Computer Homework2

K_Nearest neighbors

Mohammad Mehdi Salehi

40330481

KNN classifier:

The KNN (K Neighbor Nearest) category calculates the distance between the sample and all the available data and then labels the sample that is closest to its data.

In fact, in this method, K data is found close to the sample, and the sample belongs to the one that has the highest value in the K of the data close to the sample.

To calculate the distance between the sample and the data, different methods such as the cosine function or the Euclidean distance can be used:

$$consine = \cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

$$Euclidean\ distance = d(q, p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

Data:

In this exercise, we'll use the openML 784_mnist dataset.

This dataset includes 784 pixel photos of different digits, each pixel is a binary.

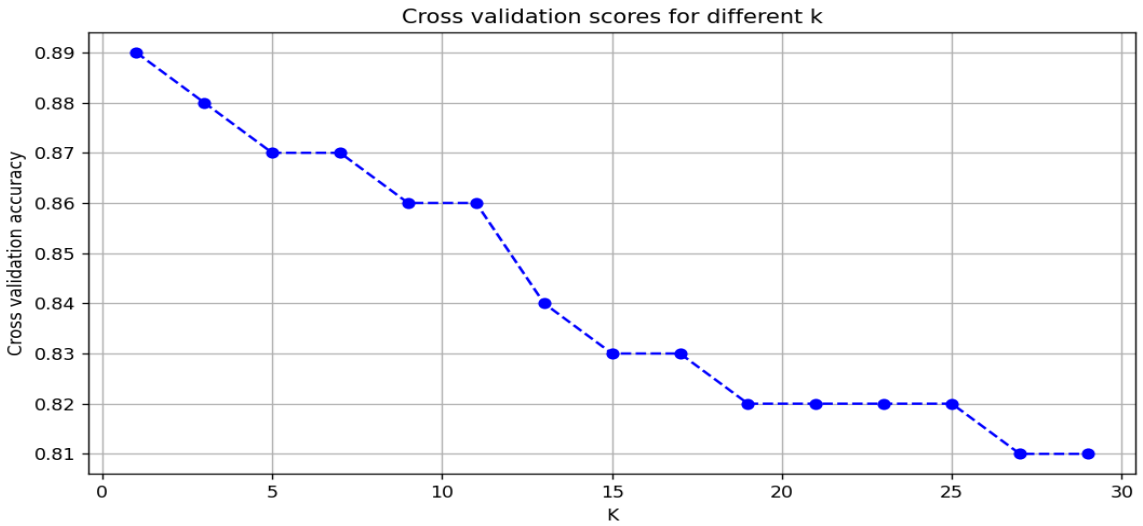
Splitting data:

First, we isolated 80% of the dataset data as train data and the remaining 20% as test data

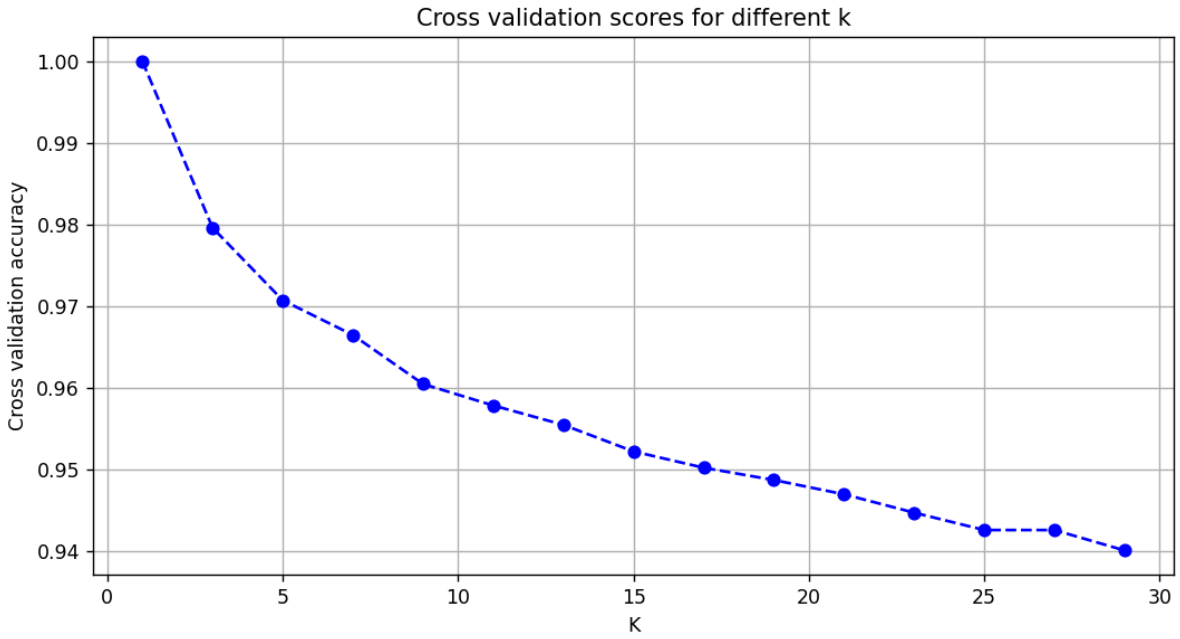
The accuracies for k 1 to 10 for 1000 data:

```
Accuracy for k=1: 0.89
Accuracy for k=2: 0.88
Accuracy for k=3: 0.88
Accuracy for k=4: 0.88
Accuracy for k=5: 0.87
Accuracy for k=6: 0.87
Accuracy for k=7: 0.87
Accuracy for k=8: 0.86
Accuracy for k=9: 0.86
Accuracy for k=10: 0.86
```

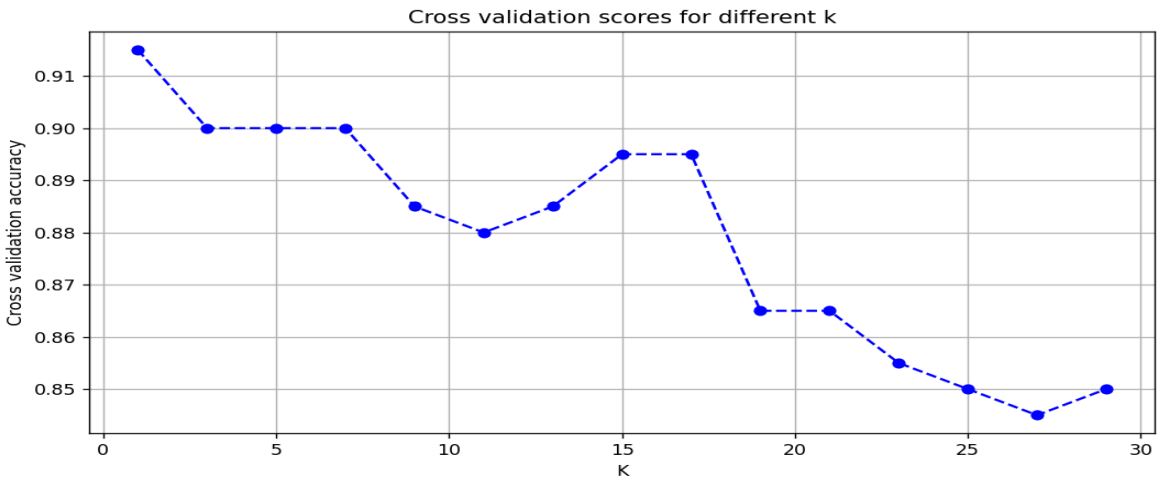
Cross validation using 1000 of data and cosine metric



Cross validation using 1000 of data and Kfold

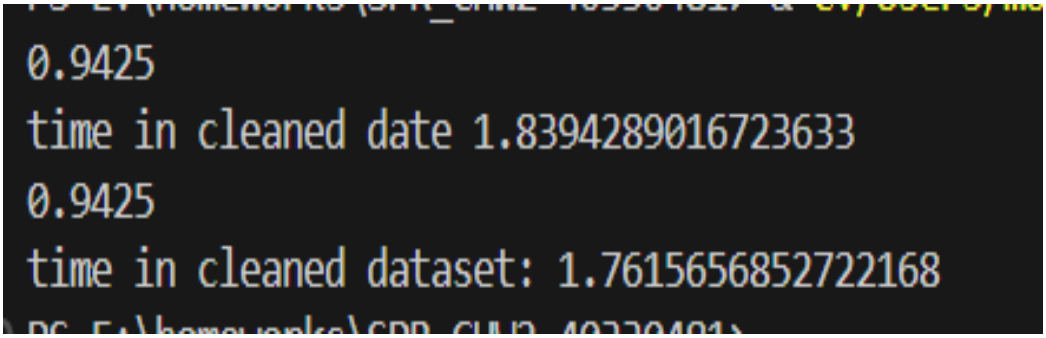


Considering the optimal K, the data that we initially put in the test category can be used to obtain the final accuracy of the model



Finally, to improve the model's temporal performance, we can use the LeaveOneOut method to remove the incorrectly categorized Train data from our datasets. And then we will regroup our test data and measure the accuracy of the model.

Using this method, the mean time for classifying new samples is significantly reduced, but this method reduces some of the accuracy of the algorithm



After running each algorithm 20 times

Time in cleaned dataset for iteration 20: 1.0440 seconds

results for each iteration:

[0.9457759857177734, 0.926263323669434, 0.906329393868408, 1.012624979019165, 0.941125869750976 6, 0.922057151794433 6, 0.938448667526245 1, 0.93062639236450 2, 0.980196714401245 1, 0.927625179290771 5, 0.945887327194213 9, 0.979690074920654 3, 0.900192260742187 5, 0.915221691131591 8, 0.906220674514770 5, 0.966530323028564 5, 1.116099596023559 6, 1.095704555511474 6, 1.028363466262817 4, 1.043999433517456]

Results for P_value:

P-value: 0.0002