1. What is the difference between supervised and unsupervised learning? Give some examples to illustrate your point.

Answer:

The main distinction between the two approaches is the use of labeled datasets. To put it simply, supervised learning uses labeled input and output data, while an unsupervised learning algorithm does not.

In supervised learning, the algorithm “learns” from the training dataset by iteratively making predictions on the data and adjusting for the correct answer. While supervised learning models tend to be more accurate than unsupervised learning models, they require upfront human intervention to label the data appropriately.

For example, a supervised learning model can predict how long your commute will be based on the time of day, weather conditions and so on. But first, we have to train it to know that rainy weather extends the driving time.

Unsupervised learning models, in contrast, work on their own to discover the inherent structure of unlabeled data. Note that they still require some human intervention for validating output variables.

For example, an unsupervised learning model can identify that online shopper often purchase groups of products at the same time. However, a data analyst would need to validate that it makes sense for a recommendation engine to group baby clothes with an order of diapers, applesauce and sippy cups.

Other key differences between supervised and unsupervised learning:

Goals: In supervised learning, the goal is to predict outcomes for new data. You know up front the type of results to expect. With an unsupervised learning algorithm, the goal is to get insights from large volumes of new data. The machine learning itself determines what is different or interesting from the dataset.

Applications: Supervised learning models are ideal for spam detection, sentiment analysis, weather forecasting and pricing predictions, among other things. In contrast, unsupervised learning is a great fit for anomaly detection, recommendation engines, customer personas and medical imaging.

Complexity: Supervised learning is a simple method for machine learning, typically calculated through the use of programs like R or Python. In unsupervised learning, you need powerful tools for working with large amounts of unclassified data. Unsupervised learning models are computationally complex because they need a large training set to produce intended outcomes.

Drawbacks: Supervised learning models can be time-consuming to train, and the labels for input and output variables require expertise. Meanwhile, unsupervised learning methods can have wildly inaccurate results unless you have human intervention to validate the output variables.

2. Mention a few unsupervised learning applications.

Answer: The main applications of unsupervised learning include clustering, visualization, dimensionality reduction, finding association rules, and anomaly detection.

* Clustering: Clustering is the process of grouping the given data into different clusters or groups. Unsupervised learning can be used to do clustering when we don’t know exactly the information about the clusters. Elements in a group or cluster should be as similar as possible, and points in different groups should be as dissimilar as possible. Unsupervised learning can be used to do clustering when we don’t know exactly the information about the clusters.
* Visualization: Visualization is the process of creating diagrams, images, graphs, charts, etc., to communicate some information. This method can be applied using unsupervised machine learning.
* Dimensionality reduction: Dimensionality reduction is the process of reducing the number of random variables under consideration by getting a set of principal variables.
* Finding Association Rules: This is the process of finding associations between different parameters in the available data. It discovers the probability of the co-occurrence of items in a collection, such as people that buy X also tend to buy Y.

In association rule learning, the algorithm will deep dive into large amounts of data and find some interesting relationships between attributes.

* Anomaly Detection: Anomaly detection is the identification of rare items, events, or observations, which brings suspicions by differing significantly from the normal data.

3. What are the three main types of clustering methods? Briefly describe the characteristics of each.

Answer: These are some of the commonly used clustering algorithms:

* K-Means
* Expectation Maximization
* Hierarchical Cluster Analysis (HCA)

K-means is probably one of the mostly known and frequently used. K-means uses an iterative refinement method to produce its final clustering based on the number of clusters defined by the user (represented by the variable K) and the dataset. For example, if you set K equal to 3 then your dataset will be grouped in 3 clusters, if you set K equal to 4 you will group the data in 4 clusters, and so on.

Expectation-maximization algorithm is an approach for performing maximum likelihood estimation in the presence of latent variables. It does this by first estimating the values for the latent variables, then optimizing the model, then repeating these two steps until convergence

Hierarchical cluster analysis (HCA), also known as hierarchical clustering, is a popular method for cluster analysis in big data research and data mining aiming to establish a hierarchy of clusters (1-3). As such, HCA attempts to group subjects with similar features into clusters.

4. Explain how the k-means algorithm determines the consistency of clustering.

Answer: K-means is probably one of the mostly known and frequently used. K-means uses an iterative refinement method to produce its final clustering based on the number of clusters defined by the user (represented by the variable K) and the dataset. For example, if you set K equal to 3 then your dataset will be grouped in 3 clusters, if you set K equal to 4 you will group the data in 4 clusters, and so on.

K-means starts off with arbitrarily chosen data points as proposed means of the data groups, and iteratively recalculates new means in order to converge to a final clustering of the data points.

When you define the value of K you are actually telling the algorithm how many means or centroids you want (if you set K=3 you create 3 means or centroids, which accounts for 3 clusters). A centroid is a data point that represents the center of the cluster (the mean), and it might not necessarily be a member of the dataset.

5. With a simple illustration, explain the key difference between the k-means and k-medoids algorithms.

Answer:

K-means is a classical partitioning technique of clustering that clusters the data set of n objects into k clusters with k known a priori. A useful tool for determining k is the silhouette.

The k-medoids algorithm is a clustering algorithm related to the k-means algorithm and the medoidshift algorithm. Both the k-means and k-medoids algorithms are partitional (breaking the dataset up into groups).

K-means attempts to minimize the total squared error, while k-medoids minimizes the sum of dissimilarities between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the k-means algorithm, k-medoids chooses datapoints as centers (medoids or exemplars).

K-medoids is also a partitioning technique of clustering that clusters the data set of n objects into k clusters with k known a priori. A useful tool for determining k is the silhouette.

6. What is a dendrogram, and how does it work? Explain how to do it.

Answer:

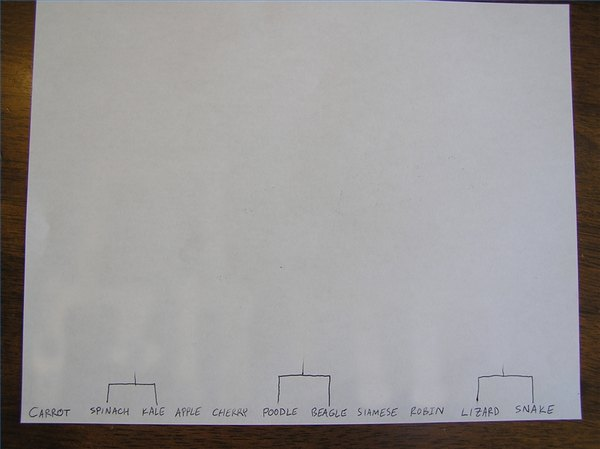
A dendrogram is a diagram that shows the attribute distances between each pair of sequentially merged classes. To avoid crossing lines, the diagram is graphically arranged so that members of each pair of classes to be merged are neighbours in the diagram.

The Dendrogram tool uses a hierarchical clustering algorithm. The program first computes the distances between each pair of classes in the input signature file. Then it iteratively merges the closest pair of classes and successively merges the next closest pair of classes and the succeeding closest until the classes are all merged. After each merging, the distances between all pairs of classes are updated. The distances at which the signatures of classes are merged are used to construct a dendrogram.

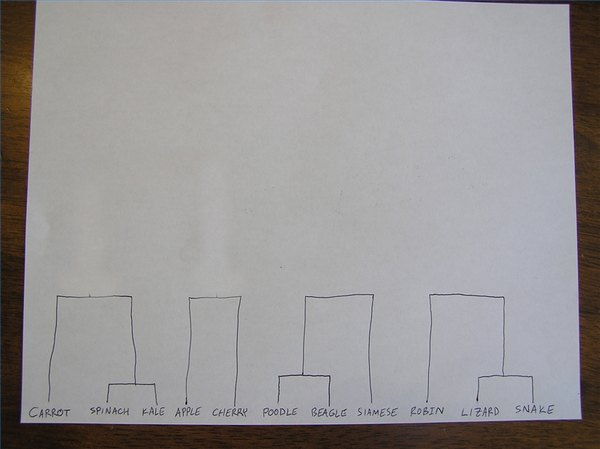
As an example, consider this list: apple, beagle, spinach, lizard, Siamese, cherry, kale, poodle, carrot, snake, and robin.

Determine how to cluster the units. The clustering method used will depend on the units being grouped together. The objective is to put like units into small groups and then similar small groups into larger groups, until the whole list is in a single cluster. Name each group. In our example, the entire list could be called "living things," while the beagle and poodle could be in a group called "dogs."

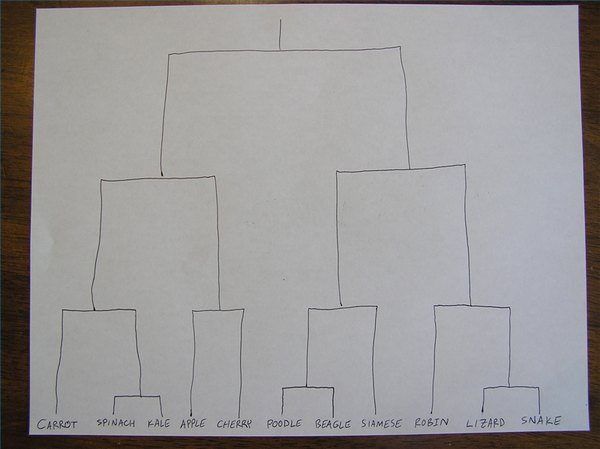
* Write the list of units across the bottom of a piece of paper. Order them so that the smallest groups are near each other.



* Draw lines to connect those units that are placed into groups of only two. Not every unit will fall into such a group.



* Draw lines to connect groups of three or four. This may involve connecting groups of two from Step 4.



* Continue connecting larger and larger groups until all units are connected. This completed chart is a dendrogram.

7. What exactly is SSE? What role does it play in the k-means algorithm?

Answer: There is no easy answer for choosing k value. One of the method is known as elbow method. First of all compute the sum of squared error (SSE) for some value of K.SSE is defined as the sum of the squared distance between centroid and each member of the cluster. Then plot a K against SSE graph. We will observe that as K increases SSE decreases as disortation will be small. So the idea of this algorithm is to choose the value of K at which the graph decrease abruptly. This sort of produces a “elbow effect” in the picture.

8. With a step-by-step algorithm, explain the k-means procedure.

Answer:

This is how the algorithm works:

* K centroids are created randomly (based on the predefined value of K)
* K-means allocates every data point in the dataset to the nearest centroid (minimizing Euclidean distances between them), meaning that a data point is considered to be in a particular cluster if it is closer to that cluster’s centroid than any other centroid
* Then K-means recalculates the centroids by taking the mean of all data points assigned to that centroid’s cluster, hence reducing the total intra-cluster variance in relation to the previous step. The “means” in the K-means refers to averaging the data and finding the new centroid
* The algorithm iterates between steps 2 and 3 until some criteria is met (e.g. the sum of distances between the data points and their corresponding centroid is minimized, a maximum number of iterations is reached, no changes in centroids value or no data points change clusters)

9. In the sense of hierarchical clustering, define the terms single link and complete link.

Answer: Hierarchical clustering treats each data point as a singleton cluster, and then successively merges clusters until all points have been merged into a single remaining cluster. A hierarchical clustering is often represented as a dendrogram.

In complete-link (or complete linkage) hierarchical clustering, we merge in each step the two clusters whose merger has the smallest diameter (or: the two clusters with the smallest maximum pairwise distance).

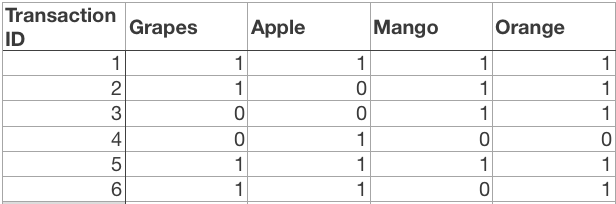
In single-link (or single linkage) hierarchical clustering, we merge in each step the two clusters whose two closest members have the smallest distance (or: the two clusters with the smallest minimum pairwise distance).

Complete-link clustering can also be described using the concept of clique. Let dn be the diameter of the cluster created in step n of complete-link clustering. Define graph G(n) as the graph that links all data points with a distance of at most dn. Then the clusters after step n are the cliques of G(n). This motivates the term complete-link clustering.

Single-link clustering can also be described in graph theoretical terms. If dn is the distance of the two clusters merged in step n, and G(n) is the graph that links all data points with a distance of at most dn, then the clusters after step n are the connected components of G(n). A single-link clustering also closely corresponds to a weighted graph's minimum spanning tree.

10. How does the apriori concept aid in the reduction of measurement overhead in a business basket analysis? Give an example to demonstrate your point.

Answer: Apriori algorithm assumes that any subset of a frequent itemset must be frequent. Its the algorithm behind Market Basket Analysis. Say, a transaction containing {Grapes, Apple, Mango} also contains {Grapes, Mango}. So, according to the principle of Apriori, if {Grapes, Apple, Mango} is frequent, then {Grapes, Mango} must also be frequent. Here is a dataset consisting of six transactions. Each transaction is a combination of 0s and 1s, where 0 represents the absence of an item and 1 represents the presence of it.



In order to find out interesting rules out of multiple possible rules from this small business scenario, we will be using the following matrices:

1. Support: Its the default popularity of an item. In mathematical terms, the support of item A is nothing but the ratio of transactions involving A to the total number of transactions.

Support(Grapes) = (Transactions involving Grapes)/(Total transaction)

Support(Grapes) = 0.666

1. Confidence: Likelihood that customer who bought both A and B. Its divides the number of transactions involving both A and B by the number of transactions involving B.

Confidence(A => B) = (Transactions involving both A and B)/(Transactions involving only A).

Confidence({Grapes, Apple} => {Mango}) = Support(Grapes, Apple, Mango)/Support(Grapes, Apple) = 2/6 / 3/6 = 0.667

1. Lift : Increase in the sale of A when you sell B.

Lift(A => B) = Confidence(A, B) / Support(B)

Lift ({Grapes, Apple} => {Mango}) = 1

So, likelihood of a customer buying both A and B together is ‘lift-value’ times more than the chance if purchasing alone.

* Lift (A => B) = 1 means that there is no correlation within the itemset.
* Lift (A => B) > 1 means that there is a positive correlation within the itemset, i.e., products in the itemset, A, and B, are more likely to be bought together.
* Lift (A => B) < 1 means that there is a negative correlation within the itemset, i.e., products in itemset, A, and B, are unlikely to be bought together.