Data Mining II

Quiz  
March 10th

1. In k-means clustering **name and BRIEFLY describe two strategies** for selecting the number of clusters.

A number of different ways have been proposed to determine the optimal number of clusters in a data set, two of them are,

1. Elbow

2. Silhouette

**THE ELBOW METHOD**

It is one of the popular methods, here the sum of the squares at each number of clusters is calculated. We need to keep in mind that the total within cluster sum of square needs to be as small as possible. We should choose a number of clusters such that adding another cluster doesn’t improve much better the total within cluster sum of square. The process begins by computing a clustering algorithm for the different values of k. Then calculate the total within cluster sum of square for each k. Later plot the graph of within cluster sum of square vs the number of clusters k. At some point in the plotted graph we can see a bend and that point is considered to be chosen as the appropriate number of clusters. This elbow type bend indication cannot always be identified clearly, so this method is very subjective and not so reliable.

**A picture containing text

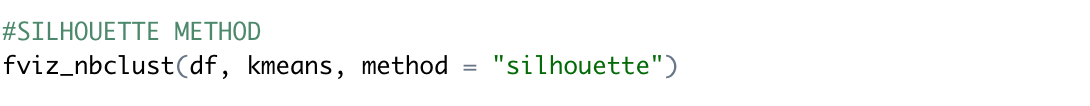
Description automatically generated**

**Chart, line chart

Description automatically generated**

**THE SILHOUETTE METHOD**

It is another visualization method to help determine the optimal number of clusters. Silhouette method helps measure quality, it helps make us understand how well each object lies within its cluster. It computes the average silhouette of observations for different values of k. The one that maximizes the average silhouette over a range of possible values for k is chosen to be the optimal number of clusters k. The process begins by computing a clustering algorithm for the different values of k. Then calculate the average silhouette of observations. Later a curve of average silhouette vs number of clusters k is plotted. The point where it maximizes the average silhouette over the range of possible values of k is considered to be chosen as the appropriate number of clusters.

****

Chart, line chart

Description automatically generated

1. From ISLR:

Table

Description automatically generated

A picture containing chart

Description automatically generated

a)

Chart, scatter chart

Description automatically generated

Graphical user interface, text

Description automatically generated

b)

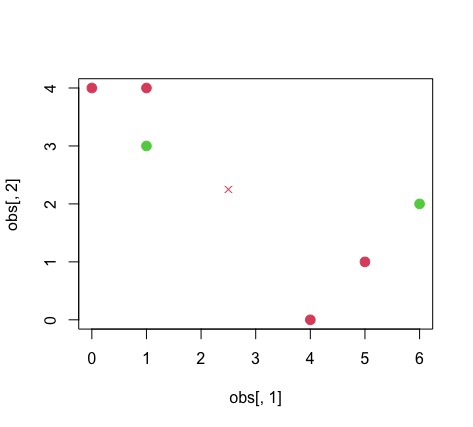
Chart, scatter chart

Description automatically generated

Text

Description automatically generated

c)

Chart, scatter chart

Description automatically generated

Graphical user interface, text

Description automatically generated

d)

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

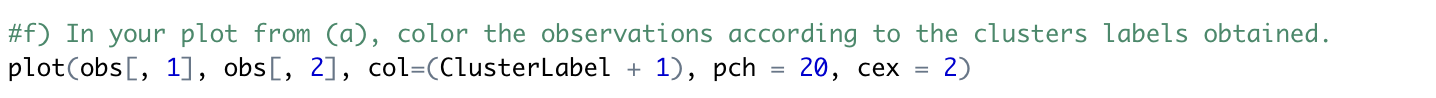
Text

Description automatically generated

e)

Chart, scatter chart

Description automatically generated



Chart, scatter chart

Description automatically generatedf)