CS440 MP3 Report

Andong Jing ajing2 - 3 credit In charge of Part 1.1/1.2/1.3

Siping Meng smeng10 - 3 credit In charge of Part 2.1

Siyu Tao siyutao2 - 3 credit In charge of Part 2.2

# 1. Naive Bayes Classifiers on Digit classification

## 1.1 Single pixels as features

## 1.2 Pixel groups as features

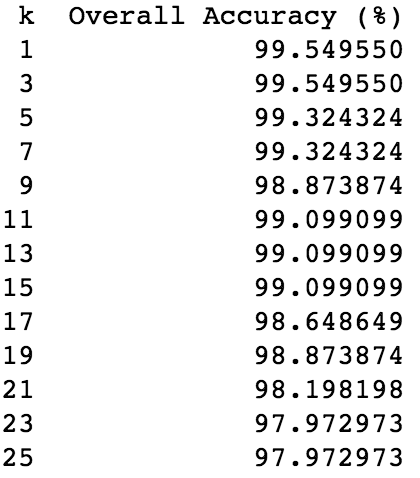
## 1.3 Face classification (Extra Credit)

# 2. Alternative models for Digit classification

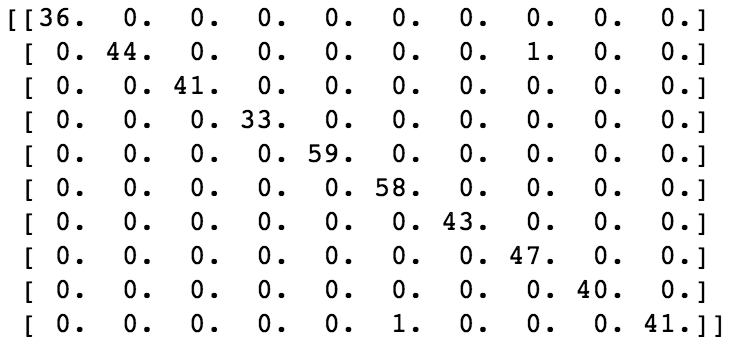
## 2.1 Digit classification with perceptrons

## 2.2 Digit classification with nearest neighbor

The team’s choice of similarity function is by evaluating the Euclidean distance between any two vectors. It yields the same output as the Manhattan distance, but is more efficient and thus quick to compute. Based on the team’s research, Euclidean distance is also the mostly widely used similarity function in the implementation of nearest neighbor classification. The overall accuracy decreases as n increases, as shown by the following figure. Based on the team’s speculation, a large k introduces bias, which may explain the relatively poor performance of large k’s. However, small k’s are prone to outliers in the training datasets. Luckily, the good performance of the model when k is small indicates the overall quality of the dataset is very good.



Since both k=1 and k=3 yields the highest accuracies and a model with k=1 is more efficient. The team constructs the confusion matrix with k=1. It is shown below.



To account for the fluctuation in the running time of individual queries, the team takes the total running time of all 444 pieces of data in the test sets and divide it by 444. The running time for a single query is about 196 μs. It is very efficient, since we utilized Euclidean distance as our similarity function. To further improve its performance, we may consider eliminating choices of potential labels if the distance between the query in the test dataset and several points in the training set far exceeds the distance with the current nearest neighbor. This process of elimination saves running costs from evaluating labels of tiny possibilities.

Finally, compare your nearest-neighbor accuracy to the accuracies you got with Naive Bayes and Perceptron.

## 2.3 Extra Credit

### 2.3.1 Visualization

### 2.3.2 Differentiable perceptron

### 2.3.3 Other learning algorithms

The team referenced the svm classifier (by scikit-learn) tutorial which could be found at the following URL.

[http://scikit-learn.org/stable/auto\_examples/classification/plot\_digits\_classification.html](# http://scikit-learn.org/stable/auto_examples/classification/plot_digits_classification.html)

The team used it to classify the digits supplied by this assignment. The svm model with default parameters yields an accuracy of less than 97% accuracy on the test dataset. However, after experimenting with different gamma values, the team found that a gamma=0.01 yields only 3 errors out of the 444 data, which is an accuracy of 99.32%. The confusion matrix is shown in the following figure.

