**ISOM 2600 Introduction to Business Analytics**

**Weekly Exercise 3**

Q1. The standardized data below are produced by standardizing the original data (use sample standard deviation here) below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Original Data: | | |  | Standardized Data: | | |
| Index |  |  |  | Index |  |  |
| 1 | 1 | 3 |  | 1 | -1.1209 | 1.0911 |
| 2 | 5 | 1 |  | 2 | a | -0.2182 |
| 3 | 4 | 0 |  | 3 | 0.3203 | b |

Given that . What are the values of a and b respectively?

Solution:

Q2. The elbow plot of a k-means clustering model is provided below:

形状

描述已自动生成

What is the optimal number of clusters?

1. 1
2. 2
3. 3
4. 4

Solution: From the elbow plot, we can see that the WSS decreases non-significantly after .

Q3-Q5. A dataset with the labels clustered by a clustering method is given below:

|  |  |  |  |
| --- | --- | --- | --- |
| Index |  |  | Label |
| 1 | 1 | -1 | 1 |
| 2 | 0 | 1 | 1 |
| 3 | 2 | 2 | 0 |

What is the coordinate of the centroid for label 1? ( means the value in is 1 and the value in is 0.)

Solution: Coordinate:

Q4. What is the Euclidean distance between centroid for label 1 and 2?

1. 1.5
2. 2
3. 2.5
4. 3

Solution: distance:

Q5. If there are two new points, (0, 0) and (1, 1), what label do they most likely have?

1. Label 0 for (0, 0), Label 0 for (1, 1)
2. Label 0 for (0, 0), Label 1 for (1, 1)
3. Label 1 for (0, 0), Label 0 for (1, 1)
4. Label 1 for (0, 0), Label 1 for (1, 1)

Solution: The Euclidean distance from (0, 0) to centroids of cluster 0 and 1 are:

The Euclidean distance from (1, 1) to centroids of cluster 0 and 1 are:

The closer cluster will be the label for two new points, i.e., label 0 for (0,0), label 1 for (1,1)

Q6. Which of the following statement is **FALSE**?

1. Clustering belongs to the unsupervised learning model
2. When using K-means algorithm, the number of clusters should be pre-determined
3. After fitting a K-means model, points in the input dataset will be selected as the centroids of clusters
4. Euclidean distance is used as the metric to measure the distance between two points in K-means

Solution: C. The centroids need not to be the points in the dataset