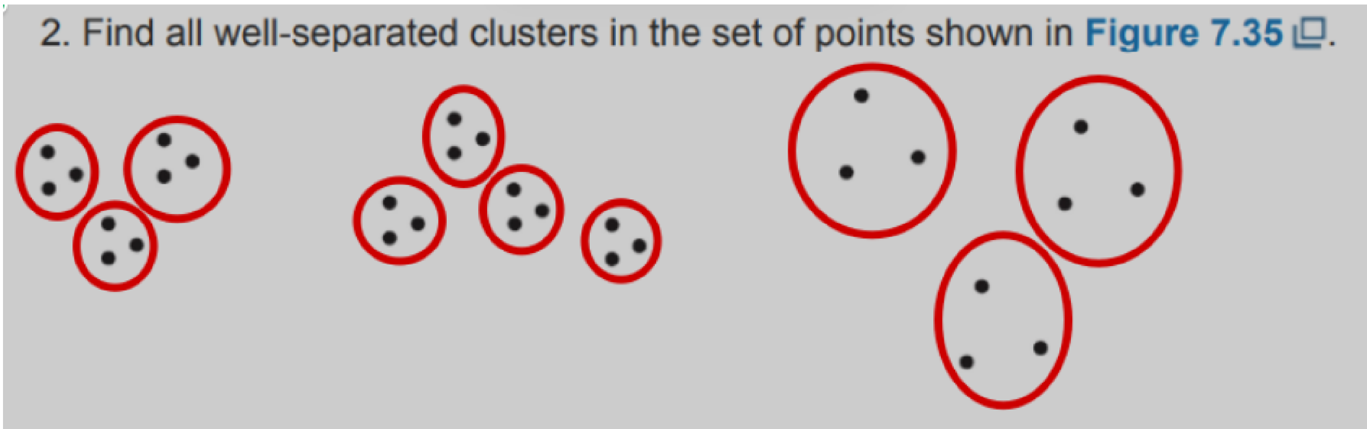
CS 422 Homework8

Xingli Li, Illinois Institute of Technology

1. Exercises

1.1-2

Answer:



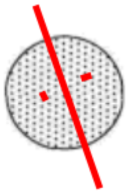
2. Find all well-separated clusters in the set of points shown in Figure 7.35.

1.1-6

For the following sets of two-dimensional points, (1) provide a sketch of how they would be split into clusters by K-means for the given number of clusters and (2) indicate approximately where the resulting centroids would be. As sume that we are using the squared error objective function. If you think that there is more than one possible solution, then please indicate whether each solution is a global or local minimum. Note that the label of each diagram in Figure 8.4 matches the corresponding part of this question, e.g., Figure 8.4(a) goes with part (a).

(a)

Answer:

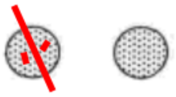


(a)

```
k=2,

In theory, there are countless ways to split a circle into two clusters, just a
straight line that divides
the circle equally. All of these solutions have the same global minimum error.
```

(b)

Answer:

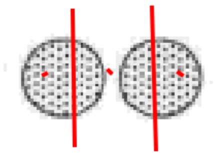


(b)

```
k=3,

In theory, there are countless ways to split a circle into two clusters, the
bisectors of first circle
(or second circle) can have any angle. All of these solutions have the same
global minimum error.
```

(c)

Answer:

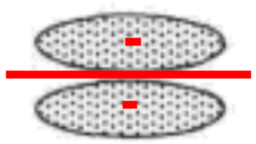


(c)

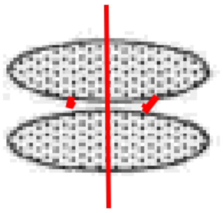
```
k =3,

The two straight line divides the circle, and making the cluster into three
parts.
```

(d)

Answer:



(d)

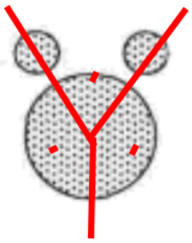


(d)

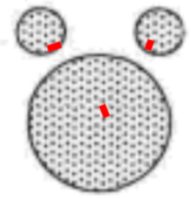
```
k=2,

According to the principle that the sum of the distances from the centroid to
other points in his cluster
is the smallest. In the first case, the two clusters are only local minimum, In
the second case, the
clusters is a global minimal solutions.
```

(e)

Answer:



(e)



(e)

```
k=3,

According to the principle that the sum of the distances from the centroid to
other points in his cluster
is the smallest. In the first case, the three clusters are only local minimum,
In the second case,
the clusters is a global minimal solutions.
```

1.1-11

Total SSE is the sum of the SSE for each separate attribute. What does it mean if the SSE for one variable is low for all clusters? Low for just one cluster? High for all clusters? High for just one cluster? How could you use the per variable SSE information to improve your clustering?

（a）

Answer:

If the SSE for one variable is low for all clusters, then the variable is almost a constant, which is little use when dividing the data into groups.

（b）

Answer:

If the SSE for one variable is low for just one clusters, then the variable is useful for defining the cluster.

（c）

Answer：

If the SSE for one variable is high for all clusters, then this probably means that the variable is noise.

（d）

Answer：

If the SSE for one variable is high for just one clusters, then this means that this variable does not help define the cluster. It may just be that the cluster defined by this variable is different from the clusters defined by other variables.

（e）

Answer：

Our aim is to have high/low SSE for all clusters, to eliminate the attributes distinguishing between clusters, as they are useless for the clustering. If the SSE of all clusters has a relatively high SSE relative to other attributes, the attributes of all clusters are particularly troublesome, because they introduce a lot of noise into the calculation of the whole SSE.