CS 422 – Data Mining

Homework 3

5 April 2017

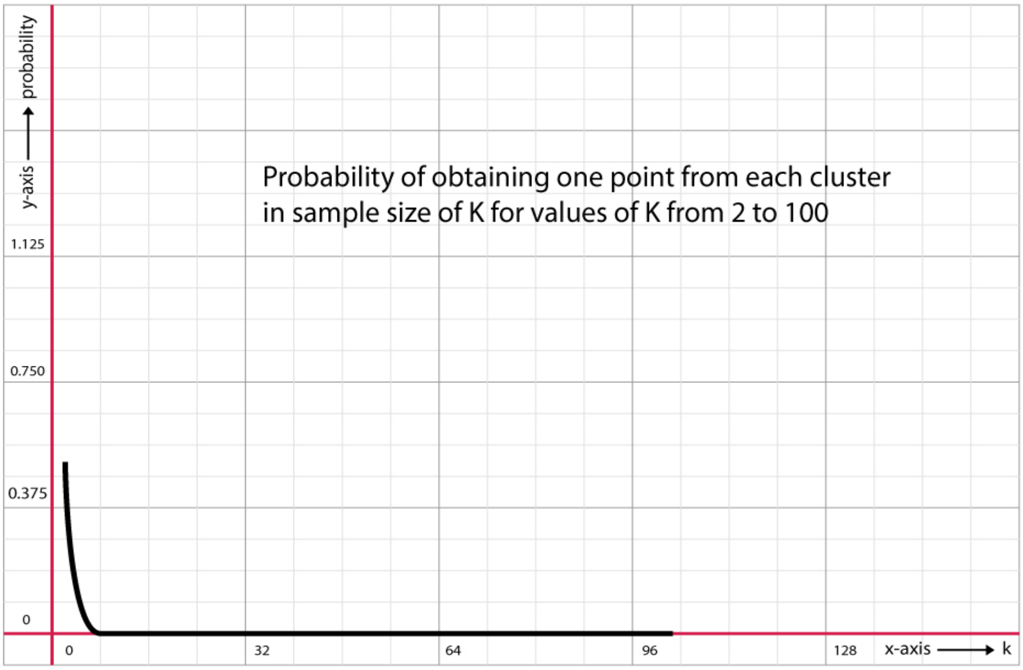
Mayank Bansal

A20392482

Recitation Problems

Problem 4

a)



b) Probability that a point doesn't come from a particular cluster = $\left(1-\frac{1}{k}\right)$

Probability that all 2k points don't come from a particular cluster = $\left(1-\frac{1}{k}\right)^{2k}$

Probability that at least 1 point comes from a particular cluster = $1 - \left(1-\frac{1}{k}\right)^{2k}$

Probability that all clusters are represented in the final sample = $\left(1 - \left(1-\frac{1}{k}\right)^{2k}\right)^{k}$

For **k = 10**,

Probability = $\left(1 - \left(1-\frac{1}{10}\right)^{20}\right)^{10}$ = 0.273

For **k = 100**,

Probability = $\left(1 - \left(1-\frac{1}{100}\right)^{200}\right)^{100}$ = $5.65 * 10^{-7}$

For **k = 1000**,

Probability = $\left(1 - \left(1-\frac{1}{1000}\right)^{2000}\right)^{1000}$ = $8.23 * 10^{-64}$

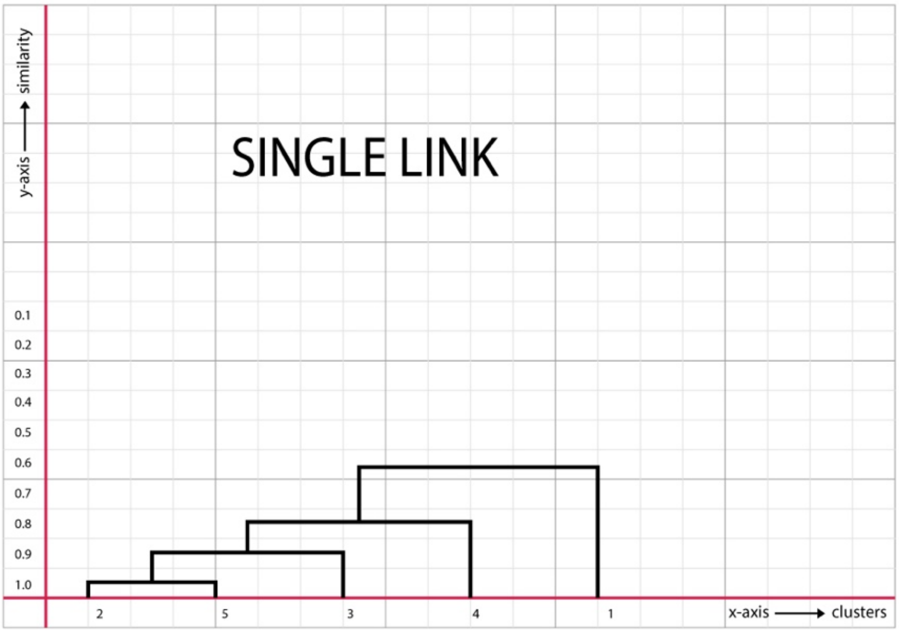
Problem 7

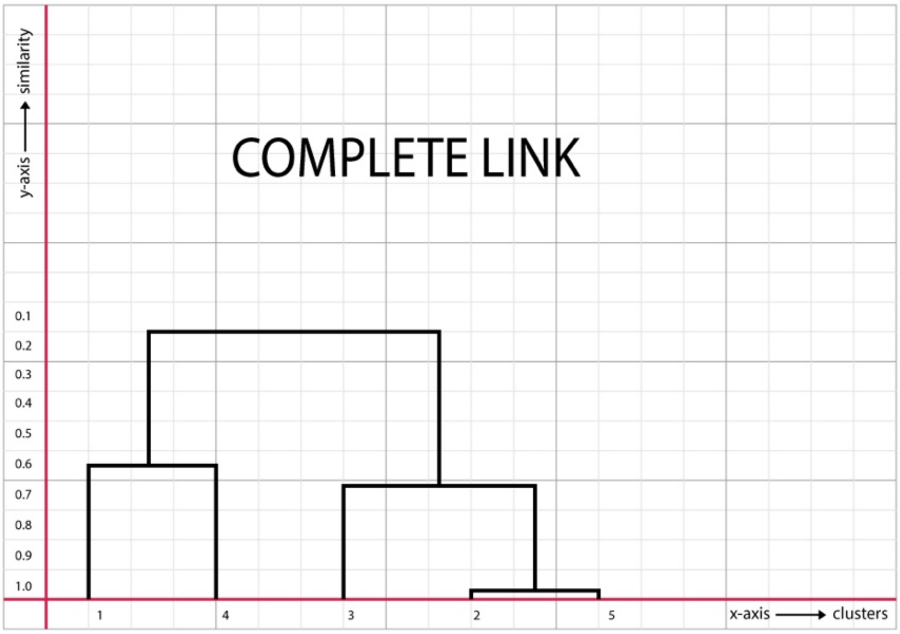
Denser regions should be allocated more centroids.

Problem 11

- If the SSE of one attribute is low for all clusters, then it is possible that the attribute is a constant and cannot help with clustering.
- If the SSE of an attribute is low for just one cluster, then that attribute helps defines the cluster.
- The attribute could just be noise if the SSE is high for all clusters.
- An attribute doesn't help define the cluster when the SSE of that attribute is high for just one cluster.
- We can omit an attribute in our clustering if the SSE is low or high for all clusters using that attribute, simply because they tend to change the cluster centroids, possibly making them useless as the attribute could be a constant or noise.

Problem 16





Problem 17

a) Clusters using the centroids

i) {18,45}

First cluster = {6, 12, 18, 24, 30}

Sum Squared Error = $(18-6)^2 + (18 - 12)^2 + (18 - 18)^2 + (18 - 24)^2 + (18 - 30)^2$ = **360**

Second cluster = {42, 48}

Sum Squared Error = $(45 - 42)^2 + (45 - 48)^2$ = **18**

Total Error = 360 + 18 = **378**

ii) 15, 40}

First cluster = {6, 12, 18, 24}

Sum Squared Error = $(15 - 6)^2 + (15 - 12)^2 + (15 - 18)^2 + (15 - 24)^2$ = **180**

Second cluster = {30, 42, 48}

Sum Squared Error = $(40 - 30)^2 + (40 - 42)^2 + (40 - 48)^2$ = **168**

Total Error = 180 + 168 = **348**

b) Stable solutions for both centroids

c) Clusters produced are {6, 12, 18, 24, 30} & {42, 48}

d) Single link produces most natural clustering.

e) This clustering corresponds to center-based and contiguous clustering.

f) For finding clusters of different sizes, k-means is not good method. This can be seen as most natural clustering has a higher sum squared error than unnatural clustering.

Problem 21

Entropy of Cluster 1 = $-\left\{\left(\left(\frac{1}{693}\right) log_2\left(\frac{1}{693}\right)\right) + \left(\left(\frac{1}{693}\right) log_2\left(\frac{1}{693}\right)\right) + \left(\left(\frac{0}{693}\right) log_2\left(\frac{0}{693}\right)\right) + \left(\left(\frac{11}{693}\right) log_2\left(\frac{11}{693}\right)\right) + \left(\left(\frac{4}{693}\right) log_2\left(\frac{4}{693}\right)\right) + \left(\left(\frac{676}{693}\right) log_2\left(\frac{676}{693}\right)\right)\right\}$ = **0.2**

Purity of Cluster 1 = $\frac{676}{693}$ = **0.975**

Entropy of Cluster 2 = $-\left\{\left(\left(\frac{27}{1562}\right) log_2\left(\frac{27}{1562}\right)\right) + \left(\left(\frac{89}{1562}\right) log_2\left(\frac{89}{1562}\right)\right) + \left(\left(\frac{333}{1562}\right) log_2\left(\frac{333}{1562}\right)\right) + \left(\left(\frac{827}{1562}\right) log_2\left(\frac{827}{1562}\right)\right) + \left(\left(\frac{253}{1562}\right) log_2\left(\frac{253}{1562}\right)\right) + \left(\left(\frac{33}{1562}\right) log_2\left(\frac{33}{1562}\right)\right)\right\}$ = **1.840**

Purity of Cluster 2 = $\frac{827}{1562}$ = **0.529**

Entropy of Cluster 3 = $-\left\{\left(\left(\frac{326}{949}\right) log_2\left(\frac{326}{949}\right)\right) + \left(\left(\frac{465}{949}\right) log_2\left(\frac{465}{949}\right)\right) + \left(\left(\frac{8}{949}\right) log_2\left(\frac{8}{949}\right)\right) + \left(\left(\frac{105}{949}\right) * log_2\left(\frac{105}{949}\right)\right) + \left(\left(\frac{16}{949}\right) log_2\left(\frac{16}{949}\right)\right) + \left(\left(\frac{29}{949}\right) log_2\left(\frac{29}{949}\right)\right)\right\}$ = **1.696**

Purity of Cluster 3 = **0.49**

Entropy of Total = $\left(\frac{693}{3204} * 0.2\right) + \left(\frac{1562}{3204} * 1.8407\right) + \left(\frac{949}{3204} * 1.696\right)$ = **1.443**

Purity of Total = $\left(\frac{693}{3204} * 0.975\right) + \left(\frac{1562}{3204} * 0.53\right) + \left(\frac{949}{3204} * 0.49\right)$ = **0.614**

Problem 22

a) The uniformly spaces set of points will have regions of small and large densities whereas the uniformly distributed set of points will have a uniform density.

b) For k=10, the set of uniformly spaced points will have a lower SSE.

c) DBSCAN merges all points of the uniformly distributed set of points as a cluster or label them as noise. DBSCAN might find useful clusters for the uniformly distributed points as it has different density regions.

Problem 23

Let $\boldsymbol{avg}$ indicate average distance to other points in the cluster.
Let $\boldsymbol{avg_{min}}$ be the minimum average distance to points in other clusters.

$$Point\ P_1:\ Silhouette\ Coefficient = 1 - \frac{avg}{avg_{min}} = 1 - \frac{0.1}{\frac{0.65+0.55}{2}} = \mathbf{0.833}$$

$$Point\ P_2:\ Silhouette\ Coefficient = 1 - \frac{avg}{avg_{min}} = 1 - \frac{0.1}{\frac{0.70+0.60}{2}} = \mathbf{0.846}$$

$$Point\ P_3:\ Silhouette\ Coefficient = 1 - \frac{avg}{avg_{min}} = 1 - \frac{0.3}{\frac{0.65+0.70}{2}} = \mathbf{0.555}$$

$$Point\ P_4:\ Silhouette\ Coefficient = 1 - \frac{avg}{avg_{min}} = 1 - \frac{0.3}{\frac{0.60+0.55}{2}} = \mathbf{0.478}$$

$$Average\ Silhouette\ Coefficient\ (Cluster\ 1) = \frac{0.833+0.846}{2} = \mathbf{0.839}$$

$$Average\ Silhouette\ Coefficient\ (Cluster\ 2) = \frac{0.555+0.478}{2} = \mathbf{0.516}$$

$$Overall\ Silhouette\ Coefficient = \frac{0.8395+0.5165}{2} = \mathbf{0.678}$$

Practicum Problems

Problem 2.1

Cluster assignment and class labels show no defining relationship.

Probabilities for Class 1:

$Cluster\ 1 = 0.482 \pm 0.062$

$Cluster\ 2 = 0.261 \pm 0.055$

$Cluster\ 3 = 0.257 \pm 0.054$

Probabilities for Class 2:

$Cluster\ 1 = 0.957 \pm 0.047$

$Cluster\ 2 = 0.043 \pm 0.047$

$Cluster\ 3 = 0$

Probabilities for Class 3:

$Cluster\ 1 = 1$

$Cluster\ 2 = 0$

$Cluster\ 3 = 0$

No change in result if we change the max depth.

Problem 2.2

Silhouette scoring for

$k = 2\ is\ 0.59$

$k = 3\ is\ 0.52$

$k = 4\ is\ 0.48$

$k = 5\ is\ 0.23$

$\boldsymbol{k = 2}$ has the best score.

Centroids for $\boldsymbol{k = 2}$ are

$Centroid\ 1 = \{2.597, 0.805, 0.946, 0.844, 1.619, 0.849, 1.606, 0.793, 0.620\}$

$Centroid\ 2 = \{6.700, 6.360, 6.289, 5.286, 4.988, 7.509, 5.624, 5.541, 2.108\}$

Distance between Centroids = **13.87**

Problem 2.3

We get the highest silhouette score for $\boldsymbol{k = 3}$ at 0.7234

Homogeneity scores the % of points that are in the same class label clustered in the same cluster. Only 12.45% (Homogeneity = 0.1245) of the points with the same class label are clustered together.

Completeness scores the % of points that are clustered together and have the same class label. Only 18.21% (Completeness = 0.1821) of the points that were clustered together have the same class label.

The mean values for all the features of each of the clusters do not differ from the centroid co-ordinates for the clusters.