Assignment 24 Solutions

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

Ans: Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.

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

Ans: K-means clustering is the most commonly used clustering algorithm. It's a centroid-based algorithm and the simplest unsupervised learning algorithm. This algorithm tries to minimize the variance of data points within a cluster. It's also how most people are introduced to unsupervised machine learning.

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

Ans: K-means clustering algorithm can be significantly improved by using a better initialization technique, and by repeating (re-starting) the algorithm. When the data has overlapping clusters, k-means can improve the results of the initialization technique.

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

Ans: Backpropagation (backward propagation) is an important mathematical tool for improving the accuracy of predictions in data mining and machine learning. Essentially, backpropagation is an algorithm used to calculate derivatives quickly.

The backpropagation algorithm works by computing the gradient of the loss function with respect to each weight by the chain rule, computing the gradient one layer at a time, iterating backward from the last layer to avoid redundant calculations of intermediate terms in the chain rule.

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

Ans: The most commonly used algorithm in clustering are partitioning, hierarchical, grid based, density based, and model based algorithms. A review of clustering and its different techniques in data mining is done considering the criteria's for big data.

6. Can you think of a scenario in which constructive learning will be

Ans: In Constructivist learning, the traditional classroom learning procedure is flipped. Instead of a teacher informing a child about a subject and constructing a meaning for them, in constructivist learning, children construct their own meanings. Now, this may not be the best approach for every learning environment, but it can be tremendously helpful at home.

The home environment is where children primarily develop a sense of self and is the perfect setting for children to practice taking initiative in their learning via a Constructivist approach.

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

Ans: In "novelty detection", you have a data set that contains only good data, and you're trying to determine whether new observations fit within the existing data set. In "outlier detection", the data may contain outliers, which you want to identify.

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

Ans: A Gaussian Mixture is a function that is comprised of several Gaussians, each identified by $k \in \{1,\ldots, K\}$, where K is the number of clusters of our dataset. Each Gaussian k in the mixture is comprised of the following parameters: A mean μ that defines its centre. A covariance Σ that defines its width.

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

Ans: An approach is to find the clusters using soft clustering methods and then see if they are gaussian. If they are then you can apply a GMM model which represents the whole dataset.