CS 622 – Homework 2

Shen Chan Huang

February 26 2023

1 Python Version and Libraries Used

- Python Version: 3.8.13
- Pandas
- Numpy
- Statistics

2 Datasets

MNIST data. The training set has 949 rows and the testing set has 50 rows. I merged them using

– X_full = pd.concat([pd.read_csv(f) for f in csv_to_merge], ignore_index=True)

3 Description

To run the script, go to cmd and use "python -u Huang_HW2_CS622.py"

In this assignment, I implemented KNN (K-nearest neighbor) from scratch using Python. In addition, since I am in CS 622, I have to merge the given csv files and perform 5-fold cross validation.

First, I merged the two datasets. I wrote functions to perform the following:

1. Split the full dataset into training and testing set given an N for N-fold cross validation.
2. Calculate the Minkowski Distance.
3. Calculate the Cosine Similarity.

4. Find the K-Nearest Neighbor and gives the majority vote.

My main body of code first shuffles the dataset because the merged dataset is too well ordered. In order to ensure best training results, I had to shuffle it. Then pass it to (1.) to split and obtain a training and testing set.

Then inside KNN function, can choose to use either Minkowski distance (commented out in the code) or Cosine Similarity (used it by default). I used numpy's vector subtraction and element-wise exponentiation. Saved the result to an array and sort (ascending if Minkowski distance, descending if . After sorting, I used the mode function from the statistics library to find the most occurring class labels in the first K nearest neighbors.

For a fun thing to do, I also plotted for various K values.

4 Experimental Results

Experiment	Accuracy
Experiment 1	0.885
Experiment 2	0.93
Experiment 3	0.865
Experiment 4	0.86
Experiment 5	0.874
Average Accuracy	0.883

Table 1: Average accuracy with 5 fold cross validation with $K = 10$

Experiment	Accuracy
Experiment 1	0.93
Experiment 2	0.88
Experiment 3	0.88
Experiment 4	0.88
Experiment 5	0.89
Experiment 6	0.82
Experiment 7	0.89
Experiment 8	0.93
Experiment 9	0.84
Experiment 10	0.91
Average Accuracy	0.885

Table 2: Cosine Similarity: Average accuracy with 10 fold cross validation with $K = 10$

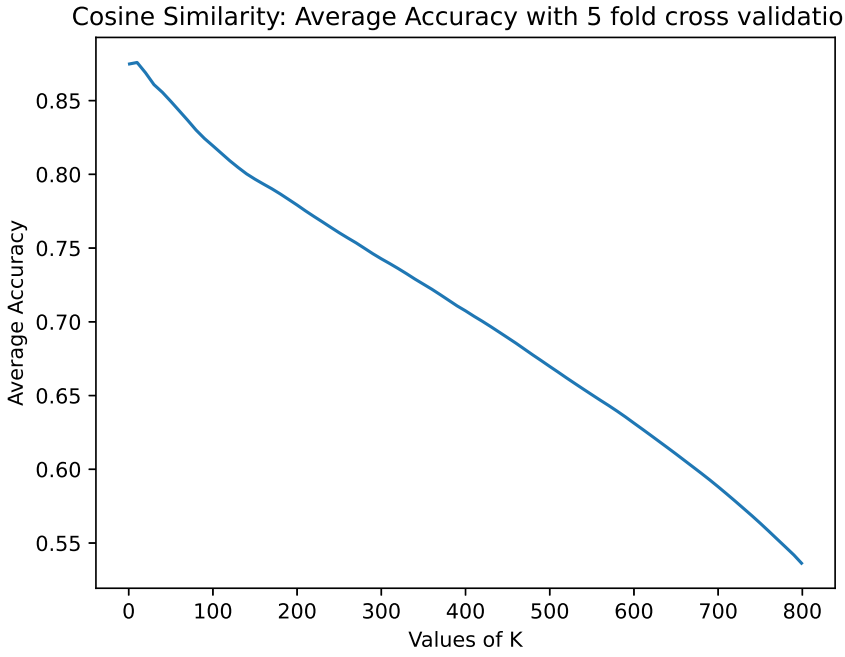

Figure 1: Cosine Similarity: 5-fold cross validation over various K values from 1 to 799

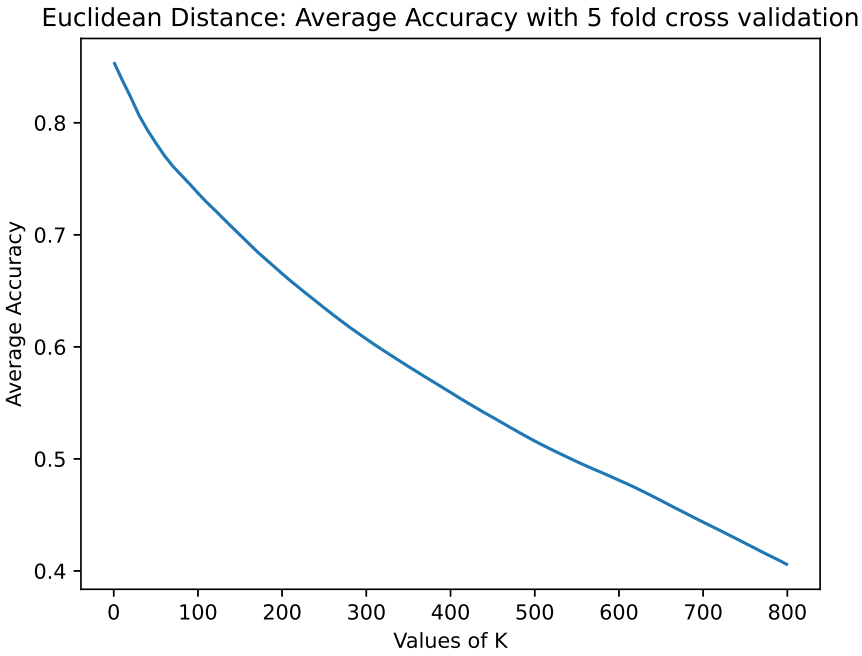

Figure 2: Cosine Similarity: 10-fold cross validation over various K values from 1 to 799