1. What is your definition of clustering? What are a few clustering algorithms you might think of?

Answer: 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.

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.

2. What are some of the most popular clustering algorithm applications?

Answer: Clustering technique is used in various applications such as market research and customer segmentation, biological data and medical imaging, search result clustering, recommendation engine, pattern recognition, social network analysis, image processing, etc.

Some practical applications which use k-means clustering are sensor measurements, activity monitoring in a manufacturing process, audio detection and image segmentation.

3. When using K-Means, describe two strategies for selecting the appropriate number of clusters.

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. K-means attempts to minimize the total squared error. A useful tool for determining k is the silhouette. 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.

Two methods that can be useful to find this mysterious k in k-Means.

These methods are:

* The Elbow Method
* The Silhouette Method

The Elbow Method:

This is probably the most well-known method for determining the optimal number of clusters. It is also a bit naive in its approach.

Calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an elbow.

Within-Cluster-Sum of Squared Errors sounds a bit complex. Let’s break it down:

* The Squared Error for each point is the square of the distance of the point from its representation i.e. its predicted cluster center.
* The WSS score is the sum of these Squared Errors for all the points.
* Any distance metric like the Euclidean Distance or the Manhattan Distance can be used.

The Silhouette Method:

The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation).

The range of the Silhouette value is between +1 and -1. A high value is desirable and indicates that the point is placed in the correct cluster. If many points have a negative Silhouette value, it may indicate that we have created too many or too few clusters.

The Silhouette score can be easily calculated in Python using the metrics module of the sklearn library

4. What is mark propagation and how does it work? Why would you do it, and how would you do it?

Answer: Backpropagation is a supervised learning algorithm, for training Multi-layer Perceptrons (Artificial Neural Networks).

While designing a Neural Network, in the beginning, we initialize weights with some random values or any variable for that fact.

Now obviously, we are not superhuman. So, it’s not necessary that whatever weight values we have selected will be correct, or it fits our model the best.

Okay, fine, we have selected some weight values in the beginning, but our model output is way different than our actual output i.e. the error value is huge.

Now, how will you reduce the error?

Basically, what we need to do, we need to somehow explain the model to change the parameters (weights), such that error becomes minimum.

Let’s put it in another way, we need to train our model. One way to train our model is called as Backpropagation.

Steps to follow:

* Calculate the error — How far is your model output from the actual output.
* Error minimum — Check whether the error is minimized or not.
* Update the parameters — If the error is huge then, update the parameters (weights and biases). After that again check the error. Repeat the process until the error becomes minimum.
* Model is ready to make a prediction — Once the error becomes minimum, you can feed some inputs to your model and it will produce the output.

The Backpropagation algorithm looks for the minimum value of the error function in weight space using a technique called the delta rule or gradient descent. The weights that minimize the error function is then considered to be a solution to the learning problem.

5. Provide two examples of clustering algorithms that can handle large datasets. And two that look for high-density areas?

Answer:

One approach to clustering large data sets is to adapt clustering algorithms suitable for small data sets to much larger data sets. There are two popular methods used to adapt clustering algorithms to large data sets. The first technique is to extract a sub-sample of the data, such that the sub-sample is small enough to fit into available memory and be clustered. Other techniques to accelerate the clustering process are often applied at the same time. Once a clustering is obtained, the remaining data points can be assigned to the clusters with the closest centroid. The major drawbacks to sampling are that

it can be difficult to know if a given subsample is a representative sample and therefore provides an accurate clustering, and that the outliers will usually be ignored.

The second technique commonly used to adapt clustering algorithms to large data sets, as in [5, 6, 21], is to approximate a given data item by assigning it to a single representative vector. One representative vector may take the place of an arbitrary number of data items. Once a data item has been assigned to a representative, it is no longer possible to differentiate it from any other data item assigned to the same vector. Thus, the resolution of any clustering of the data is limited by the granularity of the representatives.

6. Can you think of a scenario in which constructive learning will be advantageous? How can you go about putting it into action?

Answer:

Constructivist learning is a theory about how people learn. It states that learning happens when learners construct meaning by interpreting information in the context of their own experiences. In other words, learners construct their own understandings of the world by reflecting on their experiences.

Constructivism is the theory that says learners construct knowledge rather than just passively take in information. As people experience the world and reflect upon those experiences, they build their own representations and incorporate new information into their pre-existing knowledge (schemas).

7. How do you tell the difference between anomaly and novelty detection?

Answer:

Outlier detection: The training data contains outliers which are defined as observations that are far from the others. Outlier detection estimators thus try to fit the regions where the training data is the most concentrated, ignoring the deviant observations.

Novelty detection: The training data is not polluted by outliers and we are interested in detecting whether a new observation is an outlier. In this context an outlier is also called a novelty.

Outlier detection and novelty detection are both used for anomaly detection, where one is interested in detecting abnormal or unusual observations. Outlier detection is then also known as unsupervised anomaly detection and novelty detection as semi-supervised anomaly detection. In the context of outlier detection, the outliers/anomalies cannot form a dense cluster as available estimators assume that the outliers/anomalies are located in low density regions. On the contrary, in the context of novelty detection, novelties/anomalies can form a dense cluster as long as they are in a low density region of the training data, considered as normal in this context. The scikit-learn project provides a set of machine learning tools that can be used both for novelty or outlier detection.

8. What is a Gaussian mixture, and how does it work? What are some of the things you can do about it?

Answer:

Gaussian Mixture Models (GMMs) assume that there are a certain number of Gaussian distributions, and each of these distributions represent a cluster. Hence, a Gaussian Mixture Model tends to group the data points belonging to a single distribution together.

For examples, we have three Gaussian distributions– GD1, GD2, and GD3. These have a certain mean (μ1, μ2, μ3) and variance (σ1, σ2, σ3) value respectively. For a given set of data points, our GMM would identify the probability of each data point belonging to each of these distributions.

Gaussian Mixture Models are probabilistic models and use the soft clustering approach for distributing the points in different clusters. Gaussian Distributions (or the Normal Distribution) has a bell-shaped curve, with the data points symmetrically distributed around the mean value.

9. When using a Gaussian mixture model, can you name two techniques for determining the correct number of clusters?

Answer:

In real life, many datasets can be modelled by Gaussian Distribution (Univariate or Multivariate). So, it is quite natural and intuitive to assume that the clusters come from different Gaussian Distributions. Or in other words, it is tried to model the dataset as a mixture of several Gaussian Distributions. This is the core idea of this model.

In one dimension the probability density function of a Gaussian Distribution is given by



where mu and sigma^2 are respectively mean and variance of the distribution.

For Multivariate (let us say d-variate) Gaussian Distribution, the probability density function is given by



Here mu is a d dimensional vector denoting the mean of the distribution and Sigma is the d X d covariance matrix.

Two basic steps of the Expectation-Maximization (EM) algorithm

E Step or Expectation Step or Estimation Step and M Step or Maximization Step.

Estimation step:

* initialize \mu\_k, \Sigma\_k and \pi\_k by some random values, or by K means clustering results or by hierarchical clustering results.
* Then for those given parameter values, estimate the value of the latent variables (i.e \gamma\_k)

Maximization Step:

* Update the value of the parameters( i.e. \mu\_k, \Sigma\_k and\pi\_k) calculated using ML method.