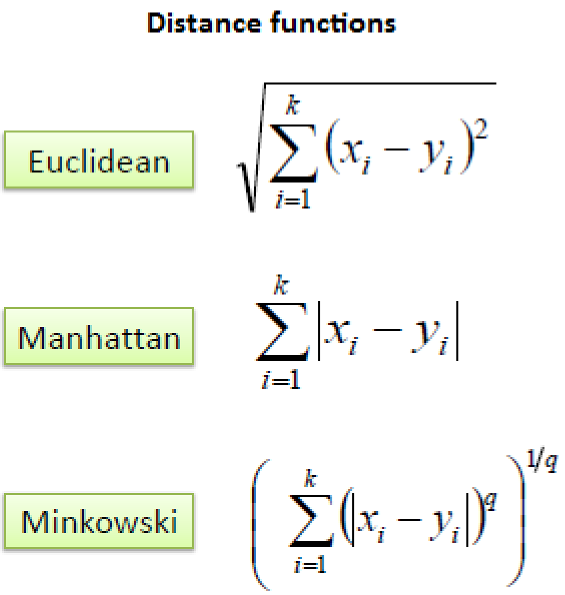
We notice that gini and cross-entropy look incredibly similar. Furthermore, while we often make the distinction between using miss-classification error versus gini or cross-entropy when growing trees, we don't often hear good reasons to use gini over cross-entropy, or vice versa (at least I haven't). I've come across people who like the interpretation of one over the other, but in practice it seems like we often try both (or just use one) and see which one gives us better results.

Why the two formulas, then? It appears that they arose from the same motivations, but from two different fields (Gini from statistics and cross-entropy from computer science/information theory). This paper discusses a little of how the two arose, but probably more importantly gives some empirical evidence to support the idea that one isn't really better than the other.

35. Write the algorithm of K-nearest neighbor classification.

Ans: The algorithm of K-nearest neighbor:

A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.



Distance functions

Euclidean

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Manhattan

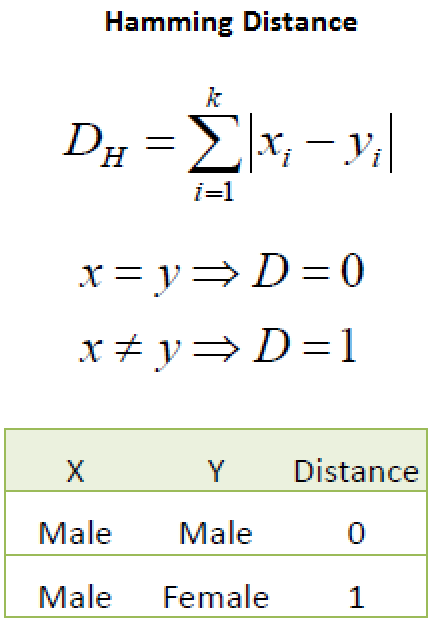
$$\sum_{i=1}^{k}\left|x_i - y_i\right|$$

Minkowski

$$\left(\sum_{i=1}^{k}(\left|x_i - y_i\right|)^q\right)^{1/q}$$

It should also be noted that all three distance measures are only valid for continuous variables. In the instance of categorical variables the Hamming distance must be used. It also brings up the issue of standardization of the numerical variables between 0 and 1 when there is a mixture of numerical and categorical variables in the dataset.

36. Write the limitations of KNN.



Hamming Distance

$$D_H = \sum_{i=1}^{k} |x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

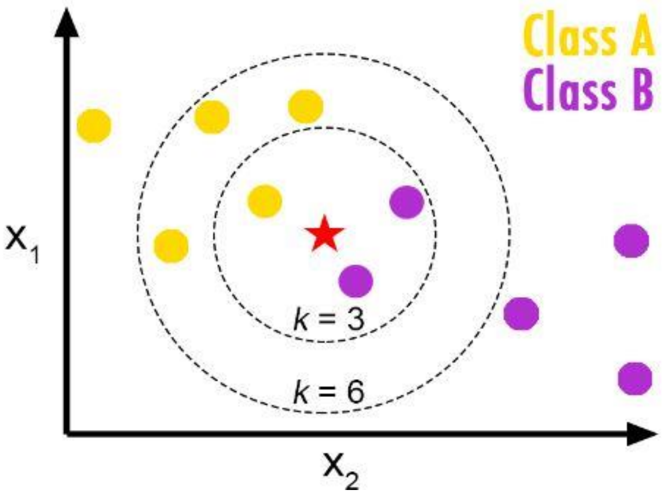
$$x \neq y \Rightarrow D = 1$$

X	Y	Distance
Male	Male	0
Male	Female	1

Choosing the optimal value for K is best done by first inspecting the data. In general, a large K value is more precise as it reduces the overall noise but there is no guarantee. Cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value. Historically, the optimal K for most datasets has been between 3-10. That produces much better results than 1NN.

Ans:

KNN is a non-parametric, lazy learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point.



Even though K-NN has several advantages but there are certain very important disadvantages or limitations of K-NN. Below are listed few cons of K-NN.

1. **K-NN slow algorithm**: K-NN might be very easy to implement but as dataset grows efficiency or speed of algorithm declines very fast.
2. **Curse of Dimensionality:** KNN works well with small number of input variables but as the numbers of variables grow K-NN algorithm struggles to predict the output of new data point.
3. **K-NN needs homogeneous features**: If you decide to build k-NN using a common distance, like Euclidean or Manhattan distances, it is completely necessary that features have the same scale, since absolute differences in features weight the same, i.e., a given distance in feature 1 must means the same for feature 2.
4. **Optimal number of neighbors**: One of the biggest issues with K-NN is to choose the optimal number of neighbors to be consider while classifying the new data entry.
5. **Imbalanced data causes problems**: k-NN doesn't perform well on imbalanced data. If we consider two classes, A and B, and the majority of the training data is labeled as A, then the model will ultimately give a lot of preference to A. This might result in getting the less common class B wrongly classified.
6. **Outlier sensitivity:** K-NN algorithm is very sensitive to outliers as it simply chose the neighbors based on distance criteria.
7. **Missing Value treatment:** K-NN inherently has no capability of dealing with missing value problem.

37. What is K-nearest neighbor classification and k-means clustering?

Ans:

<u>K-NEAREST NEIGHBORS (K-NN) CLASSIFICATION :</u>

The **k-Nearest-Neighbors (kNN)** method of classification is one of the simplest methods in machine learning.In k-Nearest Neighbor classification, the training dataset is used to classify each member of a "target" dataset. The structure of the data is that there is a classification (categorical) variable of interest ("buyer," or "non-buyer," for example), and a number of additional predictor variables (age, income, location...). Generally speaking, the algorithm is as follows:

- For each row (case) in the target dataset (the set to be classified), locate the k closest members (the k nearest neighbors) of the training dataset. A Euclidean Distance measure is used to calculate how close each member of the training set is to the target row that is being examined.
- Examine the k nearest neighbors - which classification (category) do most of them belong to? Assign this category to the row being examined.
- Repeat this procedure for the remaining rows (cases) in the target set.

- Additional to this XLMiner also lets the user select a maximum value for k, builds models parallelly on all values of k upto the maximum specified value and scoring is done on the best of these models.

There are two important decisions that must be made before making classifications. One is the value of **k** that will be used; this can either be decided arbitrarily, or you can try **cross-validation** to find an optimal value. The next, and the most complex, is the **distance metric** that will be used.

There are many different ways to compute distance, as it is a fairly ambiguous notion, and the proper metric to use is always going to be determined by the data-set and the classification task. Two popular ones, however, are **Euclidean distance** and **Cosine similarity**.

Euclidean distance is probably the one that you are most familiar with; it is essentially the magnitude of the vector obtained by subtracting the training data point from the point to be classified.

$$E(x, y) = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$$

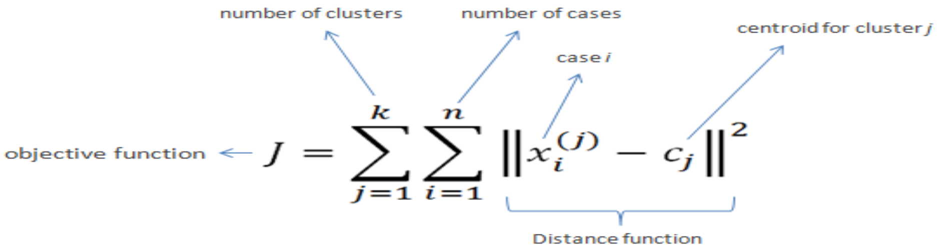
General formula for Euclidean distance

Another common metric is Cosine similarity. Rather than calculating a magnitude, Cosine similarity instead uses the difference in direction between two vectors.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{||A|| \, ||B||}$$

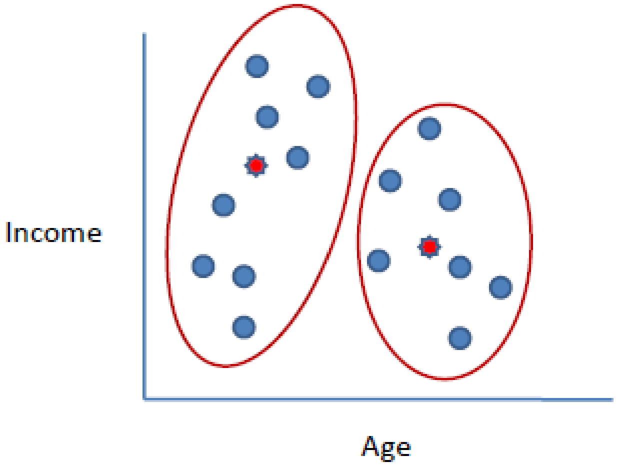
General formula for Cosine similarity

k-Means clustering intends to partition *n* objects into *k* clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly *k* different clusters of greatest possible distinction. The best number of clusters *k* leading to the greatest separation (distance) is not known as a priori and must be computed from the data. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function:



Algorithm

1. Clusters the data into *k* groups where *k* is predefined.
2. Select *k* points at random as cluster centers.
3. Assign objects to their closest cluster center according to the *Euclidean distance* function. 6.
4. Calculate the centroid or mean of all objects in each cluster.
5. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

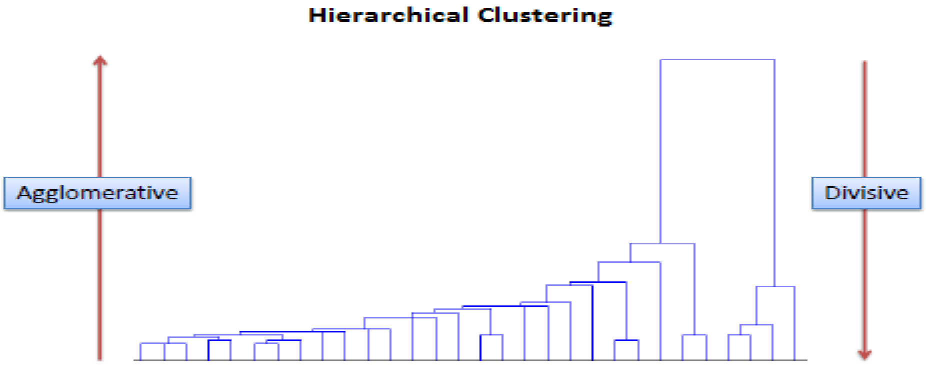


38. Define hierarchical clustering with example.

Ans:

<u>Hierarchical clustering :</u>Hierarchical clustering, also known as *hierarchical cluster analysis*, is an algorithm that groups similar objects into groups called *clusters*. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. For example, all files and folders on the hard disk are organized in a hierarchy. There are two types of hierarchical clustering, *Divisive* and *Agglomerative*.

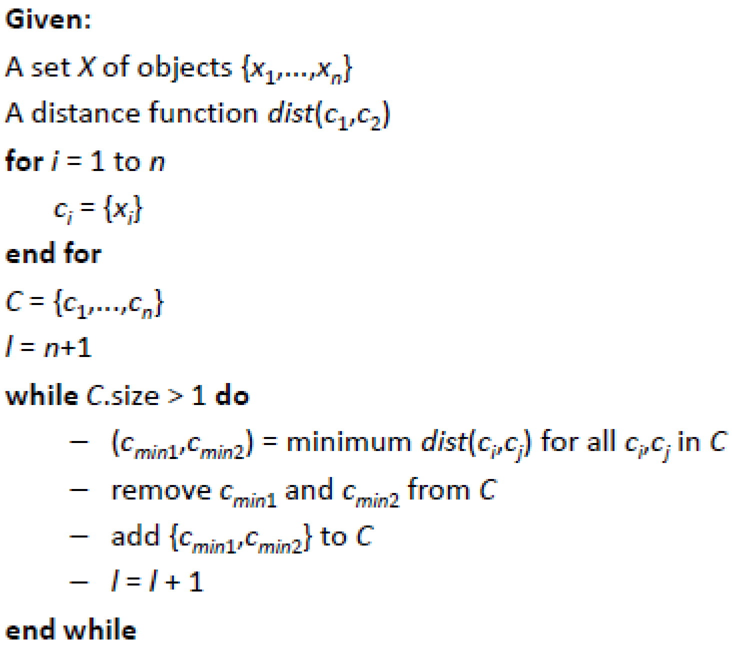


Divisive method

In *divisive* or *top-down clustering* method we assign all of the observations to a single cluster and then partition the cluster to two least similar clusters. Finally, we proceed recursively on each cluster until there is one cluster for each observation. There is evidence that divisive algorithms produce more accurate hierarchies than agglomerative algorithms in some circumstances but is conceptually more complex.

Agglomerative method

In *agglomerative* or *bottom-up clustering* method we assign each observation to its own cluster. Then, compute the similarity (e.g., distance) between each of the clusters and join the two most similar clusters. Finally, repeat steps 2 and 3 until there is only a single cluster left. The related algorithm is shown below.

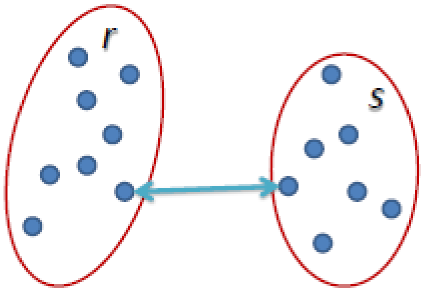


```
Given:
A set $X$ of objects $\{x_1,\ldots,x_n\}$
A distance function $dist(c_1,c_2)$
for $i$ = 1 to $n$
    $c_i = \{x_i\}$
end for
$C = \{c_1,\ldots,c_n\}$
$l = n+1$
while $C$.size > 1 do
    – $(c_{min1},c_{min2})$ = minimum $dist(c_i,c_j)$ for all $c_i,c_j$ in $C$
    – remove $c_{min1}$ and $c_{min2}$ from $C$
    – add $\{c_{min1},c_{min2}\}$ to $C$
    – $l = l + 1$
end while
```

Before any clustering is performed, it is required to determine the proximity matrix containing the distance between each point using a distance function. Then, the matrix is updated to display the distance between each cluster. The following three methods differ in how the distance between each cluster is measured.

Single Linkage

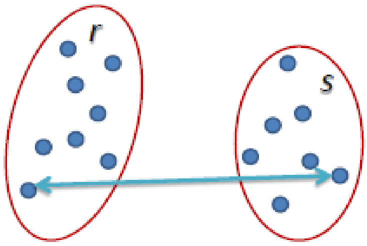
In single linkage hierarchical clustering, the distance between two clusters is defined as the *shortest* distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two closest points.



$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

Complete Linkage

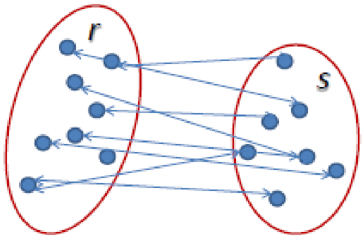
In complete linkage hierarchical clustering, the distance between two clusters is defined as the *longest* distance between two points in each cluster. For example, the distance between clusters "r" and "s" to the left is equal to the length of the arrow between their two furthest points.



$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

Average Linkage

In average linkage hierarchical clustering, the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster. For example, the distance between clusters "r" and "s" to the left is equal to the average length each arrow between connecting the points of one cluster to the other.



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

39. Describe the steps of K-means clustering.

Ans:

Algorithmic steps for k-means clustering:

Let $X = \{x_1, x_2, x_3, \ldots\ldots.., x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots., v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

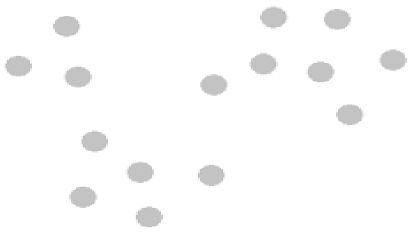
where, *'c_i'* represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

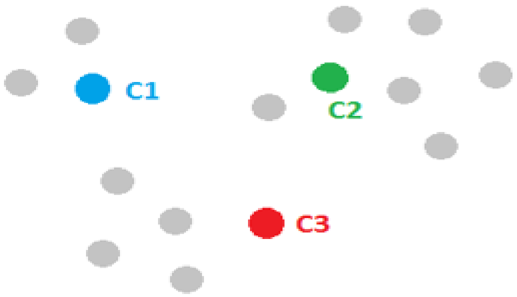
Among all the unsupervised learning algorithms, clustering via k-means might be one of the simplest and most widely used algorithms. Briefly speaking, k-means clustering aims to find the set of k clusters such that every data point is assigned to the closest center, and the sum of the distances of all such assignments is minimized.

Let's walk through a simple 2D example to better understand the idea. Imaging we have these gray points in the following figure and want to assign them into three clusters. K-means follows the four steps listed below.

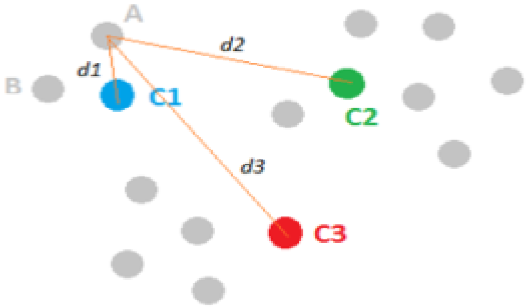


Step one: Initialize cluster centers

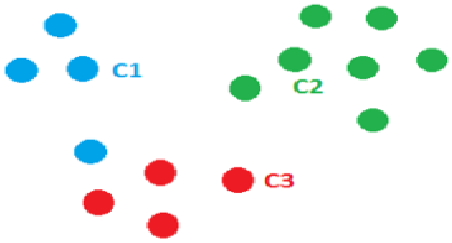
We randomly pick three points C1, C2 and C3, and label them with blue, green and red color separately to represent the cluster centers.



Step two: Assign observations to the closest cluster center

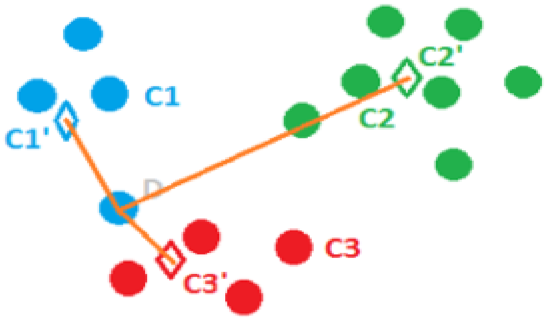


Once we have these cluster centers, we can assign each point to the clusters based on the minimum distance to the cluster center. For the gray point A, compute its distance to C1, C2 and C3, respectively. And after comparing the lengths of *d1*, *d2* and *d3*, we figure out that *d1* is the smallest, therefore, we assign point A to the blue cluster and label it with blue. We then move to point B and follow the same procedure. This process can assign all the points and leads to the following figure.



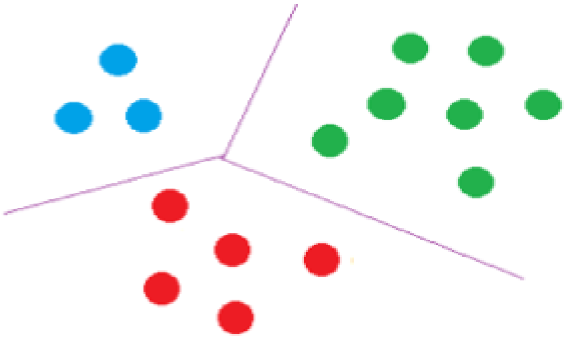
Step three: Revise cluster centers as mean of assigned observations

Now we've assigned all the points based on which cluster center they were closest to. Next, we need to update the cluster centers based on the points assigned to them. For instance, we can find the center mass of the blue cluster by summing over all the blue points and dividing by the total number of points, which is four here. And the resulted center mass C1', represented by a blue diamond, is our new center for the blue cluster. Similarly, we can find the new centers C2' and C3' for the green and red clusters.



Step four: Repeat step 2 and step 3 until convergence

The last step of k-means is just to repeat the above two steps. For example, in this case, once C1', C2' and C3' are assigned as the new cluster centers, point D becomes closer to C3' and thus can be assigned to the red cluster. We keep on iterating between assigning points to cluster centers, and updating the cluster centers until convergence. Finally, we may get a solution like the following figure. Well done!



40. What are the different types of clustering?

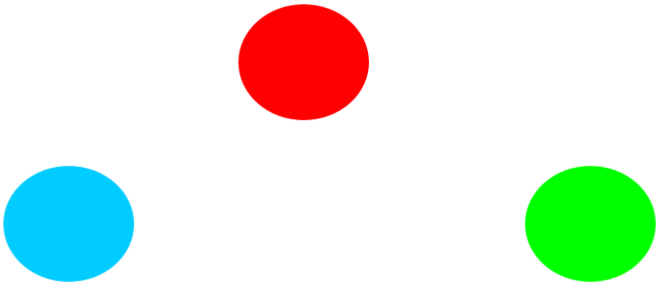
Ans:

<u>Clustering methods</u>: are used to identify groups of similar objects in a multivariate data sets collected from fields such as marketing, bio-medical and geo-spatial. Clustering methods can be classified into the following categories –

- Well-separated clusters
- Center-based clusters
- Contiguous clusters

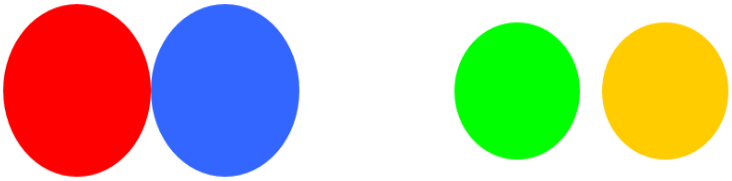
- Density-based clusters
- Property or Conceptual
- Described by an Objective Function

Well-Separated Clusters: A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.



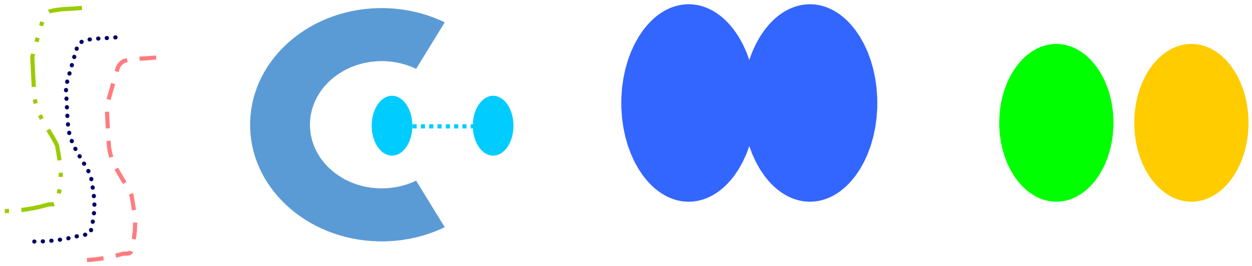
3 well-separated clusters

Center-based: A cluster is a set of objects such that an object in a cluster is closer (more similar) to the "center" of a cluster, than to the center of any other cluster. The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most "representative" point of a cluster



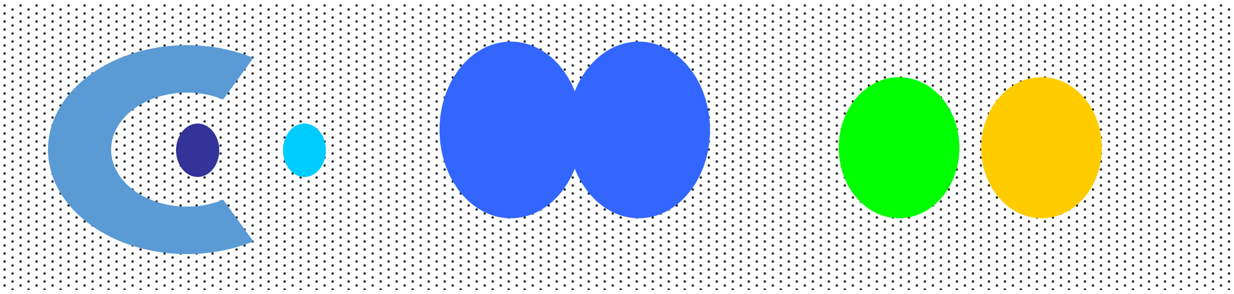
4 center-based clusters

Contiguous Cluster (Nearest neighbor or Transitive): A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.



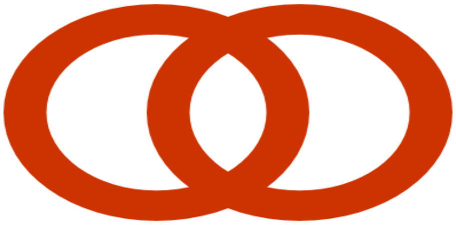
8 contiguous clusters

<u>Density-based:</u> A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density. Used when the clusters are irregular or intertwined, and when noise and outliers are present.



6 density-based clusters

<u>Shared Property or Conceptual Clusters:</u> Finds clusters that share some common property or represent a particular concept



2 Overlapping Circles

<u>Clusters Defined by an Objective Function:</u>

- Finds clusters that minimize or maximize an objective function.
- Enumerate all possible ways of dividing the points into clusters and evaluate the `goodness' of each potential set of clusters by using the given objective function. (NP Hard)
- Can have global or local objectives.

- Hierarchical clustering algorithms typically have local objectives
- Partitional algorithms typically have global objectives

– A variation of the global objective function approach is to fit the data to a parameterized model.

- Parameters for the model are determined from the data.
- Mixture models assume that the data is a ‘mixture' of a number of statistical distributions.

Map the clustering problem to a different domain and solve a related problem in that domain

– Proximity matrix defines a weighted graph, where the nodes are the points being clustered, and the weighted edges represent the proximities between points

– Clustering is equivalent to breaking the graph into connected components, one for each cluster.

– Want to minimize the edge weight between clusters and maximize the edge weight within clusters

41. What are the processes of density based clustering?

Ans:

Density-based Method:

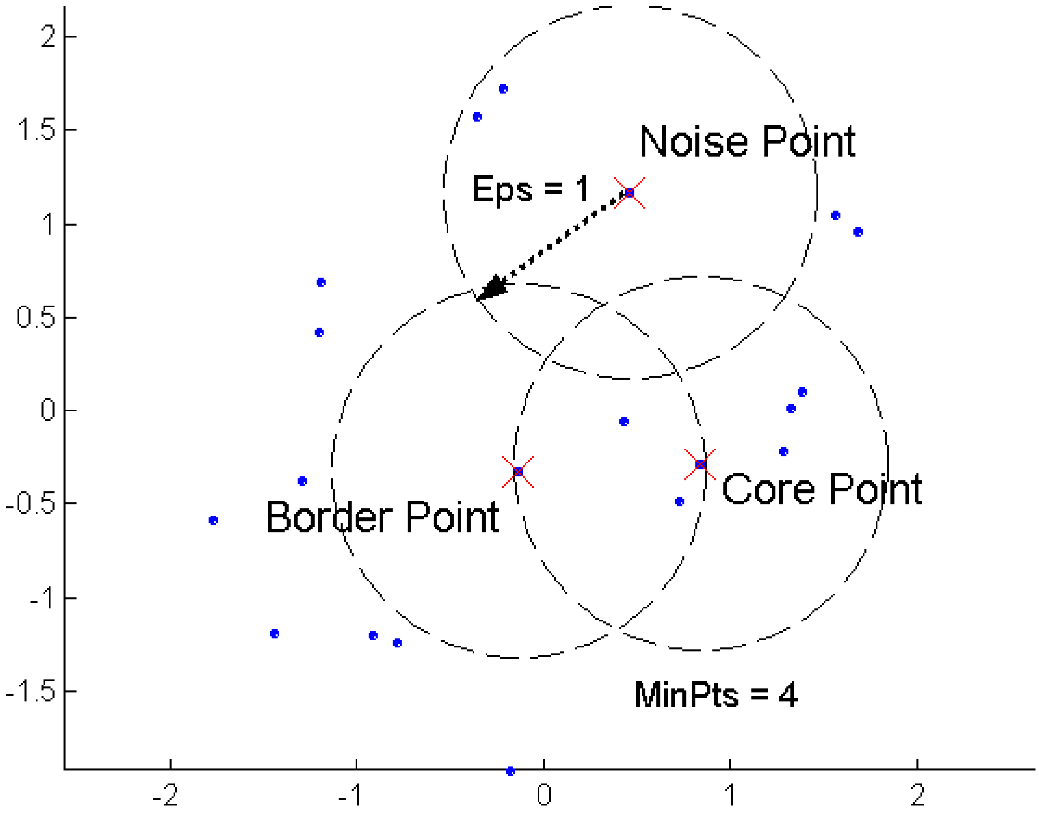
This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

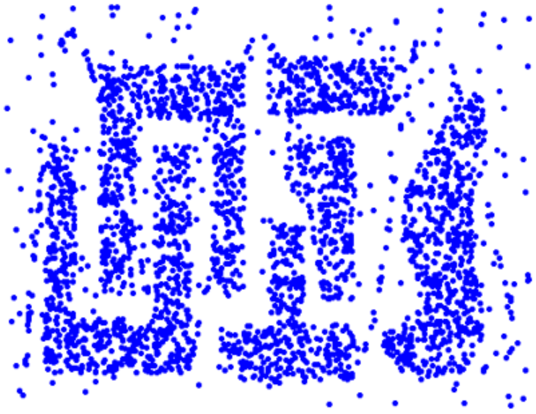
1. **Density-based spatial clustering of applications with noise** (**DBSCAN**) is a data clustering algorithm DBSCAN begins with an arbitrary starting data point that has not been visited. The neighborhood of this point is extracted using a distance epsilon ε (All points which are within the ε distance are neighborhood points).
2. If there are a sufficient number of points (according to minPoints) within this neighborhood then the clustering process starts and the current data point becomes the first point in the new cluster. Otherwise, the point will be labeled as noise (later this noisy point might become the part of the cluster). In both cases that point is marked as “visited”.

3. For this first point in the new cluster, the points within its ε distance neighborhood also become part of the same cluster. This procedure of making all points in the ε neighborhood belong to the same cluster is then repeated for all of the new points that have been just added to the cluster group.

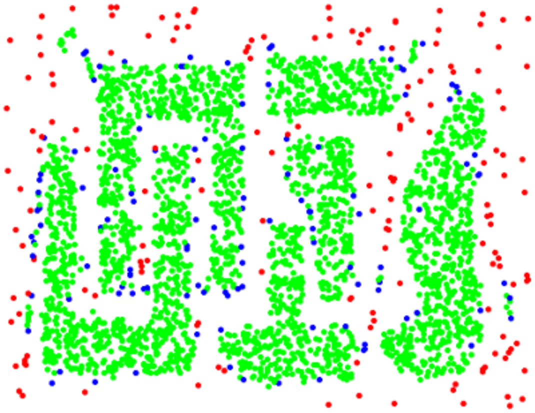
4. This process of steps 2 and 3 is repeated until all points in the cluster are determined i.e all points within the ε neighborhood of the cluster have been visited and labelled.

5. Once we're done with the current cluster, a new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise. This process repeats until all points are marked as visited. Since at the end of this all points have been visited, each point well have been marked as either belonging to a cluster or being noise.



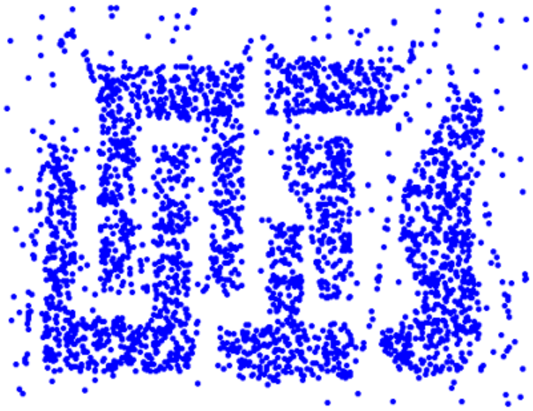


Original Points

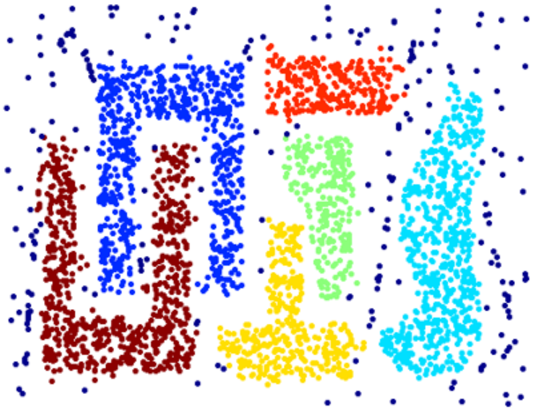


Point types: core, border and noise

Eps = 10, MinPts = 4



Original Points



Clusters

- Resistant to Noise
- Can handle clusters of different shapes and sizes

42. What do you mean by centeroid in k-means clustering

Ans:

The *k*-means clustering algorithm attempts to split a given anonymous data set (a set containing no information as to class identity) into a fixed number (*k*) of clusters.

Initially *k* number of so called *centroids* are chosen. A *centroid* is a data point (imaginary or real) at the center of a cluster. In Practical at each centroid is an existing data point in the given input data set, picked at random, such that all *centroids* are unique (that is, for all *centroids* c_i and c_j, $c_i \neq c_j$). These *centroids* are used to train a kNN classifier. The resulting classifier is used to classify (using $k = 1$) the data and thereby produce an initial randomized set of clusters. Each *centroid* is thereafter set to the arithmetic mean of the cluster it defines. The process of classification and *centroid* adjustment is repeated until the values of the *centroids* stabilize. The final *centroids* will be used to produce the final classification/clustering of the input data, effectively turning the set of initially anonymous data points into a set of data points, each with a class identity.

43. What do you mean by clustering? What are the different types of clustering?

Ans:

Clustering: Clustering is the process of making a group of abstract objects into classes of similar objects.Clustering methods can be classified into the following categories −

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

Partitioning Method:

Suppose we are given a database of ‘n’ objects and the partitioning method constructs ‘k’ partition of data. Each partition will represent a cluster and k ≤ n. It means that it will classify the data into k groups, which satisfy the following requirements −

- Each group contains at least one object.
- Each object must belong to exactly one group.

Points to remember −

- For a given number of partitions (say k), the partitioning method will create an initial partitioning.
- Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.

Hierarchical Methods

This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed. There are two approaches here −

- Agglomerative Approach
- Divisive Approach

Agglomerative Approach

This approach is also known as the bottom-up approach. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keep on doing so until all of the groups are merged into one or until the termination condition holds.

Divisive Approach

This approach is also known as the top-down approach. In this, we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone.

Approaches to Improve Quality of Hierarchical Clustering

Here are the two approaches that are used to improve the quality of hierarchical clustering −

- Perform careful analysis of object linkages at each hierarchical partitioning.
- Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters.

Density-based Method

This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

Grid-based Method

In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.

Advantages

- The major advantage of this method is fast processing time.
- It is dependent only on the number of cells in each dimension in the quantized space.

Model-based methods

In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points.

This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.

Constraint-based Method

In this method, the clustering is performed by the incorporation of user or application-oriented constraints. A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement.

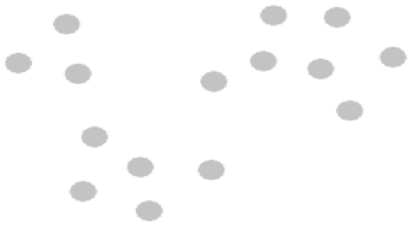
44. What is clustering? Describe the steps of K-means clustering.

Ans:

Clustering: Clustering is the process of making a group of abstract objects into classes of similar objects.

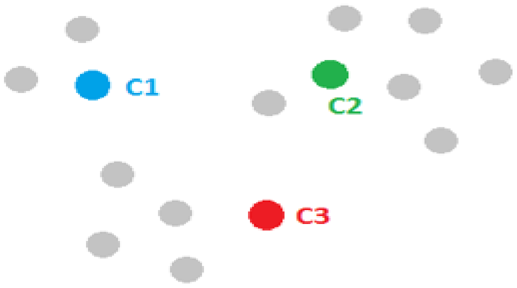
Among all the unsupervised learning algorithms, clustering via k-means might be one of the simplest and most widely used algorithms. Briefly speaking, k-means clustering aims to find the set of k clusters such that every data point is assigned to the closest center, and the sum of the distances of all such assignments is minimized.

Let’s walk through a simple 2D example to better understand the idea. Imaging we have these gray points in the following figure and want to assign them into three clusters. K-means follows the four steps listed below.

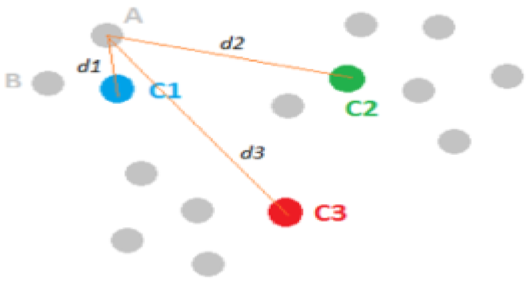


Step one: Initialize cluster centers

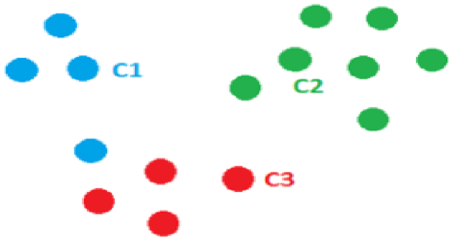
We randomly pick three points C1, C2 and C3, and label them with blue, green and red color separately to represent the cluster centers.



Step two: Assign observations to the closest cluster center

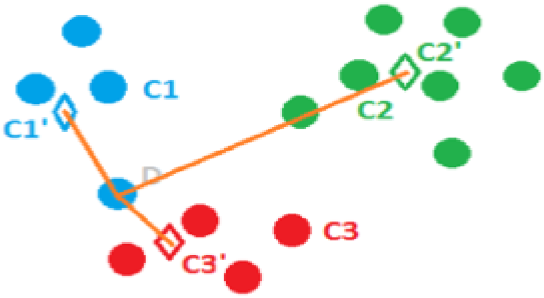


Once we have these cluster centers, we can assign each point to the clusters based on the minimum distance to the cluster center. For the gray point A, compute its distance to C1, C2 and C3, respectively. And after comparing the lengths of *d1*, *d2* and *d3*, we figure out that *d1* is the smallest, therefore, we assign point A to the blue cluster and label it with blue. We then move to point B and follow the same procedure. This process can assign all the points and leads to the following figure.



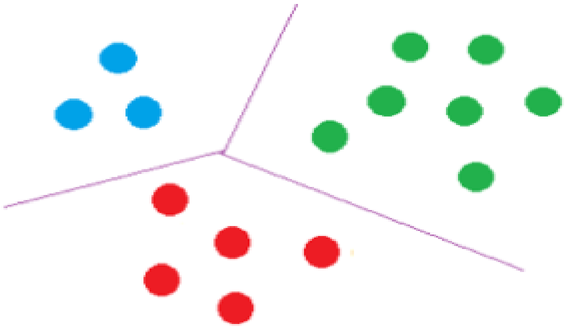
Step three: Revise cluster centers as mean of assigned observations

Now we've assigned all the points based on which cluster center they were closest to. Next, we need to update the cluster centers based on the points assigned to them. For instance, we can find the center mass of the blue cluster by summing over all the blue points and dividing by the total number of points, which is four here. And the resulted center mass C1', represented by a blue diamond, is our new center for the blue cluster. Similarly, we can find the new centers C2' and C3' for the green and red clusters.

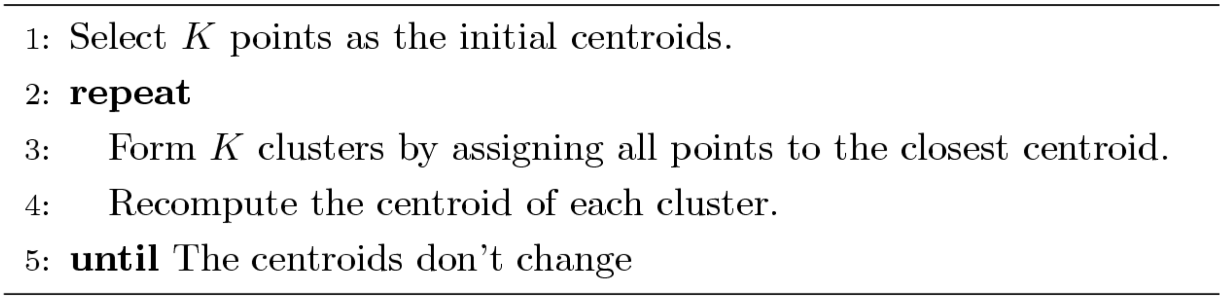


Step four: Repeat step 2 and step 3 until convergence

The last step of k-means is just to repeat the above two steps. For example, in this case, once C1', C2' and C3' are assigned as the new cluster centers, point D becomes closer to C3' and thus can be assigned to the red cluster. We keep on iterating between assigning points to cluster centers, and updating the cluster centers until convergence. Finally, we may get a solution like the following figure. Well done!



The basic algorithm is very simple



```
Select $K$ points as the initial centroids.
repeat
    Form $K$ clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change
```

45. What is the procedure of validating k-means clustering?

Ans:

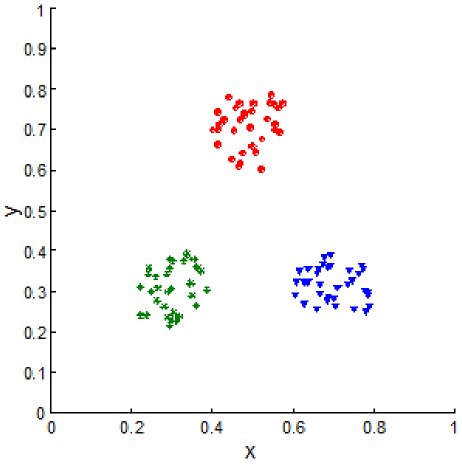
Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.

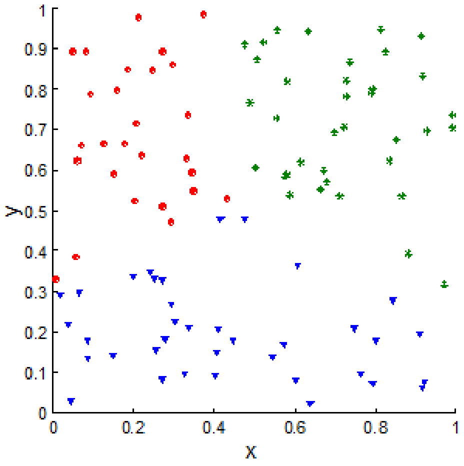
- External Index: Used to measure the extent to which cluster labels match externally supplied class labels.
    - Entropy
- Internal Index: Used to measure the goodness of a clustering structure *without* respect to external information.
    - Sum of Squared Error (SSE)
- Relative Index: Used to compare two different clusterings or clusters.
    - Often an external or internal index is used for this function, e.g., SSE or entropy

<u>Measuring Cluster Validity Via Correlation</u>

- Two matrices
    - Proximity Matrix
    - "Incidence" Matrix
        - One row and one column for each data point
        - An entry is 1 if the associated pair of points belong to the same cluster
        - An entry is 0 if the associated pair of points belongs to different clusters
- Compute the correlation between the two matrices
    - Since the matrices are symmetric, only the correlation between n(n-1) / 2 entries needs to be calculated.
- High correlation indicates that points that belong to the same cluster are close to each other.
- Not a good measure for some density or contiguity based clusters.

- Correlation of incidence and proximity matrices for the K-means clusterings of the following two data sets.



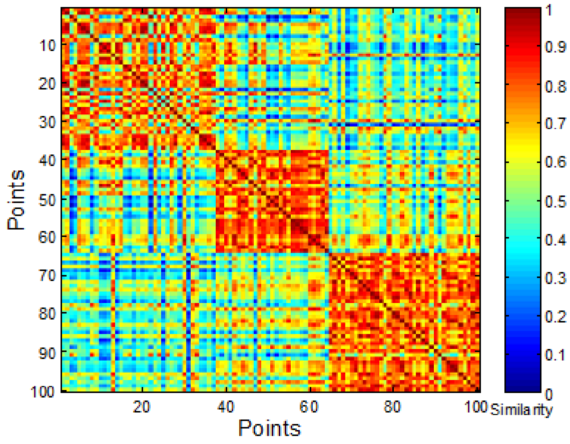


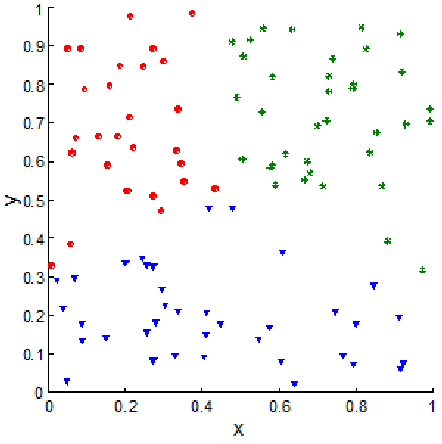
Corr = -0.9235

Corr = -0.5810

Using Similarity Matrix for Cluster Validation

- Clusters in random data are not so crisp





K-means

46. Write the steps of k-means clustering?

Ans:

Steps of k-means clustering:

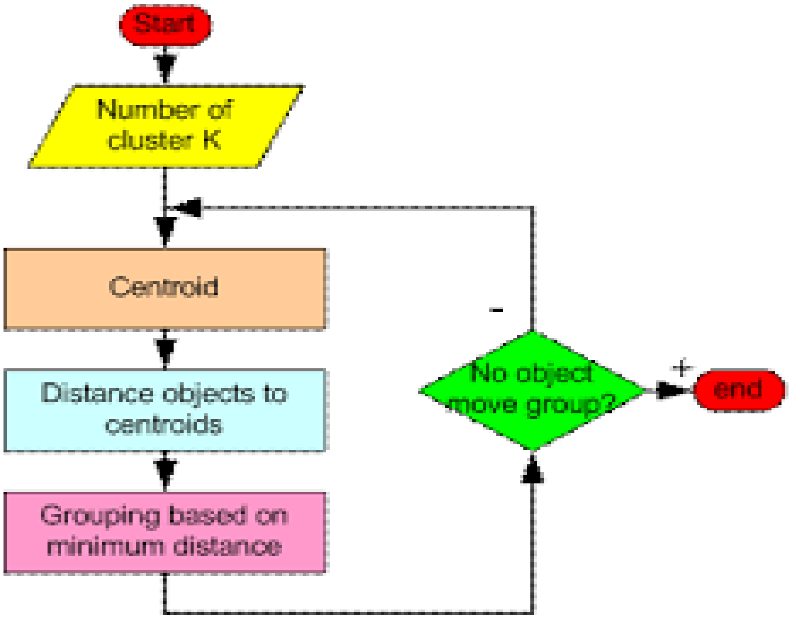
The basic step of k-means clustering is simple. In the beginning we determine number of cluster K and we assume the centroid or center of these clusters. We can take any random objects as the initial centroids or the first K objects in the sequence can also serve as the initial centroids.

Then the K means algorithm will do the three steps below until convergence:

Iterate until *stable* (=no object move group):

1. Determine the centroid coordinate
2. Determine the distance of each object to the centroids
3. Group the object based on minimum distance

The algorithm looks like this:


Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
-
No object move group?
+
end

48. Write two limitations of k-means clustering. How can you minimize these limitations?

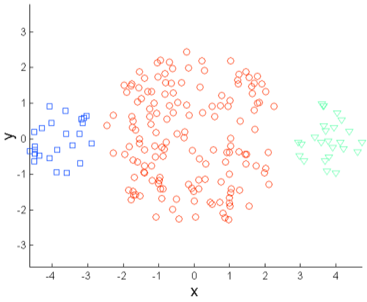
Ans:

Limitations of k-means clustering:

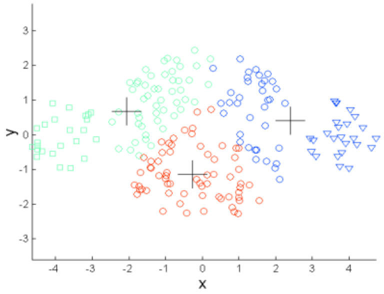
Simple k-means is one of the most known and used algorithms for clustering. One of the biggest advantages of k-means is that it is really easy to implement and even more important—most of the time we don’t even have to implement it ourselves. For most of the common programming languages used in data science an efficient implementation of k-means already exists. But this simplicity also comes with some limitations. Two of them are:

1. K-means has problems when clusters are of differing

   – Sizes

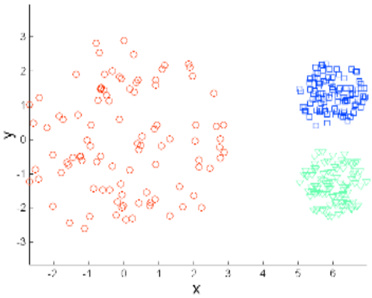


Original points

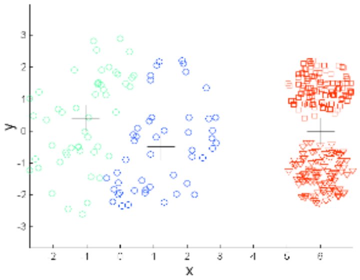


K-means (3 Clusters)

   – Densities

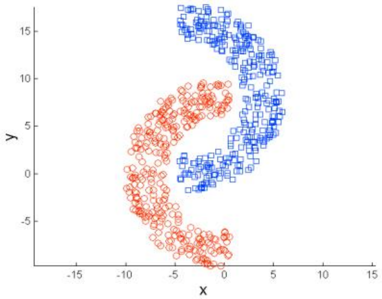


Original points

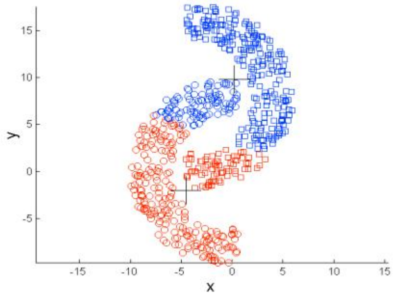


K-means (3 Clusters)

   – Non-globular shapes



Original points

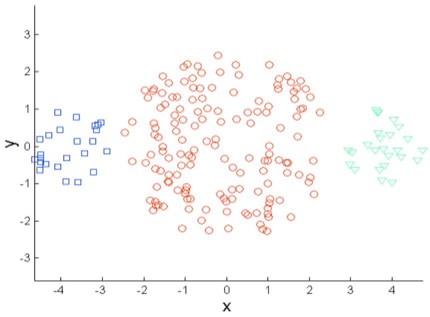


K-means (2 Clusters)

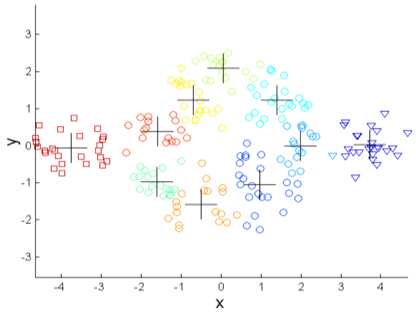
2. K-means has problems when the data contains outliers.

Minimization of limitations:

We can minimize the limitations by using many clusters.



Original Points



K-means Clusters

49. Explain with a pictorial example of Core Point, Noise Point and Border Point.

Ans:

Core Point:

A point is a core point if it has more than a specified number of points (MinPts) within Epsilon. These are points that are at the interior of a cluster. In order for a point to be considered a “core” point, it must contain the *minimum number of points* within epsilon distance.

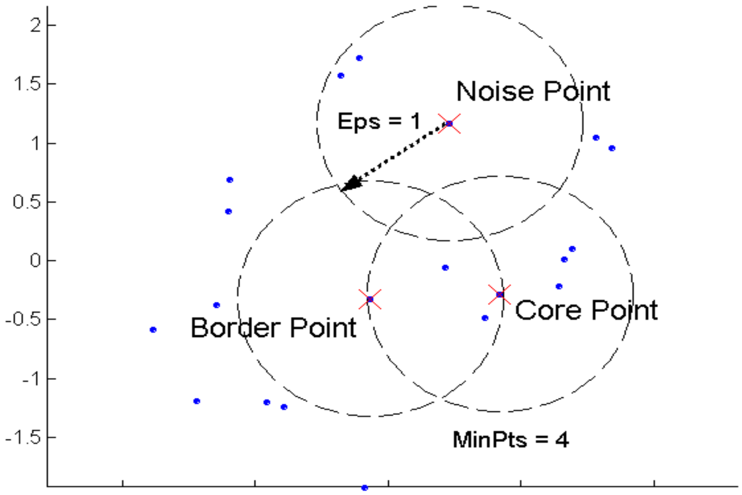
Border Point:

Border points are points that are part of a cluster, but not dense themselves (i.e. every cluster member that is *not* a core point). A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point.

Noise Point:

All points not reachable from any other point are noise points. In other word, noise point is any point that is not a core point or a border point.

A pictorial is given below:



50. In density based clustering, how do you select epsilon and distance? What are the logic of this process?
Ans:

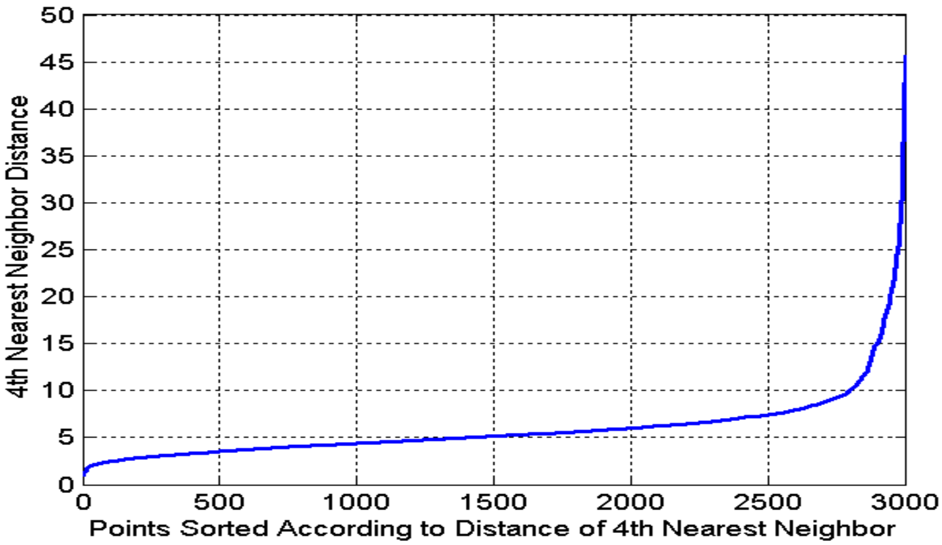
Epsilon selection:

The value for epsilon can be chosen by using a k-distance graph, plotting the distance to the *k* = *minPts*-1 nearest neighbor ordered from the largest to the smallest value. Good values of epsilon are where this plot shows an "elbow". If epsilon is chosen much too small, a large part of the data will not be clustered; whereas for a too high value of epsilon, clusters will merge and the majority of objects will be in the same cluster. In general, small values of ε are preferable and as a rule of thumb only a small fraction of points should be within this distance of each other. Alternatively, an OPTICS plot can be used to choose epsilon, but then the OPTICS algorithm itself can be used to cluster the data.

Distance selection:
The choice of distance function is tightly coupled to the choice of epsilon, and has a major impact on the results. In general, it will be necessary to first identify a reasonable measure of similarity for the data set, before the parameter epsilon can be chosen. There is no estimation for this parameter, but the distance functions need to be chosen appropriately for the data set. For example, on geographic data, the great-circle distance is often a good choice.

In summary-

- Idea is that for points in a cluster, their kth nearest neighbors are at roughly the same distance
- Noise points have the kth nearest neighbor at farther distance
- So, plot sorted distance of every point to its kth nearest neighbor
- Good values of epsilon are where this plot shows a strong bend



51. What is the basic principle of DBSCAN clustering?

Ans:

DBSCAN is a density-based algorithm, where “Density” is the number of points within a specified radius (Eps). In DBSCAN clusters that are dense in regions in the data space are separated by regions of lower density of points. The DBSCAN algorithm is based on this intuitive notion of “clusters” and “noise”. The basic principle of this clustering is to:

- Eliminate noise points.
- Perform clustering on the remaining points.

Two parameters:

1. maximum radius of the neighborhood→Eps

2. minimum number of points in an Eps neighborhood of a point→MinPts

$N_{Eps}(p)$:{q∈D s.t.dist(p,q)≤Eps}

52. What is the process of validating a density based clustering?
Ans:

- Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.
  - External Index: Used to measure the extent to which cluster labels match externally supplied class labels.
    - Entropy
  - Internal Index: Used to measure the goodness of a clustering structure *without* respect to external information.

        - ◆ Sum of Squared Error (SSE)
    - Relative Index: Used to compare two different clusterings or clusters.
        - ◆ Often an external or internal index is used for this function, e.g., SSE or entropy
- Sometimes these are referred to as criteria instead of indices
    - However, sometimes criterion is the general strategy and index is the numerical measure that implements the criterion.

53. When we need to use density based clustering?
Ans:
Density-based clustering is a well-known data clustering method that is commonly used in data mining and machine learning. Based on a set of points, density based clustering groups together points that are close to each other based on a distance measurement (usually Euclidean distance) and a minimum number of points. It also marks as outliers the points that are in low-density regions.

When to use:

- When the clusters are irregular or intertwined. Density based clustering can find arbitrarily shaped clusters. It can even find a cluster completely surrounded by (but not connected to) a different cluster.
- When noise and outliers are present.
- When we need to find non-linear shapes structure based on the density.
- When we want to avoid specifying the number of clusters in the data as priori.
- When we perform region queries.

55. Write a situation where DBSCAN clustering is appropriate.

Ans:
The DBSCAN clustering is a clustering method that is used to separate region with higher density than of the lower one. This clustering method works well when the clusters are irregular and when there is noise present. It is used to find associations and structures in data that are hard to find manually but that can be relevant and useful to find patterns and predict trends.

A situation where DBSCAN clustering is appropriate:

Suppose we have an e-commerce and we want to improve our sales by recommending relevant products to our customers. We don’t know exactly what our customers are looking for but based on a data set we can predict and recommend a relevant product to a specific customer. We can apply the DBSCAN to our data set (based on the e-commerce database) and find clusters based on the products that the users have bought. Using this clusters we can find similarities between customers, for example, the customer A have bought 1 pen, 1 book and 1 scissors and the customer B have bought 1 book and 1 scissors, then we can recommend 1 pen to the customer B. This is just a little example of use of DBSCAN, but it can be used in a lot of applications in several areas.

56. Write one application of DBSCAN clustering.

Ans: Density-based spatial clustering of applications with noise (DBSCAN) is a well-known data clustering algorithm that is commonly used in data mining and machine learning.

Application:

- Parameter estimation:

The parameter estimation is a problem for every data mining task. To choose good parameters we need to understand how they are used and have at least a basic previous knowledge about the data set that will be used.

eps: if the eps value chosen is too small, a large part of the data will not be clustered. It will be considered outliers because don't satisfy the number of points to create a dense region. On the other hand, if the value that was chosen is too high, clusters will merge and the majority of objects will be in the same cluster. The eps should be chosen based on the distance of the dataset (we can use a k-distance graph to find it), but in general small eps values are preferable.

minPoints: As a general rule, a minimum minPoints can be derived from a number of dimensions (D) in the data set, as minPoints ≥ D + 1. Larger values are usually better for data sets with noise and will form more significant clusters. The minimum value for the minPoints must be 3, but the larger the data set, the larger the minPoints value that should be chosen.

57. Data may affected by various kind of reasons. These reasons we may define as data quality problems. Answer the following questions: Explain these reasons with small examples.

Ans: Data is impacted by numerous processes, most of which affect its quality to a certain degree.

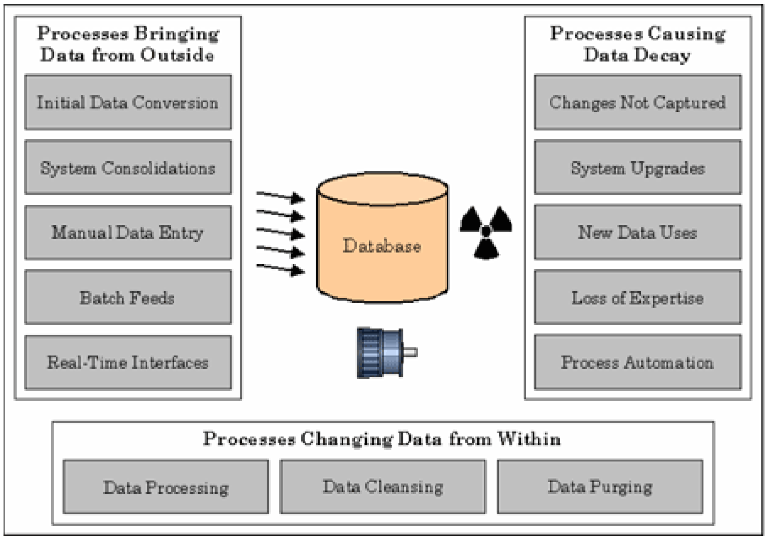


Figure 1 shows 13 categories of processes that cause the data problems which is defined as data quality, grouped into three high-level categories.

Here,I will explain some of the reasons of data quality problems with example:

- **Initial data conversion**

When I think of data conversion, my first association is with the mass extinction of dinosaurs. For 150 million years, dinosaurs ruled the earth. Then one day – BANG – a meteor came crashing down. Many animals died on impact, others never recovered and slowly disappeared in the ensuing darkness. It took millions of years for flora and fauna to recover. In the end, the formerly dominant dinosaurs were completely wiped out and replaced by the little furry creatures that later evolved into rats, lemurs, and the strange apes who find nothing better to do than write data quality books.

the data conversion is no different. Millions of unsuspecting data elements quietly do their daily work until – BANG – data conversion comes hurling at them. Much data never makes it to the new database. many of the lucky ones mutate so much in transition that they simply die out slowly in the aftermath. Most companies live with the consequences of bad data conversions for years or even decades.

- **System consolidations**

Database consolidations are the most common occurrence in the information technology landscape. They take place regularly when old systems are phased out or combined. consolidation adds the whole new dimension of complexity. First of all, the data is often merged into an existing non-empty database, whose structure can be changed little or none whatsoever. There are duplicates, there are overlaps in subject populations and data histories, and there are numerous data conflicts.

For instance, date of birth will be taken from System A if present, from System B otherwise, and from System C if it is missing in both A and B. This rarely works because it assumes that data on System A is always correct – a laughable assumption. To mitigate the problem, the winner-loser matrix is usually transformed into a complex conditional hierarchy.

- **Manual data entry**

Despite high automation, much data is typed into the databases by people through various forms and interfaces. The most common source of data inaccuracy is that the person manually entering the data just makes a mistake.

A common data entry problem is handling missing values. Users may assign the same blank value to various types of missing values. When "blank" is not allowed, users often enter meaningless value substitutes. Default values in data entry forms are often left untouched. The first entry in any list box is selected more often than any other entry.

- **Real-time interfaces**

More and more data is exchanged between the systems through real-time (or near real-time) interfaces. As soon as the data enters one database, it triggers procedures necessary to send transactions to other downstream databases. The advantage is immediate propagation of data to all relevant databases.

The basic problem is that data is propagated too fast. There is little time to verify that the data is accurate. At best, the validity of individual attributes is usually checked. Even if a data problem can be identified, there is often nobody at the other end of the line to react. The transaction must be either accepted or rejected (whatever the consequences). If data is rejected, it may be lost forever!

- **Data processing**

Data processing is at the heart of all operational systems. It comes in many shapes and forms – from regular transactions triggered by users to end-of-the-year massive calculations and adjustments. A subtle problem is when processing is accidentally done at the wrong time. Then the correct program may yield wrong results because the data is not in the state it is supposed to be. someone could then analyze the data quality implications of any changes in code, processes, data structure, or data collection procedures and thus eliminate unexpected data errors.For that reason, regular data processing inside the database will always be a cause of data problems.

- **Data cleansing**

The data quality topic has caught on in recent years, and more and more companies are attempting to cleanse the data. In the old days, cleansing was done manually and was rather safe. Data cleansing is dangerous mainly because data quality problems are usually complex and interrelated. Fixing one problem may create many others in the same or other related data elements. For instance, employment history is tightly linked with position history, pay rate history, and many other employment data attributes. Making corrections to any one of these data categories will make the data inconsistent with all other categories.

58. Explain different types of data that we face in data mining.

Ans:

data mining is not specific to one type of media or data. Data mining should be applicable to any kind of information repository. However, algorithms and approaches may differ when applied to different types of data. Indeed, the challenges presented by different types of data vary significantly.

Here are some examples in more detail:

- **Flat files:** Flat files are actually the most common data source for data mining algorithms, especially at the research level. Flat files are simple data files in text or binary format with a structure known by the data mining algorithm to be applied. The data in these files can be transactions, time-series data, scientific measurements, etc.
- **Relational Databases**: Briefly, a relational database consists of a set of tables containing either values of entity attributes, or values of attributes from entity relationships. Tables have columns and rows, where columns represent attributes and rows represent tuples. A tuple in a relational table corresponds to either an object or a relationship between objects and is identified by a set of attribute values representing a unique key. The most commonly

used query language for relational database is SQL, which allows retrieval and manipulation of the data stored in the tables, as well as the calculation of aggregate functions such as average, sum, min, max and count.

- **Data Warehouses**: A data warehouse as a storehouse, is a repository of data collected from multiple data sources (often heterogeneous) and is intended to be used as a whole under the same unified schema. A data warehouse gives the option to analyze data from different sources under the same roof. Let us suppose that OurVideoStore becomes a franchise in North America. Many video stores belonging to OurVideoStore company may have different databases and different structures. If the executive of the company wants to access the data from all stores for strategic decision-making, future direction, marketing, etc., it would be more appropriate to store all the data in one site with a homogeneous structure that allows interactive analysis. In other words, data from the different stores would be loaded, cleaned, transformed and integrated together.

- **Transaction Databases**: A transaction database is a set of records representing transactions, each with a time stamp, an identifier and a set of items. Associated with the transaction files could also be descriptive data for the items.

- **Multimedia Databases**: Multimedia databases include video, images, audio and text media. They can be stored on extended object-relational or object-oriented databases, or simply on a file system. Multimedia is characterized by its high dimensionality, which makes data mining even more challenging. Data mining from multimedia repositories may require computer vision, computer graphics, image interpretation, and natural language processing methodologies.

- **Spatial Databases**: Spatial databases are databases that, in addition to usual data, store geographical information like maps, and global or regional positioning. Such spatial databases present new challenges to data mining algorithms.

- **Time-Series Databases**: Time-series databases contain time related data such stock market data or logged activities. These databases usually have a continuous flow of new data coming in, which sometimes causes the need for a challenging real time analysis. Data mining in such databases commonly includes the study of trends and correlations between evolutions of different variables, as well as the prediction of trends and movements of the variables in time.

- **World Wide Web**: The World Wide Web is the most heterogeneous and dynamic repository available. A very large number of authors and publishers are continuously contributing to its growth and metamorphosis, and a massive number of users are accessing its resources daily. Data in the World Wide Web is organized in inter-connected documents. These documents can be text, audio, video, raw data, and even applications.

59. Explain, how do you discretize a numeric attribute? i.e. Income

Ans:

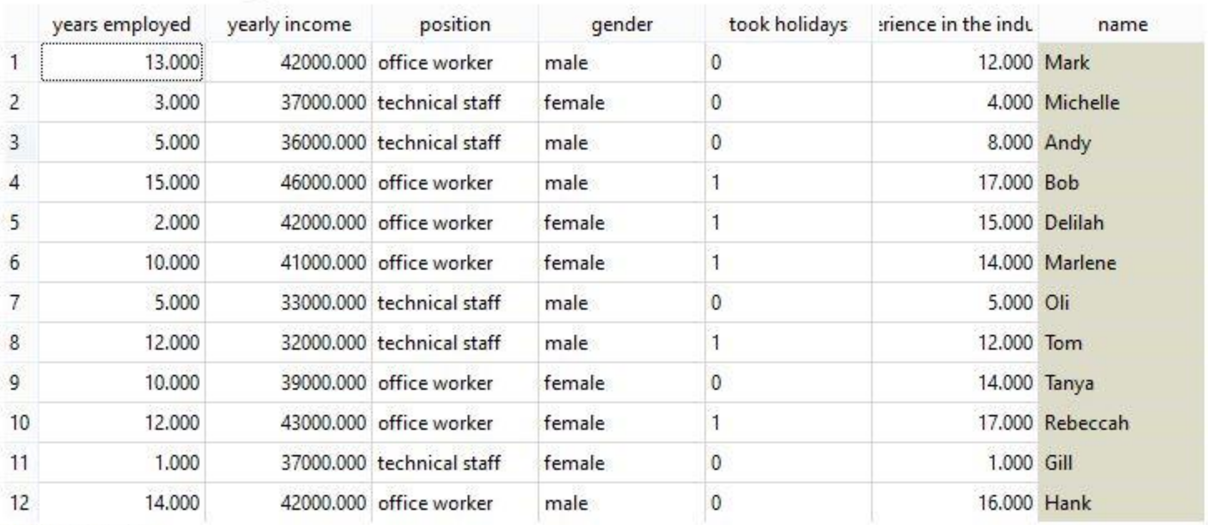
<u>Three types of attributes:</u>

Nominal — values from an unordered set
Ordinal — values from an ordered set
Continuous — real numbers

- Discretization: Divide the range of a continuous attribute into intervals because some data mining algorithms only accept categorical attributes.
- Some techniques: Binning methods – equal-width, equal-frequency
  Entropy based

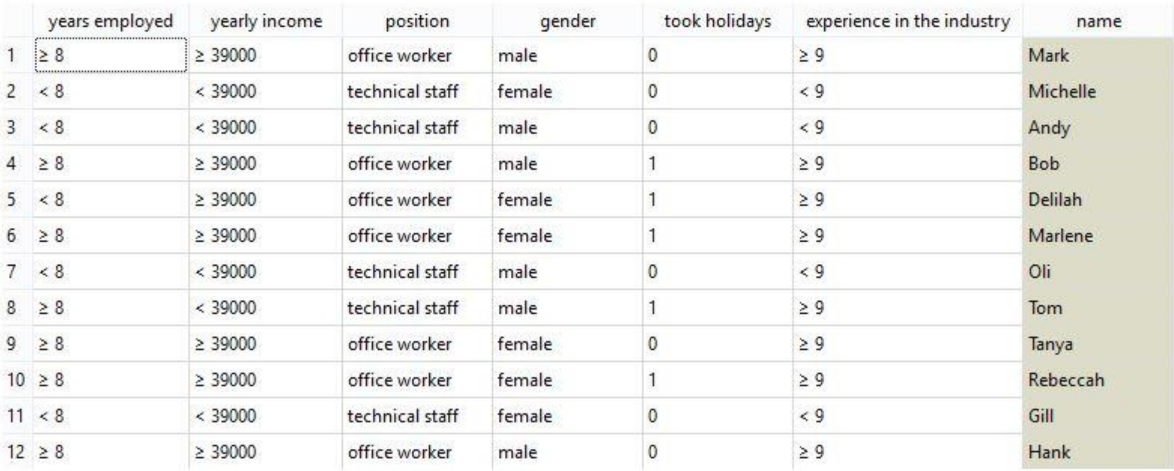
<u>Example:</u>

For certain tasks if I want to resort to binning, which is what Discretize does. It effectively distributes your continuous values into a selected number of bins, thus making the variable discrete-like.



	years employed	yearly income	position	gender	took holidays	erience in the indu	name
1	13.000	42000.000	office worker	male	0	12.000	Mark
2	3.000	37000.000	technical staff	female	0	4.000	Michelle
3	5.000	36000.000	technical staff	male	0	8.000	Andy
4	15.000	46000.000	office worker	male	1	17.000	Bob
5	2.000	42000.000	office worker	female	1	15.000	Delilah
6	10.000	41000.000	office worker	female	1	14.000	Marlene
7	5.000	33000.000	technical staff	male	0	5.000	Oli
8	12.000	32000.000	technical staff	male	1	12.000	Tom
9	10.000	39000.000	office worker	female	0	14.000	Tanya
10	12.000	43000.000	office worker	female	1	17.000	Rebeccah
11	1.000	37000.000	technical staff	female	0	1.000	Gill
12	14.000	42000.000	office worker	male	0	16.000	Hank

original data



	years employed	yearly income	position	gender	took holidays	experience in the industry	name
1	≥ 8	≥ 39000	office worker	male	0	≥ 9	Mark
2	< 8	< 39000	technical staff	female	0	< 9	Michelle
3	< 8	< 39000	technical staff	male	0	< 9	Andy
4	≥ 8	≥ 39000	office worker	male	1	≥ 9	Bob
5	< 8	≥ 39000	office worker	female	1	≥ 9	Delilah
6	≥ 8	≥ 39000	office worker	female	1	≥ 9	Marlene
7	< 8	< 39000	technical staff	male	0	< 9	Oli
8	≥ 8	< 39000	technical staff	male	1	≥ 9	Tom
9	≥ 8	≥ 39000	office worker	female	0	≥ 9	Tanya
10	≥ 8	≥ 39000	office worker	female	1	≥ 9	Rebeccah
11	< 8	< 39000	technical staff	female	0	< 9	Gill
12	≥ 8	≥ 39000	office worker	male	0	≥ 9	Hank

Discretized data with 'years employed' lower or higher then/equal to 8 (same for 'yearly income' and 'experience in the industry'.

60. How can we detect problems with the data?

Ans: We can detect problems with data if we encounter the following events in practice:

- **Missing Events**

Even if your data imported without any errors, there may still be problems with the data. For example, one typical problem is missing data. One type of missing data that you might encounter is missing events.

We can identify missing events in two ways.

❖ **Gaps in the timeline**

Check the timeline in the 'Events over time' chart to verify that there are no unusual gaps in the amount of events that occur over your log timeframe.

❖ **Unexpected amount of data**

We should have an idea about (roughly) how many rows or cases of data you are importing. Take a look at the Overview Statistics to see whether they match up with what you expect.

- **Missing Attribute Values**

One should have an idea of the kind of attributes that are expected in data. If one requests the data for all call center service requests for the Netherlands, Germany, and France from one month, but the volumes suggest that the data he got is mostly from the Netherlands that means data has missing attribute values.

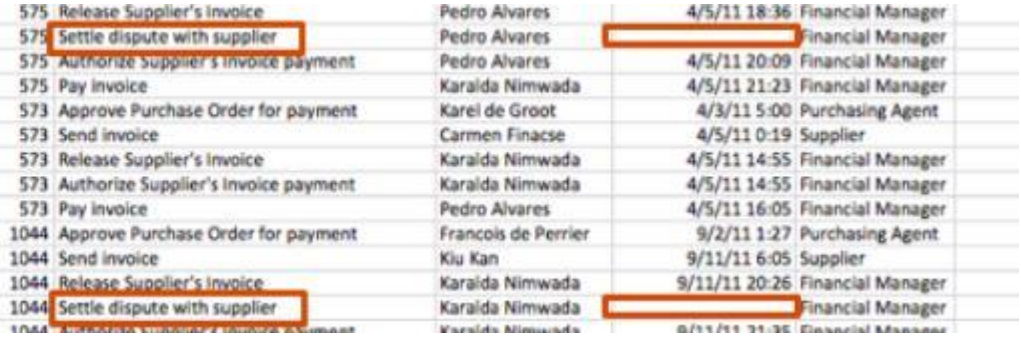
- **Missing Activities**

Some activities in your process may not be recorded in the data. For example, there may be manual activities (like a phone call) that people perform at their desk. These activities occur in the process but are not visible in the data.

- **Missing Timestamps**

In some situations, you may have information about whether or not an activity has occurred but you simply don’t have a timestamp.

For example, take a look at the data snippet from an invoice handling process in Figure 1. We can see that in some of the cases an activity ‘Settle dispute with supplier’ was performed. In contrast to all the other activities, this activity has no timestamp associated. It simply might not have been recorded by the system, or the information about this activity comes from a different system.



575	Release Supplier's Invoice	Pedro Alvares	4/5/11 18:36	Financial Manager
575	Settle dispute with supplier	Pedro Alvares		Financial Manager
575	Authorize Supplier's Invoice payment	Pedro Alvares	4/5/11 20:09	Financial Manager
575	Pay invoice	Karalda Nimwada	4/5/11 21:23	Financial Manager
573	Approve Purchase Order for payment	Karel de Groot	4/3/11 5:00	Purchasing Agent
573	Send invoice	Carmen Finacse	4/5/11 0:19	Supplier
573	Release Supplier's Invoice	Karalda Nimwada	4/5/11 14:55	Financial Manager
573	Authorize Supplier's Invoice payment	Karalda Nimwada	4/5/11 14:55	Financial Manager
573	Pay invoice	Pedro Alvares	4/5/11 16:05	Financial Manager
1044	Approve Purchase Order for payment	Francois de Perrier	9/2/11 1:27	Purchasing Agent
1044	Send invoice	Kiu Kan	9/11/11 6:05	Supplier
1044	Release Supplier's Invoice	Karalda Nimwada	9/11/11 20:26	Financial Manager
1044	Settle dispute with supplier	Karalda Nimwada		Financial Manager
1044	Authorize Supplier's Invoice payment	Karalda Nimwada	9/11/11 21:35	Financial Manager

62. If you have missing data and noise exist in your data then what are the steps you should take?

Ans:

There are several categories of missing data such as:

1.Missing at Random (MAR): Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data

2.Missing Completely at Random (MCAR): The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.

3. Missing not at Random (MNAR): Two possible reasons are that the missing value depends on the hypothetical value (e.g. People with high salaries generally do not want to reveal their incomes in surveys) or missing value is dependent on some other variable’s value.

Data missing handling process:

- **Mean imputation:**

In this method, mean of values of an attribute that contains missing data is used to fill in the missing values. In case of categorical attribute, the mode, which is the most frequent value, is used instead of mean.

- **Hot-Deck imputation:**

In this method, for each example that contains missing values, the most similar example is found, and the missing values are imputed from that example. The procedure is repeated until all missing values are successfully imputed or entire database is searched.

- **Regression Methods:**

This method assumes that the value of one variable changes in some linear way with other variables. The missing data are replaced by a linear regression function instead of replacing all missing data with a statistics.

- **K-Nearest Neighbor Imputation:**

This method uses k-nearest neighbor algorithms to estimate and replace missing data. The main advantages of this method are that: a) it can estimate both qualitative attributes and quantitative attributes; b) It is not necessary to build a predictive model for each attribute with missing data, even does not build visible models.

Noisy Data

Noise is a random error or variance in a measured variable. Noisy Data may be due to faulty data collection instruments, data entry problems and technology limitation.

How to Handle Noisy Data:

Binning:

Binning methods sorted data value by consulting its “neighbor- hood,” that is, the values around it. The sorted values are distributed into a number of “buckets,” or bins.

For example

Price = 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin a: 4, 8, 15

Bin b: 21, 21, 24

Bin c: 25, 28, 34

In this example, the data for price are first sorted and then partitioned into equal-frequency bins of size 3**.**

Smoothing by bin means:

Bin a: 9, 9, 9

Bin b: 22, 22, 22

Bin c: 29, 29, 29

In smoothing by bin means, each value in a bin is replaced by the mean value of the bin.

Smoothing by bin boundaries:

Bin a: 4, 4, 15

Bin b: 21, 21, 24

Bin c: 25, 25, 34

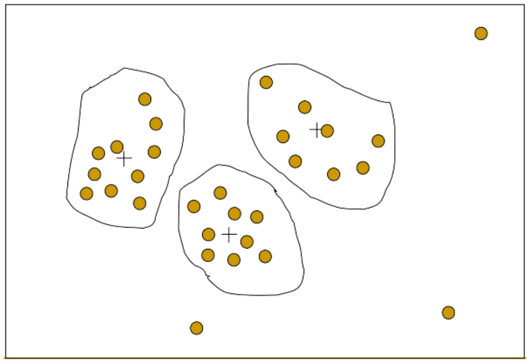
In smoothing by bin boundaries, each bin value is replaced by the closest boundary value.

Regression

Data can be smoothed by fitting the data into a regression functions.

Clustering:

Outliers may be detected by clustering, where similar values are organized into groups, or “clusters. Values that fall outside of the set of clusters may be considered outliers.



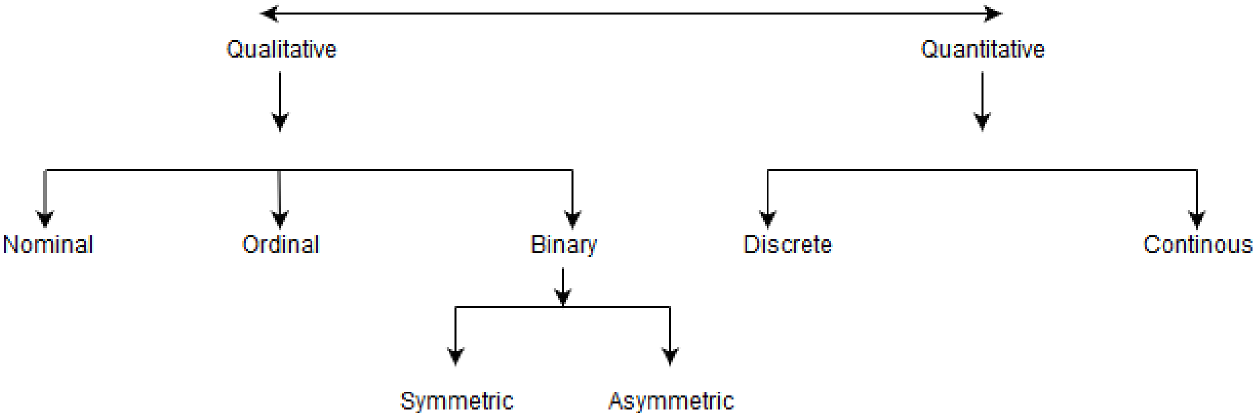
63. List different types of attributes with their general properties.

Ans:

It can be seen as a data field that represents characteristics or features of a data object. For a customer object attributes can be customer Id, address etc.

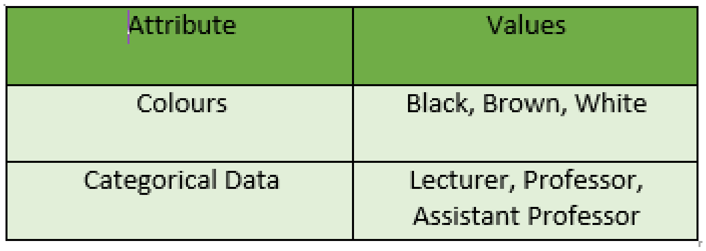
Type of attributes :

Here is description of attribute types.
1. Qualitative (Nominal (N), Ordinal (O), Binary(B)).
2. Quantitative (Discrete, Continuous)



Qualitative Attributes

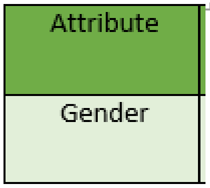
- **Nominal Attributes – related to names:**



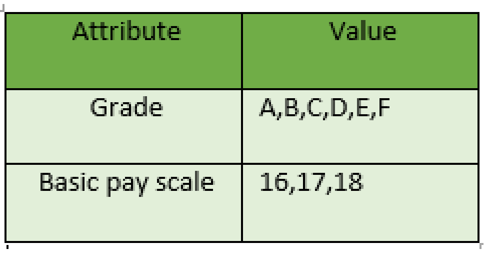
Attribute	Values
Colours	Black, Brown, White
Categorical Data	Lecturer, Professor, Assistant Professor

- **Binary Attributes :** Binary data has only 2 values/states:

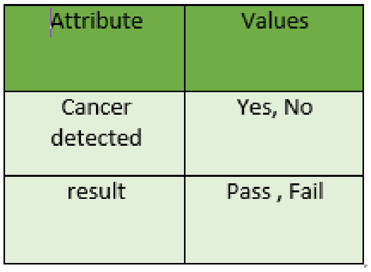
i) **Symmetric:** Both values are equally important (Gender).
ii) **Asymmetric :** Both values are not equally important (Result).



Attribute
Gender



Attribute	Value
Grade	A,B,C,D,E,F
Basic pay scale	16,17,18

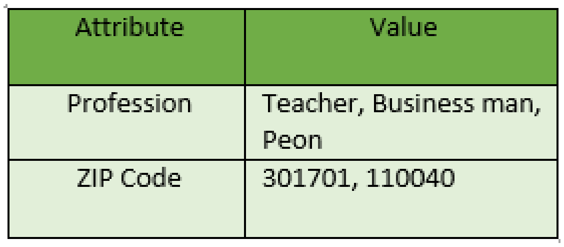


Attribute	Values
Cancer detected	Yes, No
result	Pass , Fail

- **Ordinal Attributes**

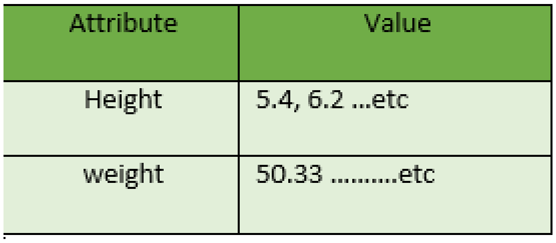
Quantitative Attributes

- **Discrete :** Discrete data have finite values it can be numerical and can also be in categorical form. These attributes has finite or countably infinite set of values. Example:



Attribute	Value
Profession	Teacher, Business man, Peon
ZIP Code	301701, 110040

- **<u>continuous</u>** : Continuous data have infinite no of states. Continuous data is of float type. There can be many values between 2 and 3. Example :



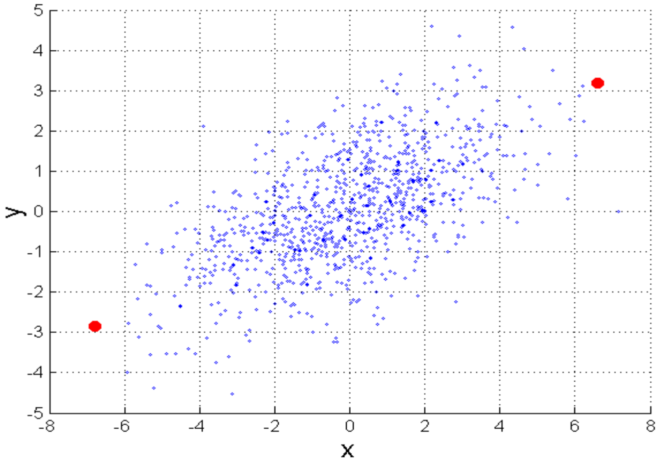
Attribute	Value
Height	5.4, 6.2 ...etc
weight	50.33 ....etc

64. To calculate dissimilarity between two data objects you can use Euclidian Distance and Mahalanobis Distance. Which one will you prefer and why?

Ans:

The most well-known distance used for numerical data is probably the Euclidean distance. This is a special case of the Minkowski distance when m = 2. Euclidean distance performs well when deployed to datasets that include compact or isolated clusters . Although Euclidean distance is very common in clustering, it has a drawback: if two data vectors have no attribute values in common, they may have a smaller distance than the other pair of data vectors containing the same attribute values.

Mahalanobis distance is a data-driven measure in contrast to Euclidean distances that are independent of the related dataset to which two data points belong . A regularized Mahalanobis distance can be used for extracting hyper ellipsoidal clusters . It is useful for detecting outliers



For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6

65. To calculate similarity or dissimilarity between two data objects which formulas you can use? Explain their differentials.

Ans:

Similarity and dissimilarity are important because they are used by a number of data mining techniques, such as clustering nearest neighbor classification and anomaly detection.

Definitions:

The **similarity** between two objects is a numeral measure of the degree to which the two objects are alike. Consequently, similarities are *higher* for pairs of objects that are more alike. Similarities are usually non-negative and are often between 0 (no similarity) and 1(complete similarity).

The **dissimilarity** between two objects is the numerical measure of the degree to which the two objects are different. Dissimilarity is *lower* for more similar pairs of objects.

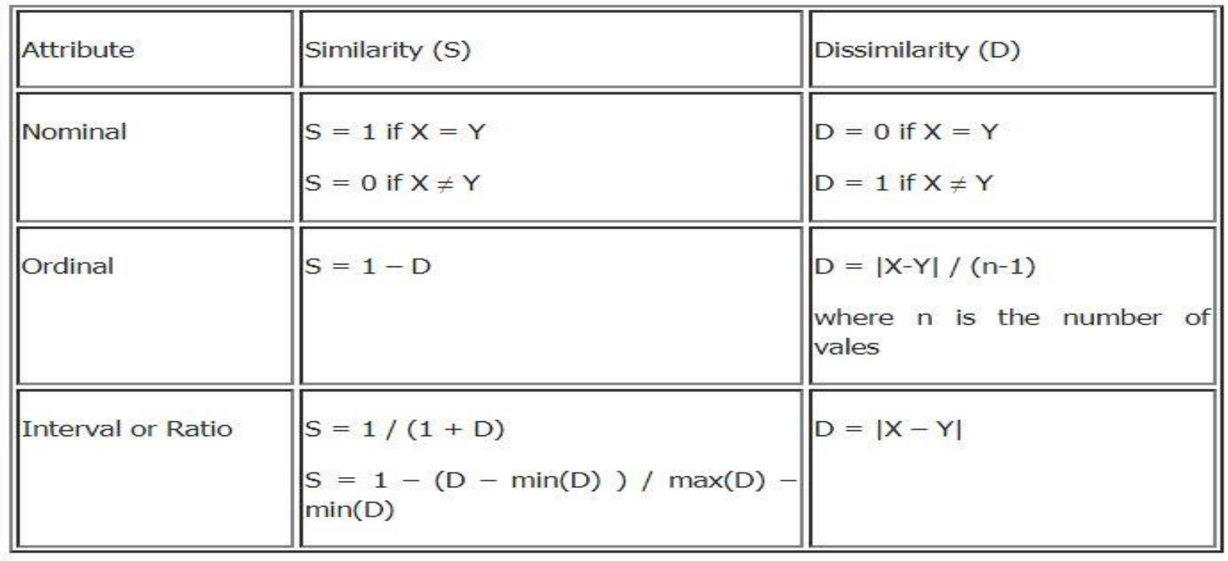
Let's consider simple objects with single attribute and discuss similarity, dissimilarity measures.

Similarity measures – a score that describe how much object are similar to each other. Similarity are measure that range from 0 to 1 [0,1]

Dissimilarity measures – a score that describe how much objects are dissimilarity to each other. Dissimilarity is measures that range from 0 to INF [0, Infinity]

Today there are variety of formulas for computing similarity and dissimilarity for simple objects and the choice of distance measures formulas that need to be used is determined by the type of attributes (Nominal, Ordinal, Interval or Ration) in the objects.

Below table summarizes the similarity and dissimilarity formulas for data objects.



Attribute	Similarity (S)	Dissimilarity (D)
Nominal	$S = 1$ if $X = Y$ $S = 0$ if $X \neq Y$	$D = 0$ if $X = Y$ $D = 1$ if $X \neq Y$
Ordinal	$S = 1 - D$	D = |X-Y| / (n-1) where n is the number of vales
Interval or Ratio	$S = 1 / (1 + D)$ $S = 1 - (D - \min(D)) / \max(D) - \min(D)$	D = |X – Y|

66. What are the different methods of calculating similarity and dissimilarity?

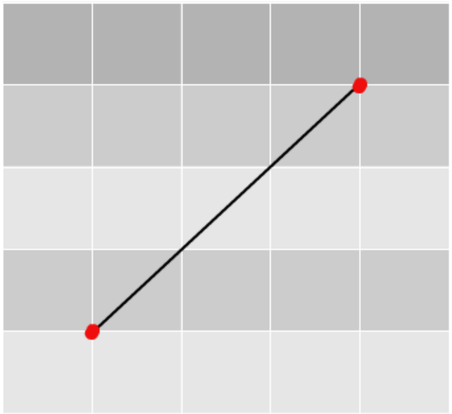
Ans:

Euclidean distance – Euclidean distance is a classical method helps compute distance between two objects A and B in Euclidean space (1- or 2- or n- dimension space). In Euclidean geometry, the distance between the points can be found by traveling along the line connecting the points. Inherently in the calculation you use the Pythagorean Theorem to compute the distance.

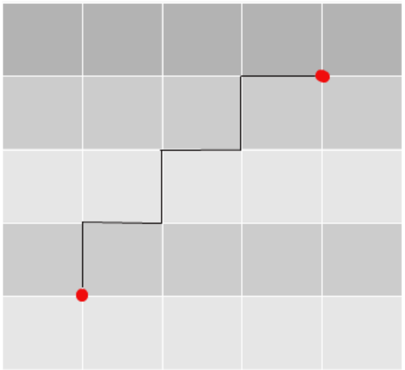
$$dist = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$

Taxicab or Manhattan distance – Similar to Euclidean distance between point A and B but only difference is the distance is calculated by traversing the vertical and horizontal line in the grid base system. Example, Manhattan distance used to calculate distance between two points that are geographically separated by the building blocks in the city.

$$d_t = |x_2 - x_1| + |y_2 - y_1|$$



Euclidean distance



Taxicab or Manhattan distance

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

Where r is a parameter.

When r =1 Minkowski formula tend to compute Manhattan distance.

When r =2 Minkowski formula tend to compute Euclidean distance.

When r =∞ Minkowski formula tend to compute Supremum.

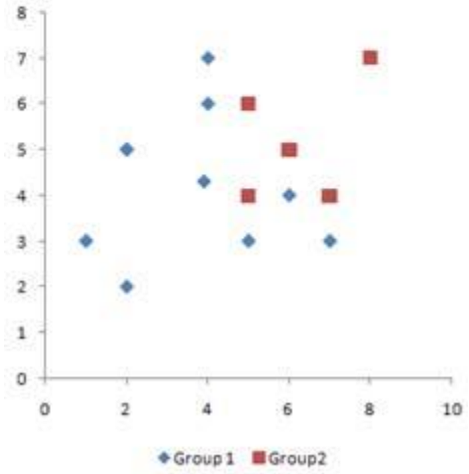
Cosine similarity

The cosine similarity between two vectors (or two documents on the Vector Space) is a measure that calculates the cosine of the angle between them. This metric is a measurement of orientation and not magnitude; it can be seen as a comparison between documents in terms of angle between them.

$$\text{similarity}(\vec{A}, \vec{B}) = \cos(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| * \|\vec{B}\|}$$

Mahalanobis distance

Mahalanobis distance is one such measure that used to measure the distance between the two groups of object. The idea of distance measure between two groups of objects can be represented graphically for better understanding.



Given with the data depicts the above picture, Mahalanobis distance can calculate distance between the Group1 and Group2. This type distance measure is helpful in classification and clustering.

$$D_{ij} = \sqrt{(X_i - X_j)' S^{-1} (X_i - X_j)}$$

66. What are the different methods of calculating similarity and dissimilarity?

Ans: Distance or similarity measures are essential in solving many pattern recognition problems such as classification and clustering. Various distance/similarity measures are available in the literature to compare two data distributions.

Similarity Measure

Numerical measure of how alike two data objects often fall between 0 (no similarity) and 1 (complete similarity).

Dissimilarity Measure

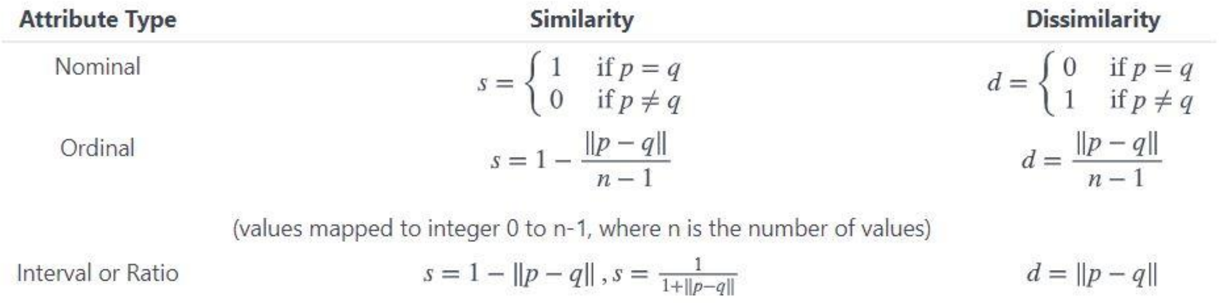
Numerical measure of how different two data objects are range from 0 (objects are alike) to ∞ (objects are different).

Proximity

Refers to a similarity or dissimilarity

Similarity/Dissimilarity for Simple Attributes

Here, *p* and *q* are the attribute values for two data objects.



Attribute Type	Similarity	Dissimilarity
Nominal	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$
Ordinal	$s = 1 - \frac{\lVert p-q \rVert}{n-1}$	$d = \frac{\lVert p-q \rVert}{n-1}$
	(values mapped to integer 0 to n-1, where n is the number of values)	
Interval or Ratio	$s = 1 - \lVert p-q \rVert$, $s = \frac{1}{1+\lVert p-q \rVert}$	$d = \lVert p-q \rVert$

Common Properties of Dissimilarity Measures

- $d(p, q) \geq 0$ for all *p* and *q*, and $d(p, q) = 0$ if and only if $p = q$,
- $d(p, q) = d(q,p)$ for all *p* and *q*,
- $d(p, r) \leq d(p, q) + d(q, r)$ for all *p*, *q*, and r, where $d(p, q)$ is the distance (dissimilarity) between points (data objects), *p* and *q*.

A distance that satisfies these properties is called a **metric**. Following is a list of several common distance measures to compare multivariate data. We will assume that the attributes are all continuous.

Euclidean Distance: The Euclidean distance between the *i*th and *j*th objects is

$$d_E(i,j) = \left(\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2 \right)^{\frac{1}{2}}$$

Minkowski Distance: The Minkowski distance is a generalization of the Euclidean distance. With the measurement, xik,i=1,…,N,k=1,…,p, the Minkowski distance is

$$d_M(i,j) = \left(\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^{\lambda} \right)^{\frac{1}{\lambda}}$$

Mahalnobis Distance: The Mahalnobis distance is

$$d_{MH}(i,j) = \left(\left(x_i - x_j \right)^T \Sigma^{-1} \left(x_i - x_j \right) \right)^{\frac{1}{2}}$$

Common Properties of Similarity Measures

- $s(p, q) = 1$ (or maximum similarity) only if $p = q$,
- $s(p, q) = s(q, p)$ for all p and q, where $s(p, q)$ is the similarity between data objects, p and q.

Similarity Between Two Binary Variables

The above similarity or distance measures are appropriate for continuous variables. However, for binary variables a different approach is necessary.

	q=1	**q=0**
p=1	$n_{1,1}$	$n_{1,0}$
p=0	$n_{0,1}$	$n_{0,0}$

Simple Matching and Jaccard Coefficients

- Simple matching coefficient =(n1,1+n0,0)/(n1,1+n1,0+n0,1+n0,0).
- Jaccard coefficient =n1,1/(n1,1+n1,0+n0,1).

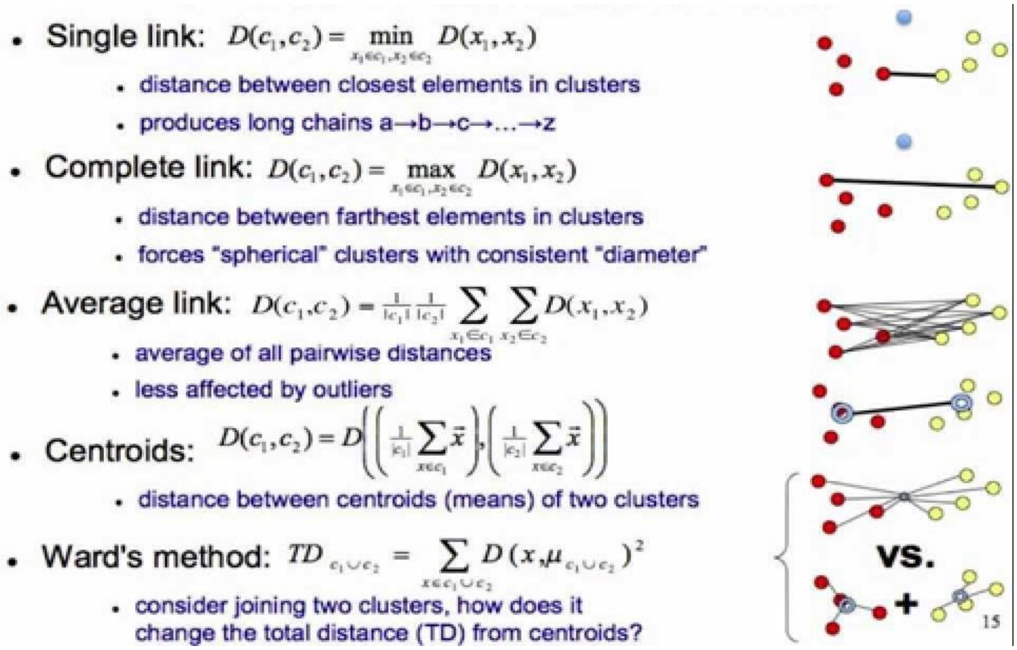
67. What are the different types of data set available? Give an example of each type.

Ans:
A dataset can be one of several different types. Dataset type is distinguished on the basis of data storage and structure.

Dataset Type	**Characteristics**	**Example**
File	a single file	AutoCAD DXF
Folder	a set of files in a single folder	Esri Shapefile
Database	a database	Oracle Spatial
Web	an Internet site	Web Feature Service (WFS)

68. How do you calculate distance between two clusters?

Ans: There are several method to calculate distance between two clusters. But the dominant methods are:



- Single link: $D(c_1,c_2) = \min_{x_1\in c_1, x_2\in c_2} D(x_1,x_2)$
  - distance between closest elements in clusters
  - produces long chains a→b→c→…→z
- Complete link: $D(c_1,c_2) = \max_{x_1\in c_1, x_2\in c_2} D(x_1,x_2)$
  - distance between farthest elements in clusters
  - forces "spherical" clusters with consistent "diameter"
- Average link: $D(c_1,c_2) = \frac{1}{|c_1|}\frac{1}{|c_2|}\sum_{x_1\in c_1}\sum_{x_2\in c_2} D(x_1,x_2)$
  - average of all pairwise distances
  - less affected by outliers
- Centroids: $D(c_1,c_2) = D\left(\left(\frac{1}{|c_1|}\sum_{x\in c_1}\vec{x}\right),\left(\frac{1}{|c_2|}\sum_{x\in c_2}\vec{x}\right)\right)$
  - distance between centroids (means) of two clusters
- Ward's method: $TD_{c_1\cup c_2} = \sum_{x\in c_1\cup c_2} D(x,\mu_{c_1\cup c_2})^2$
  - consider joining two clusters, how does it change the total distance (TD) from centroids?

70. Write procedure of hierarchical clustering with a data example.
Ans:

It presents an example of how to run a cluster analysis of the basketball superstars' data. The data are found in the BBall dataset. Here is the procedure of this data example.

1. Open the BBall dataset.
   - From the File menu of the Data window, select Open Example Data.
   - Click on the file BBall.
   - Click Open.
2. Open the Hierarchical Clustering / Dendrograms window.
   - Using the Analysis menu or the Procedure Navigator, find and select the Hierarchical Clustering / Dendrograms procedure.
   - On the menus, select File, then New Template. This will fill the procedure with the default template.
3. Specify the variables.
   - On the Hierarchical Clustering / Dendrograms window, select the Variables tab.
   - Double-click in the Interval Variables box. This will bring up the variable selection window.
   - Select Height, FgPct, Points, Rebounds from the list of variables and then click Ok. "Height, FgPct, Points, Rebounds" will appear in the Interval Variables box.
   - Double-click in the Label Variable box. This will bring up the variable selection window.
   - Select Player from the list of variables and then click Ok. "Player" will appear in the Label Variable box.
4. Specify the report.
   - On the Hierarchical Clustering / Dendrograms window, select the Reports tab.
   - Check the Distance Report. All reports should be selected.
5. Run the procedure.

- From the Run menu, select Run Procedure. Alternatively, just click the green Run button.

71. Write some applications of hierarchical clustering?
Ans:

- Recommendation engines
- Market segmentation
- Social network analysis
- Medical imaging
- Tracking viruses through polygenetic trees
- Charting evaluation through polygenetic trees.
- Image segmentation
- Anomaly detection
- Clustering of gene expression data
- Useful for seeing hierarchical structure, for relatively small data sets.

72. Write the algorithm of hierarchical clustering.
Ans: Hierarchical clustering algorithm is of two types:

i. Agglomerative Hierarchical clustering algorithm or AGNES (agglomerative nesting) and
ii. Divisive Hierarchical clustering algorithm or DIANA (divisive analysis).

Both this algorithm are exactly reverse of each other.

<u>Agglomerative Hierarchal Clustering Algorithm</u>

```
1. Compute the proximity matrix
2. Let each data point be a cluster
3. Repeat
4.         Merge the two closest clusters
5.         Update the proximity matrix
6. Until only a single cluster remains
```

<u>Divisive Hierarchical clustering Algorithm</u>

```
1. Compute a minimum spanning tree for the proximity graph.
2. Repeat
3. Create a new cluster by breaking the link corresponding to the largest
   distance (smallest similarity).
4. Until only singleton clusters remain
```

73. Write the process of hierarchical clustering in your own words.
Ans:

Hierarchical clustering can be done in two ways agglomerative and divisive hierarchical clustering which are exactly reverse of each other. Here I have written the procedure of agglomerative hierarchical clustering.
Let $X = \{x_1, x_2, x_3, ..., x_n\}$ be the set of data points.

1) Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.

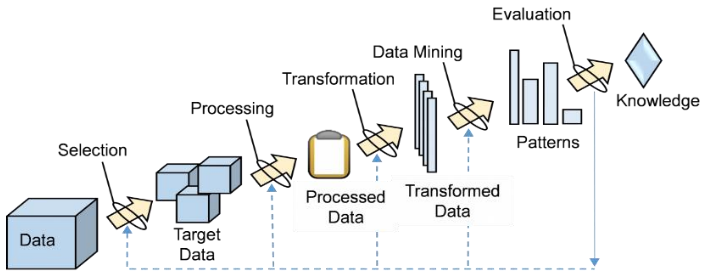
2) Find the least distance pair of clusters in the current clustering, say pair (r), (s), according to d[(r),(s)] = min d[(i),(j)] where the minimum is over all pairs of clusters in the current clustering.
3) Increment the sequence number: m = m +1.Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to L(m) = d[(r),(s)].
4) Update the distance matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The distance between the new cluster, denoted (r,s) and old cluster(k) is defined in this way: d[(k), (r,s)] = min (d[(k),(r)], d[(k),(s)]).
5) If all the data points are in one cluster then stop, else repeat from step 2).

74. Give an example of data where Data Mining techniques need to apply to extract hidden and unknown information.
Ans:
In commercial point of view there are many area where data mining techniques need to apply. Some of are:

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

Or: Consider a marketing head of telecom service provides who wants to increase revenues of long distance services. For high ROI on his sales and marketing efforts customer profiling is important. He has a vast data pool of customer information like age, gender, income, credit history, etc. But its impossible to determine characteristics of people who prefer long distance calls with manual analysis. Using data mining techniques, he may uncover patterns between high long distance call users and their characteristics.
For example, he might learn that his best customers are married females between the age of 45 and 54 who make more than $80,000 per year. Marketing efforts can be targeted to such demographic.

75. Define Data Mining? There are two types of Data mining techniques: Predictive and descriptive data mining- give example of these two.
Ans:
<u>Data Mining:</u>
Data mining is a non-trivial extraction of implicit, previously unknown and potentially useful information from data. In other words it is exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns.



Predictive data mining:
Use some variables to predict unknown or future values of other variables.
Example: Fraud Detection
–Goal: Predict fraudulent cases in credit card transactions.
–Approach:

- Use credit card transactions and the information on its account-holder as attributes.
  -When does a customer buy, what does he buy, how often he pays on time, etc
- Label past transactions as fraud or fair transactions. This forms the class attribute.
- Learn a model for the class of the transactions.
- Use this model to detect fraud by observing credit card transactions on an account.

Descriptive data mining:
Find human-interpretable patterns that describe the data.
Example: Document Clustering:
–Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
–Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
–Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

76. Define the term data mining. Give an example of predictive data mining.
Ans:
Data Mining:
Data mining is a non-trivial extraction of implicit, previously unknown and potentially useful information from data. In other words it is exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns.
Predictive data mining:
Use some variables to predict unknown or future values of other variables.
Example: Fraud Detection
–Goal: Predict fraudulent cases in credit card transactions.
–Approach:

- Use credit card transactions and the information on its account-holder as attributes.
  -When does a customer buy, what does he buy, how often he pays on time, etc
- Label past transactions as fraud or fair transactions. This forms the class attribute.
- Learn a model for the class of the transactions.
- Use this model to detect fraud by observing credit card transactions on an account.

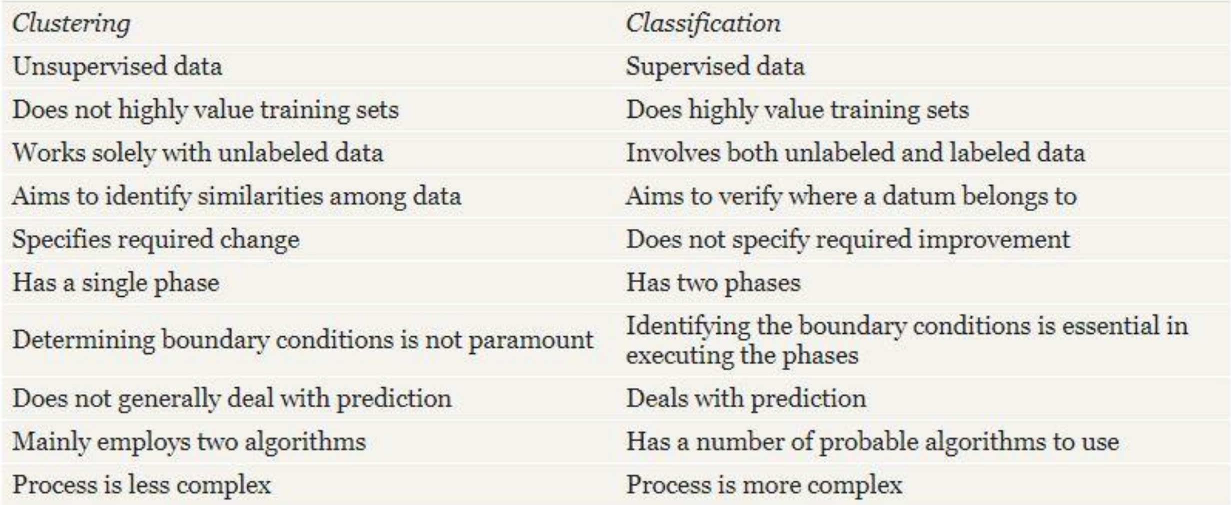
77. Non-trivial extraction of implicit, previously unknown and potentially useful information from data is called Data Mining. There are several tasks that we employ for mining; both classification and clustering. Answer the following questions: Give some examples of data where Data Mining techniques need to apply to extract hidden and unknown information.
Ans.

Applications	**Usage**
Communications	Data mining techniques are used in communication sector to predict customer behavior to offer highly targeted and relevant campaigns.
Insurance	Data mining helps insurance companies to price their products profitable and promote new offers to their new or existing customers.
Education	Data mining benefits educators to access student data, predict achievement levels and find students or groups of students which need extra attention. For example, students who are weak in math subject.
Manufacturing	With the help of Data Mining Manufacturers can predict wear and tear of production assets. They can anticipate maintenance which helps them reduce them to minimize downtime.
Banking	Data mining helps finance sector to get a view of market risks and manage regulatory compliance. It helps banks to identify probable defaulters to decide whether to issue credit cards, loans, etc.
Retail	Data Mining techniques help retail malls and grocery stores identify and arrange most sellable items in the most attentive positions. It helps store owners to come up with the offer which encourages customers to increase their spending.
Service Providers	Service providers like mobile phone and utility industries use Data Mining to predict the reasons when a customer leaves their company. They analyze billing details, customer service interactions, complaints made to the company to assign each customer a probability score and offers incentives.
E-Commerce	E-commerce websites use Data Mining to offer cross-sells and up-sells through their websites. One of the most famous names is Amazon, who uses Data mining techniques to get more customers into their eCommerce store.
Super Markets	Data Mining allows supermarkets develop rules to predict if their shoppers were likely to be expecting. By evaluating their buying pattern, they could find woman customers who are most likely pregnant. They can start targeting products like baby powder, baby shop, and diapers and so on.
Crime Investigation	Data Mining helps crime investigation agencies to deploy police workforce (where is a crime most likely to happen and when?), who to search at a border crossing etc.
Bioinformatics	Data Mining helps to mine biological data from massive datasets gathered in biology and medicine.

78. What are the difference between classification and clustering?

Ans:

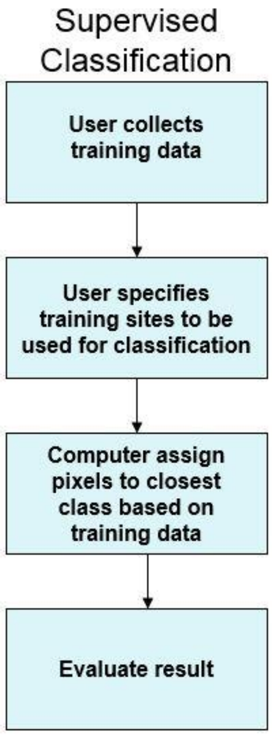
Classification and Clustering are the two types of learning methods which characterize objects into groups by one or more features. These processes appear to be similar, but there is a difference between them in context of data mining. The prior difference between classification and clustering is that classification is used in supervised learning technique where predefined labels are assigned to instances by properties, on the contrary, clustering is used in unsupervised learning where similar instances are grouped, based on their features or properties. The other differences are:



Clustering	*Classification*
Unsupervised data	Supervised data
Does not highly value training sets	Does highly value training sets
Works solely with unlabeled data	Involves both unlabeled and labeled data
Aims to identify similarities among data	Aims to verify where a datum belongs to
Specifies required change	Does not specify required improvement
Has a single phase	Has two phases
Determining boundary conditions is not paramount	Identifying the boundary conditions is essential in executing the phases
Does not generally deal with prediction	Deals with prediction
Mainly employs two algorithms	Has a number of probable algorithms to use
Process is less complex	Process is more complex

79. What do you mean by supervised and unsupervised classification?

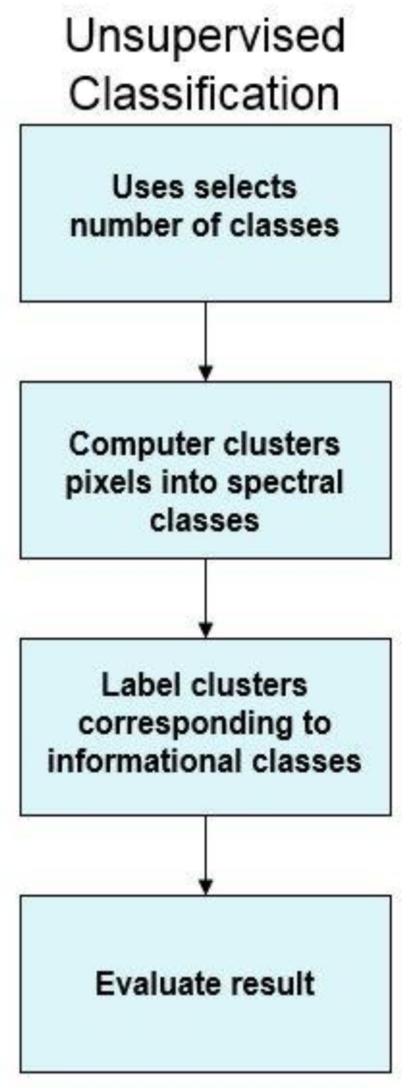
Ans:


Supervised Classification
User collects training data
User specifies training sites to be used for classification
Computer assign pixels to closest class based on training data
Evaluate result

Supervised classification:

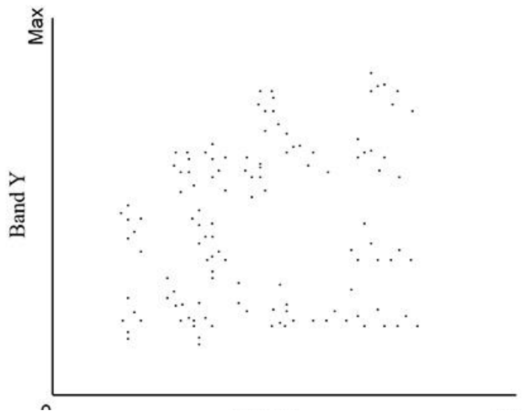
Supervised classification is the technique most often used for the quantitative analysis of remote sensing image data. At its core is the concept of segmenting the spectral domain into regions that can be associated with the ground cover classes of interest to a particular application. In practice those regions may sometimes overlap. A variety of algorithms is available for the task.

n supervised classification the user or image analyst "supervises" the pixel classification process. The user specifies the various pixels values or spectral signatures that should be associated with each class. This is done by selecting representative sample sites of known cover type called **Training Sites or Areas**. The computer algorithm then uses the spectral signatures from these training areas to classify the whole image. Ideally the classes should not overlap or should only minimally overlap with other classes.


Unsupervised Classification
Uses selects number of classes
Computer clusters pixels into spectral classes
Label clusters corresponding to informational classes
Evaluate result

Unsupervised classification:

Unsupervised classification is a form of pixel based classification and is essentially computer automated classification. The user specifies the number of classes and the spectral classes are created solely based on the numerical information in the data (the pixel values for each of the bands or indices). Clustering algorithms are used to determine the natural, statistical grouping of the data. The pixels are grouped together into based on their spectral similarity. The computer uses feature space to analyze and group the data into classes. Roll over the below image to see how the computer might use feature space to group the data into ten classes.



While the process is basically automated, the user has control over certain inputs. This includes the Number of Classes, the Maximum Iterations, (which is how many times the classification algorithm runs) and the Change Threshold %, which specifies when to end the classification procedure. After the data has been classified the user has to interpret, label and color code the classes accordingly.

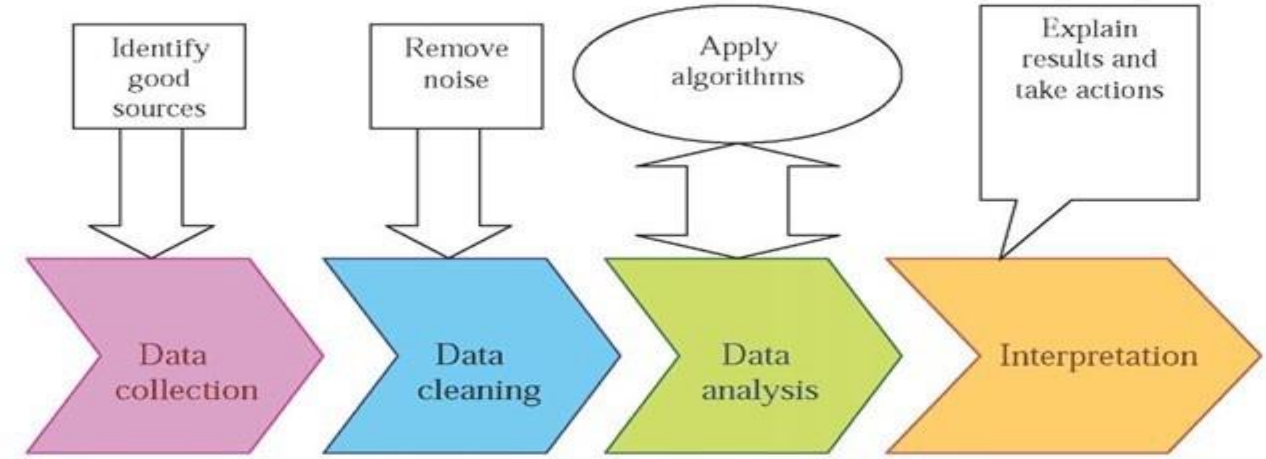
80. What is data mining and why is it an important discipline?
Ans:
Data mining:
Data mining is defined as a process used to extract usable data from a larger set of any raw data. It implies analysing data patterns in large batches of data using one or more software. Data mining has applications in multiple fields, like science and research. As an application of data mining, businesses can learn more about their customers and develop more effective strategies related to various business functions and in turn leverage resources in a more optimal and insightful manner. This helps businesses be closer to their objective and make better decisions.

Data mining involves effective data collection and warehousing as well as computer processing. For segmenting the data and evaluating the probability of future events, data mining uses sophisticated mathematical algorithms. Data mining is also known as Knowledge Discovery in Data (KDD).Data mining is not a simple task. It takes a certain amount of time and it requires a special procedure as well.

Data Mining Steps

The basic steps of data mining are follows

1. Data Collection
2. Data Cleaning
3. Data Analysis
4. Interpretation



Why an important discipline:

Data mining is an important process to discover knowledge about your customer behavior towards your business offerings. It explores the unknown credible patterns those are significant for business success.

Data mining has often misunderstood; people think that it includes only processing of data but is actually far more than this i.e. it covers advanced tools and technologies.

According to Doug Alexander of the University of Texas it is actually defined as the "computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of data".

With data mining, Business Organizations are able to make more accurate business decisions and incur more profits. From business, marketing advertising and introduction of new products or services, and everything in between. Data mining draws the results to:

- Improve customer loyalty

- Find hidden profitability

- Reduce Client Churn

Data mining has benefited most of the companies with products need to sell or not; medical researchers use the facts that are helpful with vaccines required to develop by analyzing recent disease patterns; assist engineers with highways need to be build & much more.

Data mining services has become integral process of every business to gain competitive edge in the business.

81. Why do we divide data in two parts before data mining starts?
Ans:
<u>Why:</u>
Before starting data mining data are divided into two parts called "Training set" and "Testing set". Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. Analysis Services randomly samples the data to help ensure that the testing and training sets are similar. By using similar data for training and testing, you can minimize the effects of data discrepancies and better understand the characteristics of the model.

After a model has been processed by using the training set, you test the model by making predictions against the test set. Because the data in the testing set already contains known values for the attribute that you want to predict, it is easy to determine whether the model's guesses are correct.

82. Write the list of predictive data mining. How anomaly detection is one kind of data mining?
Ans:
<u>Predictive Data Mining:</u>
Predictive data mining is a process that uses data mining and probability to forecast outcomes. Each model is made up of a number of predictors, which are variables that are likely to influence future results.
Following is the list of predictive data mining.

- Classification
- Regression
- Anomaly Detection

Anomaly detection is one kind of data mining because the goal of anomaly detection is to identify cases that are unusual within data that is seemingly homogeneous. Anomaly detection identifies data points atypical of a given distribution. In other words, it finds the outliers. Though simpler data analysis techniques than full-scale data mining can identify outliers, data mining anomaly detection techniques identify much more subtle attribute patterns and the data points that fail to conform to those patterns.

Anomaly detection is an important tool for detecting fraud, network intrusion, and other rare events that may have great significance but are hard to find.

Anomaly detection can be used to solve problems like the following:

A law enforcement agency compiles data about illegal activities, but nothing about legitimate activities. How can suspicious activity be flagged?

The law enforcement data is all of one class. There are no counter-examples.

An insurance agency processes millions of insurance claims, knowing that a very small number are fraudulent. How can the fraudulent claims be identified?

The claims data contains very few counter-examples. They are outliers.

83. What is Data Mining? Why is data mining important in our daily life?

Ans:
Data Mining:
We can simply define data mining as a process that involves searching, collecting, filtering and analyzing the data. It is important to understand that this is not the standard or accepted definition. But the above definition caters to the whole process.

A large amount of data can be retrieved from various websites and databases. It can be retrieved in form of data relationships, co-relations, and patterns. With the advent of computers, internet, and large databases it is possible to collect large amounts of data. The data collected may be analyzed steadily and help identify relationships and find solutions to the existing problems.

Governments, private companies, large organizations and all businesses are after a large volume of data collection for the purposes of business and research development. The data collected can be stored for future use. Storage of information is quite important whenever it is required. It is important to note that it may take a long time for finding and searching for information from websites, databases and other internet sources.

Why important:

The below-mentioned points explain why data mining is required.

1. Data mining is the procedure of capturing large sets of data in order to identify the insights and visions of that data. Nowadays, the demand of data industry is rapidly growing which has also increased the demands for Data analysts and Data scientists.
2. Since with this technique, we analyze the data and then convert that data into meaningful information. This helps the business to take accurate and better decisions in an organization.
3. Data mining helps to develop smart market decision, run accurate campaigns, predictions are taken and many more.
4. With the help of Data mining, we can analyze customer behaviors and their insights. This leads to great success and data-driven business.

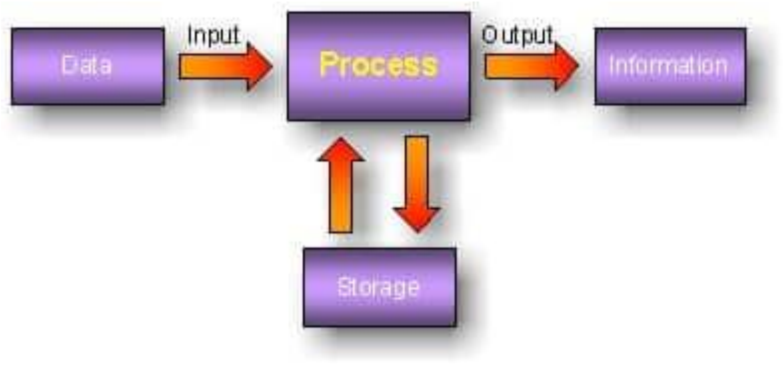
84. Before applying data mining techniques, data processing techniques need to apply. Explain some data processing techniques.

Ans:
Data processing is the conversion of data into usable and desired form. This conversion or “processing” is carried out using a predefined sequence of operations either manually or automatically. Most of the data processing is done by using computers and thus done automatically. The output or “processed” data can be obtained in different forms like image, graph, table, vector file, audio, charts or any other desired format depending on the software or method of data processing used.

Data Processing Techniques:

1. **Manual data processing -** In this method data is processed manually without the use of a machine, tool or electronic device. Data is processed manually, and all the calculations and logical operations are performed manually on the data.
2. **Mechanical data processing –** Data processing is done by use of a mechanical device or very simple electronic devices like calculator and typewriters. When the need for processing is simple, this method can be adopted.
3. **Electronic data processing –** Electronic data processing or EDP is the modern technique to process data. The data is processed through a computer; Data and set of instructions are given to the computer as input, and the computer automatically processes the data according to the given set of instructions. The computer is also known as electronic data processing machine. This method of processing data is very fast and accurate. For example, in a computerized education environment results of students are prepared through a computer; in banks, accounts of customers are maintained (or processed) through computers, etc.
4. **Batch Processing -** Batch Processing is a method where the information to be organized is sorted into groups to allow for efficient and sequential processing. Online Processing is a method that utilizes Internet connections and equipment directly attached to a computer. It is used mainly for information recording and research. Real-Time Processing is a technique that can respond almost immediately to various signals, to acquire and process information. Distributed Processing is commonly utilized by remote workstations connected to one big central workstation or server. ATMs are good examples of this data processing method.
5. **Online Processing -** This is a method that utilizes Internet connections and equipment directly attached to a computer. This allows for the data stored in one place and being used at an altogether different place. Cloud computing can be considered as an example which uses this type of processing. It is used mainly for information recording and research.
6. **Real-Time Processing -** This technique can respond almost immediately to various signals to acquire and process information. These involve high maintenance and upfront cost attributed to very advanced technology and computing power. Time saved is higher in this case as the output is seen in real time. For example in banking transactions
7. **Distributed Processing -** This method is commonly utilized by remote workstations connected to one big central workstation or server. ATMs are good examples of this data

processing method. All the end machines run on fixed software located at a particular place and make use of exactly same information and sets of instruction.



85. Explain different distance measures.
Ans:
For interval data the most common distance measure used is the Euclidean distance.

Euclidean distance:

- In general, if you have p variables X1,X2, . . . ,Xp measured on a sample of n subjects, the observed data for subject i can be denoted by xi1, xi2, . . . , xip and the observed data for subject j by xj1, xj2, . . . , xjp. The Euclidean distance between these two subjects is given by

$$dij(xi1-xj1)2+(xi2-xj2)2+....+(xip-xjp)2$$

Hierarchical agglomerative methods:

- **Nearest neighbour method (single linkage method):**In this method the distance between two clusters is defined to be the distance between the two closest members, or neighbours.
- **Furthest neighbour method (complete linkage method):**In this case the distance between two clusters is defined to be the maximum distance between members — i.e. the distance between the two subjects that are furthest apart.
- **Average (between groups) linkage method (sometimes referred to as UPGMA):** The distance between two clusters is calculated as the average distance between all pairs of subjects in the two clusters.
- **Centroid method:**Here the centroid (mean value for each variable) of each cluster is calculated and the distance between centroids is used. Clusters whose centroids are closest together are merged.
- **Ward’s method:**In this method all possible pairs of clusters are combined and the sum of the squared distances within each cluster is calculated. This is then summed over all clusters. The combination that gives the lowest sum of squares is chosen.

86. What is OLAP? Why do we need OLAP? Define the term “Slicing” and “Dicing”.
Ans:

OLAP : Online analytical processing, or OLAP, is an approach to answer multi-dimensional analytical (MDA) queries swiftly in computing. OLAP is part of the broader category of business intelligence, which also encompasses relational databases, report writing and data mining. Typical applications of OLAP include business reporting for sales, marketing, management reporting, business process management (BPM), budgeting and forecasting, financial reporting and similar areas, with new applications emerging, such as agriculture. The term *OLAP* was created as a slight modification of the traditional database term online transaction processing (OLTP)

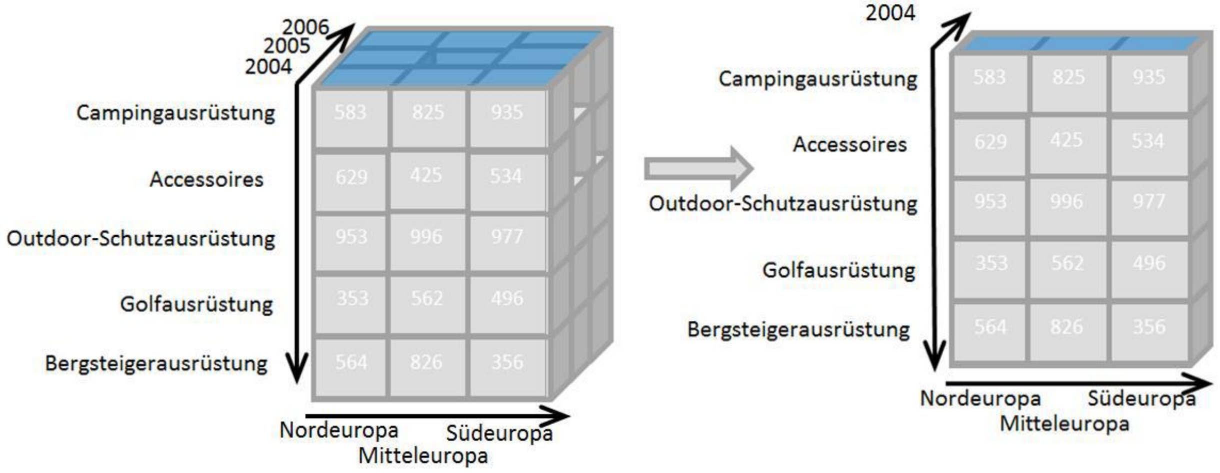
Why we need OLAP:

An effective OLAP solution solves problems for both business users and IT departments. For business users, it enables fast and intuitive access to centralized data and related calculations for the purposes of analysis and reporting. For IT, an OLAP solution enhances a data warehouse or other relational database with aggregate data and business calculations. In addition, by enabling business users to do their own analyses and reporting, OLAP systems reduce demands on IT resources.

- OLAP offers five key benefits:
- Business-focused multidimensional data
- Business-focused calculations
- Trustworthy data and calculations
- Speed-of-thought analysis
- Flexible, self-service reporting

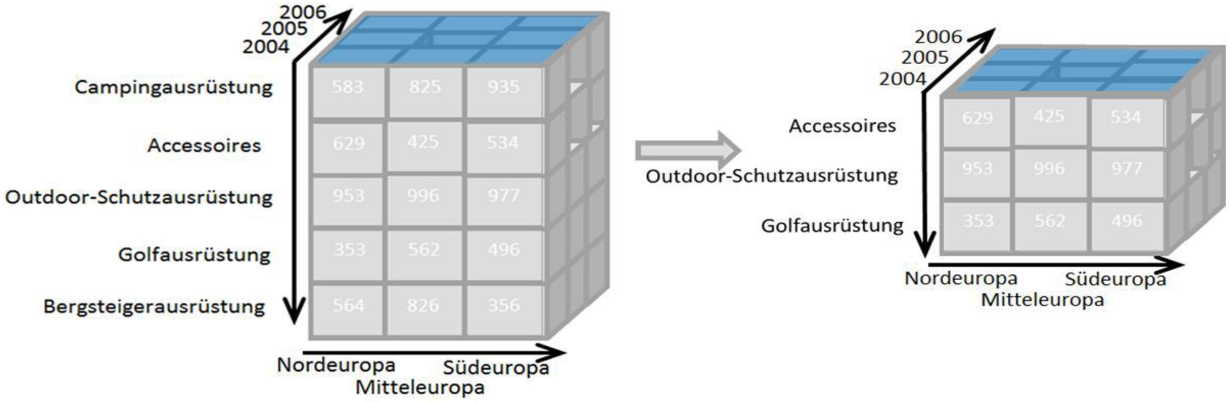
Slicing*:*

Slicing is the act of picking a rectangular subset of a cube by choosing a single value for one of its dimensions, creating a new cube with one fewer dimension.[5] The picture shows a slicing operation: The sales figures of all sales regions and all product categories of the company in the year 2005 and 2006 are "sliced" out of the data cube.



<u>Dicing:</u>

The dice operation produces a subcube by allowing the analyst to pick specific values of multiple dimensions The picture shows a dicing operation: The new cube shows the sales figures of a limited number of product categories, the time and region dimensions cover the same range as before.



87. When do we need to use discretization?

Ans:

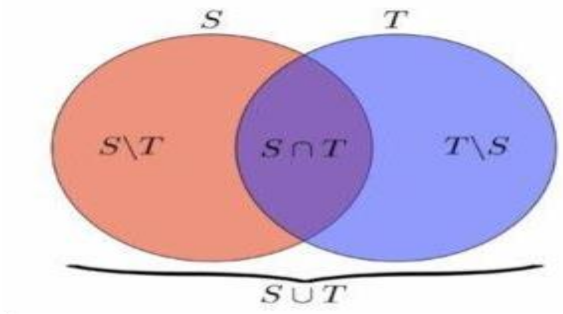
<u>When to use discretization:</u>

Often data are given in the form of continuous values. If their number is huge, model building for such data can be difficult. Moreover, many datamining algorithms operate only in discrete search or variable space. For instance, decision trees typically divide the values of a variable into two parts according to an appropriate threshold value. Many techniques apply computation of various criteria, for example, mutual information, or data mining algorithms that assume discrete values.

We use discretization to reduce the number of values a continuous variable assumes by grouping them into a number, b, of intervals or bins.

88. When will you use Jaccard Coefficient and Cossine Similarity Index?

Ans: Jaccard similarity and **cosine similarity** are two very common measurements while comparing item similarities and today, Similarity measures are used in various ways, examples include in plagiarism



Cosine Similarity

Cosine similarity measures the similarity between two vectors by taking the cosine of the angle the two vectors make in their dot product space. If the angle is zero, their similarity is one, the larger the angle is, the smaller their similarity. The measure is independent of vector length (the two vectors can even be of different length), which makes it a commonly used measure for high-dimensional spaces.

$$sim(\mathbf{x}_a, \mathbf{x}_b) = cos(\theta) = \frac{\mathbf{x}_a \cdot \mathbf{x}_b}{\|\mathbf{x}_a\|\|\mathbf{x}_b\|} = \frac{\sum_{i=1}^{d} \mathrm{x}_a^i \times \mathrm{x}_b^i}{\sqrt{\sum_{i=1}^{d} (\mathrm{x}_a^i)^2} \times \sqrt{\sum_{i=1}^{d} (\mathrm{x}_b^i)^2}}$$

Jaccard similarity

Jaccard similarity measures the similarity between two nominal attributes by tak- ing the intersection of both and divide it by their union. In terms of the above definitions this gives

$$J = \frac{A_{11}}{A_{01} + A_{10} + A_{11}}$$

A11= total number of binary values where both vectors have the value 1.
A01 = total number of binary values where first vector has value 1, other has value 0.
A10 = total number of binary values where first vector has value 0, other has value
A00 = total number of binary values where both vectors have the value 0

89. Why do we apply aggregation on data?

Ans:

Why: Data aggregation is any process in which information is gathered and expressed in a summary form, for purposes such as statistical analysis. A common aggregation purpose is to get more information about particular groups based on specific variables such as age, profession, or income. The information about such groups can then be used for Web site personalization to choose content and advertising likely to appeal to an individual belonging to one or more groups for which data has been collected. For example, a site that sells music CDs might advertise certain CDs based on the age of the user and the data aggregate for their age group. Online analytic processing (OLAP) is a simple type of data aggregation in which the marketer uses an online reporting mechanism to process the information.