

Student Number:

Student Name:

Date

CS5483 Lecture Review Question 9

What is supervised clustering and what is unsupervised clustering? How do you compare their difference with respect to performance? Illustrate the strength and weakness of k-means in comparison with the k-medoids algorithm.

Model Answer:

Clustering is similar to classification in that data are grouped. Supervised clustering means that users input are expected to specify k (clusters) and initial centers to complete the clustering process. Unsupervised clustering means that let computer process derives k (clusters) to complete the clustering process.

However, clustering algorithms like k-means and k-medoids need users to specify k as number of partition before clustering process. They also need to set up initial seed points for deriving mean value and medoid object. Suppose we decide to utilize the K-Means algorithm for the evaluation. The clusters formed by an application of the K-Means algorithm are highly affected by the initial selection of cluster means. In this case, supervised clustering can help to speed up clustering process by setting up good initial mean value and medoid object. Furthermore, a supervised learner capable of generating production rules is particularly appealing for purposes of explanation.

K-means:

- Strength

- Relatively efficient: $O(knt)$, where n is # objects, k is # clusters, and t is # iterations. Normally, $k, t \ll n$.
- Often terminates at a local optimum. The global optimum may be found using techniques such as: deterministic annealing and genetic algorithms

- Weakness

- Applicable only when mean is defined, then what about categorical data?
- Need to specify k , the number of clusters, in advance
- Unable to handle noisy data and outliers
- Not suitable to discover clusters with non-convex shapes

K-medoids:

The k-medoids method is more robust than k-means in the presence of noise and outliers, because a medoid is less influenced by outliers or other extreme values than a mean. However, its processing is more costly than the k-means method.