Homework 4
Karsen Diepholz
CS 422-Data Mining

Chapter 7

4. B.

K=10→ $\frac{2x10!}{10^{10}} = .000728$

K=100→ $\frac{2x100!}{100^{100}} = 1.87x10^{-42}$

K=1000→ $\frac{2x1000!}{1000^{1000}} = 0$

7. The answer is B: More centroids should be allocated to the less dense region. This is because as more dense regions have more points in a small area, a centroid in that area would have less distance from that point and thus the calculation can be more accurate with less points.

11. If it is low for all clusters, the variable is technically a constant and does not help splitting data into groups. If it is low for just one cluster, then it helps define that cluster accordingly. If it is high for all clusters, it can mean the variable is noise. If it is high for a single cluster, then it is odd with the information given by the variables with low SSE (which defines the cluster). The per variable SSE information can help us improve clustering by eliminating variables which have bad distinguishing powers between certain clusters.

21.

Entropy:

$E_1 = -((\frac{1}{693} * log_2(\frac{1}{693}) + (\frac{1}{693} * log_2 \frac{1}{693}) + (\frac{11}{693} * log_2 \frac{11}{693}) + (\frac{4}{693} * log_2 \frac{4}{693}) = .165$

$E_2 =$

$-((\frac{27}{1562} * log_2 \frac{27}{1562}) + (\frac{89}{1562} * log_2 \frac{89}{1562}) + (\frac{333}{1562} * log_2 \frac{333}{1562}) + (\frac{827}{1562} * log_2 \frac{827}{1562}) + (\frac{253}{1562} * log_2 \frac{253}{1562}) + (\frac{33}{1562} * log_2(\frac{33}{1562}) = 1.84$

E_3= Same method as above… 1.6964

$E = \frac{693}{3204} * .165 + \frac{1562}{3204} * 1.84 + \frac{949}{3204} * 1.6964 =$ **1.43505**

Purity:

$P_1 = \frac{676}{693} = .975$

$P_2 = \frac{827}{1562} = .529$

$P_3 = \frac{465}{949} = .489$

$P = \frac{693}{3204} * .975 + \frac{1562}{3204} * .529 + \frac{949}{3204} * .489 =$ **.614**

22. A. Yes, there is a difference between these two sets of points.

B. Smaller SSE would go to the uniformly distributed cluster, as it is uniformly spaced, so the SSE for K=10 clusters would typically be found in that set of points.

C. The behavior in the random dataset would be difficult, as DBSCAN would rarely search clusters in there because it would not have any differentiation in the density system. In the uniform points, DBSCAN would combine points in the dataset which is uniform into a single clusters or make them complete as noise supporting on the threshold.