Data Mining Important Questions with Solutions

Tajim Md. Niamat Ullah Akhund
Lecturer, CSE, DIU
MSc, BSc, IIT-JU

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

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 **s**eparated 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 sci kit learn.

KNN classifier-

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

Two limitations of KNN classification:

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

2. Define the term True Positive and False Negative.

Ans:

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.

Ans:

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

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

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

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

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

Key Factors:

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

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

3. Steps Involved

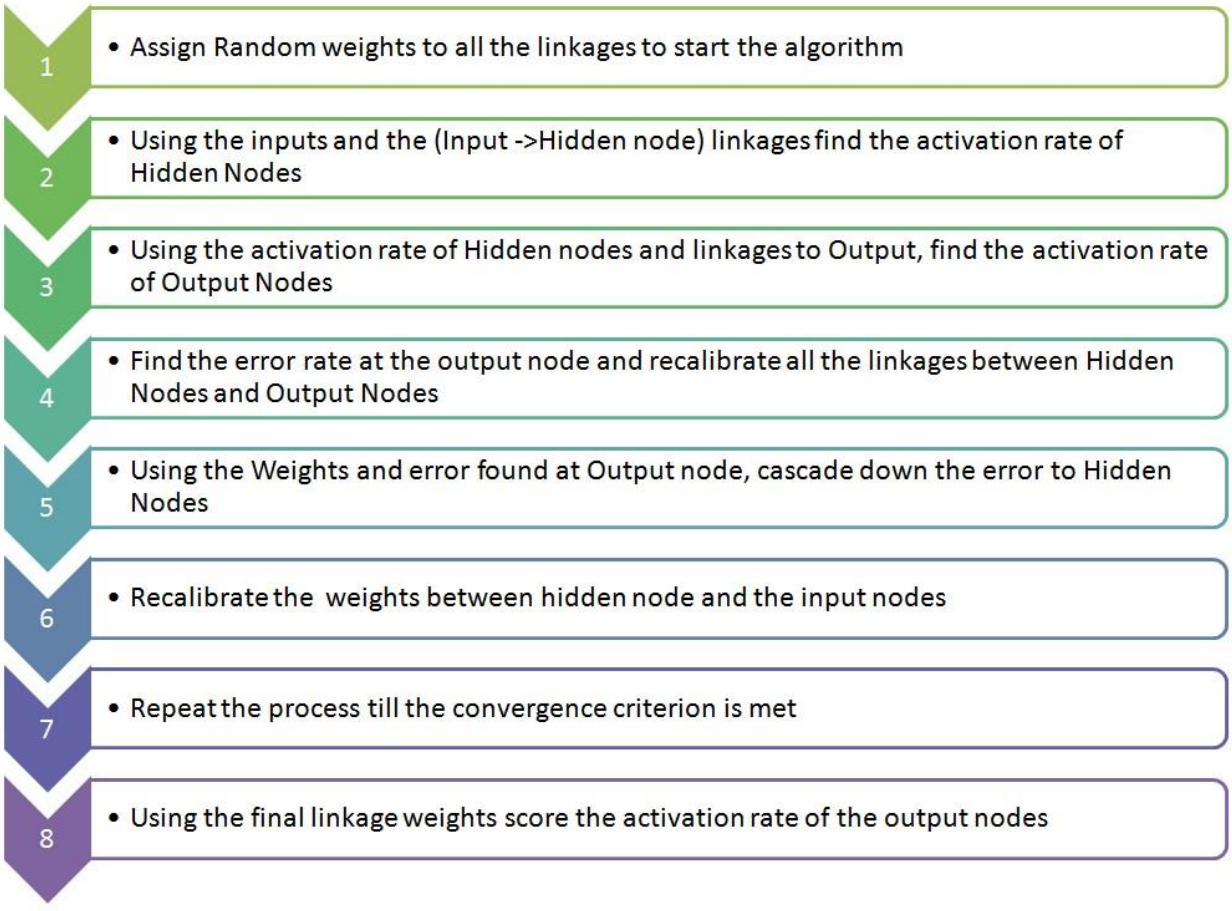
Step 1: Calculate entropy of the target.

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

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

6. How ANN classifier works?

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



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, ,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) □ Can we estimate P(C| A_1, A_2,…,A_n) directly from data?
- Approach:
  - compute the posterior probability P(C | A_1, A_2, …, A_n) for all values of C using the Bayes theorem
  - Choose value of C that maximizes P (C | A_1, A_2, …, A_n)
  - Equivalent to choosing value of C that maximizes P (A_1, A_2, …, 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).

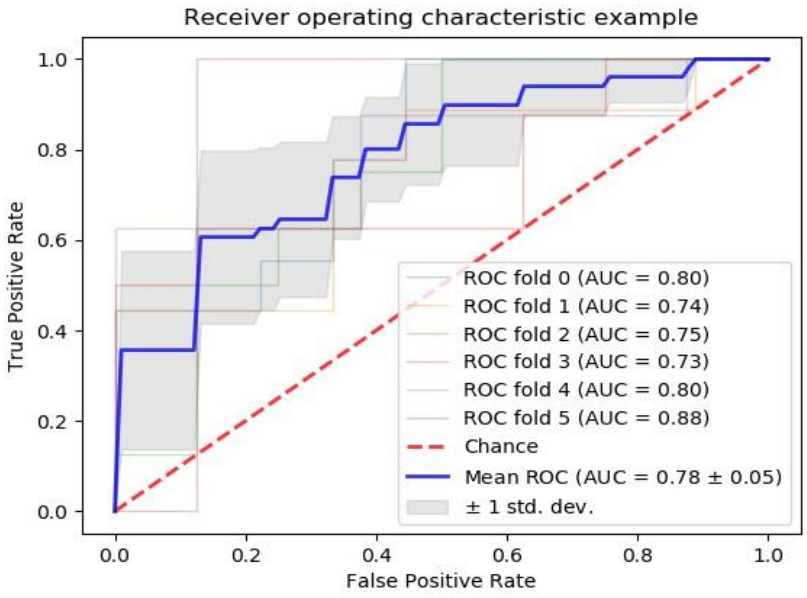
10. What are the functions of ROC curve in validation?

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.

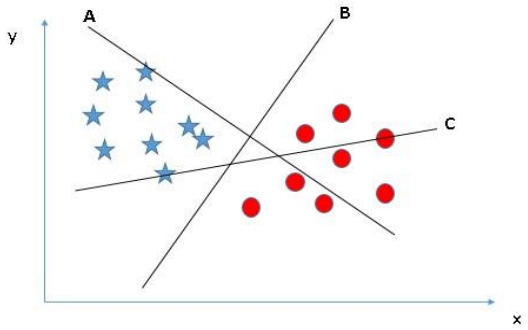
This example shows the ROC response of different datasets, created from K-fold cross-validation. Taking all of these curves, it is possible to calculate the mean area under curve, and see the variance of the curve when the training set is split into different subsets. This roughly shows how the classifier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



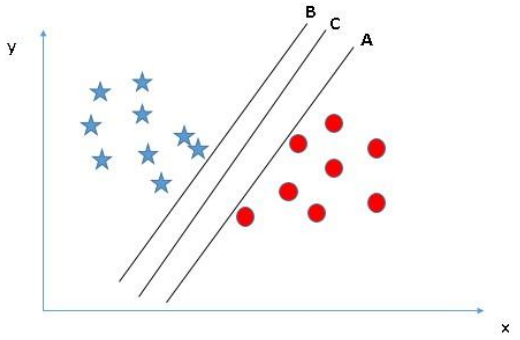
11. How SVM classifier works?

Ans:

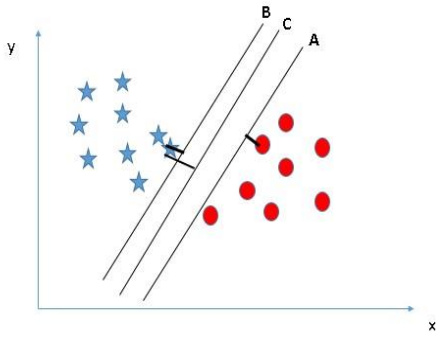
- **Identify the right hyper-plane (Scenario-1):** Here, we have three hyper-planes (A, B and C). Now, identify the right hyper-plane to classify star and circle.



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



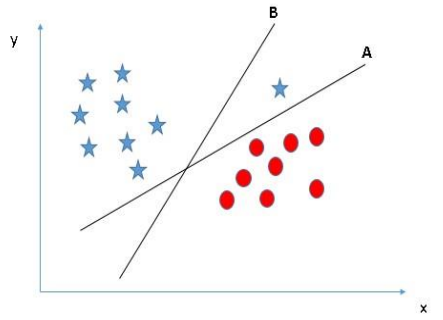
Here, maximizing the distances between nearest data point (either class) and hyper-plane will help us to decide the right hyper-plane. This distance is called as **Margin**. Let's look



Above, you can see that the margin for hyper-plane C is high as compared to both A and

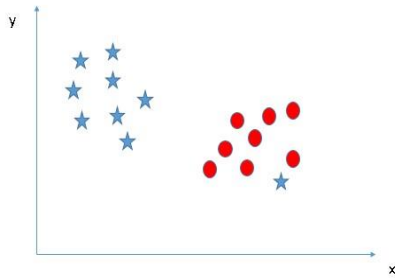
B. Hence, we name the right hyper-plane as C. Another lightning reason for selecting the hyper-plane with higher margin is robustness. If we select a hyper-plane having low margin then there is high chance of miss-classification.

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

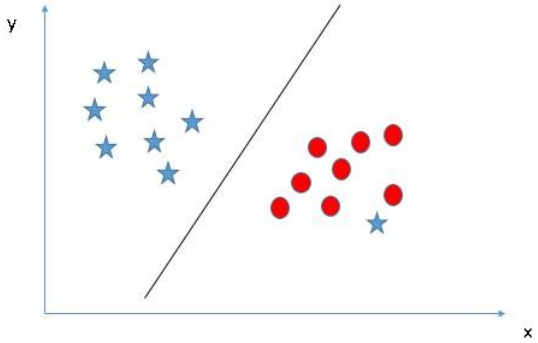


Some of you may have selected the hyper-plane **B** as it has higher margin compared to **A.** But, here is the catch, SVM selects the hyper-plane which classifies the classes accurately prior to maximizing margin. Here, hyper-plane B has a classification error and A has classified all correctly. Therefore,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

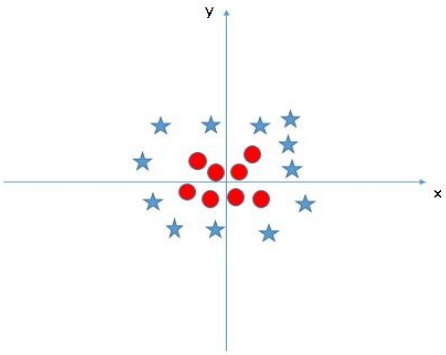
- **Can we classify two classes (Scenario-4)?:** Below, I am unable to segregate the two classes using a straight line, as one of star lies in the territory of other(circle) class as an outlier.



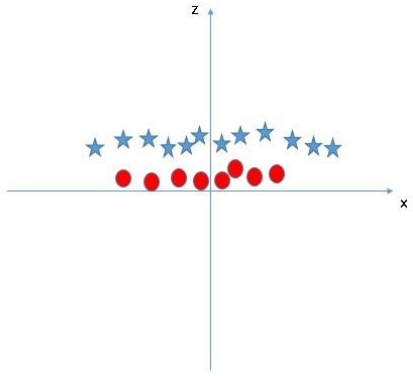
- As I have already mentioned, one star at other end is like an outlier for star class. SVM has a feature to ignore outliers and find the hyper-plane that has maximum margin. Hence, we can say, SVM is robust to outliers.



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



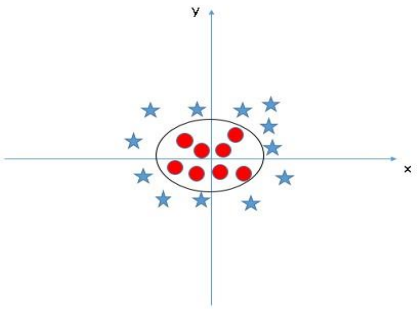
- SVM can solve this problem. Easily! It solves this problem by introducing additional feature. Here, we will add a new feature z=x^2+y^2. Now, let's plot the data points on axis x and z:



In above plot, points to consider are:

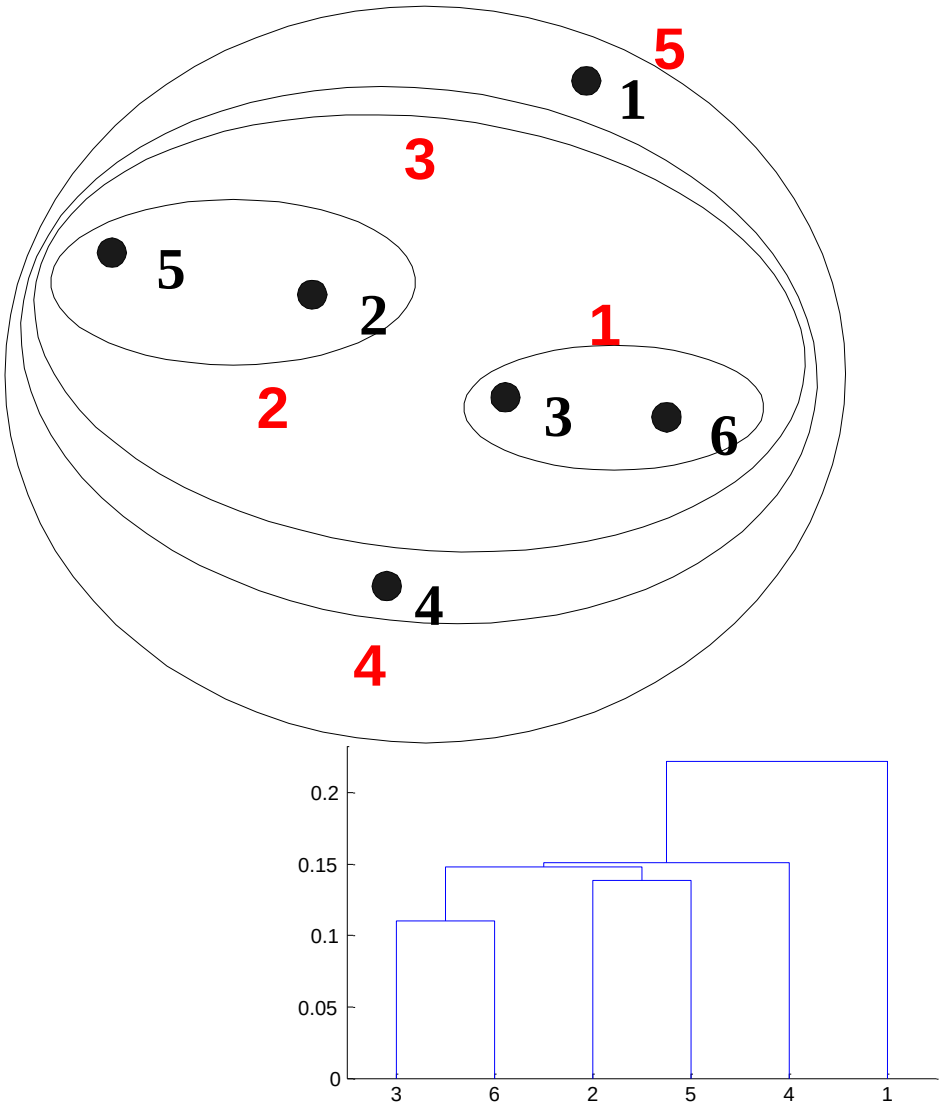
- All values for z would be positive always because z is the squared sum of both x and y
- In the original plot, red circles appear close to the origin of x and y axes, leading to lower value of z and star relatively away from the origin result to higher value of z.

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



12. Math on Draw dendrogram for hierarchical clustering.

Ans: Dendrogram for hierarchical clustering

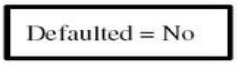


13. **If you have a data set with class attribute and a new data without class attribute. You want to predict the value of class attribute of new data. Write the process of classifying this new data using decision tree classification (Hunt's Algorithm).**

Ans: Let D be the set of training records that reach a node t t

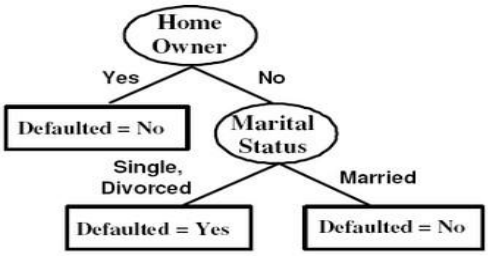
General Procedure:

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



(a) Step 1

(b) Step 2



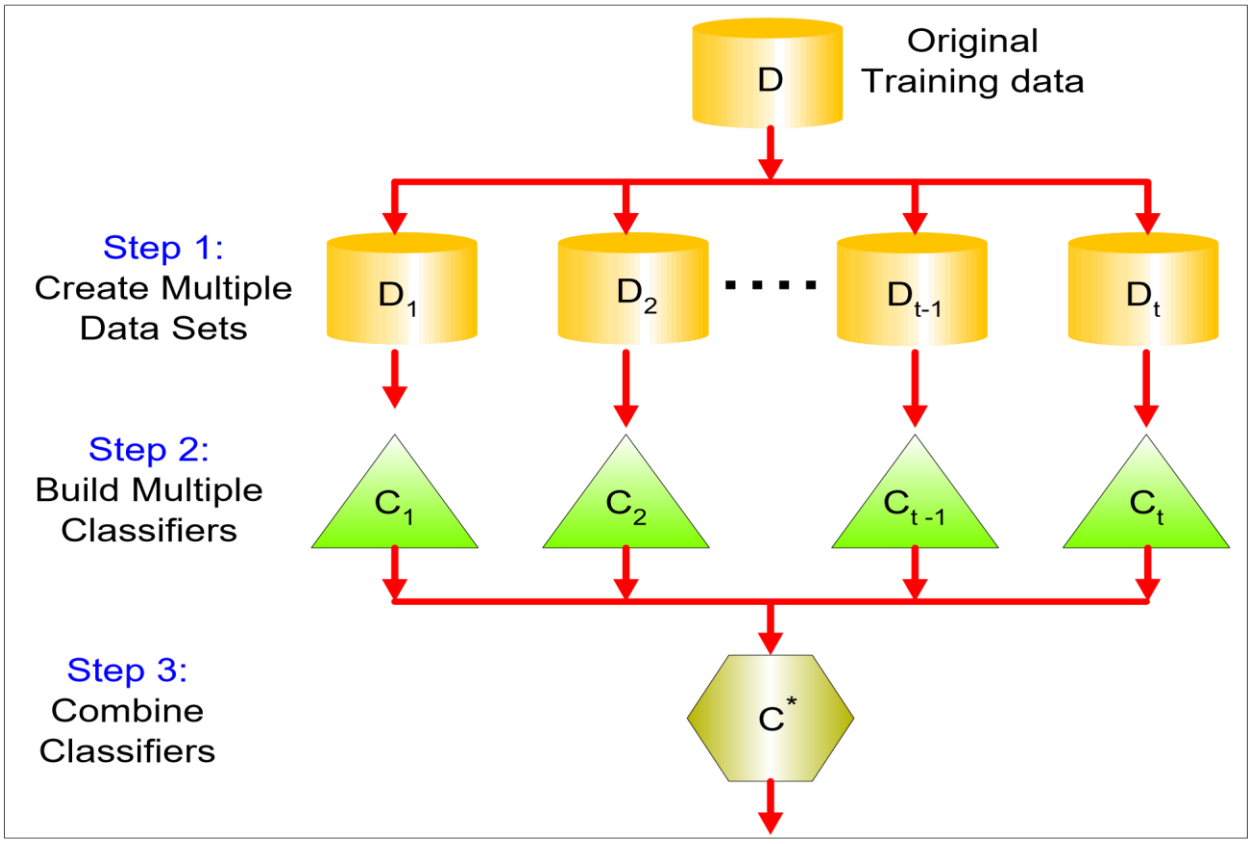
(c) Step 3

(d) Step 4

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

Ans:

- Construct a classifier of training data
- Predict class label of previously unseen records by aggregating predictions made by multiple classifiers



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

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

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

but other measures can be more suitable for a given setting and include the Manhattan, Chebyshev and Hamming distance.
More formally, given a positive integer K, an unseen observation x and a similarity metric d, KNN classifier performs the following two steps:

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

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

$$P(y=j \mid X=x)=1/K$$

$$\sum_{i\in A} I(y^{(i)}=j)$$

Finally, our input x gets assigned to the class with the largest probability. An alternate way of understanding KNN is by thinking about it as calculating a decision boundary (i.e. boundaries for more than 2 classes) which is then used to classify new points.

16. 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.

17. 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:
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:

Gini (t) = 1-$\sum[p(j \mid t)]$^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-$\sum[p(j \mid t)]$^2

C1	0
C2	6

P(C1)=0/6 = 0 P(C2) = 6/6 = 1
Gini = 1 - P(C1)² - P(C2)² = 1-0-1 =0

C1	1
C2	5

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

C1	2

C2	4

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

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

20. What are the advantage of tree-based classification?

Ans: <u>Advantage of tree-based classification:</u>

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

21. What are the main principal of Bayesian Classification?

Ans: The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often perform better in many complex real-world situations. Requires a small amount of training data to estimate the parameters.

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

P(A) : Prior probability of hypothesis A

P(B) : Prior probability of training data B

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

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

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

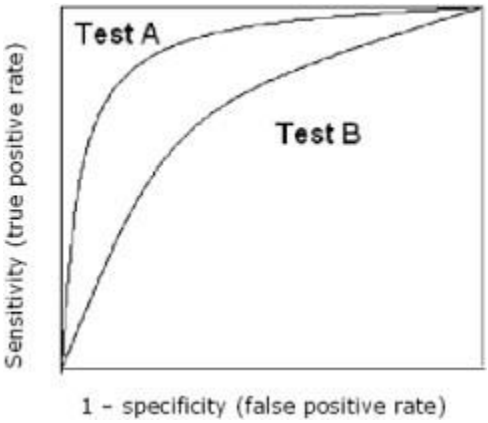
22. What are the sequential steps in doing classification of a set of data?

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: Support Vector Machine (SVM): "Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for either classification or regression challenges. However, it is mostly used in classification problems. It uses a technique called the kernel trick to transform user data. SVM is capable of doing both classification and regression. SVM also works very well with high-dimensional data and avoids the curse of dimensionality problem.

Uses of SVM:

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

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

Ans: 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 cutoff 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.

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**.

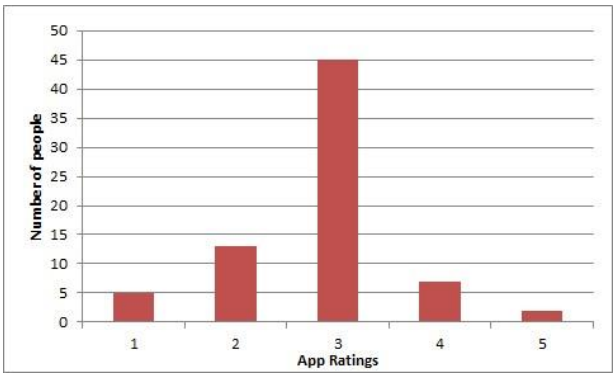
Ensemble methods help to minimize these factors. These methods are designed to improve the stability and the accuracy of Machine Learning algorithms. **Simple Ensemble techniques**

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)

Main principal of Gini Index-

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

2. *Scale of independence*
   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 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.

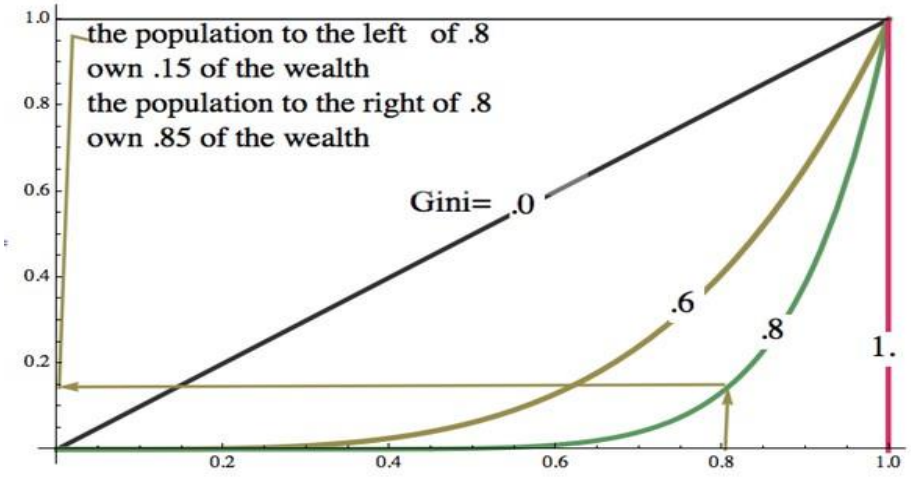
33. Why and when do we use Gini coefficient or entropy?

Ans:

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

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

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

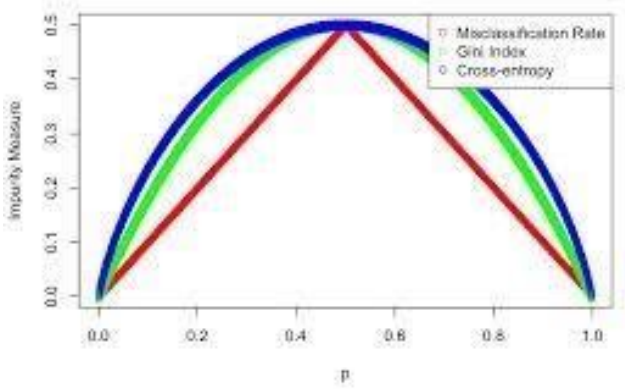


the y axis is the fraction of total wealth owned by the fraction on the x axis.

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

34. Why entropy is used instead of Gini index?

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



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

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

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

Ans: The algorithm of K-nearest neighbor:

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

Distance functions

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

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

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

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

36. Write the limitations of KNN.
Ans*:*

Limitations:

- 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 nonparametric 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.

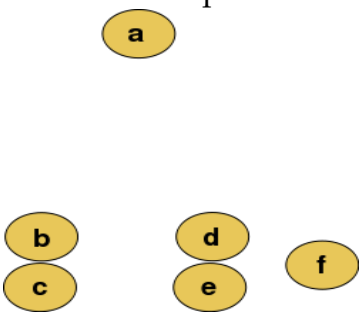
The algorithm:

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.

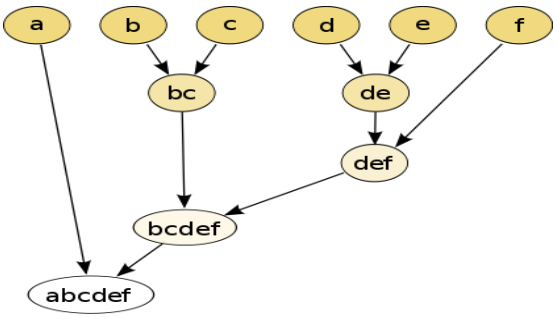
Each node in the cluster tree contains a group of similar data; Nodes group on the graph next to other, similar nodes. Clusters at one level join with clusters in the next level up, using a degree of similarity; The process carries on until all nodes are in the tree, which gives a visual snapshot of 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 $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ 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.

40. What are the different types of clustering?

Ans:

1. Partitional Clustering:

– A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset.

2. Hierarchical clustering:

- A set of nested clusters organized as a hierarchical tree.

3. 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.

4. 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.

5. 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.

6. 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.

7. Shared Property or Conceptual Clusters:

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

40. What are the different types of clustering?

Clustering methods are used to identify groups of similar objects in a multivariate data sets collected from fields such as marketing, bio-medical and geo-spatial. They are different **types of clustering** methods, including:

- Partitioning methods
- Hierarchical clustering
- Fuzzy clustering
- Density-based clustering
- Model-based clustering
-

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

Ans:

Clustering, in the context of databases, refers to the ability of several servers or instances to connect to a single database. An instance is the collection of memory and processes that interacts with a database, which is the set of physical files that actually store data.

Clustering offers two major advantages, especially in high-volume database environments:

- Fault tolerance: Because there is more than one server or instance for users to connect to, clustering offers an alternative, in the event of individual server failure.
- Load balancing: The clustering feature is usually set up to allow users to be automatically allocated to the server with the least load.

2. Types of Clustering

Broadly speaking, clustering can be divided into two sub-groups:

- **Hard Clustering:** In hard clustering, each data point either belongs to a cluster completely or not. For example, in the above example each customer is put into one group out of the 10 groups.
- **Soft Clustering**: In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned. For example, from the above scenario each costumer is assigned a probability to be in either of 10 clusters of the retail store.

3. Types of clustering algorithms

Since the task of clustering is subjective, the means that can be used for achieving this goal are plenty. Every methodology follows a different set of rules for defining the 'similarity' among data points. In fact, there are more than 100 clustering algorithms known. But few of the algorithms are used popularly, let's look at them in detail:

- **Connectivity models:** As the name suggests, these models are based on the notion that the data points closer in data space exhibit more similarity to each other than the data points lying farther away. These models can follow two approaches. In the first approach, they start with classifying all data points into separate clusters & then aggregating them as the distance decreases. In the second approach, all data points are classified as a single cluster and then partitioned as the distance increases. Also, the choice of distance function is subjective. These models are very easy to interpret but lacks scalability for handling big datasets. Examples of these models are hierarchical clustering algorithm and its variants.

- **Centroid models:** These are iterative clustering algorithms in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters. K-Means clustering algorithm is a popular algorithm that falls into this category. In these models, the no. of clusters required at the end have to be mentioned beforehand, which makes it important to have prior knowledge of the dataset. These models run iteratively to find the local optima.

- **Distribution models:** These clustering models are based on the notion of how probable is it that all data points in the cluster belong to the same distribution (For example: Normal, Gaussian). These models often suffer from overfitting. A popular example of these models is Expectation-maximization algorithm which uses multivariate normal distributions.

- **Density Models:** These models search the data space for areas of varied density of data points in the data space. It isolates various different density regions and assign the data points within these regions in the same cluster. Popular examples of density models are DBSCAN and OPTICS.

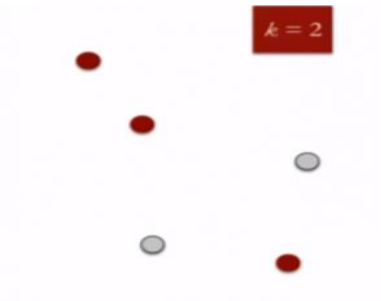
44. What is clustering? Describe the steps of K-means clustering.

Ans:

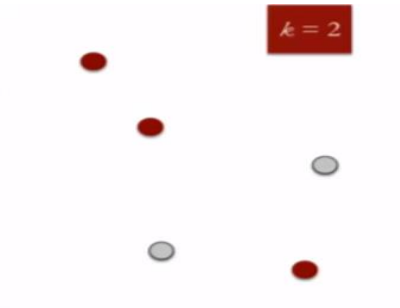
4. K Means Clustering

K means is an iterative clustering algorithm that aims to find local maxima in each iteration. This algorithm works in these 5 steps:

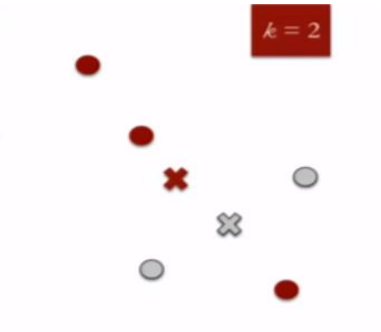
1. Specify the desired number of clusters K : Let us choose k=2 for these 5 data points in 2-D space.



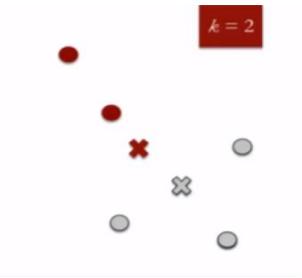
2. Randomly assign each data point to a cluster: Let's assign three points in cluster 1 shown using red color and two points in cluster 2 shown using grey color.



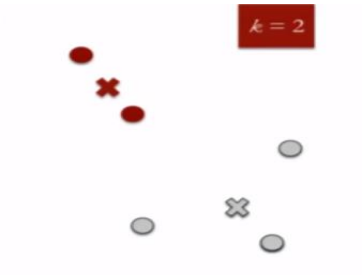
3. Compute cluster centroids: The centroid of data points in the red cluster is shown using red cross and those in grey cluster using grey cross.



4. Re-assign each point to the closest cluster centroid: Note that only the data point at the bottom is assigned to the red cluster even though it's closer to the centroid of grey cluster. Thus, we assign that data point into grey cluster



5. Re-compute cluster centroids: Now, re-computing the centroids for both the clusters.



6. Repeat steps 4 and 5 until no improvements are possible: Similarly, we'll repeat the 4th and 5th steps until we'll reach global optima. When there will be no further switching of data points between two clusters for two successive repeats. It will mark the termination of the algorithm if not explicitly mentioned.

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

The term **cluster validation** is used to design the procedure of evaluating the goodness of clustering algorithm results. This is important to avoid finding patterns in a random data, as well as, in the situation where you want to compare two clustering algorithms.

Generally, clustering validation statistics can be categorized into 3 classes (Charrad et al. 2014,Brock et al. (2008), Theodoridis and Koutroumbas (2008)):

1. **Internal cluster validation**, which uses the internal information of the clustering process to evaluate the goodness of a clustering structure without reference to external information. It can be also used for estimating the number of clusters and the appropriate clustering algorithm without any external data.
2. **External cluster validation**, which consists in comparing the results of a cluster analysis to an externally known result, such as externally provided class labels. It measures the

extent to which cluster labels match externally supplied class labels. Since we know the "true" cluster number in advance, this approach is mainly used for selecting the right clustering algorithm for a specific data set.

3. **Relative cluster validation**, which evaluates the clustering structure by varying different parameter values for the same algorithm (e.g.,: varying the number of clusters k). It's generally used for determining the optimal number of clusters.

46. Write the steps of k-means clustering?

Disavantages

• Difficult to predict the number of clusters (K---Value)

• IniHal seeds have a strong impact on the final results

• The order of the data has an impact on the final results

• Pensive to scale: rescaling your datasets (normalizing or standardizing) will completely

Change results. While this itself is not bad, not realizing that you have to spend extra a4en (onto scaling your data might be bad.

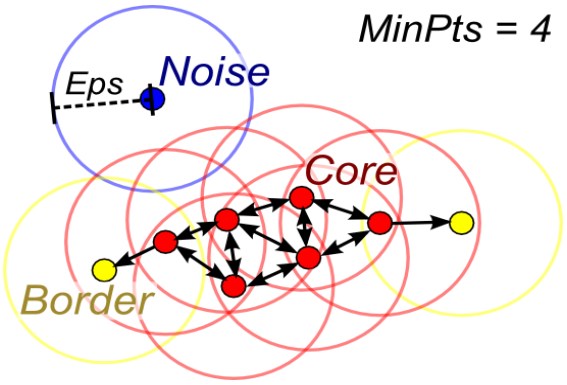
49. Explain with a pictorial example of Core Point, Noise Point and Border Point.

Ans:

The points here are classified as core points, border points or noise.

- A point p is a core point if at least min Pts points are within distance ε of it, and those points are said to be directly reachable from p. No points are directly reachable from a non-core point.
- A point q is reachable from p if there is a path p1, ..., pn with p1 = p and pn = q, where each pi+1 is directly reachable from pi (so all the points on the path must be core points, with the possible exception of q).
- All points not reachable from any other point are outlier

Consider the following image:

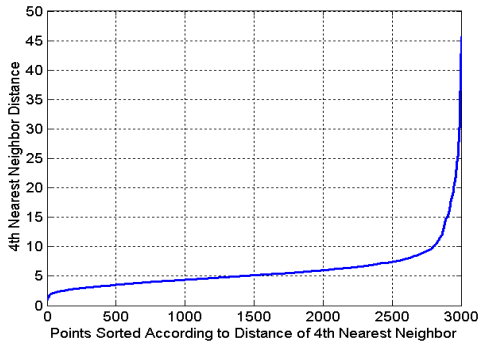


50. In density-based clustering, how do you select epsilon and distance? What is the logic of this process?

Ans:

Selecting epsilon and distance:

- 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 principal of DBSCAN clustering?

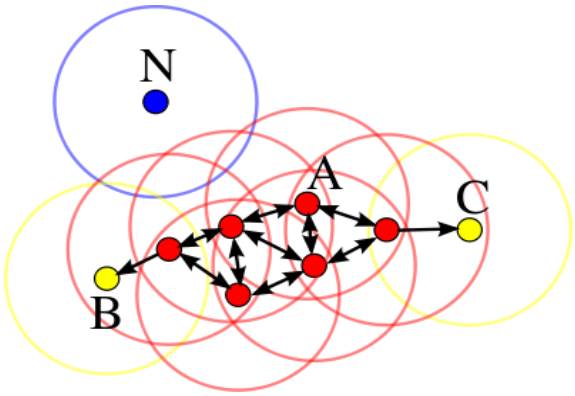
Ans:

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:

- For supervised classification we have a variety of measures to evaluate how good our model is
  - Accuracy, precision, recall
  - For cluster analysis, the analogous question is how to evaluate the "goodness" of the resulting clusters?
- But "clusters are in the eye of the beholder"!
- Then why do we want to evaluate them?
  - To avoid finding patterns in noise
  - To compare clustering algorithms
  - To compare two sets of clusters
  - To compare two clusters
- Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.
  - External Index: Used to measure the extent to which cluster labels match externally supplied class labels.
    - Entropy

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

- Sometimes these are referred to as criteria instead of indices
  - However, sometimes criterion is the general strategy and index is the numerical measure that implements the criterion.

53. When we need to use density-based clustering?

Ans:

When to use:

- 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?

Ans:

- 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.

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 clusters, we can find similarities between customers, for example, the customer A have bought 1 pen, 1 book and 1 scissors and the customer B have bought 1 book and 1 scissors, then we can recommend 1 pen to the customer B. This is just a little example of use of DBSCAN, but it can be used in a lot of applications in several areas.

56. Write one application of DBSCAN clustering.

Ans:

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

57. Data may affect 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. Inaccurate Customer and Contact Data

The ability to efficiently and systematically reach current and potential customers is crucial to any business. The challenge is that customer contact data touches nearly every aspect of an organization, and contact data flows through each phase of the customer lifecycle. This includes data collected at the time of purchase/order placement to the moment cross-sell offers are auto-generated and social media marketing campaigns are launched.

2. Big data makes it hard to focus

Big data is often a challenge for CIOs, IT departments, and marketers because of the nature of traditional relational databases. These complex enterprise data systems often struggle to keep pace with the volume of data collected and the disparity of information sources. When bad data enters these systems— even when it's a small error such as an incorrect courtesy title (Ms. vs Mr.) —there's a compounded effect as data is compiled, analyzed, and filtered throughout the organization's CRM, ERP, billing, and other enterprise transactional systems.

3. Duplicate or Obsolete Data

Since contact data is rarely static, obsolete data and duplicate records are common data quality challenges. Every day, people move, marry, and change their names and contact preferences, resulting in the business need for effective data verification methods at each collection point. This is particularly true for organizations that collect information at multiple stages during the customer lifecycle and from multiple channels (i.e. call centers, websites, and retail locations). Duplicate or obsolete records make it nearly impossible for companies to effectively communicate with prospects and customers.

4. Compliance issues

Data security and compliance requirements come from various sources and may include corporate requirements in addition to industry and government mandates such as HIPAA or PCI Data Security Standards (PCI DSS). Failure to meet these rules can lead to heavy fines, and, perhaps even more costly, loss of customer loyalty. Often, requirements outlined by mandates such as HIPAA and PCI make a strong case for implementing a comprehensive data quality management program.

5. Reconciling data quality issues during a merger

During any merger or acquisition there is the need to integrate multiple disparate source systems into a centralized or extended data warehouse environment. After integration, the data is generally filtered downstream into a comprehensive business analytics solution, or a series of BI solutions for each business segment. The major data quality challenge here is that with multiple data source systems, transactions cannot be properly attributed to a single source.

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

Data mining can be performed on following types of data

- Relational databases
- Data warehouses
- Advanced DB and information repositories
- Object-oriented and object-relational databases
- Transactional and Spatial databases
- Heterogeneous and legacy databases
- Multimedia and streaming database
- Text databases
- Text mining and Web mining

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:

Data for process mining can come from many different places. One of the big advantages of process mining is that it is not specific to some kind of system. Any workflow or ticketing system, ERPs, data warehouses, click-streams, legacy systems, and even data that was collected manually in Excel, can be analyzed as long as a Case ID, an Activity name, and a Timestamp column can be identified (see Data Requirements).

But most of that data was not originally collected for process mining purposes. And especially data that has been manually entered can always contain errors. How do you make sure that errors in the data will not jeopardize your analysis results?

Data quality is an important topic for *any* data analysis technique: If you base your analysis results on data, then you have to make sure that the data is sound and correct. Otherwise, your results will be wrong! If you show your analysis results to a business

user and they turn out to be incorrect due to some data problems, then you can lose their trust into process mining forever.

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 missings 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:

The different types of attributes are as follows

- Single valued attributes
- Multi valued attributes
- Compound /Composite attributes
- Simple / Atomic attributes
- Stored attributes
- Derived attributes

- Complex attributes
- Key attributes
- Non key attributes
- Required attributes
- Optional/ null value attributes

The detailed explanation of all the attributes is as follows:

Single Valued Attributes: It is an attribute with only one value.

- Example: Any manufactured product can have only one serial no. , but the single valued attribute cannot be simple valued attribute because it can be subdivided. Likewise in the above example the serial no. can be subdivided on the basis of region, part no. etc.

Multi Valued Attributes: These are the attributes which can have multiple values for a single or same entity.

- Example: Car's colors can be divided into many colors like for roof, trim.
- The notation for multi valued attribute is:

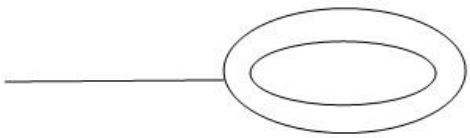


Fig3: Multi value attribute notation

Compound / Composite attributes: This attribute can be further divided into more attributes.

- The notation for it is:

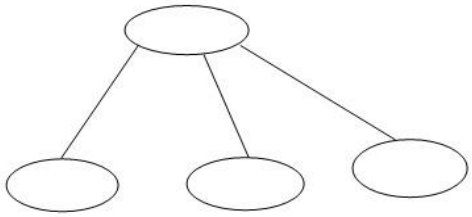


Fig 4: Compound / Composite attribute notation

- Example: Entity Employee Name can be divided into sub divisions like FName, MName, LName.

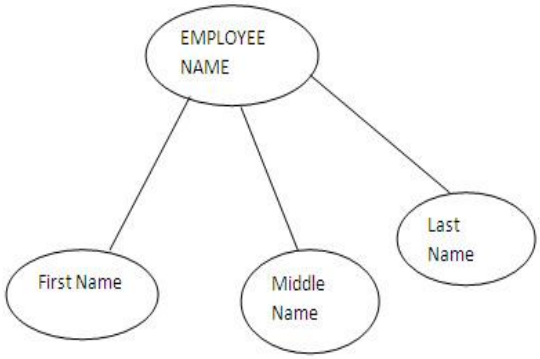


Fig 5: Sample of compound / composite attribute

Simple / Atomic Attributes: The attributes which cannot be further divided are called as simple / atomic attributes.

- Example: The entities like age, marital status cannot be subdivided and are simple attributes.

Stored Attributes: Attribute that cannot be derived from other attributes are called as stored attributes.

- Example: Birth date of an employee is a stored attribute.

Derived Attributes: These attributes are derived from other attributes. It can be derived from multiple attributes and also from a separate table.

- Example: Today's date and age can be derived. Age can be derived by the difference between current date and date of birth.

- The notation for the derived attribute is:

Fig 6: Notation of derived attribute

Complex Attributes: For an entity, if an attribute is made using the multi valued attributes and composite attributes then it is known as complex attributes.

- Example: A person can have more than one residence; each residence can have more than one phone.

Key Attributes: This attribute represents the main characteristic of an entity i.e. primary key. Key attribute has clearly different value for each element in an entity set.

- Example: The entity student ID is a key attribute because no other student will have the same ID.

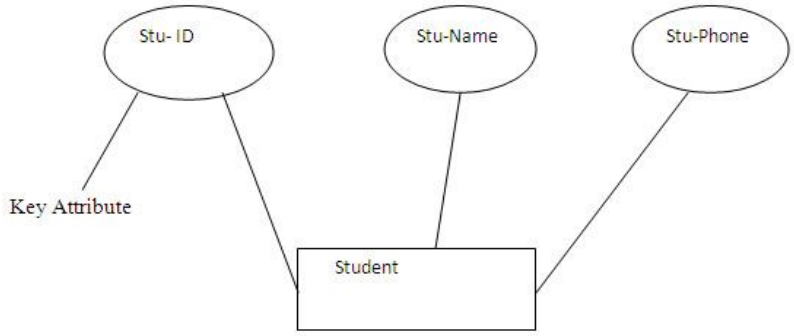


Fig 7: Sample of key attribute

Non- Key Attributes: Excluding the candidate key attributes in an entity set are the non key attributes.

- Example: First name of a student or employee is a non-key attribute as it does not represent main characteristic of an entity.

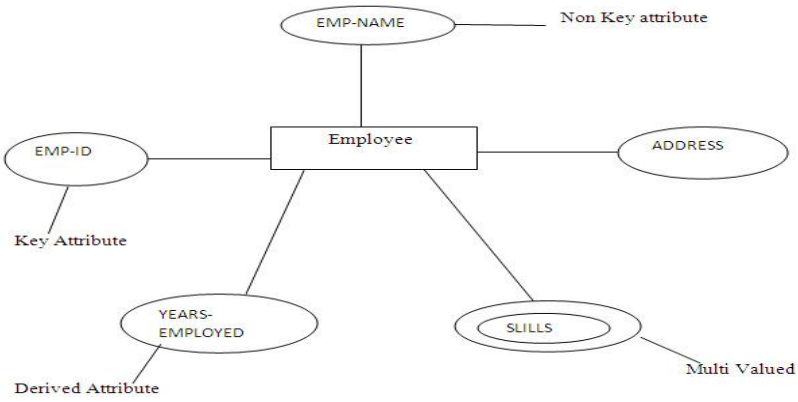


Fig 8: Sample of non-key attribute

Required Attributes: Required attribute must have a data because they describe the vital part of entity.

- Example: Taking the example of a college, there the student's name is a vital thing.

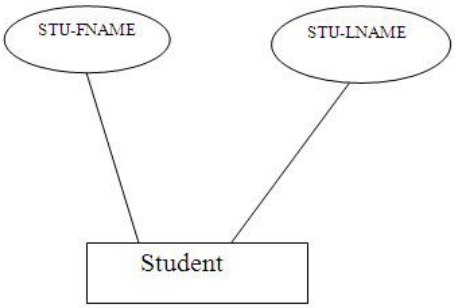


Fig 9: sample of required attribute

Optional / Null value Attributes: It does not have a value and can be left blank, it's optional can be filled or cannot be.

- Example: Considering the entity student there the student's middle name and the email ID is optional.

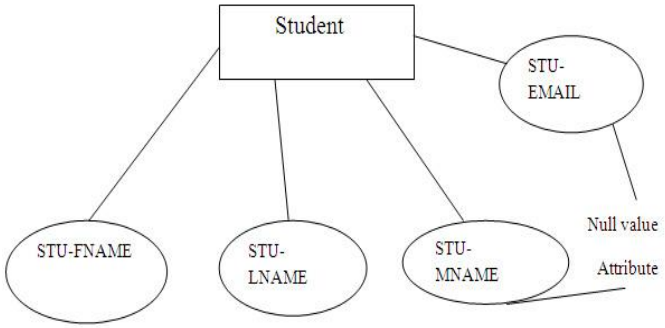


Fig 10: Sample of optional / null attribute

These are different types of attributes and they all play a vital role in the database management system. Generally an oval is used to represent an attribute as shown below:

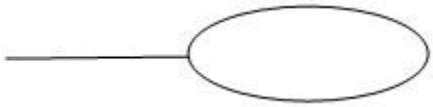


Fig 11: Notation of an attribute

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

Ans:

I will prefer MahalaNobis Distance and the reason is here-

Euclidean distance only makes sense when all the dimensions have the same units (like meters), since it involves adding the squared value of them. When you are dealing with probabilities, a lot of times the features have different units. For example: I have a model for men and women, based on their weight [Kg] and height [m]. I know the mean and covariance for each. Now I get a new measurement set of weight and height and I try to decide if it's a man or a woman. I can use the Mahalanobis distance from the models of both men and women to decide which is closer, meaning which is more probable.

The Mahalnobis distance transforms the random vector into a zero-mean vector with an identity matrix for covariance. In that space, the Euclidean distance is safely applied.

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

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.

This page then contains a brief discussion of several important similarity metric.

Euclidean Distance

Euclidean Distance between two points is given by Minkowski distance metric. It can be used in one-, tow-, or higher-dimensional space. The formula of Euclidean distance is as following. [3]

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

where n is the number of dimensions. It measures the numerial difference for each corresponding attributes of point p and point q. Then it combines the square of differencies in each dimension into an overal 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 represents 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. [4]

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 coefficent. Tanimoto coefficent 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:

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

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.

This page then contains a brief discussion of several important similarity metric.

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. [3]

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

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 data set (or dataset) is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question.

In general, there are many types of variable that can be used to measure the properties of an object. A lack of understanding of the differences between the various types can lead to problems with any form of data analysis. At least six main types of variable can be distinguished.

Nominal Dataset:

A variable used to put objects into categories, e.g. the name or color of an object. A nominal variable may be numerical in form, but the numerical values have no mathematical interpretation. For example we might label 10 people as numbers 1,2,3,...,10, but any arithmetic with such values, e.g. 1 + 2 = 3 would be meaningless. They are simply labels. A classification can be viewed as a nominal variable which has been designated as of particular importance.

Binary Dataset:

A binary variable is a special case of a nominal variable that takes only two possible values: true or false, 1 or 0 etc.

Ordinal Dataset:

Ordinal variables are similar to nominal variables, except that an ordinal variable has values that can be arranged in a meaningful order, e.g. small, medium, large.

Integer Dataset:

Integer variables are ones that take values that are genuine integers, for example 'number of children'. Unlike nominal variables that are numerical in form, arithmetic with integer variables is meaningful (1 child + 2 children = 3 children etc.).

Interval-scaled Dataset:

Interval-scaled variables are variables that take numerical values which are measured at equal intervals from a zero point or origin. However, the origin does not imply a true absence of the measured characteristic. Two well-known examples of interval-scaled variables are the Fahrenheit and Celsius temperature scales. To say that one temperature measured in degrees Celsius is greater than another or greater than a constant value such as 25 is clearly meaningful, but to say that one temperature measured in degrees Celsius is twice another is meaningless. It is true that a temperature of 20 degrees is twice as far from the zero value as 10 degrees, but the zero value has been selected arbitrarily and does not imply 'absence of temperature'. If the temperatures are converted to an equivalent scale, say degrees Fahrenheit, the 'twice' relationship will no longer apply.

Ratio-scaled Dataset:

Ratio-scaled variables are similar to interval-scaled variables except that the zero point does reflect the absence of the measured characteristic, for example Kelvin temperature and molecular weight. In the former case the zero value corresponds to the lowest possible temperature 'absolute zero', so a temperature of 20 degrees Kelvin is twice one of 10 degrees Kelvin. A weight of 10 kg is twice one of 5 kg, a price of 100 dollars is twice a price of 50 dollars etc.

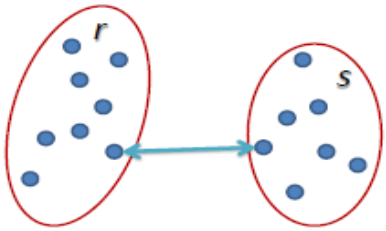
68. How do you calculate distance between two clusters?

Ans:

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

Single Linkage

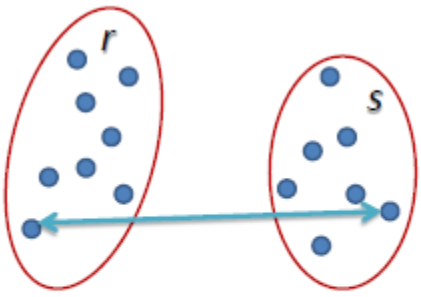
In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster. For example, the distance between clusters "r" and "s" to the left is equal to the length of the arrow between their two closest points.



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

Complete Linkage

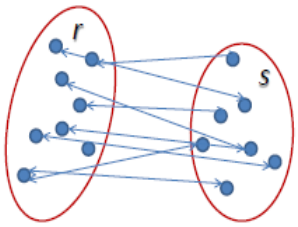
In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster. For example, the distance between clusters "r" and "s" to the left is equal to the length of the arrow between their two furthest points.



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

Average Linkage

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



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

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

Ans:

Hierarchical Clustering

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.

Answer:

Procedure of hierarchical clustering:

1. Start by assigning each item to a 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 the same as the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)

Step 3 can be done in different ways, which is what distinguishes single-linkage from complete-linkage and average-linkage clustering.

In single-linkage clustering (also called the connectedness or minimum method), we consider the distance between one cluster and another cluster to be equal to the shortest distance from any member of one cluster to any member of the other cluster. If the data consist of similarities, we consider the similarity between one cluster and another cluster to be equal to the greatest similarity from any member of one cluster to any member of the other cluster.

In complete-linkage clustering (also called the diameter or maximum method), we consider the distance between one cluster and another cluster to be equal to the greatest distance from any member of one cluster to any member of the other cluster.

In average-linkage clustering, we consider the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster.

A variation on average-link clustering is the UCLUS method of uses

the median distance, which is much more outlier-proof than the average distance.

Example:

Clustering starts by computing a distance between every pair of units that you want to cluster. A distance matrix will be symmetric (because the distance between x and y is the same as the distance between y and x) and will have zeroes on the diagonal (because every item is distance zero from itself). The table below is an example of a distance matrix. Only the lower triangle is shown, because the upper triangle can be filled in by reflection.

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

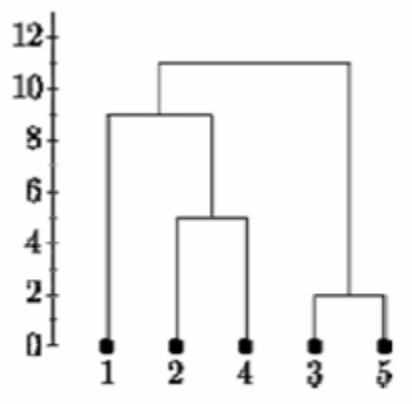
Now lets start clustering. The smallest distance is between three and five and they get linked up or merged first into the cluster '35'.

To obtain the new distance matrix, we need to remove the 3 and 5 entries, and replace it by an entry "35". Since we are using complete linkage clustering, the distance between "35" and every other item is the maximum of the distance between this item and 3 and this item and 5. For example, d(1,3)= 3 and d(1,5)=11. So, D(1,"35")=11. This gives us the new distance matrix. The items with the smallest distance get clustered next. This will be 2 and 4.

	35	1	2	4
35	0			
1	11	0		
2	10	9	0	
4	9	6	5	0

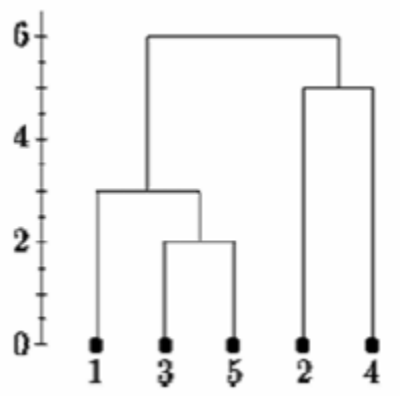
Continuing in this way, after 6 steps, everything is clustered. This is summarized below. On this plot, the y-axis shows the distance between the objects at the time they

were clustered. This is called the cluster height. Different visualizations use different measures of cluster height.



Complete Linkage

Below is the single linkage dendrogram for the same distance matrix. It starts with cluster "35" but the distance between "35" and each item is now the minimum of d(x,3) and d(x,5). So c(1,"35")=3.

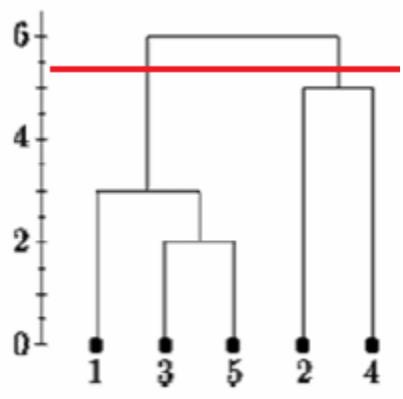


Single Linkage

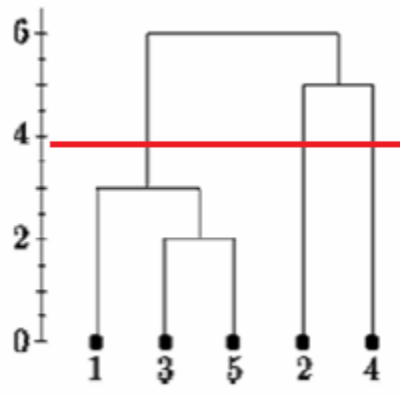
Determining clusters

One of the problems with hierarchical clustering is that there is no objective way to say how many clusters there are.

If we cut the single linkage tree at the point shown below, we would say that there are two clusters.

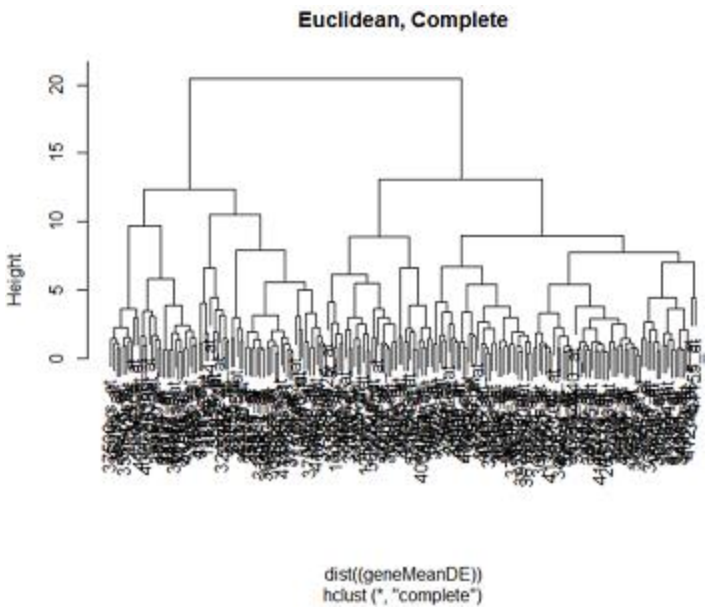


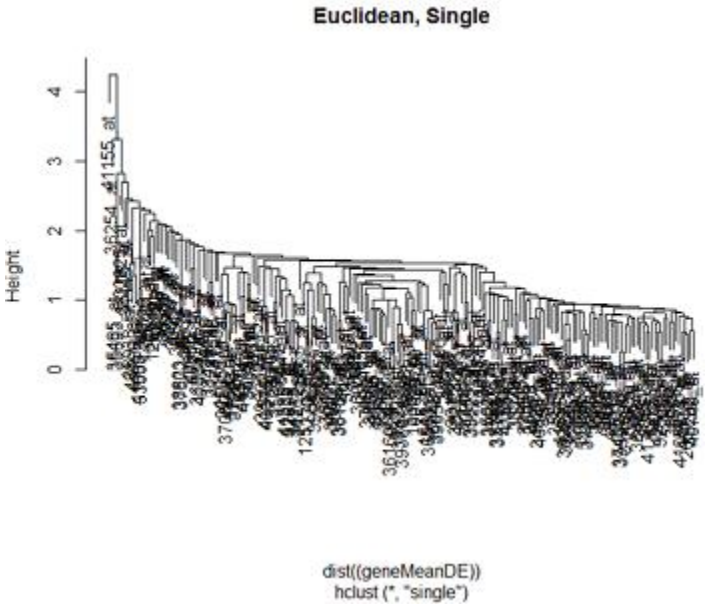
However, if we cut the tree lower we might say that there is one cluster and two singletons.

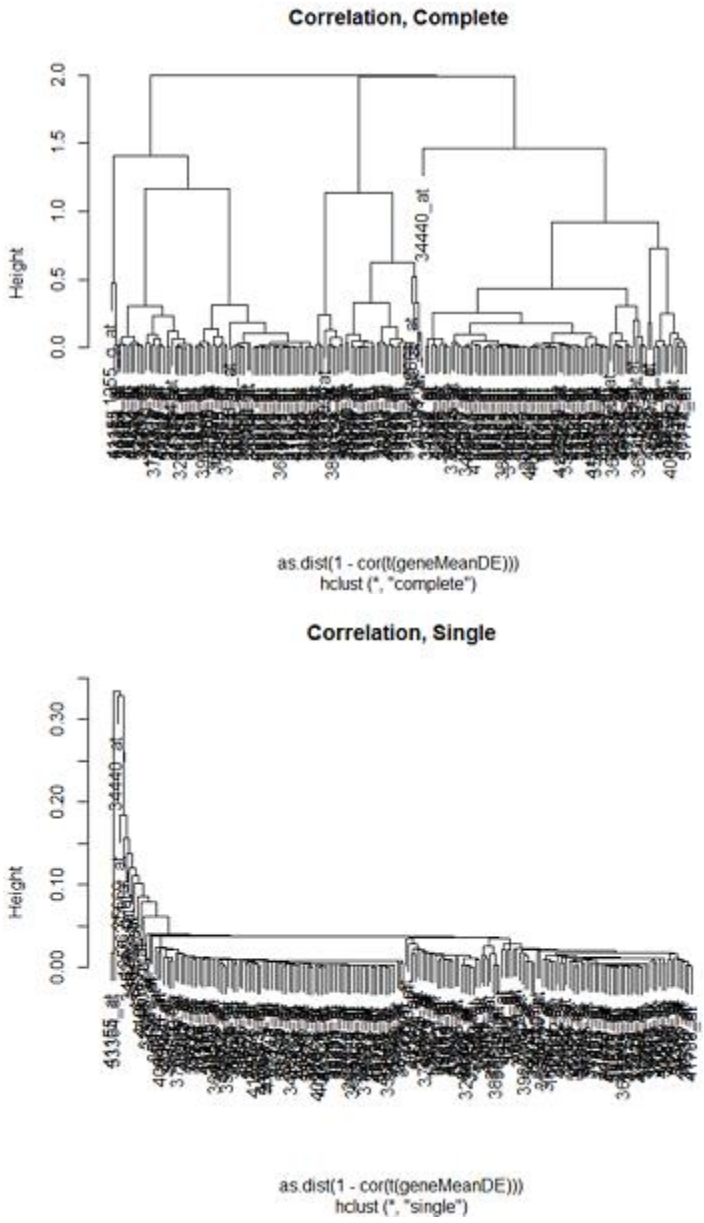


There is no commonly agreed-upon way to decide where to cut the tree.

Let's look at some real data. In homework 5 we consider gene expression in 4 regions of 3 human and 3 chimpanzee brains. The RNA was hybridized to Affymetrix human gene expression microarrays. We normalized the data using RMA and did a differential expression analysis using LIMMA. Here we selected the 200 most significantly differentially expressed genes from the study. We cluster all the differentially expressed genes based on their mean expression in each of the 8 species by brain region treatments Here are the clusters based on Euclidean distance and correlation distance, using complete and single linkage clustering.


Euclidean, Complete
Height
0
5
10
15
20
dist((geneMeanDE))
hclust (*, "complete")


Euclidean, Single
Height
0
1
2
3
4
dist((geneMeanDE))
hclust (*, "single")



We can see that the clustering pattern for complete linkage distance tends to create compact clusters of clusters, while single linkage tends to add one point at a time to the cluster, creating long stringy clusters. As we might expect from our discussion of distances, Euclidean distance and correlation distance produce very different dendrograms.

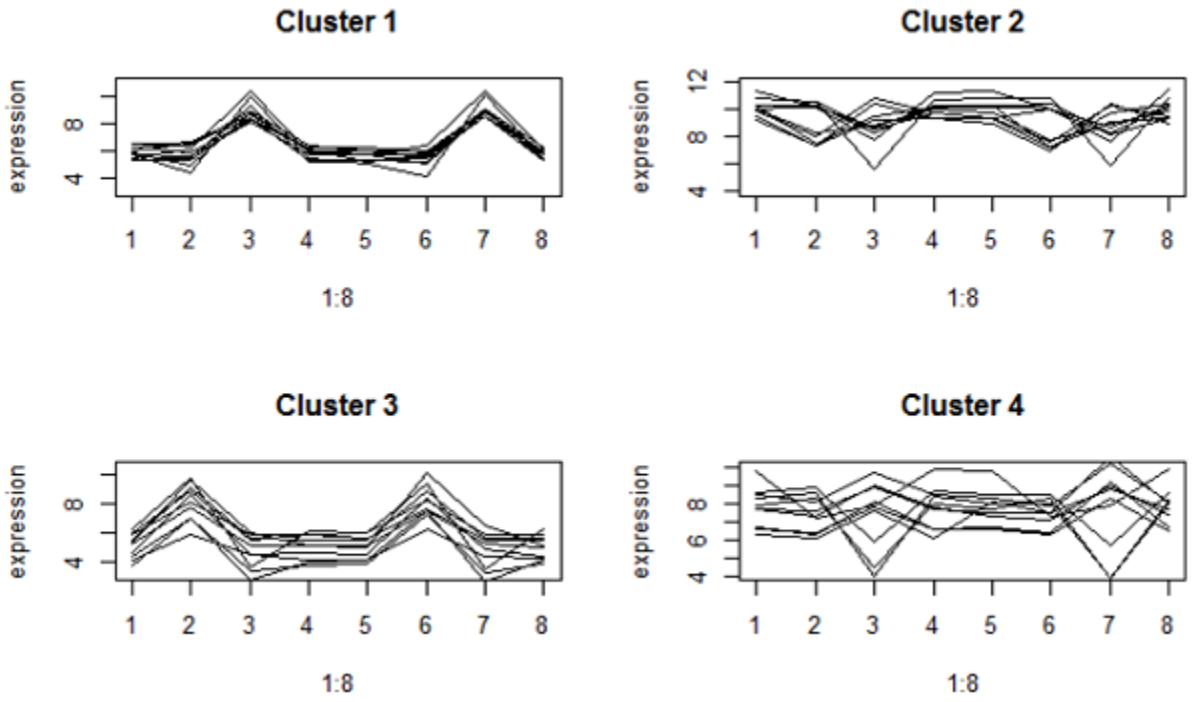
Hierarchical clustering does not tell us how many clusters there are, or where to cut the dendrogram to form clusters. In R there is a function **cutttree** which will cut a tree into clusters at a specified height. However, based on our visualization, we might prefer to cut the long branches at different heights. In any case, there is a fair amount of

subjectivity in determining which branches should and should not be cut to form separate clusters.

Understanding the clusters

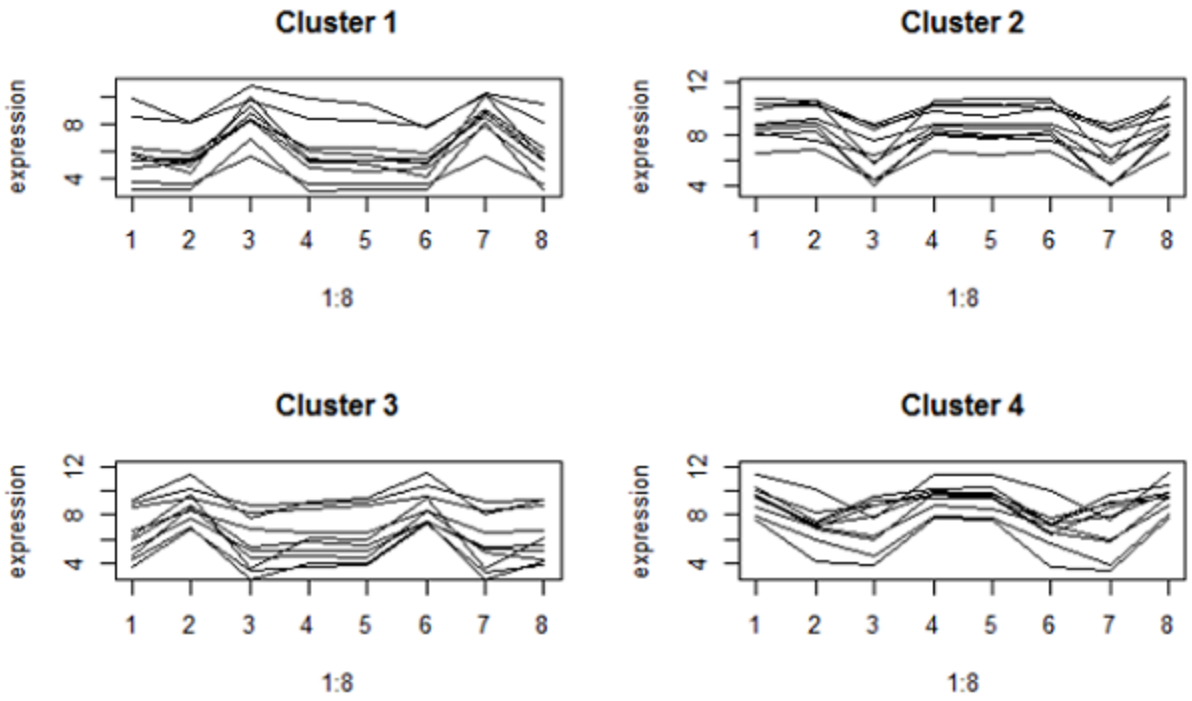
To understand the clusters we usually plot the log2(expression) values of the genes in the cluster, or in other words, plot the gene expressions over the samples. (The numbering in these graphs are totally arbitrary.) Even though the treatments are unordered, I usually connect the points coming from a single feature to make the pattern clearer. These are called profile plots.

Here is some of the profile plots from complete linkage clustering when we used Euclidean distance:



These look very tightly packed. However, clusters 2 and 4 have genes with different up and down patterns, because they have about the same mean expression. Cluster 2 are very highly expressed genes.

Here's what we got when we use correlation distance:



These are much looser on the y-axis because correlation focuses on the expression pattern, not the mean. However, all the genes in the same cluster have a peak or valley in the same treatments (which are brain regions by species combinations). Clusters 1 and 2 are genes that are respectively higher or lower in the cerebellum compared to other brain regions in both species.

71. Write some applications of hierarchical clustering?

Answer:

Some application of hierarchical clustering:

1) US Senator Clustering through Twitter:

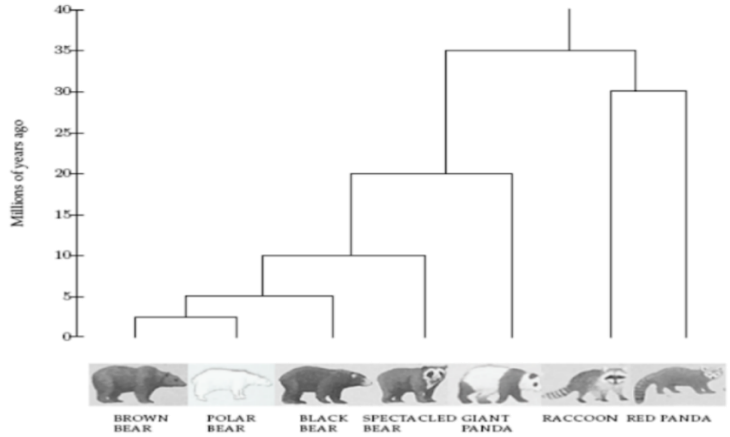
2) Charting Evolution through Phylogenetic Trees

How can we relate different species together?

In the decades before DNA sequencing was reliable, the scientists struggled to answer a seemingly simple question: Are giant pandas closer to bears or raccoons?

Nowadays, we can use DNA sequencing and hierarchical clustering to find the phylogenetic tree of animal evolution:

1. Generate the DNA sequences
2. Calculate the edit distance between all sequences.
3. Calculate the DNA similarities based on the edit distances.
4. Construct the phylogenetic tree.



As a result of this experiment, the researchers were able to place the giant pandas closer to bears.

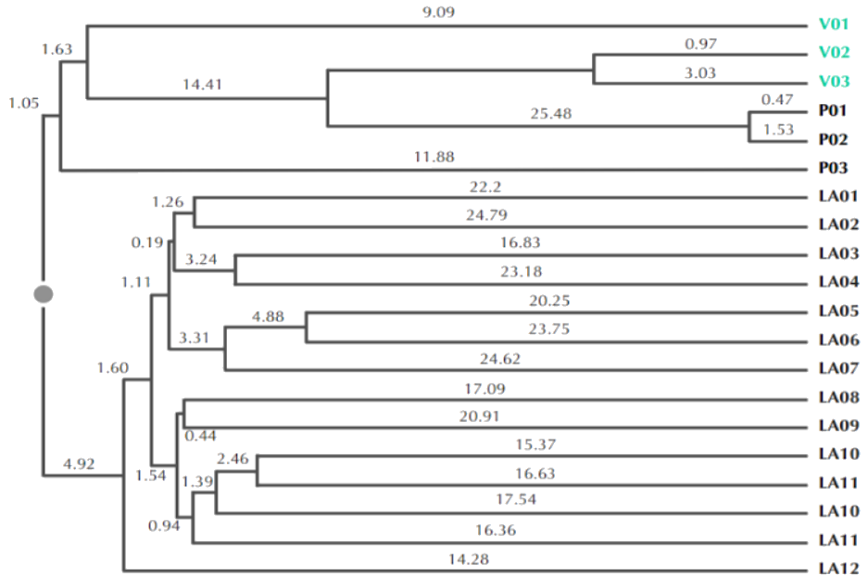
3) Tracking Viruses through Phylogenetic Trees

Can we find where a viral outbreak originated?

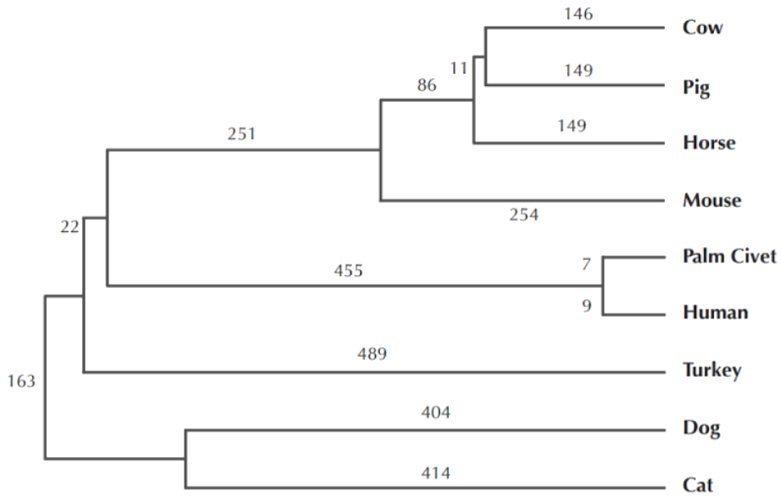
Tracking viral outbreaks and their sources is a major health challenge. Tracing these outbreaks to their source can give scientists additional data as to why and how the outbreak began, potentially saving lives.

Viruses such as HIV have high mutation rates, which means the similarity of the DNA sequence of the same virus depends on the time since it was transmitted. This can be used to trace paths of transmission.

This method was used as evidence in a court case, wherein the victim's strand of HIV was found to be more similar to the accused patient's strand, compared to a control group.



V1–3 are victim's strands, P1–3 are accused patient's, and LA1–12 are the control group



So humans got the SARS virus from palm civets… right?

"With the data at hand, we see how the virus used different hosts, moving from **bat to human to civet**, in that order. So the civets actually got SARS **from humans**." — Science Daily.

72. Write the algorithm of hierarchical clustering.

Answer:

Hierarchical clustering algorithm is of two types:

i) Agglomerative Hierarchical clustering algorithm or AGNES (agglomerative nesting) and

ii) Divisive Hierarchical clustering algorithm or DIANA (divisive analysis).

Both this algorithm are exactly reverse of each other. So we will be covering Agglomerative Hierarchical clustering algorithm in detail.

Agglomerative Hierarchical clustering -This algorithm works by grouping the data one by one on the basis of the nearest distance measure of all the pairwise distance between the data point. Again distance between the data point is recalculated but which distance to consider when the groups has been formed? For this there are many available methods. Some of them are:

1) single-nearest distance or single linkage.

2) complete-farthest distance or complete linkage.

3) average-average distance or average linkage.

4) centroid distance.

5) ward's method - sum of squared euclidean distance is minimized.

This way we go on grouping the data until one cluster is formed. Now on the basis of dendogram graph we can calculate how many number of clusters should be actually present.

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.

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

Answer:

A bank wants to search new ways to increase revenues from its credit card operations. They want to check whether usage would double if fees were halved.

Bank has multiple years of record on average credit card balances, payment amounts, credit limit usage, and other key parameters. They create a model to check the impact of the proposed new business policy. The data results

show that cutting fees in half for a targetted customer base could increase revenues by $10 million.

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

Answer:

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.

Data Mining Techniques

Data mining is highly effective, so long as it draws upon one or more of these techniques:

1. Tracking patterns. One of the most basic techniques in data mining is learning to recognize patterns in your data sets. This is usually a recognition of some aberration in your data happening at regular intervals, or an ebb and flow of a certain variable over time. For example, you might see that your sales of a certain product seem to spike just before the holidays, or notice that warmer weather drives more people to your website.

2. Classification. Classification is a more complex data mining technique that forces you to collect various attributes together into discernable categories, which you can then use to draw further conclusions, or serve some function. For example, if you're evaluating data on individual customers' financial backgrounds and purchase histories, you might be able to classify them as "low," "medium," or "high" credit risks. You could then use these classifications to learn even more about those customers.

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

Ans:

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.

Example of predictive data mining:

- Sky Survey Cataloging
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).

    3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.
    - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

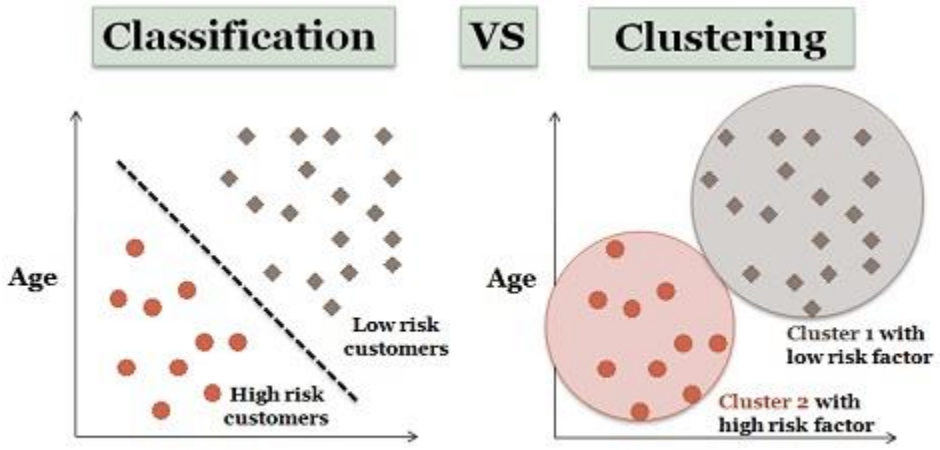
77. Non-trivial extraction of implicit, previously unknown and potentially useful information from data is called Data Mining. There are several tasks that we employ for mining; both classification and clustering. Answer the following questions: Give some examples of data where Data Mining techniques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



Risk classification for the loan payees on the basis of customer salary

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

clustering is used in unsupervised learning where similar instances are grouped, based on their features or properties.

Classification	Clustering
A Supervised Learning technique	An Unsupervised Learning technique
Finite set of classes	Finite set of clusters
Goal of assigning new input to a class	Goal of finding similarities within a given dataset
Infinite set of input data	Finite set of data

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 an correct outcome from labeled data.

For instance, suppose you are given an 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.

Unsupervised learning classified into two categories of algorithms:

- **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
- **Association**: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

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

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.

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**[1]) 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 sums 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 $\lambda > 0$, 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 Cossine 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.

f 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).