CSCI-720-Big Data Analytics

HW01 Report

Student Name: Guo, Zizhun

Email: zg2808@cs.rit.edu

Before heading to the reports, here offers a quick statement about the validation of homework's coding details. The code for this homework is worked purely individually with references on lecture pseudocode specifically towards the Otsu's realization. There might be several lines of codes that works for handling Pandas package's DataFrame and Series are similar to other's work, which is originated from my discussion with another student Martin Qian. I.e. `data[data < start]` for selecting data under the given condition. So just be advised for this.

Languages & Tools

Language: Python 3.8.1 32bit

IDE: VSCode, Jupyter Notebook, Anaconda

Packages: Numpy, Pandas, Matplotlib

Mystery Data Exploration Result

```
35                          17
36                          18
37                          22
38                          10
39                          18
(before) Average =             18.0
(before) Standard Deviation = 8.408329203831164
(after) Average =              18.0
(after) Standard Deviation =  8.408329203831164
```

Fig 1: The results for section 1

1. Exploratory Data Analysis

The average of data set: 18.0. The standard deviation of data set: 8.408 (around)

Since the last value is 18.0, there re-computed average is 18.0. The average stays the same. Since the last value equals to the average, so there would be no difference on average and standard deviation. Mathematically, this makes sense.

Snowfolks' Data Exploration Result based on Ages

2. 1D Clustering using Otsu's method on the age.

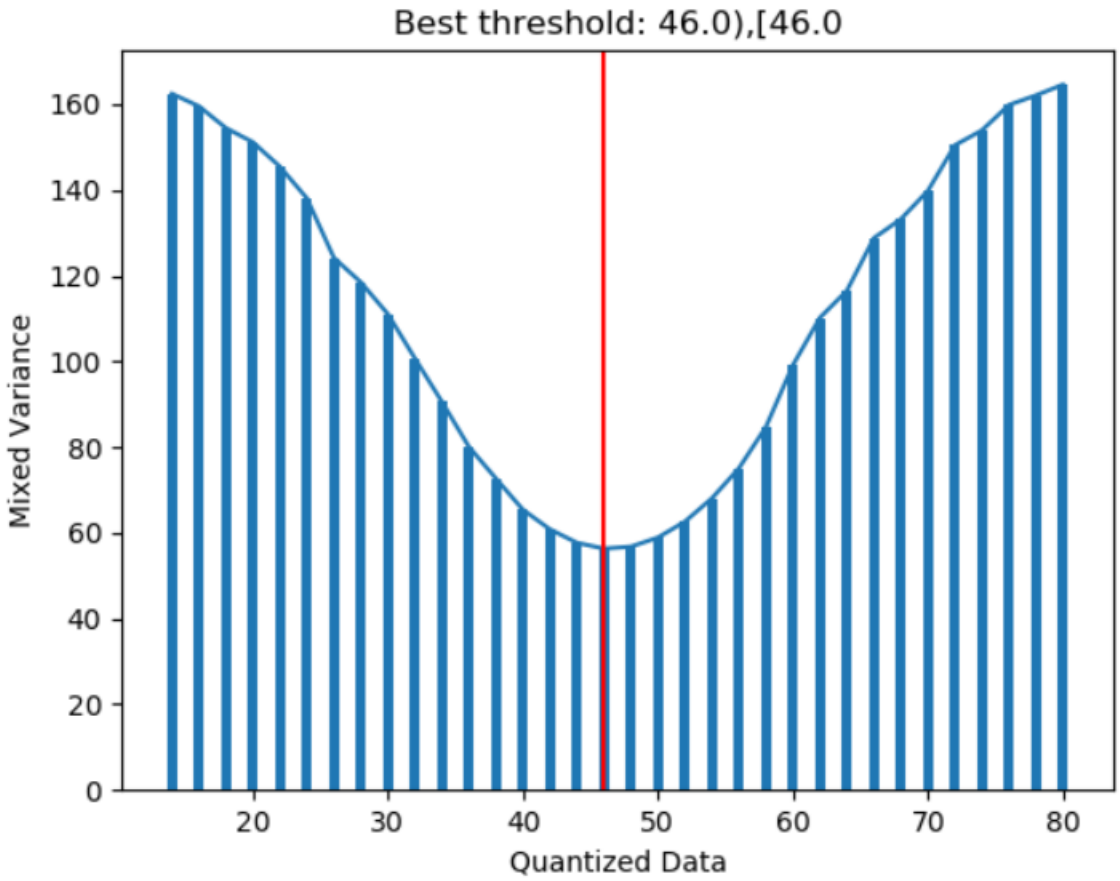
a. 44 is the best splitting age for the separation. As the result showed from these two intervals: [0, 46),[46, 80]

b. The minimal mixed variance is 56.309054955624774

c. My program has the first condition of "age ≤ each_threshold" for making decision if has two duplicates occurred. The program would choose the later one as the result. This situation does happen if we keep limited number of digits of result for comparing, which if happen too often, would cause result prone to be bigger than it supposed to be. However, if there is no digit's limitation, the duplicates rarely happen.

d. For quantizing the data, we can set different size of bins to quantize it. As for regressing data value to integer type, there are three options: floor, ceil and round. Both methods have different impacts on the data set, it depends on how the data set is distributed.

3. Graphing

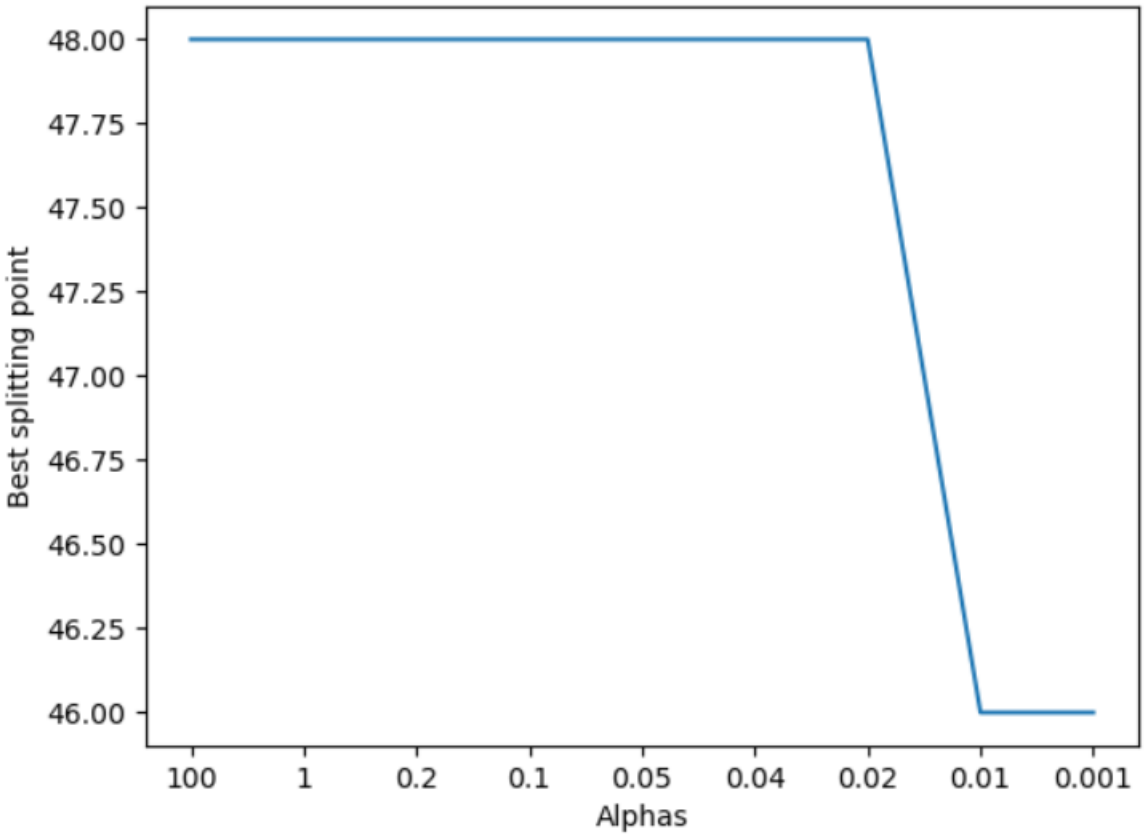


This is a graph of distribution of mixed variance based on different ages of snowfolks. The red line indicates the minimal mixed variance.

4. Exploring Regularization

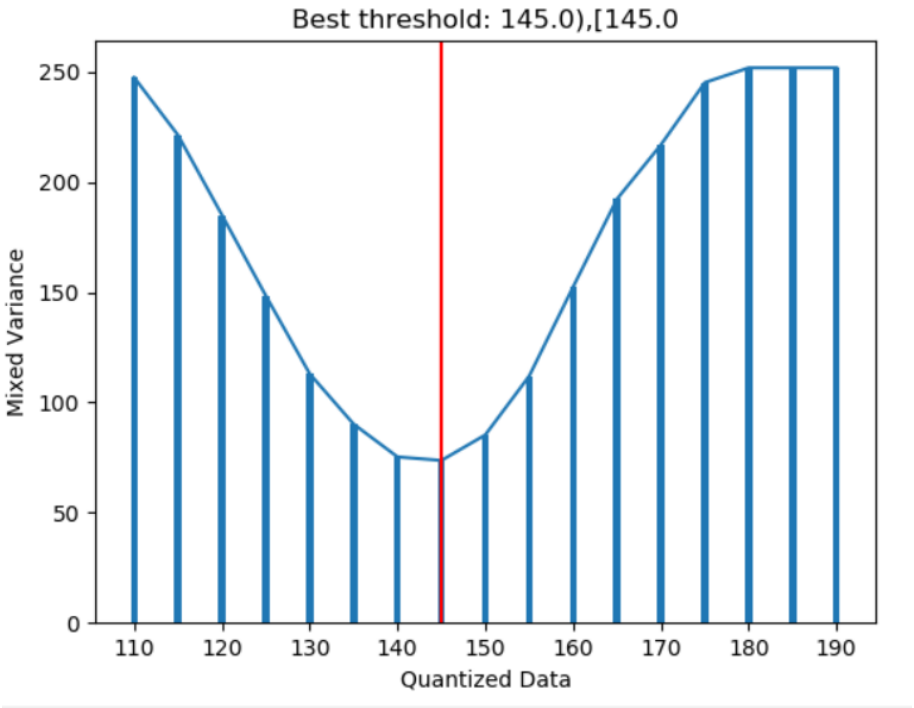
[100, 1, 1/5, 1/10, 1/20, to 1/25, to 1/50, and 1/100, 1/1000].

Which value of alpha causes the "best" splitting point to change?



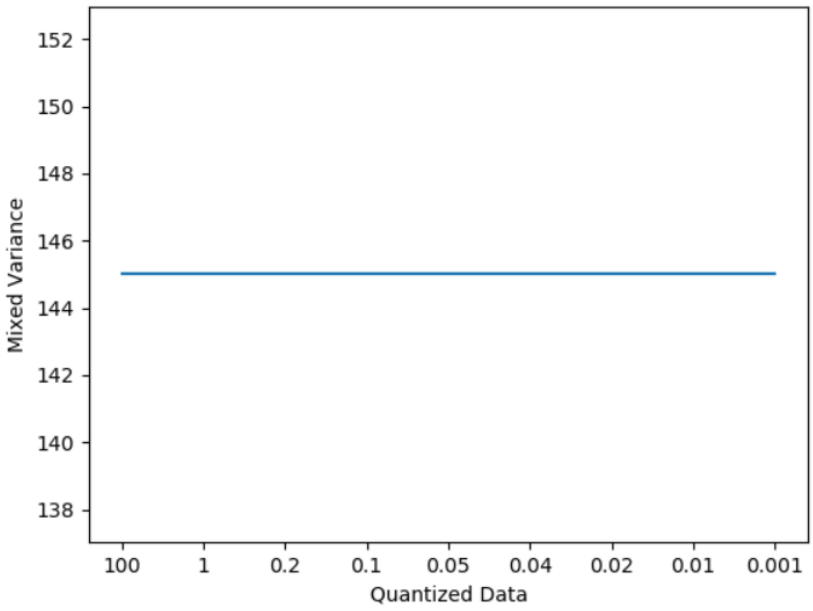
This graph shows the distribution of best splitting points under each alpha. There has only results changed while alpha equals to 1/100 or 1/1000, which implies if alpha is greater than these two values, the regularization would somehow dominant the object function. Since, the results changed does not change so much (it only decreases 1 unit bin), so all alphas are good to use for this project case that can be ideally divided.

Snowfolks' Data Exploration Result based on Heights



Under the same procedure on the heights dataset under the bin size changed to 5, the best selecting point is 145.

The minimal mixed variance is 73.71080858228834



For alphas, the final minimal cost unchanged.

Conclusion

- Through this homework, I gained more intuition on the data sets as how it is distributed, what the possible consequences would be if remove some values. As in the snowfolks dataset on heights, there is an outlier that is too big than the rest of others (The one equals to 190), so if remove this value, the average and standard deviation would change.
- I also learned how to bin the data using floor, ceil and round.
- For Otsu Method, through experiment, I learned by using different regularization formula, there would or would not have help on division. It depends on the alpha value that decides its weights or depends on how the regularization is made up. However, it is dependent on how our purpose is.
- Through development, I enhanced my familiarity on multiple tools like packages of Pandas, Numpy and Matplotlib, so that it makes more efficient on future homework and as well the future jobs.
- Last but not the least, this homework enhances the concept that use most efficient codes and methods to finish tasks. Make the codes clear and followed the conventions which make the world a lot easier and beautiful.