DATA 604, Winter, 2021, Project 1

Due on January 12th, 2020, at 6PM.

1) Download the dataset of handwritten digits collected by USPS.

2) (15 pts) Study the role of the split of the training vs testing data for classification. Specifically:

- Divide each class into two sets: the training set consisting of N examples for each of the digits 0 through 9, and the testing set consisting of $1100 - N$ examples of each of the digits 0 through 9, where N ranges from 100 to 1000. Propose and describe a selection algorithm for choosing N out of 1100 images for any integer value of N.
- Design and implement a method to test and estimate your computer's capability to perform numerical computations. Estimate the computational cost of this Project and use it to decide on a reasonable number of experiments you can perform. Quantify this information and use it to guide your personal numerical goals, such as the number of different values of N that you will test.
- Use the k nearest neighbors classification scheme in the standard Euclidean metric with fixed $k = 20$ to verify the global success rate of your classifications for each chosen value of N.
- Draw conclusions about the impact of the size of the training set on the performance of the classification scheme. To do this provide a method for choosing an optimal size of the training set. Describe what is the notion of optimality that you choose. Substantiate your conclusions with numerical evidence.

3) (15 pts) Study the role of the structure of the split of the training vs testing data for classification. Specifically:

- Propose a new method of selecting the training set, which is different from the one proposed in Part 2. Describe the new selection algorithm.

- Divide each digit class (using this new split method) into two sets: the training set consisting of N examples for each of the digits 0 through 9, and the testing set consisting of $1100 - N$ examples of each of the digits 0 through 9, where N is the optimal value chosen in Part 2.
- Use the same k nearest neighbors classification scheme in the standard Euclidean metric with fixed $k = 20$ to verify the success rate of your classification for the chosen optimal value of N with the new split of training vs testing data.
- Draw conclusions about the impact of the structure of the training vs testing split based on comparison of results of Part 2 and Part 3.

4) (10 pts) Choose the best training set selection method that you proposed with the optimal value of N from Part 2, and analyze the role of the metric in the classification process. For this purpose compare the distance induced by the $\| \cdot \|_1$ norm, with the Euclidean distance. Draw conclusions.

5) (10 pts) Revisit the best classification results from previous parts, and this time analyze the success percentages for each compressed digit separately, as well as globally. Analyze the performance differences between individual digits in each case. Draw conclusions.