NetID: smarty
STUID: 2283420

Homework 3

CS 470
February 27, 2020

Collaboration Statement

The LaTeXtemplate I am using can be found on *overleaf.com*. I used the textbook extensively to design my algorithm. I discussed potential approaches with Jack Frumkes. I used descriptions and datasets from the UCI database to test my implementation of K-Means.

1 Algorithm

This implementation of K-Means is written in Python, using numpy and pandas. The code is properly commented. K-Means is written as a self-contained class, capable of fitting a distribution, then ouputting SSE and Silhouette coefficients as needed. The code in contained in a file named `kmeans.py`, which also contains a main method to test the class. I used the generalized euclidean distance as a distance metric.

2 Datasets

I used two datasets from the UCI database. The first dataset was the Iris dataset provided. The second dataset was named `Sonar:  Mines vs Rock`.

2.1 Iris

This dataset contains measurements of different species of Iris. The attributes seemed to have very close ranges, so I did not bother with normalizing them (there was no risk of one attribute skewing the distance metric). In order to apply K-Means to this dataset, I had to remove its last attribute, which contained the actual species of Iris the row belonged to. Below are the results for varying k

k	Silhouette	SSE
2	0.6808	152.37
3	0.5509	78.94
4	0.415	71.34
5	0.4928	46.56
6	0.341	47.79

We get the best Silhouette coefficient when `k = 2`. This is unexpected, as there are actually 3 species in the dataset. K-Means might need a greater breadth of attributes to correctly differentiate species of Iris. The `SSE` coefficient decreases as `k` increases, which might suggest that it is a weaker measure of the quality of K-Means.

2.2 Sonar: Mines vs Rock

The dataset was created to train a neural network to differentiate mines and rocks from a sonar signal. The data set contains patterns obtained from a variety of different aspect angles, spanning 90 degrees around a metal cylinder and 180 degrees around a rock. Each pattern is a set of 60 numbers in the range 0.0 to 1.0. Each number represents the energy within a particular frequency band, integrated over a certain period of time. Just like the Iris dataset, the Sonar dataset contained an attribute that represented the actual type (Rock, Mine) of object in the row, which I had to remove. Numeric attributes were provided already normalized. Here are the results for varying k.

k	Silhouette	SSE
2	0.19	280
3	0.19	235
4	0.165	220
5	0.172	207
6	0.133	197

The K-Means algorithm has a much harder time classifying this dataset. Overall, the silhouette coefficients seem best for `k = 2` and `k = 3` (The actual expected number of categories is 2). The SSE is very high, even though the data was provided normalized.

3 Conclusion

This experiment has helped me better understand the limitations of the K-Means algorithm. It seems to perform well when clusters are well defined and there are plenty of attributes to justify that. As soon as some attributes start to overlap, like in the Sonar dataset, K-Means has a more difficult time differentiating the data points.