Institute of Information Technology,
Jahangirnagar University,
Savar, Dhaka-1342, Bangladesh.

ASSIGNMENT ON:
Answering Questions.

COURSE TITLE: Data Mining

COURSE CODE: IT-5110

Submitted To:

Md. Fazlul Karim Patwary
Associate professor,
Institute of Information Technology,
Jahangirnagar University.

Submitted By:

Tarek Ahmed
M.Sc. (Session: 2017-18)
Roll no.- 1107

Submission date: 24.03.2019

1. Define the term "KNN" classification. Write two limitations of this classification.

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. Just for reference, this is "where" KNN is positioned in the algorithm list of scikit learn.

Limitation:

- Need to determine value of K (number of nearest neighbors)
- Distance based learning is not clear which type of distance to use and which attributes to use to produce the best results.

2. Define the term True Positive and False Negative.

Ans:

True Positive: A **true positive** test result is one that detects the condition when the condition is present.

False Negative: A **true negative** test result is one that does not detect the condition when the condition is absent.

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

Accuracy : Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, you have to look at other parameters to evaluate the performance of your model. For our model, we have got 0.803 which means our model is approx. 80% accurate.

Accuracy = TP+TN/TP+FP+FN+TN
Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.
Precision = TP/TP+FP

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes. The question recall answers is: Of all the passengers that truly survived, how many did we label? We have got recall of 0.631 which is good for this model as it's above 0.5.
Recall = TP/TP+FN

F1 score - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the

cost of false positives and false negatives are very different, it's better to look at both Precision and Recall. In our case, F1 score is 0.701.
F1 Score = 2*(Recall * Precision) / (Recall + Precision)

4. Define tree based and rule based classification.

Tree based learning: Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown in the diagram at right. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

Rule based classification: 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

Points to remember −

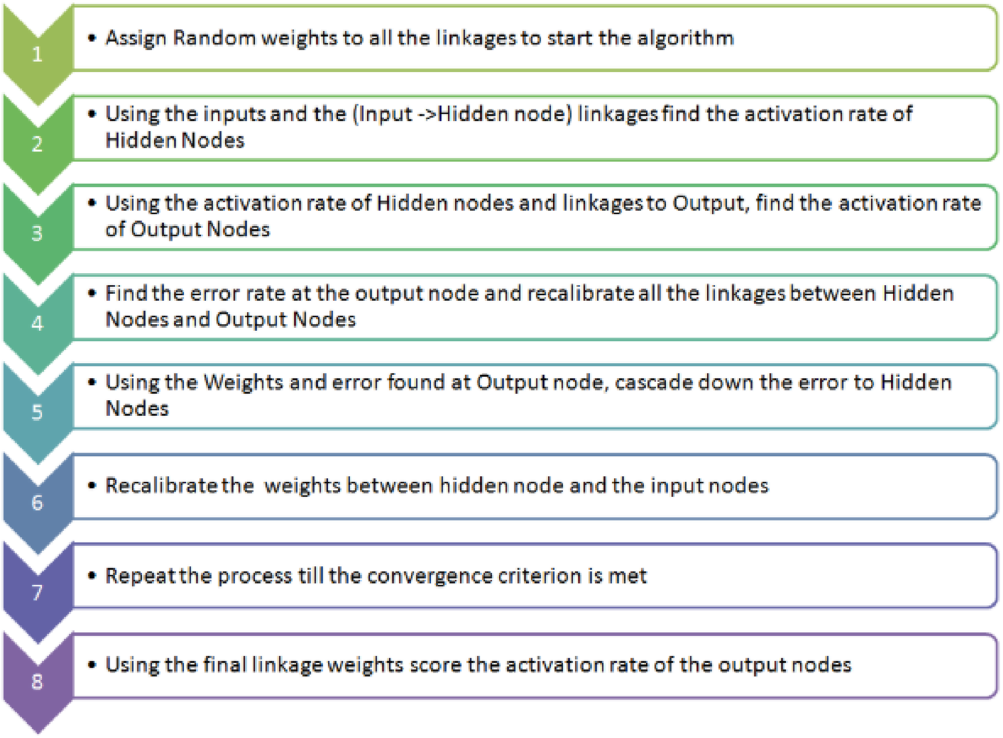
- The IF part of the rule is called **rule antecedent** or **precondition**.
- The THEN part of the rule is called **rule consequent**.
- The antecedent part the condition consist of one or more attribute tests and these tests are logically ANDed.
- The consequent part consists of class prediction.

5. Write the process of classification of any one.

Ans: Nearest neighbor classifier: • Requires three things- – The set of stored records – Distance Metric to compute distance between records – s – To classify an unknown record: – 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) • Compute distance between two points: – Euclidean distance – 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

6. How ANN classifier works?

Following is the framework in which artificial neural networks (ANN) work:



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: Bayes theorem named after Rev. Thomas Bayes. It works on conditional probability. Conditional probability is the probability that something will happen, given that something else has already occurred. Using the conditional probability, we can calculate the probability of an event using its prior knowledge.

Below is the formula for calculating the conditional probability.

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

- □ Consider each attribute and class label as random variables
- □ Given a record with attributes ($A_1, A_2,\ldots,A_n$)
    - ▣ Goal is to predict class C
    - ▣ Specifically, we want to find the value of C that maximizes $P(C| A_1, A_2,\ldots,A_n)$
    - ▣ Can we estimate $P(C| A_1, A_2,\ldots,A_n)$ directly from data?
- □ Approach:
    - ▣ compute the posterior probability $P(C | A_1, A_2, \ldots, A_n)$ for all values of C using the Bayes theorem
    - ▣ Choose value of C that maximizes $P(C | A_1, A_2, \ldots, A_n)$
    - ▣ Equivalent to choosing value of C that maximizes $P(A_1, A_2, \ldots, A_n|C) P(C)$

8. How do you perform KNN?

Ans: The steps are given below:

1. Determine K = number of nearest neighbors

2. Calculate the distance between the query-instance and all the training samples

3. Sort the distance and determine nearest neighbors based on the k-th minimum distance

4. Gather the category of the nearest neighbors

5. Use sample majority of the category of nearest neighbors as the prediction value of the query instance

9. How do you validate a classification model?

Ans: The steps are given below:

1. Choose several appropriate models/algorithms.
2. Split the training set into a smaller training set and validation set. The split depends on the problem at hand and the nature of the data.
3. Tune the parameters of each model by cross-validation on the smaller training set.
4. Choose the best parameter set for each model based on the the point above.
5. Test each model separately on the validation set.
6. Choose the best model according to your metric (accuracy is not the best in most cases, although it depends).

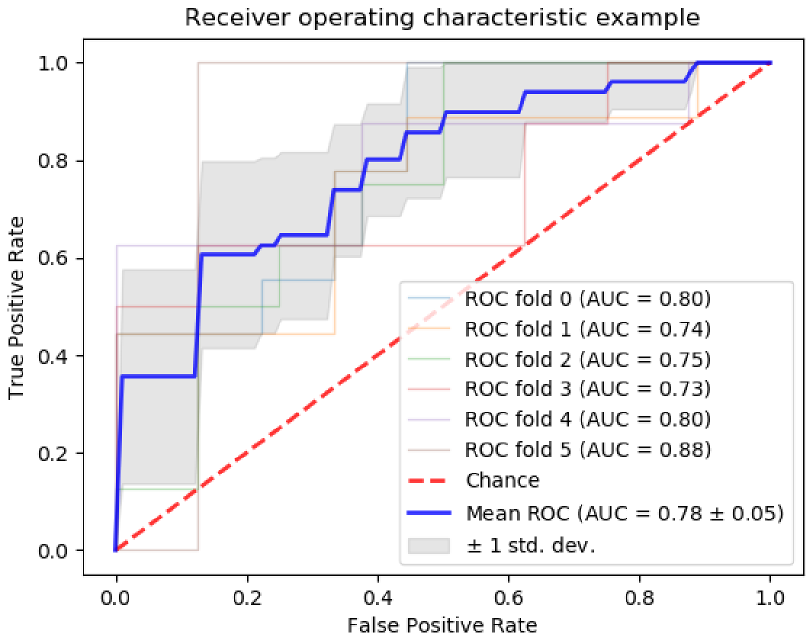
10. What are the functions of ROC curve in validation?

Ans: Example of Receiver Operating Characteristic (ROC) metric to evaluate classifier output quality using cross-validation.

ROC curves typically feature true positive rate on the Y axis, and false positive rate on the X axis. This means that the top left corner of the plot is the “ideal” point - a false positive rate of zero, and a true positive rate of one. This is not very realistic, but it does mean that a larger area under the curve (AUC) is usually better.

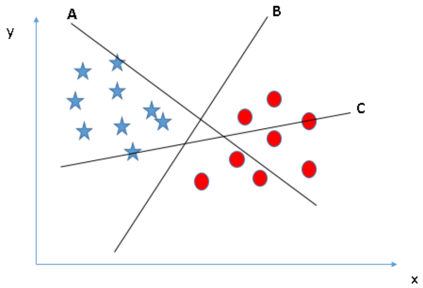
The “steepness” of ROC curves is also important, since it is ideal to maximize the true positive rate while minimizing the false positive rate.

This example shows the ROC response of different datasets, created from K-fold cross-validation. Taking all of these curves, it is possible to calculate the mean area under curve, and see the variance of the curve when the training set is split into different subsets. This roughly shows how the classifier output is affected by changes in the training data, and how different the splits generated by K-fold cross-validation are from one another.

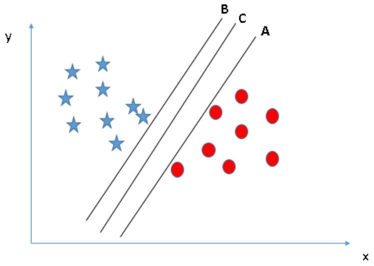


11. How SVM classifier works?

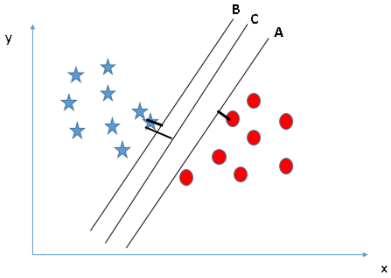
- **Ans: Identify the right hyper-plane (Scenario-1):** Here, we have three hyper-planes (A, B and C). Now, identify the right hyper-plane to classify star and circle.



- 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. Now, How can we identify the right hyper-plane?

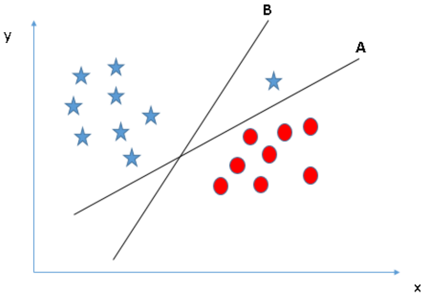


Here, maximizing the distances between nearest data point (either class) and hyper-plane will help us to decide the right hyper-plane. This distance is called as **Margin**. Let's look at the below snapshot:



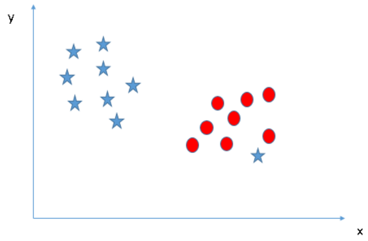
Above, you can see that the margin for hyper-plane C is high as compared to both A and B. Hence, we name the right hyper-plane as C. Another lightning reason for selecting the hyper-plane with higher margin is robustness. If we select a hyper-plane having low margin then there is high chance of miss-classification.

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

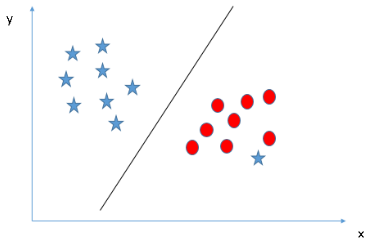


Some of you may have selected the hyper-plane **B** as it has higher margin compared to **A.** But, here is the catch, SVM selects the hyper-plane which classifies the classes accurately prior to maximizing margin. Here, hyper-plane B has a classification error and A has classified all correctly. Therefore,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

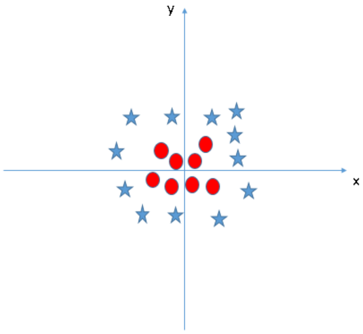
- **Can we classify two classes (Scenario-4)?:** Below, I am unable to segregate the two classes using a straight line, as one of star lies in the territory of other(circle) class as an outlier.



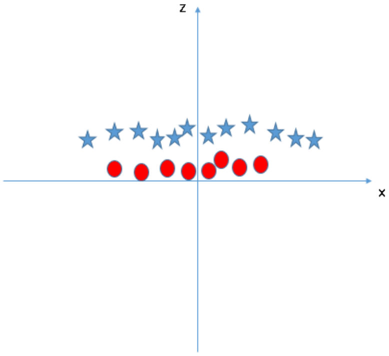
- As I have already mentioned, one star at other end is like an outlier for star class. SVM has a feature to ignore outliers and find the hyper-plane that has maximum margin. Hence, we can say, SVM is robust to outliers.



- **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:

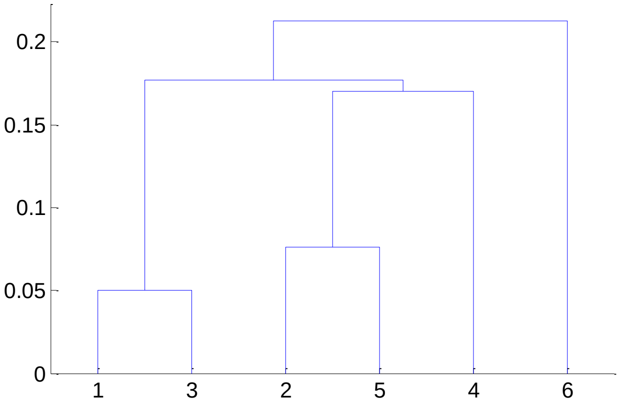


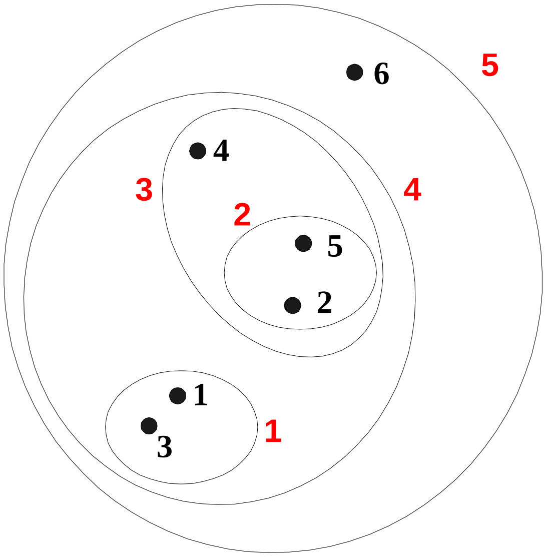
In above plot, points to consider are:

- 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.

12. Math on Draw dendrogram for hierarchical clustering.

Ans: Dendrogram for hierarchical clustering





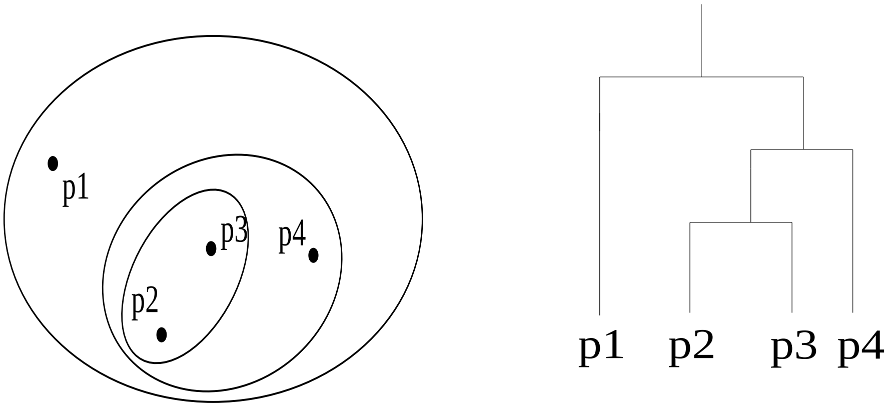
13. Write the process of classifying this new data using decision tree classification (Hunt’s Algorithm).

Ans: Hunt's algorithm grows a decision tree in a recursive fashion by partitioning the trainig records into successively purer subsets. Let Dt be the set of training records that reach a node t. The general recursive procedure is defined as below:

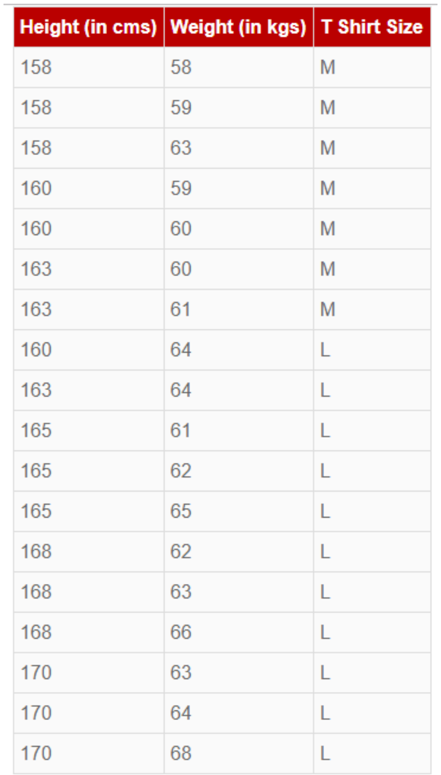
1. If Dt contains records that belong the same class yt, then t is a leaf node labeled as yt
2. If Dt is an empty set, then t is a leaf node labeled by the default class, yd
3. If Dt contains records that belong to more than one class, use an attribute test to split the data into smaller subsets.

It recursively applies the procedure to each subset until all the records in the subset belong to the same class. The Hunt's algirithm assumes that each combination of attribute sets has a unique class label during the procedure. If all the records associated with Dt have identical attribute values except for the class label, then it is not possible to split these records any future. In this case, the node is decalred a leaf node with the same class label as the majority class of training records associated with this node.

14. Write the process of classifying this new data using ensemble method.
Ans:



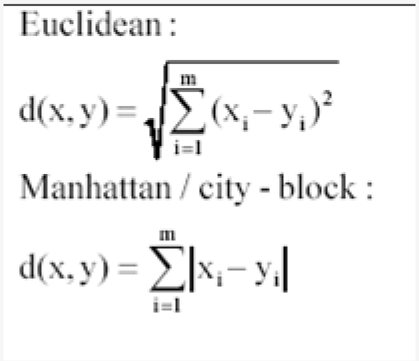
15. Write the process of classifying this new data using KNN.
Ans: Suppose we have height, weight and T-shirt size of some customers and we need to predict the T-shirt size of a new customer given only height and weight information we have. Data including height, weight and T-shirt size information is shown below



Height (in cms)	Weight (in kgs)	T Shirt Size
158	58	M
158	59	M
158	63	M
160	59	M
160	60	M
163	60	M
163	61	M
160	64	L
163	64	L
165	61	L
165	62	L
165	65	L
168	62	L
168	63	L
168	66	L
170	63	L
170	64	L
170	68	L

Step 1: Calculate similarity based onn distance function

There are many distance functions but Euclidean is the most commonly used measure. It is mainly used when data is continuous. Manhattan distance is also very common for continuous variables.



Euclidean :

$$d(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

Manhattan / city - block :

$$d(x, y) = \sum_{i=1}^{m} |x_i - y_i|$$

Distance Functions

The idea to use distance measure is to find the distance (similarity) between new sample and training cases and then finds the k-closest customers to new customer in terms of height and weight.

New customer named 'Monica' has height 161cm and weight 61kg.

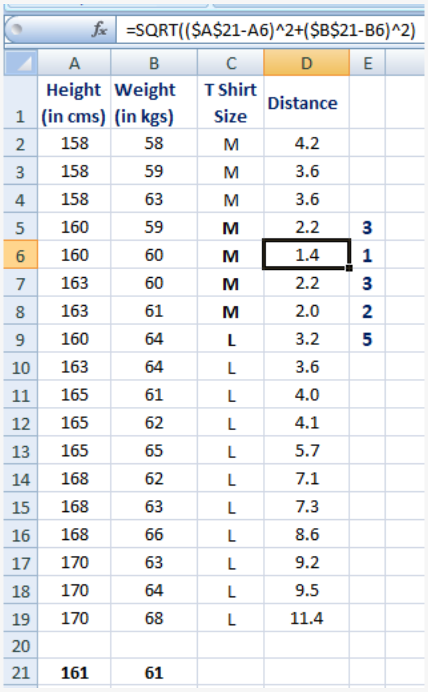
Euclidean distance between first observation and new observation (monica) is as follows -

=SQRT((161-158)^2+(61-58)^2)

Similarly, we will calculate distance of all the training cases with new case and calculates the rank in terms of distance. The smallest distance value will be ranked 1 and considered as nearest neighbor.

Step 2 : Find K-Nearest Neighbors

Let k be 5. Then the algorithm searches for the 5 customers closest to Monica, i.e. most similar to Monica in terms of attributes, and see what categories those 5 customers were in. If 4 of them had ‘Medium T shirt sizes’ and 1 had ‘Large T shirt size’ then your best guess for Monica is ‘Medium T shirt. See the calculation shown in the snapshot below -

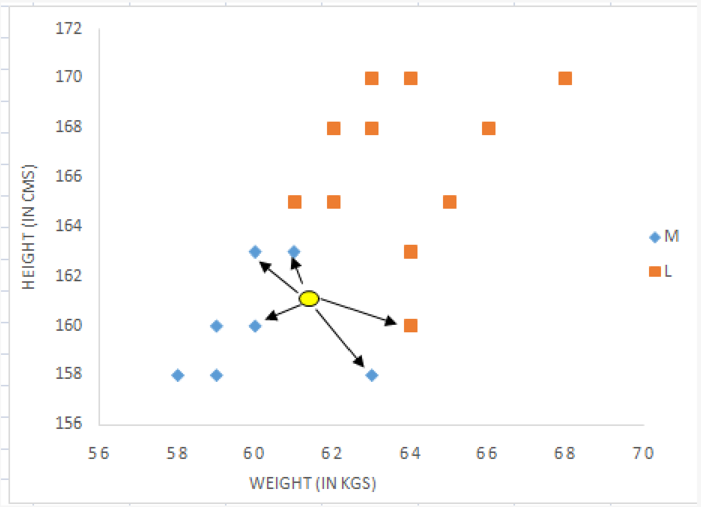


=SQRT((A21-A6)^2+(B21-B6)^2)

	A	B	C	D	E
1	Height (in cms)	Weight (in kgs)	T Shirt Size	Distance	
2	158	58	M	4.2	
3	158	59	M	3.6	
4	158	63	M	3.6	
5	160	59	**M**	2.2	**3**
6	160	60	**M**	1.4	**1**
7	163	60	**M**	2.2	**3**
8	163	61	**M**	2.0	**2**
9	160	64	**L**	3.2	**5**
10	163	64	L	3.6	
11	165	61	L	4.0	
12	165	62	L	4.1	
13	165	65	L	5.7	
14	168	62	L	7.1	
15	168	63	L	7.3	
16	168	66	L	8.6	
17	170	63	L	9.2	
18	170	64	L	9.5	
19	170	68	L	11.4	
20					
21	**161**	**61**			

Calculate KNN manually

In the graph below, binary dependent variable (T-shirt size) is displayed in blue and orange color. 'Medium T-shirt size' is in blue color and 'Large T-shirt size' in orange color. New customer information is exhibited in yellow circle. Four blue highlighted data points and one orange highlighted data point are close to yellow circle. so the prediction for the new case is blue highlighted data point which is Medium T-shirt size.



KNN: Visual Representation

16. What is validation.
Ans:
Validation: 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: determine the attribute that best classifies the training data; use this attribute at the root of the tree. Repeat this process at for each branch.

This means we are performing top-down, greedy search through the space of possible decision trees.

use the attribute with the highest **information gain** in **ID3**

In order to define information gain precisely, we begin by defining a measure commonly used in information theory, called **entropy** that characterizes the (im)purity of an arbitrary collection of examples.

When to stop:

- Stop expanding a node when all the records belong to the same class
- Stop expanding a node when all the records have similar attribute value.
- Early termination.

18. On what principal Bayesian classifier has been built?

Ans: Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.
Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

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

Now, with regards to our dataset, we can apply Bayes' theorem in following way:

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

where, y is class variable and X is a dependent feature vector (of size *n*) where:

$$X = (x_1, x_2, x_3, ..., x_n)$$

Naive assumption
Now, its time to put a naive assumption to the Bayes' theorem, which is, **independence**among the features. So now, we split **evidence** into the independent parts.
Now, if any two events A and B are independent, then

P(A,B)= p(A)p(B)

Hence, we reach to the result:

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

Now, as the denominator remains constant for a given input, we can remove that term:

$$P(y|x_1, ..., x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

Now, we need to create a classifier model. For this, we find the probability of given set of inputs for all possible values of the class variable *y* and pick up the output with maximum probability. This can be expressed mathematically as:

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

So, finally, we are left with the task of calculating P(y) and P(x_i | y).

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:

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

p(j | 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 advantages of tree based classification?

Ans*:*

- Inexpensive to construct
- Extremely fast at classifying unknown records
- Easy to interpret for small-sized trees
- Accuracy is comparable to other classification techniques for many simple data sets

21. What are the main principle of Bayesian Classification?

Ans:

- □ Consider each attribute and class label as random variables
- □ Given a record with attributes (A_1, A_2,…,A_n)
  - ◘ Goal is to predict class C
  - ◘ Specifically, we want to find the value of C that maximizes P(C| A_1, A_2,…,A_n)

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

Ans:

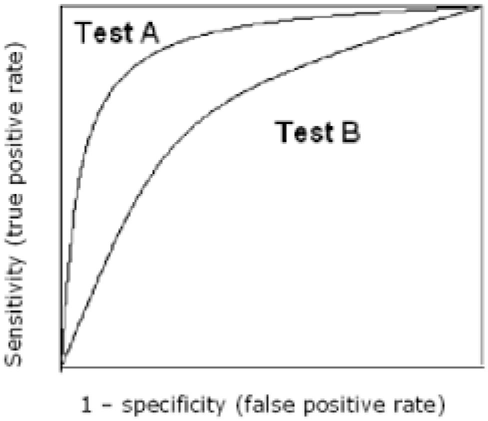
Four steps to data classification

- Step 1: Choose your target. Define your goal.
- Step 2: Map an approach and appropriate toolset. Determine the metrics you'll collect. ...
- Step 3: Gather your data and validate it. ...
- Step 4: Organize and communicate the data in a form that will lead to positive change/action.

23. What are the use of ROC curve?

Ans:

The best cut-off has the highest true positive rate together with the lowest false positive rate. As the area under a ROC curve is a measure of the usefulness of a test in general, where a greater area means a more useful test, the areas under ROC curves are used to compare the usefulness of tests.



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

Ans:

Uses of SVM:

Support Vector Machines are a type of algorithm often used in supervised learning. It allows your model to find a way to separate a labeled dataset, and thus to classify new unseen data. It is one of the most used algorithms in supervised learning, and is often used at the end of a deep convolutional neural network, to classify images based on the features extracted by the network.

Support vector machine is used for classification, like classifying an apple from an orange. There are binary and multi-class classifications, which use different algorithms to conduct classification mathematically.

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

Ans:

Precision:

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

For example, for a text search on a set of documents, precision is the number of correct results divided by the number of all returned results.

Precision takes all retrieved documents into account, but it can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system. This measure is called precision at n or P@n.

Precision is used with recall, the percent of all relevant documents that is returned by the search. The two measures are sometimes used together in the F1 Score (or f-measure) to provide a single measurement for a system.

Note that the meaning and usage of "precision" in the field of information retrieval differs from the definition of accuracy and precision within other branches of science and technology.

Recall:

In information retrieval, recall is the fraction of the relevant documents that are successfully retrieved.
For example, for a text search on a set of documents, recall is the number of correct results divided by the number of results that should have been returned.

In binary classification, recall is called sensitivity. It can be viewed as the probability that a relevant document is retrieved by the query.

It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore, recall alone is not enough but one needs to measure the number of non-relevant documents also, for example by also computing the precision.

When do we use:

It describes how good a model is at predicting the positive class. Precision is referred to as the positive predictive value. Recall is calculated as the ratio of the number of true positives divided by the sum of the true positives and the false negatives. Recall is the same as sensitivity.Precision-Recall curves should be used when there is a moderate to large class imbalance.

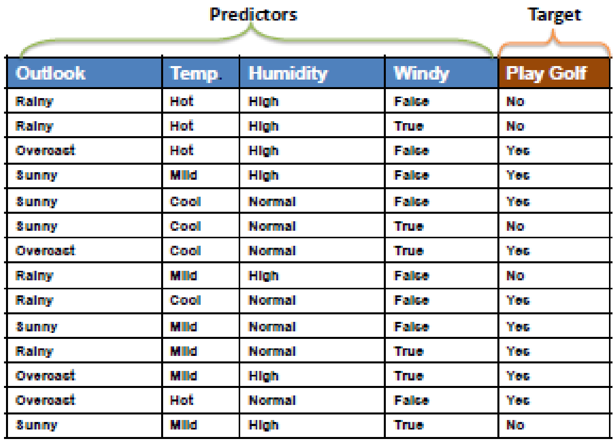
26. What do you mean by rule base classification?

Ans: The term rule-based classification can be used to refer to any classification scheme that make use of IF-THEN rules for class prediction. Rule-based classification schemes typically consist of the following components:

- Rule Induction Algorithm This refers to the process of extracting relevant IF-THEN rules from the data which can be done directly using sequential covering algorithms or indirectly from other data mining methods like decision tree building or association rule mining.
- Rule Ranking Measures This refers to some values that are used to measure the usefulness of a rule in providing accurate prediction. Rule ranking measures are often used in the rule induction algorithm to prune off unnecessary rules and improve efficiency. They are also used in the class prediction algorithm to give a ranking to the rules which will be then be utilized to predict the class of new cases.
- Class Prediction *Algorithm*Given a new record.

27. What is decision tree classification?

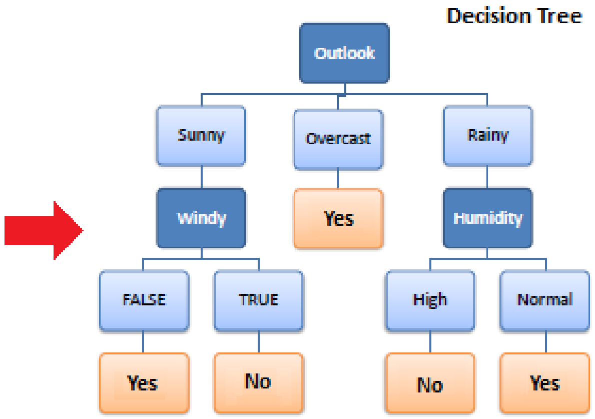
Ans: Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with **decision nodes** and **leaf nodes**. 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.



Predictors

Target

Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



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

Ans: *Ensemble models in machine learning combine the decisions from multiple models to improve the overall performance.* They operate on the similar idea as employed while buying headphones.

The main causes of error in learning models are due to **noise, bias and variance**.

Ensemble methods help to minimize these factors. These methods are designed to improve the stability and the accuracy of Machine Learning algorithms.

Simple Ensemble techniques

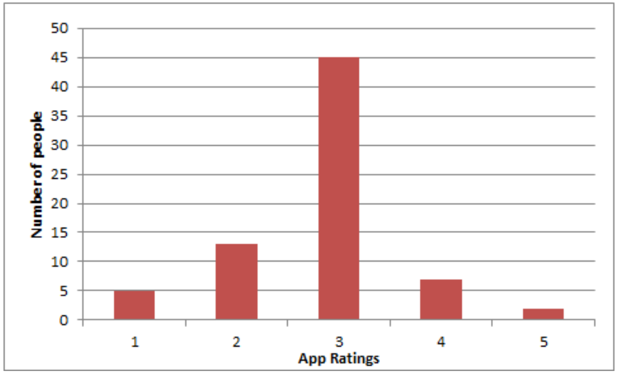
1. **Taking the mode of the results**

MODE: The mode is a statistical term that refers to the most frequently occurring number found in a set of numbers.

In this technique, multiple models are used to make predictions for each data point. The predictions by each model are considered as a separate vote. The prediction which we get from the majority of the models is used as the final prediction.

For instance: We can understand this by referring back to Scenario 2 above. I have inserted a chart below to demonstrate the ratings that the beta version of our health and fitness app got from the user community. (*Consider each person as a different model*)

Output= MODE=3, as majority people voted this



2. **Taking the average of the results**

In this technique, we take an average of predictions from all the models and use it to make the final prediction.

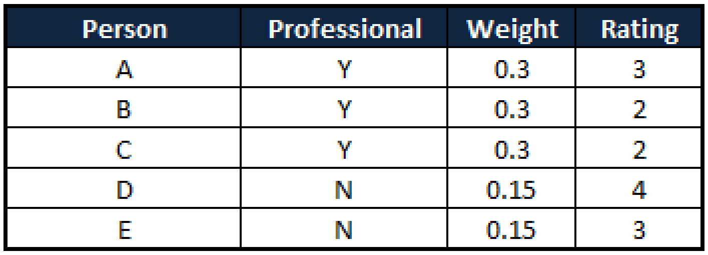
AVERAGE= sum(Rating*Number of people)/Total number of people= (1*5)+(2*13)+(3*45)+(4*7)+(5*2)/72 = 2.833 =Rounded to nearest integer would be 3

3. **Taking weighted average of the results**

This is an extension of the averaging method. All models are assigned different weights defining the importance of each model for prediction. For instance, if about 25 of your responders are professional app developers, while others have no prior experience in this field, then the answers by these 25 people are given more importance as compared to the other people.

For example: For posterity, I am trimming down the scale of the example to 5 people

WEIGHTED AVERAGE= (0.3*3)+(0.3*2)+(0.3*2)+(0.15*4)+(0.15*3) =3.15 = rounded to nearest integer would give us 3



Person	Professional	Weight	Rating
A	Y	0.3	3
B	Y	0.3	2
C	Y	0.3	2
D	N	0.15	4
E	N	0.15	3

29. What is the main principal of Gini index? – explain.

Ans: The Gini coefficient measures the inequality among values of a frequency distribution (for example levels of income). A Gini coefficient of zero expresses perfect equality where all values are the same (for example, where everyone has an exactly equal income)

<u>Main principal of Gini Index-</u>

<u>1. Anonymity:</u>

The coefficient does not disclose the identities of high-income and low-income individuals in a population.

<u>2. Scale of independence</u>

The calculation of the Gini coefficient does not depend on how large the economy is, how it is measured, or how wealthy a country is. For example, both rich and poor countries may show the same coefficient due to similar income distribution.

3. Population independence

The coefficient does not depend on the size of the population.

4. Transfer principle

The coefficient reflects situations when income is transferred from a rich to a poor individual.

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

Ans: We use the Gini Index as our cost function used to evaluate splits in the dataset. our target variable is Binary variable which means it take two values (Yes and No). There can be 4 combinations. A Gini score gives an idea of how good a split is by how mixed the classes are in the two groups created by the split.

31. Which classification technique will you use?

Ans: I will prefer decision tree-based classification, as it is-

- Inexpensive to construct.
- Extremely fast at classifying unknown records.
- Easy to interpret for small-sized trees.
- Accuracy is comparable to other classification techniques for many simple data sets.

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

Ans: Why to use:

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.

When to use:

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 classification, drug disambiguation etc. where number of features are high.

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

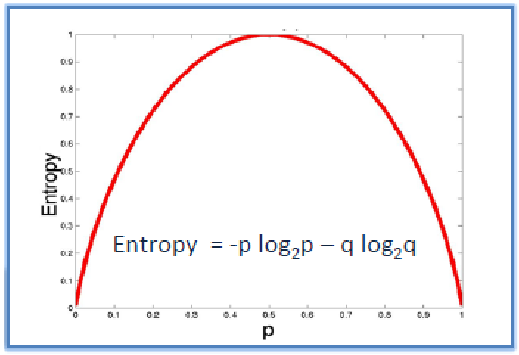
Ans:

Gini coefficient : The most common method used to measure inequality is known as the Gini coefficient. This is a mathematical measure which looks at income distribution over a whole society, not just between different pre-defined groups. By lining up the whole population from poorest to richest and calculating the percentage of income each person has, this measure can show how far a society is from a perfectly equal one.

The problem with the Gini coefficient is that while it gives you a number to indicate how much inequality there is (0 = complete equality, 100 = very very unequal!), it won't say anything about the nature of the inequality in a particular society.

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 (homogenous). ID3 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.



Entropy = $-0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$

34. Why entropy is used instead of Gini index?

Ans: Gini impurity and Information Gain Entropy are pretty much the same. And people do use the values interchangeably. Below are the formulae of both:

1. Gini:Gini(E)=1−∑cj=1p2j
2. Entropy:H(E)=−∑cj=1pjlogpj

As per parsimony principal Gini outperform entropy as of computation ease (log is obvious has more computations involved rather that plain multiplication at processor/Machine level).

But entropy definitely has an edge in some data cases involving high imbalance.

Since entropy uses log of probabilities and multiplying with probabilities of event, what is happening at background is value of lower probabilities are getting scaled up.

If your data probability distribution is exponential or Laplace (like in case of deep learning where we need probability distribution at sharp point) entropy outperform Gini.

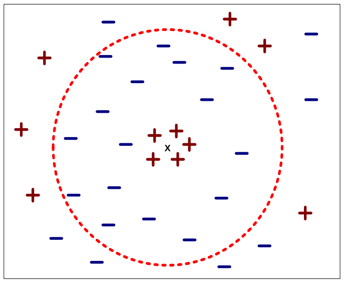
To give an example if you have 2 events one .01 probability and other .99 probability.

In Gini Prob sq will be .01^2+.99^2, .0001 + .9801 means lower probability do not play any role as everything is govern by majority probability.

Now in case of entropy .01*log(.01)+.99*log(.99)= .01*(-2)+ .99*(-.00436) = -.02-.00432 now in this case clearly seen lower probabilities are given better weight-age.

35. Write the algorithm of K-nearest neighbor classification
Ans: Algorithm:

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



36. Write the limitations of KNN.

Ans*:*

<u>Limitations:</u>

- □ 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

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

Ans: KNN neighbor: In pattern recognition, the ***k*-nearest neighbors algorithm** (***k*-NN**) is a non-parametric method used for classification and regression. In both cases, the input consists of the *k* closest training examples in the feature space. The output depends on whether *k*-NN is used for classification or regression:

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.

In *k-NN regression*, the output is the property value for the object. This value is the average of the values of *k' nearest neighbors.*

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The *k*-NN algorithm is among the simplest of all machine learning algorithms.

K-means clustering: K-means clustering is a method used for clustering analysis, especially in data mining and statistics. It aims to partition a set of observations into a number of clusters (k), resulting in the partitioning of the data into Voronoi cells. It can be considered a method of finding out which group a certain object really belongs to.

It is used mainly in statistics and can be applied to almost any branch of study. For example, in marketing, it can be used to group different demographics of people into simple groups that make it easier for marketers to target. Astronomers use it to sift through huge amounts of astronomical data; since they cannot analyze each object one by one, they need a way to statistically find points of interest for observation and investigation.

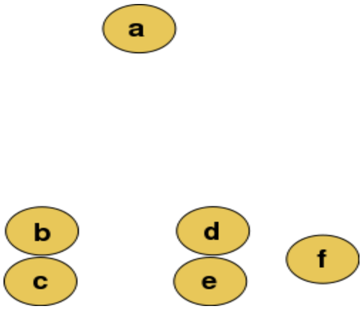
The algorithm:

1. K points are placed into the object data space representing the initial group of centroids.
2. Each object or data point is assigned into the closest k.
3. After all objects are assigned, the positions of the k centroids are recalculated.
4. Steps 2 and 3 are repeated until the positions of the centroids no longer move.

38. Define hierarchical clustering with example.

Ans: Hierarchical clustering is where you build a cluster tree (a dendrogram) to represent data, where each group (or "node") links to two or more successor groups. The groups are nested and organized as a tree, which ideally ends up as a meaningful classification scheme.
Each node in the cluster tree contains a group of similar data; Nodes group on the graph next to other, similar nodes. Clusters at one level join with clusters in the next level up, using a degree of similarity; The process carries on until all nodes are in the tree, which gives a visual snapshot

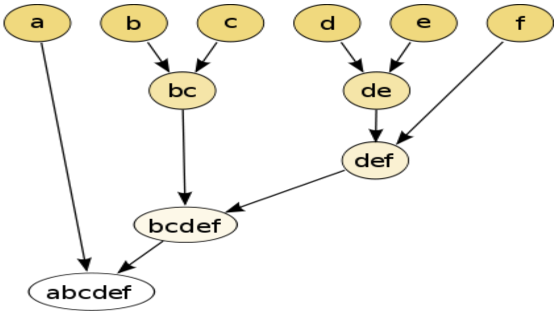
of the data contained in the whole set. The total number of clusters is *not* predetermined before you start the tree creation.



Raw data

For example, suppose this data is to be clustered, and the Euclidean distance is the distance metric.

The hierarchical clustering dendrogram would be as such:



39.Describe the steps of K-means clustering.

Ans: Let X = {x_1,x_2,x_3,……..,x_n} be the set of data points and V = {v_1,v_2,…….,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.

40. What are the different types of clustering?

- **Ans: Partitional Clustering:**
  - A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset.
- **Hierarchical clustering:**
  - A set of nested clusters organized as a hierarchical tree.
- **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.
- **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.
- **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.
- **Density-based:**
  - 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.
- **Shared Property or Conceptual Clusters:**
  - Finds clusters that share some common property or represent a particular concept.

41. What are the processes of density based clustering?

Ans: Density based clustering algorithm has played a vital role in finding non linear shapes structure based on the density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most widely used density based algorithm. It uses the concept of **density reachability** and **density connectivity**.

Let $X = \{x_1, x_2, x_3, ..., x_n\}$ be the set of data points. DBSCAN requires two parameters: ε (eps) and the minimum number of points required to form a cluster (minPts).

1) Start with an arbitrary starting point that has not been visited.

2) Extract the neighborhood of this point using ε (All points which are within the ε distance are neighborhood).

3) If there are sufficient neighborhood around this point then clustering process starts and point is marked as visited else this point is labeled as noise (Later this point can become the part of the cluster).

4) If a point is found to be a part of the cluster then its ε neighborhood is also the part of the cluster and the above procedure from step 2 is repeated for all ε neighborhood points. This is repeated until all points in the cluster is determined.

5) A new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.

6) This process continues until all points are marked as visited.

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

Ans: A *centroid* is a data point (imaginary or real) at the center of a cluster. In Praat 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 is the process of making a group of abstract objects into classes of similar objects. A *cluster* of *data* objects can be treated as one group. While doing *cluster* analysis, we first partition the set of *data* into groups based on *data* similarity and then assign the labels to the groups.

Types of clustering:

- **Partitional Clustering:**
  - A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset.
- **Hierarchical clustering**
  - A set of nested clusters organized as a hierarchical tree.
- **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.
- **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.
- **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.
- **Density-based:**
    - 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.
- Shared Property or Conceptual Clusters:

Finds clusters that share some common property or represent a particular concept.

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

Ans: Clustering is the process of making a group of abstract objects into classes of similar objects. A *cluster* of *data* objects can be treated as one group. While doing *cluster* analysis, we first partition the set of *data* into groups based on *data* similarity and then assign the labels to the groups.

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.

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

Ans: Cluster validation is an important part of any cluster analysis. External measures such as entropy, purity and mutual information are often used to evaluate K-means clustering. However, whether these measures are indeed suitable for K-means clustering remains unknown. Along this line, in this paper, we show that a data distribution view is of great use to selecting the right measures for K-means clustering. Specifically, we first introduce the data distribution view of K-means, and the resultant uniform effect on highly imbalanced data sets. Eight external measures widely used in recent data mining tasks are also collected as candidates for K-means evaluation. Then, we demonstrate that only three measures, namely the variation of information (VI), the van Dongen criterion (VD) and the Mirkin metric (M), can detect the negative uniform effect of K-means in the clustering results. We also provide new normalization schemes for these three measures, i.e., VInorm′, VDnorm′ and Mnorm′, which enables the cross-data comparisons of clustering qualities. Finally, we explore some properties such as the consistency and sensitivity of the three measures, and give some advice on how to use them in K-means practice.

1. Determining the clustering tendency of a set of data, i.e., distinguishing whether non-random structure actually exists in the data.
2. Comparing the results of a cluster analysis to externally known results, e.g., to externally given class labels.
3. Evaluating how well the results of a cluster analysis fit the data without reference to external information.

   - Use only the data
4. Comparing the results of two different sets of cluster analyses to determine which is better.
5. Determining the ‘correct’ number of clusters.

For 2, 3, and 4, we can further distinguish whether we want to evaluate the entire clustering or just individual clusters.

46. Write the steps of k-means clustering?

Ans: Let X = {x_1,x_2,x_3,……..,x_n} be the set of data points and V = {v_1,v_2,…….,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.

47. Write the steps of k-means clustering?

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

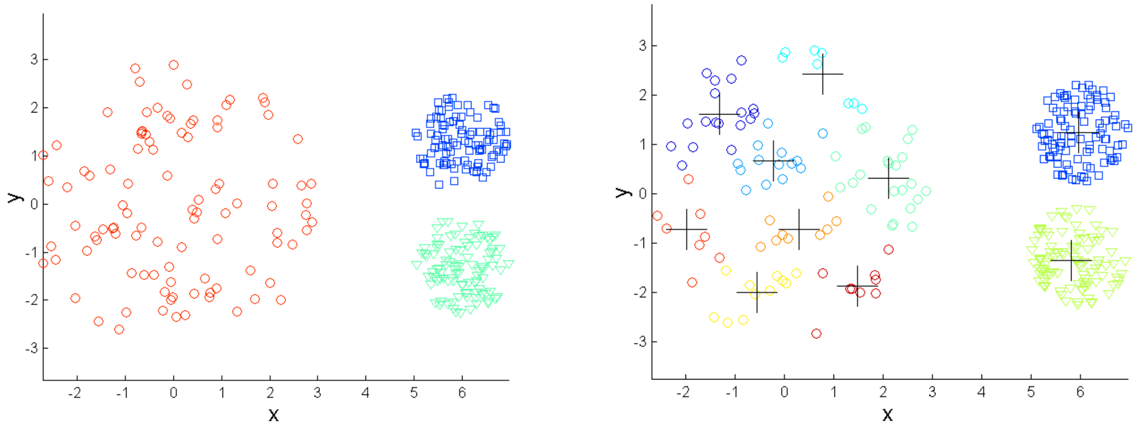
Ans: Limitations:

- K-means has problems when clusters are of differing
    - Sizes
    - Densities
    - Non-globular shapes
    - K-means has problems when the data contains outliers.

Solutions:

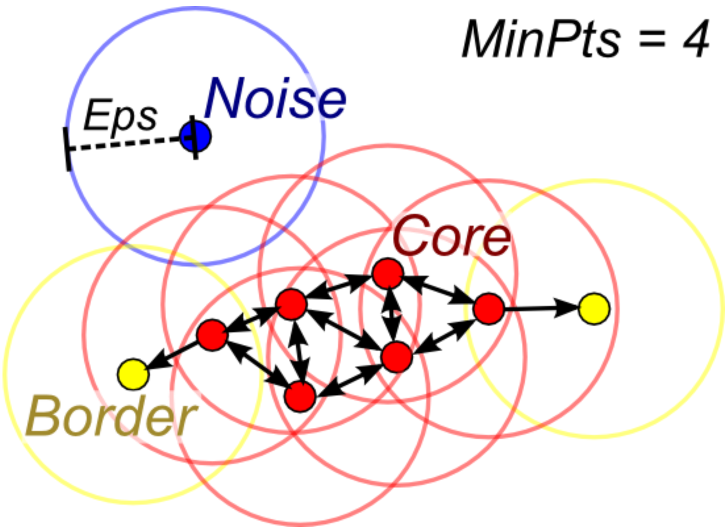
One solution is to use many clusters.

- Find parts of clusters, but need to put together.



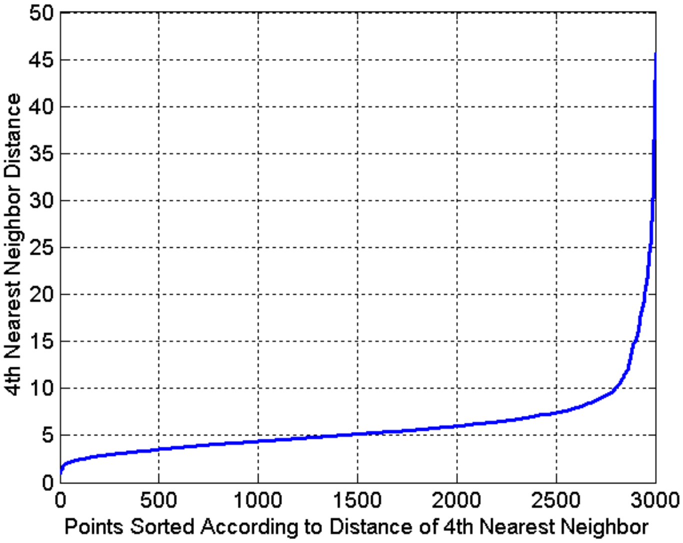
49. Explain with a pictorial example of Core Point, Noise Point and Border Point.

- **core point** if its ε-neighborhood contains at least μ points;
- **border point** if it is not a core point but belongs to ε-neighborhood of some core points (we'll call all such core points **associated** with border point);
- **noise point** if it is neither a core point nor a border point.



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

- Idea is that for points in a cluster, their k^{th} nearest neighbors are at roughly the same distance
- Noise points have the k^{th} nearest neighbor at farther distance
- So, plot sorted distance of every point to its k^{th} nearest neighbor



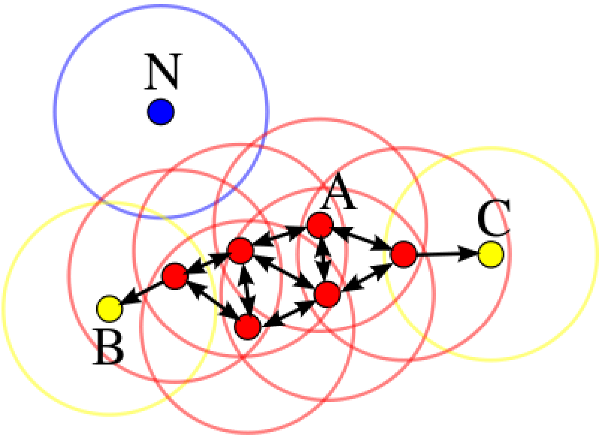
51. What is the basic principle of DBSCAN clustering?

Ans: It groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away). DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature.

Consider a set of points in some space to be clustered. Let ε be a parameter specifying the radius of a neighborhood with respect to some point. For the purpose of DBSCAN clustering, the points are classified as core points, (density-)reachable points and outliers, as follows:

- A point p is a core point if at least minPts points are within distance ε of it (including p).
- A point q is directly reachable from p if point q is within distance ε from core point p. Points are only said to be directly reachable from core points.
- A point q is reachable from p if there is a path p_1, ..., p_n with p_1 = p and p_n = q, where each p_{i+1} is directly reachable from p_i. Note that this implies that all points on the path must be core points, with the possible exception of q.
- All points not reachable from any other point are outliers or noise points.

Now if p is a core point, then it forms a cluster together with all points (core or non-core) that are reachable from it. Each cluster contains at least one core point; non-core points can be part of a cluster, but they form its "edge", since they cannot be used to reach more points.



In this diagram, minPts = 4. Point A and the other red points are core points, because the area surrounding these points in an ε radius contain at least 4 points (including the point itself). Because they are all reachable from one another, they form a single cluster. Points B and C are not core points, but are reachable from A (via other core points) and thus belong to the cluster as well. Point N is a noise point that is neither a core point nor directly-reachable.

Reachability is not a symmetric relation since, by definition, no point may be reachable from a non-core point, regardless of distance (so a non-core point may be reachable, but nothing can be reached from it). Therefore, a further notion of connectedness is needed to formally define the extent of the clusters found by DBSCAN. Two points p and q are density-connected if there is a point o such that both p and q are reachable from o. Density-connectedness is symmetric.

A cluster then satisfies two properties:

1. All points within the cluster are mutually density-connected.
2. If a point is density-reachable from any point of the cluster, it is part of the cluster as well.

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

Ans: The methodology is summarized as follows:

1. Given a dataset with known ground truth, generate n_ partitions with different properties by varying

the parameters of one or more clustering methods.

2. Compute the values of the relative and external validity criteria for each one of the n_ partitions.

3. Compute the correlation between the vectors with the n_ relative validity measure values and the n_ external validity measure values. This correlation quantifies the accuracy of the relative validity criterion w.r.t. the external validity measure (ARI).

An important aspect in the evaluation of the relative measures for density-based clustering is how to deal with noise objects, given that partitions generated with density-based clustering algorithms may contain noise. As far as we know, DBCV is the first relative validity measure capable of handling noise. Since other relative indices do not have this capability, noise has to be handled prior to their application for a fair comparison. To the best of our knowledge, there is no established procedure in the literature defining how to deal with noise objects in a partition when applying a relative validity index.

53. When we need to use density based clustering?

Ans:

- When data points are resistant to noise.
- When need to handle clusters of different shapes and sizes.
- Data of Varying densities.
- High-dimensional data.

54. When we need to use density based clustering?

55. Write a situation where DBSCAN clustering is appropriate.

Ans: The DBSCAN algorithm should be 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.

Clustering methods are usually used in biology, medicine, social sciences, archaeology, marketing, characters recognition, management systems and so on.

Let's think in a practical use of DBSCAN. 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 cluster 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: 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 cluster, 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.

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:

1. Incorrect logging

In the process mining world most people use the term “Noise” for exceptional behavior – not for incorrect logging. This means that if a process discovery algorithm is said to be able to deal with noise, then it can abstract from low-frequent behavior by only showing the main process flow. The reason is simple: It is impossible for discovery algorithms to distinguish incorrect logging from exceptional events.

Here are two true stories of incorrect data:
In an ERP system, data entries from invoice documents had been scanned automatically. However, because of a mistake in the scanning procedure the invoice ID was interpreted as the invoice date for some of the cases. As a result, activities with a timestamp of the year 2020 appeared in the log data.

2. Insufficient logging

While incorrect logging is about wrong data, insufficient logging is about missing data. The minimum requirements for process mining are a case ID, an activity name, and a timestamp per event to reconstruct the *history* of each process instance.

One problem with missing data are:
Fields in the database of the information system are simply overwritten. So, old entries are lost and the database only provides information about the current status, but not the overall history of what happened in the past.

3. Semantics

One of the biggest challenges can be to find the right information and to understand what it means.

In fact, figuring out the semantics of existing IT logs can be anything between really easy and incredibly complicated. It largely depends on how distant the logs are from the actual business logic. For example, the performed business process steps may be recorded directly with their activity name, or you might need a mapping between some kind of cryptic action code and the actual business activity.

It is best to work together with an IT specialist who helps you extract the right data and explain the meaning of the different fields. In terms of process mining it helps not to try to understand everything at once.

4. Correlation

Because process mining is based on the *history* of a process, the individual process instances need to be reconstructed from the log data. Correlation is about stitching everything together in the correct way:

Business processes often span multiple IT systems, and usually each IT system has its own local IDs. One needs to correlate these local process IDs to combine log fragments from the different systems (local ID from system No. 1 and local ID from system No. 2) in order to get a full picture of the process from start to end.

5. Timing

Precisely because process mining evaluates the history of performed process instances, the timing is very important for ordering the events within each sequence. If the timestamps are wrong or not precise enough, then it is difficult to create the correct order of events in the history.

One of the problems I have seen with timestamps are:
Timestamp resolution is too low. For example, only the date of a performed activity (but not the time) is recorded. But even if the time is recorded, it may be necessary to record it at least with millisecond accuracy if many events follow each other in automated systems.

58. Explain different types of data that we face in data mining.
Ans:
Let’s discuss what type of data can be mined:

1. Flat Files

- Flat files is defined as data files in text form or binary form with a structure that can be easily extracted by data mining algorithms.

2. Relational Databases

- A Relational database is defined as the collection of data organized in tables with rows and columns.

3. DataWarehouse

- A datawarehouse is defined as the collection of data integrated from multiple sources that will queries and decision making.

4. Transactional Databases

- Transactional databases is a collection of data organized by time stamps, date, etc to represent transaction in databases.

5. Multimedia Databases

- Multimedia databases consists audio, video, images and text media.

6. Spatial Database

- Store geographical information.

7. Time-series Databases

- Time series databases contains stock exchange data and user logged activities.

8. WWW

- WWW refers to World wide web is a collection of documents and resources like audio, video, text, etc which are identified by Uniform Resource Locators (URLs) through web browsers, linked by HTML pages, and accessible via the Internet network.

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

Ans: Some machine learning algorithms prefer or find it easier to work with discrete attributes.

For example, decision tree algorithms can choose split points in real valued attributes, but are much cleaner when split points are chosen between bins or predefined groups in the real-valued attributes.

Discrete attributes are those that describe a category, called nominal attributes. Those attributes that describe a category that where there is a meaning in the order for the categories are called ordinal attributes. The process of converting a real-valued attribute into an ordinal attribute or bins is called discretization.

You can discretize your real valued attributes in Weka using the Discretize filter.

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. ϖ 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. ϖ 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.

61. How do you discretize a numeric attribute? i.e. Age

Ans:

During data analysis, it is often super useful to turn continuous variables into categorical ones. In Stata you would do something like this:

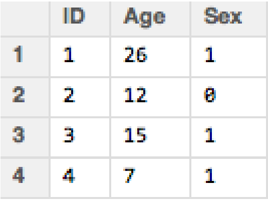
```
gen catvar=0
replace catvar=1 if contvar>0 & contvar<=3
replace catvar=2 if contvar>3 & contvar<=5
```

etc. And then you would label your values like so:

```
label define agelabel 0 “0” 1 “1-3” 2 “3-5”
label values catvar agelabel
```

How can we do this in R? There’s a great function in R called cut() that does everything at once. It takes in a continuous variable and returns a factor (which is an ordered or unordered categorical variable). Factor variables are extremely useful for regression because they can be treated as dummy variables. I’ll have another post on the merits of factor variables soon.

But for now, let’s focus on getting our categorical variable. Here is our data:

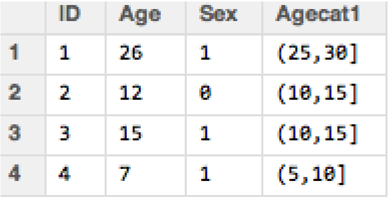


	ID	Age	Sex
1	1	26	1
2	2	12	0
3	3	15	1
4	4	7	1

And now we want to take that “Age” variable and turn in into a categorical variable. The most basic statement is like so:

```
mydata$Agecat1<-cut(mydata$Age, c(0,5,10,15,20,25,30))
```

Here the function cut() takes in as the first argument the continuous variable mydata$Age and it cuts it into chunks that are described in the second argument. So here I’ve indicated to make groups that go from 0-5, 6-10, 11-15, 16-20, etc. By default, the right side of the interval is closed while the left is open. You can change that, as we will see below. First, the output with the new “Agecat” variable:



	ID	Age	Sex	Agecat1
1	1	26	1	(25,30]
2	2	12	0	(10,15]
3	3	15	1	(10,15]
4	4	7	1	(5,10]

Now we can customize our intervals. First, in Agecat2, I show how instead of spelling out every cutoff of the interval, I can just specify a sequence using seq(0, 30, 5) – this means we start at 0 and go to 30 by intervals of 5.

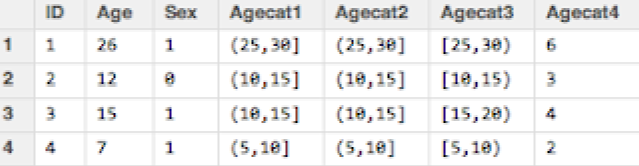
For Agecat3, I switch the default closed interval to be the left one by specifying “right=FALSE”.

Finally, for Agecat4 I add in my own labels instead of the default "(0,5]" labels that are provided by R. I want them to be numbers instead so I indicate "labels=c(1:6)". The output of all of the options are shown below.

```
mydata$Agecat2<-cut(mydata$Age, seq(0,30,5))

mydata$Agecat3<-cut(mydata$Age, seq(0,30,5), right=FALSE)

mydata$Agecat4<-cut(mydata$Age, seq(0,30,5), right=FALSE, labels=c(1:6))
```



	ID	Age	Sex	Agecat1	Agecat2	Agecat3	Agecat4
1	1	26	1	(25,30]	(25,30]	[25,30)	6
2	2	12	0	(10,15]	(10,15]	[10,15)	3
3	3	15	1	(10,15]	(10,15]	[15,20)	4
4	4	7	1	(5,10]	(5,10]	[5,10)	2

Now, if I want some summary statistics or a bivariate table, I get some nice output:

```
summary(mydata$Agecat1)

    (0,5]     (5,10]     (10,15]     (15,20]     (20,25]     (25,30]
  0    1    2    0    0    1
```

```
table(mydata$Agecat1, mydata$Sex)

                                                  0          1
            (0,5]                            0               0
               (5,10]                  0                     1
               (10,15]                 1                     1
               (15,20]                 0                     0
               (20,25]                 0                     0
 (25,30] 0 1
```

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

Ans:

Missing data is random:

For MCAR and MAR, many missing data methods have been developed in the last two decades (3). Although MCAR seems to be the least problematic mechanism, deleting cases can still reduce the power of finding an effect. It is argued that the MAR mechanism is most frequently seen in practice. An argument for this is that in most research multifactorial or multivariable problems are studied, so when data on variables are missing it is mostly related to other variables in the dataset.

Missing data is not random:

For MNAR, imputation is not sufficient, because the missing data are totally different from the available data, i.e. your complete data has become a selective group of persons. If you think your data is MNAR it might be wise to contact a statistician from EMGO+ who is willing to help you.

For MCAR and MAR, there are roughly two kinds of techniques for imputation:

1. Single imputation is possible in SPSS and is an easy way to handle missing's when just a few cases are missing (less than 5%) and you think your missing values are MCAR or MAR. However, after single imputation the cases are more similar which may result in an underestimation of the standard errors, i.e. smaller confidence intervals. This increases the chance of a type 1 error (the null hypothesis of no effect is rejected, while there is truly no effect). Therefore, this method is less adequate when you have >5% missing data. This is also the case when item scores are missing in questionnaires (4).

2. Multiple imputation is more complex, but also implemented in SPSS 17.0 and later versions. Multiple imputation takes into account the uncertainty of missing values (present in all values of variables) and is therefore more preferred than single imputation. When the amount of missing data is high (exceeds 5% in several variables and different persons), multiple imputation is more adequate. Multiple Imputation works for total scores in questionnaires as well as for item scores in questionnaires.

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.

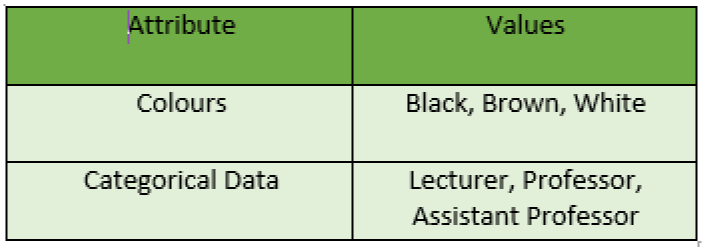
Type of attributes:

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:**



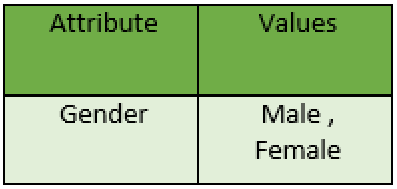
Attribute	Values
Colours	Black, Brown, White
Categorical Data	Lecturer, Professor, Assistant Professor

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

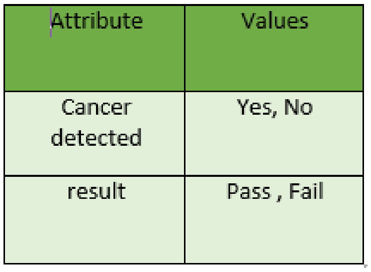
  i) **Symmetric:** Both values are equally important (Gender).

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

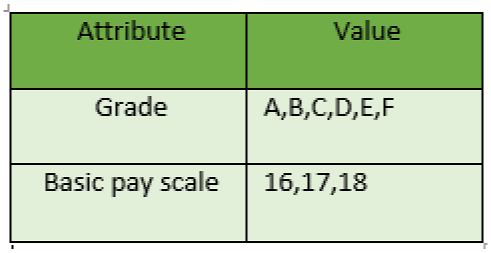
- **Ordinal Attributes**



Attribute	Values
Gender	Male , Female



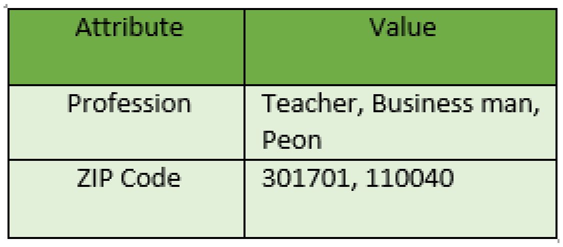
Attribute	Values
Cancer detected	Yes, No
result	Pass , Fail



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

Quantitative Attributes

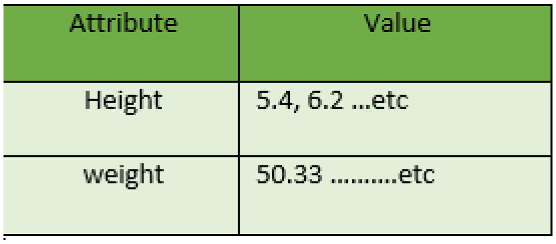
- **Discrete:** Discrete data have finite values it can be numerical and can also be in categorical form. These attributes has finite or countably infinite set of values.



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

  Example:

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



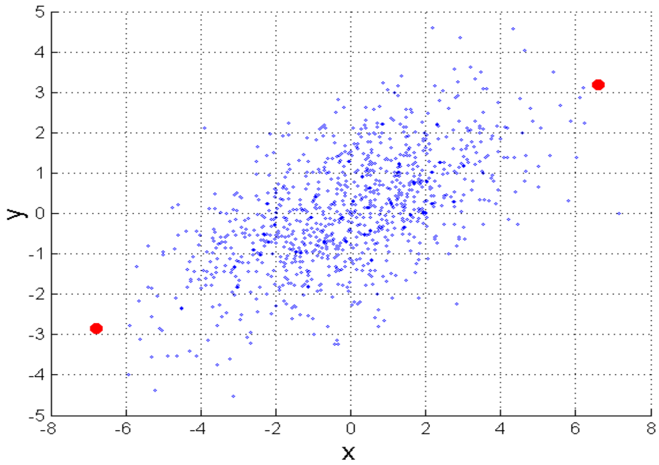
Attribute	Value
Height	5.4, 6.2 ...etc
weight	50.33 ....etc

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.

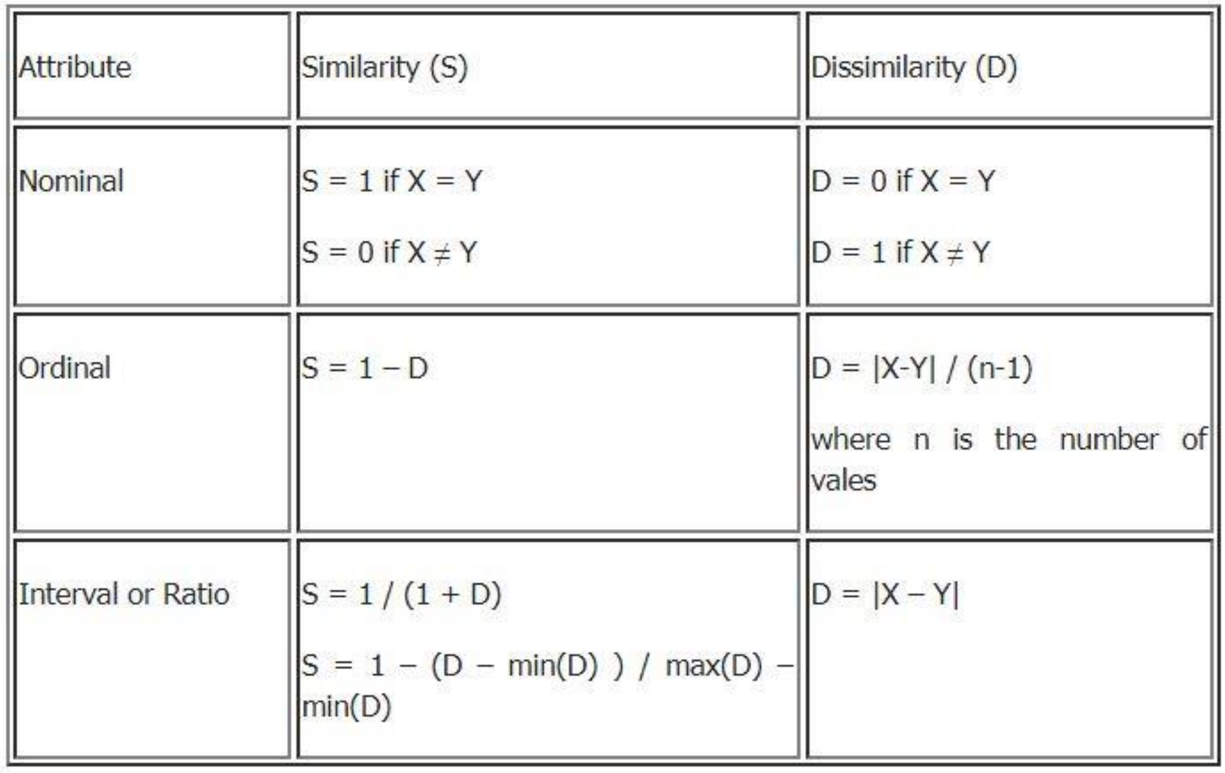
Let's consider simple objects with single attribute and discuss similarity, dissimilarity measures.

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 ≠ Y
		D = 0 if X = Y
		D = 1 if X ≠ 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:

Similarity Measurement:

Similarity metric is the basic measurement and used by a number of data mining algorithms. It measures the similarity or dissimilarity between two data objects which have one or multiple attributes. Informally, the similarity is a numerical measure of the degree to which the two objects are alike. It is usually non-negative and are often between 0 and 1, where 0 means no similarity, and 1 means complete similarity.

Considering different data type with a number of attributes, it is important to use the appropriate similarity metric to well measure the proximity between two objects. For example, Euclidean distance and correlation are useful for dense data such as time series or two-dimensional points. Jaccard and cosine similarity measures are useful for sparse data like documents, or binary data.

Euclidean Distance:

Euclidean Distance between two points is given by Minkowski distance metric. It can be used in one-, two-, or higher-dimensional space. The formula of Euclidean distance is as following.

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \sqrt{(p_1-q_1)^2 + (p_2-q_2)^2 + \cdots + (p_n-q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}.$$

where n is the number of dimensions. It measures the numerical difference for each corresponding attributes of point p and point q. Then it combines the square of differences in each dimension into an overall distance.

Pearson Correlation:

The correlation coefficient is a measure of how well two sets of data fit on a straight line. Correlation is always in the range -1 to 1. A correlation of 1 (-1) means that x and y have a perfect positive (negative) linear relationship. If the correlation is 0, then there is no linear relationship between the attributes of the two data objects. However, the two data objects might have non-linear relationships.

Pearson correlation is defined by the following equation. x and y represent two data objects.

$$corr(x,y) = \frac{covariance(x,y)}{standard\_deviation(x) \times standard\_deviation(y)}$$

Unlike the distance metric, this formula is not very intuitive, but it does tell you how much the variables change together divided by the product of how much they vary individually.

Jaccard Coefficient:

Jaccard coefficient is often used to measure data objects consisting of asymmetric binary attributes. The asymmetric binary attributes have two values 1 indicates present and 0 indicates not present. Most of the attributes of the object will have the similar value.

The Jaccard coefficient is given by the following equation:

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}.$$

where

M11 represents the total number of attributes where object A and object B both have a value of 1.

M01 represents the total number of attributes where the attribute of A is 0 and the attribute of B is 1.

M10 represents the total number of attributes where the attribute of A is 1 and the attribute of B is 0.

Tanimoto Coefficient (Extended Jaccard Coefficient):

Tanimoto coefficient is also known as extended Jaccard coefficient. It can used for handling the similarity of document data in text mining. In the case of binary attributes, it reduces to the Jaccard coefficient. Tanimoto coefficient is defined by the following equation:

$$T(A,B) = \frac{A \cdot B}{\|A\|^2 + \|B\|^2 - A \cdot B}.$$

where A and B are two document vector objects.

Cosine similarity:

The cosine similarity is a measure of similarity of two non-binary vector. The typical example is the document vector, where each attribute represents the frequency with which a particular word occurs in the document. Similar to sparse market transaction data, each document vector is sparse since it has relatively few non-zero attributes. Therefore, the cosine similarity ignores 0-0 matches like the Jaccard measure. The cosine similarity is defined by the following equation:

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

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.

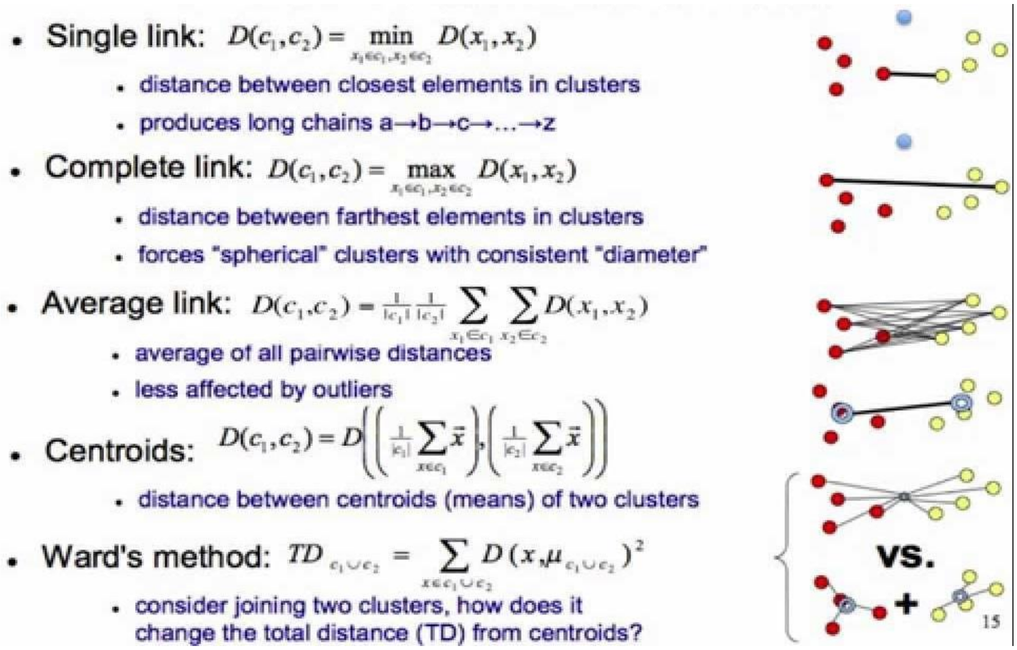
Dataset Type	**Characteristics**	**Example**
File	a single file	AutoCAD DXF

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?

69. How hierarchical clustering helps to construct other clustering techniques?

***Ans*:**

<u>Hierarchical Clustering</u>:

As mentioned before, hierarchical clustering relies using these clustering techniques to find a hierarchy of clusters, where this hierarchy resembles a tree structure, called a dendrogram.

Hierarchical clustering is the hierarchical decomposition of the data based on group similarities

Finding hierarchical clusters:

There are two top-level methods for finding these hierarchical clusters:

- **Agglomerative** clustering uses a *bottom-up* approach, wherein each data point starts in its own cluster. These clusters are then joined greedily, by taking the two most similar clusters together and merging them.
- **Divisive** clustering uses a *top-down* approach, wherein all data points start in the same cluster. You can then use a parametric clustering algorithm like K-Means to divide the cluster into two clusters. For each cluster, you further divide it down to two clusters until you hit the desired number of clusters.

Both of these approaches rely on constructing a similarity matrix between all of the data points, which is usually calculated by cosine or Jaccard distance.

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

***Ans*:**

Popular hierarchical clustering technique-

- ✓ Basic algorithm is straightforward
- ✓ Compute the proximity matrix
- ✓ Let each data point be a cluster
- ✓ Repeat
- ✓ Merge the two closest clusters
- ✓ Update the proximity matrix
- ✓ Until only a single cluster remains

- Key operation is the computation of the proximity of two clusters
  - Different approaches to defining the distance between clusters distinguish the different algorithms.

71. Write some applications of hierarchical clustering?

Ans:

- ✓ US Senator Clustering through Twitter
- ✓ Charting Evolution through Phylogenetic Trees
- ✓ Tracking Viruses through Phylogenetic Trees

72. Write the algorithm of hierarchical clustering.

Ans:

Algorithmic steps for 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).
Divisive Hierarchical clustering - It is just the reverse of Agglomerative Hierarchical approach.

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

Ans:

Given a set of N items to be clustered, and an NxN distance (or similarity) matrix, the basic process of Johnson's (1967) hierarchical clustering is this:

- ✓ Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters equal the distances (similarities) between the items they contain.
- ✓ Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.
- ✓ Compute distances (similarities) between the new cluster and each of the old clusters.
- ✓ Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

74. Give an example of data where Data Mining techniques need to apply to extract hidden and unknown information.

Ans:

Data mining is used for recognize the unknown data with the known data set. Data mining processes of the following data-

1. Flat Files
2. Relational Databases
3. Data Warehouse
4. Transactional Databases
5. Multimedia Databases
6. Spatial Databases
7. Time Series Databases
8. World Wide Web(WWW)

1. ***Flat Files***:
   - Flat files is defined as data files in text form or binary form with a structure that can be easily extracted by data mining algorithms.
   - Data stored in flat files have no relationship or path among themselves, like if a relational database is stored on flat file, then there will be no relations between the tables.
   - Flat files are represented by data dictionary. Eg: CSV file.
   - **Application**: Used in Data Warehousing to store data, Used in carrying data to and from server, etc.

2. ***Relational Databases*:**
   - A Relational database is defined as the collection of data organized in tables with rows and columns.
   - Physical schema in Relational databases is a schema which defines the structure of tables.
   - Logical schema in Relational databases is a schema which defines the relationship among tables.
   - Standard API of relational database is SQL.
   - **Application**: Data Mining, ROLAP model, etc.

3. ***Data Warehouse*:**

   - A data warehouse is defined as the collection of data integrated from multiple sources that will queries and decision making.
   - There are three types of data warehouse: **Enterprise** data warehouse, **Data Mart** and **Virtual** Warehouse.
   - Two approaches can be used to update data in Data warehouse: **Query-driven** Approach and **Update-driven** Approach.
   - **Application**: Business decision making, Data mining, etc.

4. ***Transactional Databases:***

   - Transactional database is a collection of data organized by time stamps, date, etc to represent transaction in databases.
   - This type of database has the capability to roll back or undo its operation when a transaction is not completed or committed.
   - Highly flexible system where users can modify information without changing any sensitive information.
   - Follows ACID property of DBMS.
   - **Application**: Banking, Distributed systems, Object databases, etc.

5. ***Multimedia Databases*:**

   - Multimedia databases consists audio, video, images and text media.
   - They can be stored on Object-Oriented Databases.
   - They are used to store complex information in a pre-specified format.
   - **Application**: Digital libraries, video-on demand, news-on demand, musical database, etc.

6. ***Spatial Database:***

   - Store geographical information.
   - Stores data in the form of coordinates, topology, lines, polygons, etc.
   - **Application**: Maps, Global positioning, etc.

7. ***Time-series Databases:***

   - Time series databases contains stock exchange data and user logged activities.

- Handles array of numbers indexed by time, date, etc.
- It requires real-time analysis.
- **Application**: eXtremeDB, Graphite, InfluxDB, etc.

8. ***WWW:***
   - WWW refers to World wide web is a collection of documents and resources like audio, video, text, etc which are identified by Uniform Resource Locators (URLs) through web browsers, linked by HTML pages, and accessible via the Internet network.
   - It is the most heterogeneous repository as it collects data from multiple resources.
   - It is dynamic in nature as Volume of data is continuously increasing and changing.
   - **Application**: Online shopping, Job search, Research, studying, etc.

75. Define Data Mining? There are two types of Data mining techniques: Predictive and descriptive data mining- give example of these two.

Ans:

Data Mining: Data mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. Data mining tools allow enterprises to predict future trends.

Predictive Data Mining: Predictive data mining is data mining that is done for the purpose of using business intelligence or other data to forecast or predict trends. This type of data mining can help business leaders make better decisions and can add value to the efforts of the analytics team.

One of the most common uses of predictive modeling is in online advertising and marketing. Modelers use web surfers' historical data, running it through algorithms to determine what kinds of products users might be interested in and what they are likely to click on.

Descriptive Data Mining: Descriptive analysis or statistics does exactly what the name implies they "Describe", or summarize raw data and make it something that is interpretable by humans. They are analytics that describe the past. The past refers to any point of time that an event has occurred, whether it is one minute ago, or one year ago. Descriptive analytics are useful because they allow us to learn from past behaviors, and understand how they might influence future outcomes.

76. Define the term data mining. Give an example of predictive data mining.

Ans:

Data Mining: Data mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. Data mining tools allow enterprises to predict future trends.

Predictive Data Mining: Predictive data mining is data mining that is done for the purpose of using business intelligence or other data to forecast or predict trends. This type of data mining can help business leaders make better decisions and can add value to the efforts of the analytics team.

One of the most common uses of predictive modeling is in online advertising and marketing. Modelers use web surfers' historical data, running it through algorithms to determine what kinds of products users might be interested in and what they are likely to click on.

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:

- 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)
- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data

– in Hypothesis Formation

78. What are the difference between classification and clustering?

Ans:

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

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

Ans:

Supervised learning:

Supervised learning as the name indicates a presence of supervisor as teacher. Basically, supervised learning is a learning in which we teach or train the machine using data which is well labeled that means some data is already tagged with correct answer. After that, machine is provided with new set of examples(data) so that supervised learning algorithm analyses the training data (set of training examples) and produces a correct outcome from labeled data.

For instance, suppose you are given a basket filled with different kinds of fruits. Now the first step is to train the machine with all different fruits one by one like this:

- If shape of object is rounded and depression at top having color Red then it will be labelled as –Apple.
- If shape of object is long curving cylinder having color Green-Yellow then it will be labelled as –Banana.

Now suppose after training the data, you have given a new separate fruit say Banana from basket and asked to identify it.

Since machine has already learnt the things from previous data and this time have to use it wisely. It will first classify the fruit with its shape and color, and would confirm the fruit name as BANANA and put it in Banana category. Thus, machine learns the things from training data (basket containing fruits) and then apply the knowledge to test data (new fruit).

Supervised learning classified into two categories of algorithms:

- Classification: A classification problem is when the output variable is a category, such as "Red" or "blue" or "disease" and "no disease".
- Regression: A regression problem is when the output variable is a real value, such as "dollars" or "weight".

Unsupervised learning:

Unsupervised learning is the training of machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data.

Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore, machine is restricted to find the hidden structure in unlabeled data by our-self.
For instance, suppose it is given an image having both dogs and cats which have not seen ever.

Thus, machine has no any idea about the features of dogs and cat so we can't categorize it in dogs and cats. But it can categorize them according to their similarities, patterns and differences i.e., we can easily categorize the above picture into two parts. First first may contain all pics having **dogs** in it and second part may contain all pics having **cats** in it. Here you didn't learn anything before, means no training data or examples.

80. What is data mining and why is it an important discipline?

Ans:

Data mining:

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns.

Importance:

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

81. Why do we divide data in two parts before data mining starts?

Ans:

One issue when fitting a model is how well the newly-created model behaves when applied to new data. To address this issue, the data set can be divided into multiple partitions: a training partition used to create the model, a validation partition to test the performance of the model, and a third test partition. Partitioning is performed randomly to protect against a biased partition -- according to proportions specified by the user -- or according to rules concerning the data set type. For example, when creating a time series forecast, data is partitioned by chronological order.

Training Set:

The Training Set is used to train or build a model. For example, in a linear regression, the training set is used to fit the linear regression model (i.e., to compute the regression coefficients). In a neural network model, the training set is used to obtain the network weights. After fitting the model on the Training Set, the performance of the model should be tested on the Validation Set.

Validation Set/test set:

Once a model is built using the Training Set, the performance of the model must be validated using new data. If the Training Set itself was utilized to compute the accuracy of the model fit, the result would be an overly optimistic estimate of the accuracy of the model. This is because the training or model fitting process ensures that the accuracy of the model for the training data is as high as possible, and the model is specifically suited to the training data. To obtain a more realistic estimate of how the model would perform with unseen data, we must set aside a part of the original data and not include this set in the training process. This data set is known as the Validation Set.

82. Write the list of predictive data mining. How anomaly detection is one kind of data mining?

Ans:

Here is a list of predictive data mining-

- Ordinary Least Squares.
- Generalized Linear Models (GLM)
- Logistic Regression.
- Random Forests.
- Decision Trees.
- Neural Networks.
- Multivariate Adaptive Regression Splines (MARS)

Anomaly detection:

In data mining, anomaly detection (also outlier detection) is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data.[1] Typically the anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or errors in a text. Anomalies are also referred to as outliers, novelties, noise, deviations and exceptions.

Thus, anomaly detection is one kind of data mining.

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

Ans:

Data mining:

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns.

Importance:

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.

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

Ans:

Methods of data processing:

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 –** This is the modern technique to process data. The fastest and best available method with the highest reliability and accuracy. The technology used is latest as this method used computers and employed in most of the agencies. The use of software forms the part of this type of data processing. 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.

85. Explain different distance measures.

Ans:

Euclidean distance:

Euclidean distance is the most common use of distance. In most cases when people said about distance, they will refer to Euclidean distance. Euclidean distance is also known as simply distance. When data is dense or continuous, this is the best proximity measure.

The Euclidean distance between two points is the length of the path connecting them. The Pythagorean theorem gives this distance between two points.

Manhattan distance:

Manhattan distance is a metric in which the distance between two points is the sum of the absolute differences of their Cartesian coordinates. In a simple way of saying it is the total sum of the difference between the x-coordinates and y-coordinates.

Suppose we have two points A and B if we want to find the Manhattan distance between them, just we have, to sum up, the absolute x-axis and y – axis variation means we have to find how these two points A and B are varying in X-axis and Y- axis. In a more mathematical way of saying Manhattan distance between two points measured along axes at right angles.

In a plane with p1 at (x1, y1) and p2 at (x2, y2).

Manhattan distance = |x1 – x2| + |y1 – y2|

This Manhattan distance metric is also known as Manhattan length, rectilinear distance, L1 distance or L1 norm, city block distance, Minkowski's L1 distance, taxi-cab metric, or city block distance.

Minkowski distance:

The Minkowski distance is a generalized metric form of Euclidean distance and Manhattan distance.

$$d^{MKD}(i,j) = \sqrt[\lambda]{\sum_{k=0}^{n-1} \left|y_{i,k} - y_{j,k}\right|^{\lambda}}$$

In the equation, d^MKD is the Minkowski distance between the data record i and j, k the index of a variable, n the total number of variables y and λ the order of the Minkowski metric. Although it is defined for any λ > 0, it is rarely used for values other than 1, 2 and ∞.

The way distances are measured by the Minkowski metric of different orders between two objects with three variables (In the image it displayed in a coordinate system with x, y, z-axes).

Cosine similarity:

Cosine similarity metric finds the normalized dot product of the two attributes. By determining the cosine similarity, we would effectively try to find the cosine of the angle between the two objects. The cosine of 0° is 1, and it is less than 1 for any other angle.

It is thus a judgement of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude.

Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1]. One of the reasons for the popularity of cosine similarity is that it is very efficient to evaluate, especially for sparse vectors.

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

Ans:

OLAP:

OLAP (Online Analytical Processing) is the technology behind many Business Intelligence (BI) applications. OLAP is a powerful technology for data discovery, including capabilities for limitless report viewing, complex analytical calculations, and predictive “what if” scenario (budget, forecast) planning.

Need of OLAP:

One main benefit of OLAP is consistency of information and calculations. No matter how much or how fast data is processed through OLAP software or servers, the reporting that results is presented in a consistent presentation, so analysts and executives always know what to look for where. This is especially helpful when comparing information from previous reports to information contained in new ones and projected future ones. It avoids the lengthy discussions about who has the correct information.

"What if" scenarios are some of the most popular uses of OLAP software and are made eminently more possible by multidimensional processing.

Another benefit of multidimensional data presentation is that it allows a manager to pull down data from an OLAP database in broad or specific terms. In other words, reporting can be as simple as comparing a few lines of data in one column of a spreadsheet or as complex as viewing all aspects of a mountain of data.

Also, multidimensional presentation can create an understanding of relationships not previously realized.

OLAP creates a single platform for all the information and business needs; planning, budgeting, forecasting, reporting and analysis.

Last but not least, the learning curve to use OLAP is minimal. The most used interface to analyze data stored in OLAP technology is the well known and loved spreadsheet.

And all of this, of course, can be done in the blink of an eye.

So, OLAP is necessary in Data Mining.

Slicing and dicing:

Slicing is selecting a group of cells from the entire multidimensional array by specifying a specific value for one or more dimensions.

Dicing involves selecting a subset of cells by specifying a range of attribute values. – This is equivalent to defining a subarray from the complete array.

In practice, both operations can also be accompanied by aggregation over some dimensions.

87. When do we need to use discretization and binarization?

Ans:

Discretization in data mining is the process that is frequently used and it is used to transform the attributes that are in continuous format.

On the other hand, binarization is used to transform both the discrete attributes and the continuous attributes into binary attributes in data mining.

It is often necessary to transform a continuous attribute into a categorical attribute (discretization), and both continuous and discrete attributes may need to be transformed into one or more binary attributes (binarization).

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

Ans:

Jaccard Similarity is given by $sij=pp+q+r$

where,

p = # of attributes positive for both objects
q = # of attributes 1 for i and 0 for j
r = # of attributes 0 for i and 1 for j

Whereas, cosine similarity = $A\cdot B\|A\|\|B\|$

where A and B are object vectors.

Simply put, in cosine similarity, the number of common attributes is divided by the total number of possible attributes. Whereas in Jaccard Similarity, the number of common attributes is divided by the number of attributes that exists in at least one of the two objects.

And there are many other measures of similarity, each with its own eccentricities. When deciding which one to use, try to think of a few representative cases and work out which index would give the most usable results to achieve your objective.

The Cosine index could be used to identify plagiarism, but will not be a good index to identify mirror sites on the internet. Whereas the Jaccard index, will be a good index to identify mirror sites, but not so great at catching copy pasta plagiarism (within a larger document).

When applying these indices, you must think about your problem thoroughly and figure out how to define similarity. Once you have a definition in mind, you can go about shopping for an index.

89. Why do we apply aggregation on data?

Ans:

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.

If your analysis requires aggregation, you need to consider two things:

1. **How the outcome will be structured:** Consider the new granularity—that is, what a row represents. If we're looking at voter turnout, is it at the level of political party? Political party and voting district? Political party, voting district, age bracket, and gender? The field or fields that determine what makes up a row are the grouping fields (in Tableau Prep).

2. **How we aggregate multiple values down to a single value:** For example, are we *summing* the number of shirts of each color for a total number of shirts? Are we taking the *maximum* hourly temperature reading over the course of a day and providing the daily max? Are we doing a *count distinct* of IP addresses to hit a webpage and measuring the unique pageviews?

Numeric fields can be aggregated by various mathematical operations depending on the desired outcome. See the full list here. This includes:

- Sum
- Average or Median
- Count or Count Distinct
- Minimum or Maximum
- Or various statistical operations can be performed such as variance or standard deviation.

Dates and text-based fields can be aggregated as count, count distinct, maximum, or minimum (for text, maximum and minimum are based on sort order).