BYU IDAHO Machine Learning & Data Mining | CS 450

10 Prove : Assignment

Clustering the States

I. Objective

Understand the basics of clustering using k-means and agglomerative hierarchical clustering.

II. Instructions

This week you will look at different ways to cluster data regarding the United States. Once again, we will not be implementing any algorithms this week, but will be experimenting with them in R.

The dataset is included in the datasets package:

```
library(datasets)
myData = state.x77
```

This dataset has the following attributes for each of the 50 States:

- Population
- Income
- Illiteracy
- Life Exp
- Murder
- HS Grad
- Frost

- Area

Your instructions are to play around with the kmeans function and the hclust function to learn more about this dataset and about the clustering process. Please follow the steps in the Experiment Guidelines, and then look for ways to do additional analysis on this and other datasets.

Please refer to A few helpful hints for doing clustering in R in I-Learn for some basic syntax to help get you started.

III. Experiment Guidelines

To help get you started, please follow the prescribed steps below:

AGGLOMERATIVE HIERARCHICAL CLUSTERING

01. Load the dataset
02. Use hierarchical clustering to cluster the data on all attributes and produce a dendrogram
03. Repeat the previous item with a normalized dataset and note any differences
04. Remove "Area" from the attributes and re-cluster (and note any differences)
05. Cluster only on the Frost attribute and observe the results

USING *K*-MEANS

01. Make sure to use a normalized version of the dataset.
02. Using k-means, cluster the data into 3 clusters. Note the size of each cluster and the mean values. Do you have any insight into why they were divided this way?
03. Using a for loop, repeat the clustering process for k = 1 to 25, and plot the total within-cluster sum of squares error for each k-value.

04. Evaluate the plot from the previous item, and choose an appropriate k-value using the "elbow method" mentioned in your reading. Then re-cluster a single time using that k-value. Use this clustering for the remaining questions.

05. List the states in each cluster.

06. Use "clusplot" to plot a 2D representation of the clustering.

07. Analyze the centers of each of these clusters. Can you identify any insight into this clustering?

After going through these steps, you are encouraged to continue on to analyze this and other datasets further to see if you can find other interesting things.

IV. Submission

Prepare a PDF document with graphs and discussion for each of the points above. Then, list anything additional you have done (above and beyond these requirements).

Finally, please state which category you feel best describes your assignment and give a 1-2 sentence justification for your choice:

- A) Some attempt was made
- B) Developing, but significantly deficient
- C) Slightly deficient, but still mostly adequate
- D) Meets requirements
- E) Shows creativity and excels above and beyond requirements

Upload this document to I-Learn in the space provided.