**Problem One**

The code used for this problem can be found in understanding\_data.py. It can be run using the following command:

*python understanding\_data.py*

**A.** I examined 50 images of the digit 8. I chose this digit because it is a relatively intricate number with two stacked loops, and can be misinterpreted as the digit 3, if written incompletely, or the digit 0 if written too thickly. With this intuition, I examined the set of images, and identified the following pictures to be potential difficult cases.



The above three figures demonstrate a key challenge with classifying the digit 8. When written messily, often times either one or two of the loops is not fully connected. Because the fundamental characteristic of the letter 8 is two closed loops, the digit becomes a suddenly very complex image of curling lines. One could easily misinterpret an unclosed digit 8 with a messy digit 3, which likely has a complementary issue of accidentally closing loops.



The above three figures demonstrate another key challenge when classifying the digit 8. When written extremely thickly, the loops become hard to identify. In turn, the classifier is likely to struggle when discerning a thick blob, and may assume the bottom loop is a thick line (far left), or perhaps a messy 0. With that, it is also challenging when handling finely written 8s (far right). If one loop is larger and significantly clearer, the classifier may only capture it and the thin stroke would make it increasingly difficult for the classifier to even recognize the smaller, messier loop, leading to misclassifications.

**B.**

The total number of images for digit 0: 5923

The total number of images for digit 1: 6742

The total number of images for digit 2: 5958

The total number of images for digit 3: 6131

The total number of images for digit 4: 5842

The total number of images for digit 5: 5421

The total number of images for digit 6: 5918

The total number of images for digit 7: 6265

The total number of images for digit 8: 5851

The total number of images for digit 9: 5949

The total number of images: 60000

I made the training set such that it consists of ¾ of each digit’s images, enabling the other ¼ of each digit’s images to be put in the testing set. While within each digit image set the partition is relatively random, I wanted to ensure that each digit was properly represented in both the training and testing set. As such, the classifier should be trained for all digits within the testing set, in order to appropriately assess its ability in classifying unseen data within the same realm. Moreover, I chose this distribution such that the model is exposed to a substantial number of images for each digit during training. As shown above, there is a large variation in handwriting for each given digit. As such, it is important the classifier sees a breadth of data during classification, so it can further refine its interpretation of each digit and understand the disparities.

**Problem Two**

**A.** K Nearest Neighbors is a classifier that depends on pattern recognition within a given vector space. It takes an input, and then finds k data points with feature vectors that are “nearest” to the input vector. In order to determine which data points are nearest, the system uses a distance measure to determine how closely related two vectors are, such as Euclidean distance. With this, the classifier looks at the k nearest neighbors’ labels, and ascertains the input’s label based on the majority of these neighbors’ labels. In turn, the input label is assigned basically on what the most common label is amongst its k nearest neighbors, utilizing clustering at the core of the classifier.

Support Vector Machines (SVMs) are a supervised classifier that works by mapping data to an N (or more)-dimensional space (where N is the number of features). The Support Vector Machine algorithm selects a division of the data that best divides the classes. Specifically, it finds the hyperplane that maximizes the margin to the nearest data points in each class, dubbed the “support vectors.” If the data is not linearly separable in a N-dimensional space, a kernel trick may be used to map the data to a higher dimension that it is linearly separable in. SVM can either be done with a hard-margin algorithm which requires the data to be fully linearly separable, or a soft margin algorithm that is more robust to classification error and outliers. In the soft-margin variety, misclassifications are allowed at a cost proportional to the value of the slack parameter.

To classify images, both the training set and testing set of gray scale images must be encoded such that they are a 2D array of pixel intensity values. These 2D arrays of pixel intensity values serve as each image’s feature vector, where each pixel is representative of a feature with a corresponding value.

In turn, KNN can utilize its pattern recognition algorithm to determine which sets of pixel values are most similar to the inputted set of pixel values. By leveraging Euclidean distance, KNN will determine k images that are nearest to the inputted image’s pixel intensity values. As such, taking the majority of the nearest neighbors’ labels will classify the inputted image, and essentially cast it as that majority label.

SVM, on the other hand, can find a hyperplane that linearly separates the observations possibly using a kernel trick to make the data linearly separable.

**B.** In K Nearest Neighbors, changing the hyperparamter, k, changes how many neighbors are considered during classification. Meaning, when a new data point is inputted, it finds k feature vectors that are the most similar. As such, increasing k takes more data points into account when labeling the input. Having a lower k makes the classification algorithm more susceptible to noise because fewer data points influence the label. With that, increasing k is computationally expensive, and may begin to defeat the purpose of clustering, as the number gets too high.

In the soft-margin variety of SVM, misclassifications are allowed at a cost proportional to the value of the slack parameter. It allows the optimization problem to allow misclassifications so long as the cost is below some threshold. The larger this slack parameter the more misclassification will be permitted. The lower the threshold the closer to a “hard” margin SVM it acts.

**Problem 3 + 4**

**A.** See Code.

**B.**

**KNN Training Set Size**

|  |  |
| --- | --- |
| **Testing Size** | **F1-measure** |
| **6000** | 0.93567 |
| **9602** | 0.94391 |
| **10801** | 0.94042 |
| **11400** | 0.94120 |
| **11762** | 0.96659 |

**KNN Parameters**

|  |  |
| --- | --- |
|  | **F1-measure** |
| **K=[?]** |  |
| **K=[?]** |  |

**SVM Parameters**

|  |  |
| --- | --- |
|  | **F1-measure** |
| **Slack=[?]** |  |
| **Slack=[?]** |  |

**KNN Confusion Matrix**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Predictions** | | | | | | | | | |
| **Actual** |  | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| **1** |  |  |  |  |  |  |  |  |  |
| **2** |  |  |  |  |  |  |  |  |  |
| **3** |  |  |  |  |  |  |  |  |  |
| **4** |  |  |  |  |  |  |  |  |  |
| **5** |  |  |  |  |  |  |  |  |  |
| **6** |  |  |  |  |  |  |  |  |  |
| **7** |  |  |  |  |  |  |  |  |  |
| **8** |  |  |  |  |  |  |  |  |  |
| **9** |  |  |  |  |  |  |  |  |  |

**SVM Confusion Matrix**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Predictions** | | | | | | | | | |
| **Actual** |  | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| **1** |  |  |  |  |  |  |  |  |  |
| **2** |  |  |  |  |  |  |  |  |  |
| **3** |  |  |  |  |  |  |  |  |  |
| **4** |  |  |  |  |  |  |  |  |  |
| **5** |  |  |  |  |  |  |  |  |  |
| **6** |  |  |  |  |  |  |  |  |  |
| **7** |  |  |  |  |  |  |  |  |  |
| **8** |  |  |  |  |  |  |  |  |  |
| **9** |  |  |  |  |  |  |  |  |  |

**Problem 5.**

[Insert images of misclassifications]

**Problem 6.**

**Problem 7.**