Homework 3 Report

Microl Chen

March 16, 2023

1 Collaboration Statement

For this assignment, the following resources, help, or/and collaboration was utilized:
Overleaf - for the Latex Template
Stack Overflow - minor error debugging
Debugged, diagnosed, and isolated errors with Programiz.com
Referenced class textbook for k-means optimization, math formulas

2 Evaluation and Data

iris.data is the default data set given to us in class.
Pandas was used to load and set up the data set as well as to drop the unwanted categorical variable in the end.
The data set was segregated with ','(unlike the previous hw), so I used the same method (replace ' ', with ',') to prepare the data.

Below are the average SSE and SC values for varying k values. Note that I have ran each several times and selected an output that I believe is the most frequent one:
k = 5, SSE: 53.458791 SC: 0.450930
k = 4, SSE: 57.345409 SC: 0.497228
k = 8, SSE: 36.372285 SC: 0.434414

abalone.data is a data set that was a lot bigger. It contained multiple characteristics about individual abalones like sex, size, etc.
I pre-processed the data in relatively the same manner. However, since my code only works with a sequential column index (from 0 to n-1 columns), I had to rename the columns of categorical variables that I removed via the following code:

```
df.drop(df.columns[len(df.columns) - 1], axis=1, inplace=True)
    df.drop(df.columns[0], axis=1, inplace=True)
    temp = []
    for i in range(len(df.columns)):
        temp.append(i)
    df.set_axis(temp, axis=1, inplace=True)
```

I also dropped about 3000 rows from the original abalone.data since the run time was getting too long (due to the inefficiency of my code). The results shown are from the first 999 rows.

Below are the average SSE and SC values for varying k values. Note that these have only been ran once at their respective k values because it is more computationally expensive(at least with my code).
k = 4, SSE: 35.889511 SC: 0.495009
k = 5, SSE: 24.837724 SC: 0.490997
k = 8, SSE: 13.916506 SC: 0.428923

3 Insights

3.1 Experiment

From my testing, it seems that a higher k value enables us to obtain a smaller SSE while SC remains relatively consistent. This finding largely fits what we learned in class since we are effectively providing a more diverse set of initial cluster (by random luck) the more clusters we start with. It also seems like SSE significantly drops only when many data points are utilizing the full k clusters. If individual rows are not utilizing the full k or at least near k clusters, SSE does not seem to be effected even though we initialized with k centeroids.

3.2 Personal

This assignment has really helped me improve my python skills for working with data sets. Before this class, I took QTM 100, and that was about all the knowledge of data manipulation I have had. Additionally, working with k-means has made me really appreciate the use of tuples (some thing I normally never use). More-ever, I learned, through trial and error, how to make things like loc and iloc work (still don't really understand it completely) on pandas.