Jarod Klion

CAP5771

Homework 5

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 value. When would such an approach not be desirable?**
    a. In certain cases, such as document data, only considering the presence of non-zero values instead of actual magnitudes could give a more accurate view of the objects because we could want to match the context in which words are said instead of the absolute frequency they appear. Such an approach is not desirable when performing clustering analysis where we want the true number of clusters.
2. **Describe the change in the time complexity of K-means as the number of clusters to be found increases.**
    a. As the number of clusters increases in k-means, the time complexity increases linearly because it has time complexity $O(I * K * m * n)$, where *I* is iterations required for convergence, *m* is the number of points, and *n* is number of attributes.
8. **Explain the difference between likelihood and probability.**
    a. Probability is the chance that a given event will occur after repeated experiments while likelihood is the plausibility that a particular distribution explains the given data.
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.**
    a. EM clustering is a subset of K-means clustering that works differently in that it computes the probability that each point belongs to each cluster, where the probability depends on the distance from the cluster center and cluster's variance. Therefore, the point marked by the arrow can belong to the cluster whose center is farther away than the compact cluster because the farther cluster has a higher variance than the closer cluster. This won't be possible with K-means because it only considers the distance to the closest cluster when assigning points.
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*?**
        i. If we have a set of *k* + 1 objects and an outlier that is farther from any of the objects than they are from each other, then none of the objects in the

set will have this outlier in their *k*-nearest neighbors, but the outlier will have *k* of the objects in the set in its *k*-nearest neighbor list.

b. **Suggest at least two approaches that could be used to make the sparsified proximity matrix symmetric.**
    i. To make the matrix symmetric, we can set the *ij*th entry to 0 or 1 if the *ji*th entry is 0 or 1, respectively, or vice versa.

18. **Name at least one situation in which you would not want to use clustering based on SNN similarity or density.**
    a. You would not want to use clustering based on SNN similarity or density if you wish to cluster based on absolute density or distance.

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 computation requirements of each approach, such as the time required to compute the density of each object.**
    a. K-means clustering requires time proportional to the number of points, which has time complexity of $O(m)$. On the other hand, proximity/distance-based and density-based approaches usually require computing all pairwise proximities, so they typically have time complexity $O(m^2)$. In low-dimensional spaces for distance or density based approaches, one can possibly use kd-trees, which have time complexity $O(m \log m)$.

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?**
    a. It could be appropriate to consider most of the points as anomalies if the density has an absolute meaning, but typically one would change the anomaly detection technique that would take the relative density of the regions into account.