# CSE 3212 Data Mining

## Clustering

1. Consider the following data with two attributes:

| Instance | X | Y |
|---|---|---|
| D1 | 1.0 | 1.5 |
| D2 | 1.0 | 4.5 |
| D3 | 2.0 | 1.5 |
| D4 | 2.0 | 3.5 |
| D5 | 3.0 | 2.5 |
| D6 | 5.0 | 6.0 |

   - Perform K- Means clustering with K=2 and using Manhattan and Euclidean distance measures.

   - Are the two distance measures produced the same set of clusters? Why/Why not?

2. One way to evaluate the goodness of clusters for a given data set is compute the squared error. If the squared error is small, then one can assume that the clustering is good.

   Squared Error = Summation of the squared differences between the cluster centers and the corresponding cluster instances

   Consider the following three different 2 clusters for the above data set:

   Clustering 1: {D2, D4, D6}, { D1, D3, D5}
   Clustering 2: {D1, D3}, { D2, D4, D5, D6}
   Clustering 3: {D1, D2, D3, D4, D5}, { D6 }

   - For each clustering, compute the centroid of clusters and the squared error. Comment the clustering based on the squared error values.

   - Which of the above cluster did you get by using the K-means algorithm?

   - Is the clusters discovered using the K-means algorithm the best cluster?

   - What factors of the K-means algorithm would influence the result?