Data Exploration

Rett Kinsey and Emre Guvenilir

What We Know About MNIST

- 28x28 Pixel Images where each image is classified as a digit 0-9
- 60,000 training set data points with 10,000 test set data points
- We can display the images utilizing pyplot and cmap in Jupyter
- Utilizing grayscale values

Features

- Features depends on which model type we choose to proceed with
- Individual grayscale value of each pixel (whether or not a pixel is being used)
  - This can be split into a binary value based on a cutoff or left as is.
- Relationship between those pixels
  - There are algorithms that can detect lines and curves.
    - But to use these algorithms would both be extremely taxing
    - And would likely drift out of machine learning
- Could also do some type of total brightness values
- Also possible to use reference figures to establish a base similarity.
  - Select one 2 and compare everything to that 2 to get 2 likeness.

Model Types

- K-Nearest-Neighbors
  - Feature: Grayscale value of individual pixels

Pros	Cons
❖ Easier to code ❖ Can compare to wider number of images compared to next idea	❖ Large processing time and incredibly large dataset to carry around ❖ Processing done at time of prediction making

Model Types

- Decision Tree
  - Calculating similarity to a singular example of each digit
  - The similarity would be the features of the decision tree

Pros	Cons
❖ Less processing time at time of prediction ❖ Can expand comparison to be the average across many examples ❖ Easily explainable	❖ Bias by initial choice of examples ❖ Processing done at time of prediction making

Data Processing

- We will want to be able to locate a specific pixel (i.e. [0][0],[10][10])
- For Knn we will likely want to utilize distance metrics to identify patterns over weighted voting
- If we utilize decision trees, we were likely utilize pruning in our processes
- Parallelization is something that could be used as the dataset is extremely large
  - Dividing a computational task into smaller subtasks to be executed simultaneously
  - We could leverage this to find patterns simultaneously and have a “bank of patterns” to match against

Ethical Considerations

- Data Privacy
  - These are handwritten digits, and while they are not identifiable to a person, there could be security concerns if used to identify a specific individual's handwriting
- Transparency
  - The model should be well-documented to build trust in how the model was built to ensure the interpretation of output is trusted
- Deployment Considerations
  - If such a model were to be applied for tasks such as automated grading on math assignments, many concerns appear in the accuracy and generalizability of the model