5/30/2020
CS434 Machine Learning

Implementation Assignment 4 Report

Group Members:
Lindsey Kvarfordt: kvarforl@oregonstate.edu
Aiden Nelson: nelsonai@oregonstate.edu

Workload Split: 50/50, paired programming and discussion

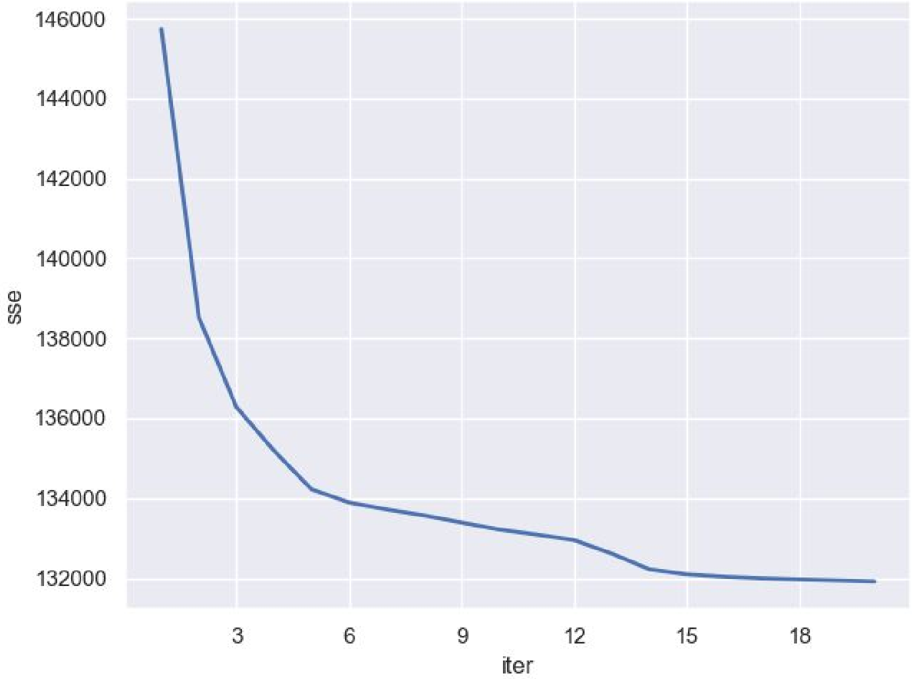
Part 1: k-means clustering

The program can run on flip using the following command from the src directory:
python3 main.py --pca=0 --kmeans=1

Make sure that the following packages are installed before running:

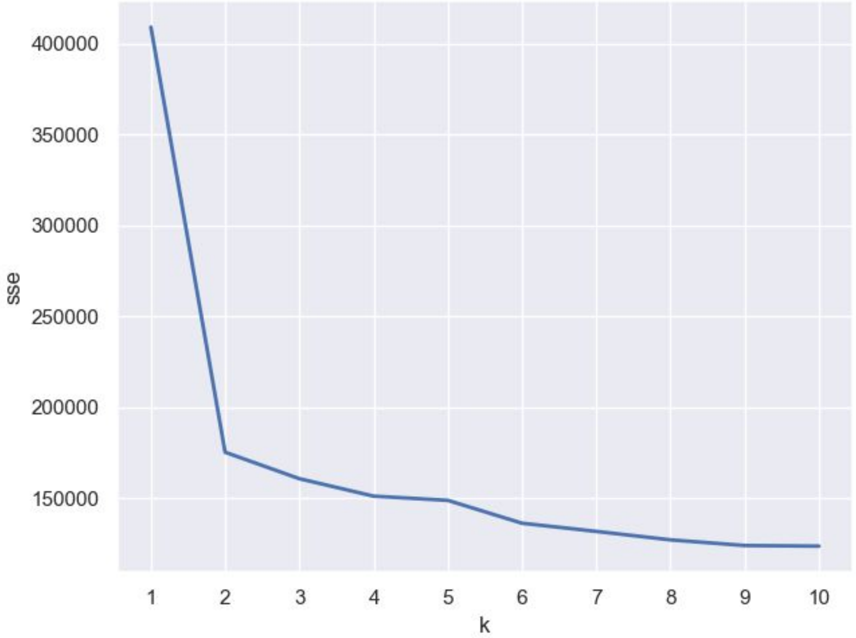
- Pandas
    - *pip3 install --user pandas*
- Seaborn
    - *pip3 install --user seaborn*

Graph: Average (over 5 runs) of SSE versus iterations for k = 6



SSE decreases as the number of iterations in kmeans increases.

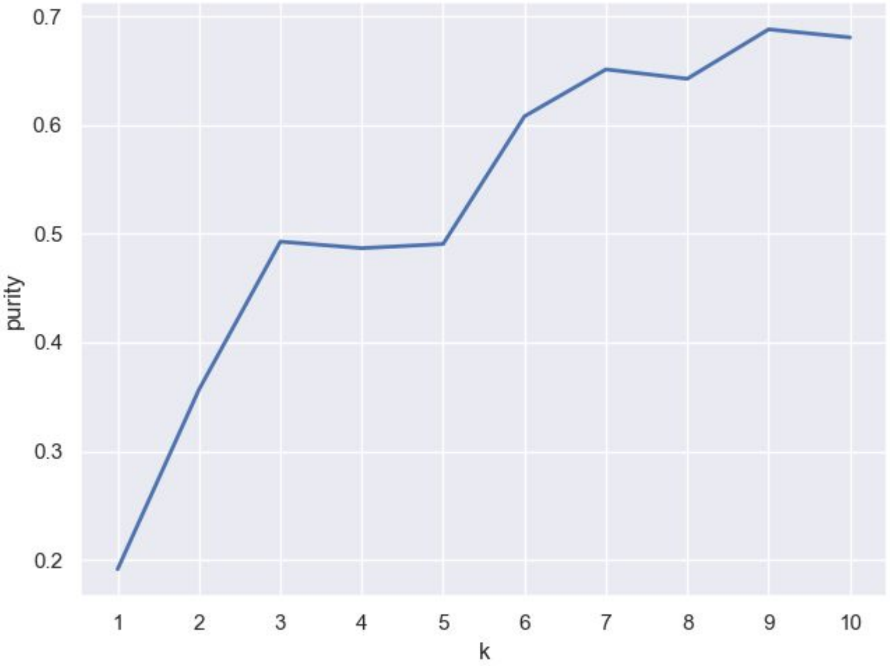
Graph: Average (over 5 runs) of the SSE versus k for k ∈ 1...10



Best k found:

The K with the lowest SSE was 10. This makes sense, because we know that we are looking for 10 clusters in our data. Forcing the same amount of data into less than 10 clusters causes the clusters to be bigger and more spread out, so our euclidean distance sum of squares error increases if there are less clusters.

Graph: Average of purity versus k for k ∈ 1 . . . 10 for the train set



Observations:

The purity of the clusters tends to increase as the number of clusters created increases. This makes sense because we tested on clusters 1-10, and we know the data should have 10 clusters. As the algorithm gets closer to having the correct number of clusters, the purity of the clusters increases because the clusters are being more correctly categorized.

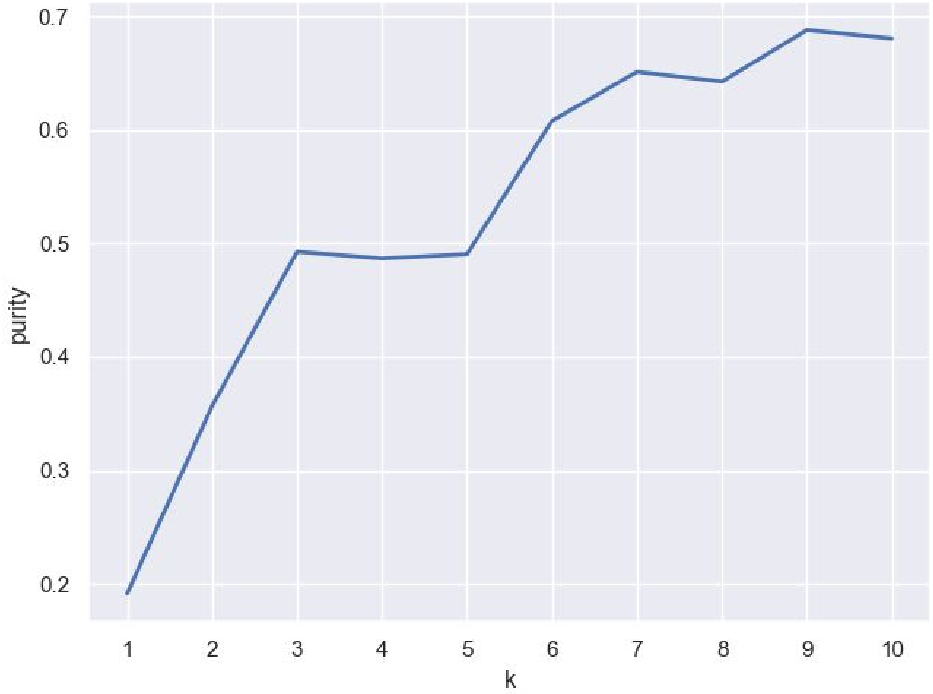
Part 2: Dimension reduction (PCA)

The program can run on flip using the following command from the src directory:

python3 main.py --pca=1 --kmeans=0

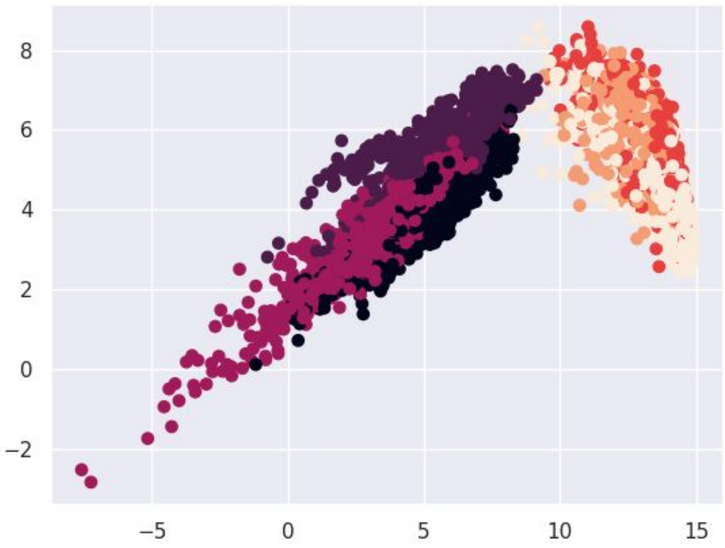
We found d to be 34 with a retain ratio of 0.9.

Graph: k-means for k ∈ 1 . . . 10 vs purity with reduce dimension for retain ratio r = 0.9



As in part 1, the purity of the dimension reduced data clusters tends to increase as the number of clusters increase.

Graph: dimension reduced training data



The figure above shows a two dimensional representation of the dimension reduced training data along the first two principal components. Because this is basically a projection or a shadow, if thinking in terms of 3 space, it makes sense that the clusters are somewhat overlapping.