CS 422

Data Mining

Homework Assignment 3

Ronit Rudra
A20379221
rrudra@hawk.iit.edu

Course Instructor
Prof. Jawahar Panchal

1 Textbook Questions

1.1 Chapter 6

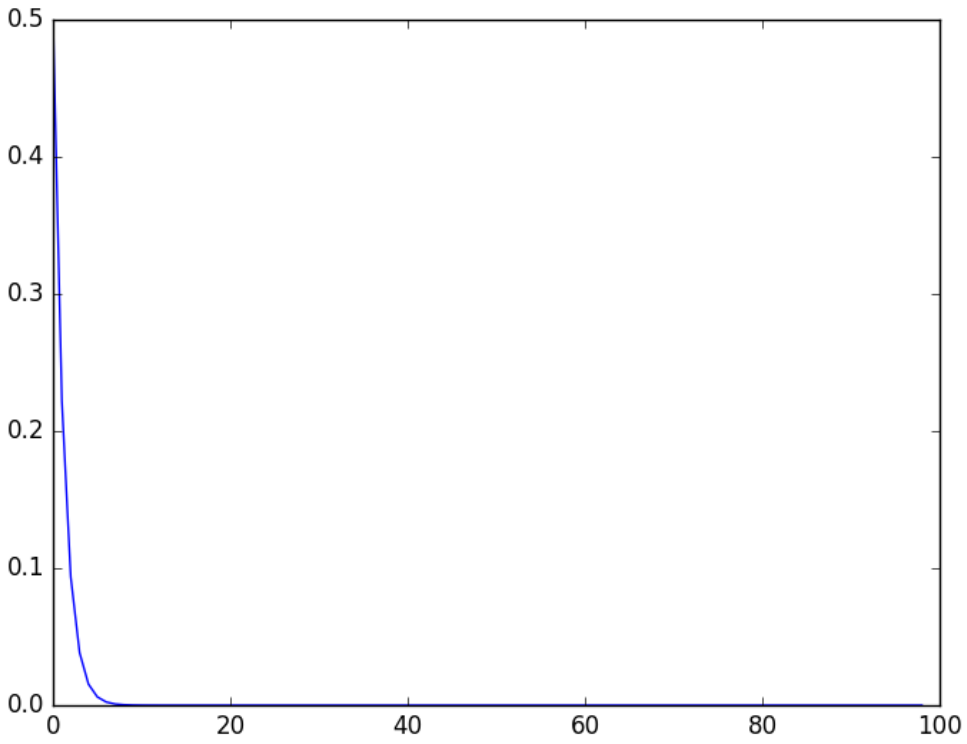
1.1.1 Question 4

(a) We are given the probability expression as:

$$p = \frac{k!}{k^k} \tag{1}$$

where, k is the number of clusters. For cluster sizes ranging from 2 to 100, the probability of selecting one point from each cluster is:

Figure 1: Probability Versus Cluster Size



(b) By analytical calculations, we have:

$$P(\textit{point belong to cluster}) = \frac{1}{K}$$
$$P(\textit{point does not belong to cluster}) = 1 - \frac{1}{K}$$
$$P(2K \textit{ points do not belong to cluster})) = (1 - \frac{1}{K})^{2K}$$

$$P(at\ least\ one\ point\ comes\ from\ a\ cluster) = 1 - (1 - \frac{1}{K})^{2K}$$
$$P(at\ least\ one\ point\ comes\ from\ each\ cluster) = (1 - (1 - \frac{1}{K})^{2K})^{K}$$

Plugging in the values of k=10,100 and 100, we get:

$$P(k = 10) = (1 - (1 - \frac{1}{10})^{20})^{1}0 = 0.2735$$
$$P(k = 100) = (1 - (1 - \frac{1}{100})^{200})^{1}00 = 5.6598 \times 10^{-7}$$
$$P(k = 10) = (1 - (1 - \frac{1}{1000})^{2000})^{1}000 = 8.236 \times 10^{-64}$$

1.1.2 Question 7

Option **(b)** would be the most appropriate solution as more centroids would be required to cluster datapoints in the less dense region as the points are further apart. For the denser region, less number of centroids can be used as the datapoints are close together and would contribute towards fewer centroids.

1.1.3 Question 11

The Sum of Squared Errors is given by:

$$SSE = \sum_{i=1}^{k} \sum_{x \epsilon c_i} dist(x, c_i)^2$$

Total SSE is the sum of the SSE for each separate attribute.

(a) If SSE of one attribute is low for all clusters it means that it is not representative of the data and can be discarded.

(b) The attribute with lower SSE for one cluster means that it is representative of the data and is a suitable definition for the cluster.

(c) For an attribute with high SSE for all clusters, there is a high chance that this is an outlier or plain noise.

(d) If the SSE is high for just one cluster, it would not be chosen to define the cluster as other attributes, which have lower SSE, would be used to define the cluster. **(e)** In order to improve clustering, attributes with low or high SSE across all clusters (see part (a) and (c)) should be discarded as they do not contribute to the cluster and may introduce high SSE resulting in spurious clusters.

1.1.4 Question 17

The One dimensional data is {6,12,18,24,30,42,48}

(a) (i) Centroids are {18,45}

$$Centroid\ 1\ distances = \{12, 6, 0, 6, 12, 24, 30\}$$
$$Centroid\ 2\ distances = \{39, 33, 27, 21, 15, 3, 3\}$$
$$Cluster\ 1 = \{6, 12, 18, 24, 30\}$$
$$Cluster\ 2 = \{42, 48\}$$

$$SSE_{cluster\ 1} = 12^2 + 6^2 + 0^2 + 6^2 + 12^2 = 360$$
$$SSE_{cluster\ 2} = 3^2 + 3^2 = 18$$
$$Total\ Error = 360 + 18 = 378$$

(ii) Centroids are {15,40}

$$Centroid\ 1\ distances = \{9, 3, 3, 9, 15, 27, 33\}$$
$$Centroid\ 2\ distances = \{34, 28, 22, 16, 10, 2, 8\}$$
$$Cluster\ 1 = \{6, 12, 18, 24\}$$
$$Cluster\ 2 = \{30, 42, 48\}$$
$$SSE_{cluster\ 1} = 9^2 + 3^2 + 3^2 + 9^2 = 180$$
$$SSE_{cluster\ 2} = 10^2 + 2^2 + 8^2 = 168$$
$$Total\ Error = 180 + 168 = 348$$

(b) Performing K-Means with these centroids we have:

$$For\ \{18, 45\}$$
$$Cluster\ 1 = \{6, 12, 18, 24, 30\}$$
$$Cluster\ 2 = \{42, 48\}$$
$$C_{1_{new}} = \frac{6 + 12 + 18 + 24 + 30}{5} = \frac{90}{5} = 18$$
$$C_{2_{new}} = \frac{42 + 48}{2} = \frac{90}{2} = 45$$
$$For\ 15, 40$$
$$Cluster\ 1 = \{6, 12, 18, 24\}$$
$$Cluster\ 2 = \{30, 42, 48\}$$
$$C_{1_{new}} = \frac{6 + 12 + 18 + 24}{4} = \frac{60}{5} = 15$$
$$C_{2_{new}} = \frac{30 + 42 + 48}{3} = \frac{120}{3} = 40$$

Thus, these solutions are stable as the centroid does not change.

(c) In Single Link, we take singleton clusters and then merge objects closest to each other to form a compund object. The Distance of a compound object to another object is the minimum distance between a member of the compound object and the other object. This is done recursively till all points are under one cluster. From the data given, we have:

$$level\ 1 : \{6, 12\}\ and\ \{42, 48\}$$
$$level\ 2 : \{\{6, 12\}, 18\}$$
$$level\ 3 : \{\{\{6, 12\}, 18\}, 24\}$$
$$level\ 4 : \{\{\{\{6, 12\}, 18\}, 24\}, 30\}$$
$$level\ 5 : \{\{\{\{\{6, 12\}, 18\}, 24\}, 30\}, \{42, 48\}\}$$

Thus, the two clusters created are $\{6, 12, 18, 24, 30\}$ and $\{42, 48\}$.

(d) Single Link produces the most natural clustering intuitively.

(e) Single link clustering produces contiguous clusters as the points in the clsuter are similar to each other than to points not in the cluster.

(f) The K-Means clustering tries to minimize the SSE resulting in formation of smaller clusters with approximately equal number of objects in each cluster. This can be seen from part (a) in which the cluster with lower SSE is unnatural.

1.1.5 Question 21

Entropy: Degree to which each cluster consists of objects of single class.

$$e_i = -\sum_{j=1}^{L} p_{ij} log_2 p_{ij} \tag{2}$$

where, L=no.of classes, i=i^{th} cluster, $p_{ij} = \frac{m_{ij}}{m_i}$, m_i=no. of objects in cluster i, m_{ij}=no. of objects of class j in cluster i.

Total entropy is the sum of entropies of each cluster weighted by the size of each cluster.

$$e = \sum_{i=1}^{K} e_i \cdot \frac{m_i}{m} \tag{3}$$

Purity: Another metric defined by:

$$Purity = P_i = max(P_{ij}) \tag{4}$$

$$Overall == \sum_{i=1}^{K} P_i \cdot \frac{m_i}{m} \tag{5}$$

From the table: **Cluster 1:**

$$\begin{aligned} e_1 = -(&\frac{1}{693} \cdot log_2(\frac{1}{693}) + \frac{1}{693} \cdot log_2(\frac{1}{693}) + \frac{0}{693} \cdot log_2(\frac{0}{693}) + \\ &\frac{4}{693} \cdot log_2(\frac{4}{693}) + \frac{11}{693} \cdot log_2(\frac{11}{693}) + \frac{676}{693} \cdot log_2(\frac{676}{693})) = 0.1998 \\ P_1 = max(&0.0014, 0.0014, 0.0158, 0.0057, 0.975) = 0.975 \end{aligned}$$

Cluster 2:

$$\begin{aligned} e_1 = -(&\frac{27}{1562} \cdot log_2(\frac{27}{1562}) + \frac{89}{1562} \cdot log_2(\frac{89}{1562}) + \frac{827}{1562} \cdot log_2(\frac{827}{1562}) + \\ &\frac{333}{1562} \cdot log_2(\frac{333}{1562}) + \frac{253}{1562} \cdot log_2(\frac{253}{1562}) + \frac{33}{1562} \cdot log_2(\frac{33}{1562})) = 1.836 \\ P_1 = max(&0.0172, 0.0569, 0.213, 0.530, 0.161, 0.021) = 0.53 \end{aligned}$$

Cluster 3:

$$\begin{aligned} e_1 = -(&\frac{326}{949} \cdot log_2(\frac{326}{949}) + \frac{465}{949} \cdot log_2(\frac{465}{949}) + \frac{8}{949} \cdot log_2(\frac{8}{949}) + \\ &\frac{105}{949} \cdot log_2(\frac{105}{949}) + \frac{16}{949} \cdot log_2(\frac{16}{949}) + \frac{29}{949} \cdot log_2(\frac{29}{949})) = 1.6954 \end{aligned}$$

$$P_1 = max(0.3435, 0.489, 0.0084, 0.1106, 0.0168, 0.0305) = 0.489$$

Total:

$$Entropy = 0.1998 \cdot \frac{693}{3204} + 1.836 \cdot \frac{1562}{3204} + 1.6954 \cdot \frac{949}{3204} = 1.44$$
$$Purity = = 0.975 \cdot \frac{693}{3204} + 0.53 \cdot \frac{1562}{3204} + 0.489 \cdot \frac{949}{3204} = 0.614$$

1.1.6 Question 22

(a) The uniformly spaced points will ahve constant density throughout the region while the points generated through the uniform distribution will have variable density.

(b) The Random points will have lower SSE due to less dense and more dense regions.

(c) For the uniform data, DBSCAN might cluster all of them in one cluster or flag them as noise but would perform relatively well on the random data as there is a variation of density.

1.1.7 Question 23

The Silhouette Coefficient is:

$$S = \frac{(b_i - a_i)}{max(b_i, a_i)} \tag{6}$$

Cluster 1

Point P1:

$$B_1 = \frac{(0.65 + 0.55)}{2} = 0.6$$
$$S_1 = \frac{(0.60.1)}{0.6} = 0.833$$

Point P2:

$$B_2 = \frac{(0.7 + 0.6)}{2} = 0.65$$
$$S_2 = \frac{(0.650.1)}{0.65} = 0.846$$

Average S for cluster 1 = $\frac{(0.833+0.846)}{2} = 0.864$

Cluster 2

Point P3:

$$B_3 = \frac{(0.65 + 0.7)}{2} = 0.675$$
$$S_3 = \frac{(0.6750.3)}{0.675} = 0.556$$

Point P4:

$$B_4 = \frac{(0.55 + 0.6)}{2} = 0.575$$
$$S_4 = \frac{(0.5750.3)}{0.575} = 0.478$$

Average S for cluster 2 = $\frac{(0.556+0.478)}{2} = 0.517$

Overall S = $\frac{(0.864+0.517)}{2} = 0.69$

1.1.8 Question 24

From the similarity matrix and the ideal similarity matrix, we get vectors from the upper triangular region:

$$A = 0.8, 0.65, 0.55, 0.7, 0.6, 0.9$$
$$B = = 1, 0, 0, 0, 0, 1$$

The Correlation R is given by:

$$R = \frac{s_{xy}}{s_y \cdot x_y}$$

$$s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y})$$

$$s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2}$$

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

Plugging in the values we get:

$$\overline{A} = 0.7$$
$$\overline{B} = 0.33$$
$$s_A = 0.1322$$
$$s_B = 0.5163$$
$$S_{AB} = 0.06$$
$$R = \frac{0.06}{0.1322 \cdot 0.5163} = 0.00409$$