Data Mining

Problem Set 6

Chapter 9

1. For sparse data, discuss why considering only the presence of non-zero values might give a more accurate view of the objects than considering the actual magnitudes of values. When would such an approach not be desirable?

2. Describe the change in the time complexity of K-means as the number of clusters to be found increases.

8. Explain the difference between likelihood and probability.

12. Figure 9.1 shows a clustering of a two-dimensional point data set with two clusters: The leftmost cluster, whose points are marked by asterisks, is somewhat diffuse, while the rightmost cluster, whose points are marked by circles, is compact. To the right of the compact cluster, there is a single point (marked by an arrow) that belongs to the diffuse cluster, whose center is farther away than that of the compact cluster. Explain why this is possible with EM clustering, but not K-means clustering.

14. One way to sparsify a proximity matrix is the following: For each object (row in the matrix), set all entries to 0 except for those corresponding to the objects k-nearest neighbors. However, the sparsified proximity matrix is typically not symmetric.
(a) If object a is among the k-nearest neighbors of object b, why is b not guaranteed to be among the k-nearest neighbors of a?
(b) Suggest at least two approaches that could be used to make the sparsified proximity matrix symmetric.

18. Name at least one situation in which you would not want to use clustering based on SNN similarity or density.

Chapter 10

6. Describe the potential time complexity of anomaly detection approaches based on the following approaches: model-based using clustering, proximity based, and density. No knowledge of specific techniques is required. Rather, focus on the basic computational requirements of each approach, such as the time required to compute the density of each object.

15. Consider a set of points, where most points are in regions of low density, but a few points are in regions of high density. If we define an anomaly as a point in a region of low density, then most points will be classified as anomalies. Is this an appropriate use of the density-based definition of an anomaly or should the definition be modified in some way?