**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

When creating the training and testing sets during our experiments, we ensured that 1/10th of each set each came from each digit. This ensures that each digit is properly represented in both the training and testing set. Through this methodology, the classifier is trained for all digits within the testing set, so that our evaluations test how it classifies unseen data within the same realm. We split data like this such that a case never arises where the classifier has a training set of all 3’s, and is tested with the number 9, or something similar.

Moreover, we 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. A kernel is also used to transform the data to be more linearly separable. Which kernel works best is an empirical question that depends on the specifics of the data set that is being used.

**Problem 3**

**A.** The code used for this problem can be found in classifier\_1.py. It can be run using the following command:

*python classifier\_1.py*

**B.**

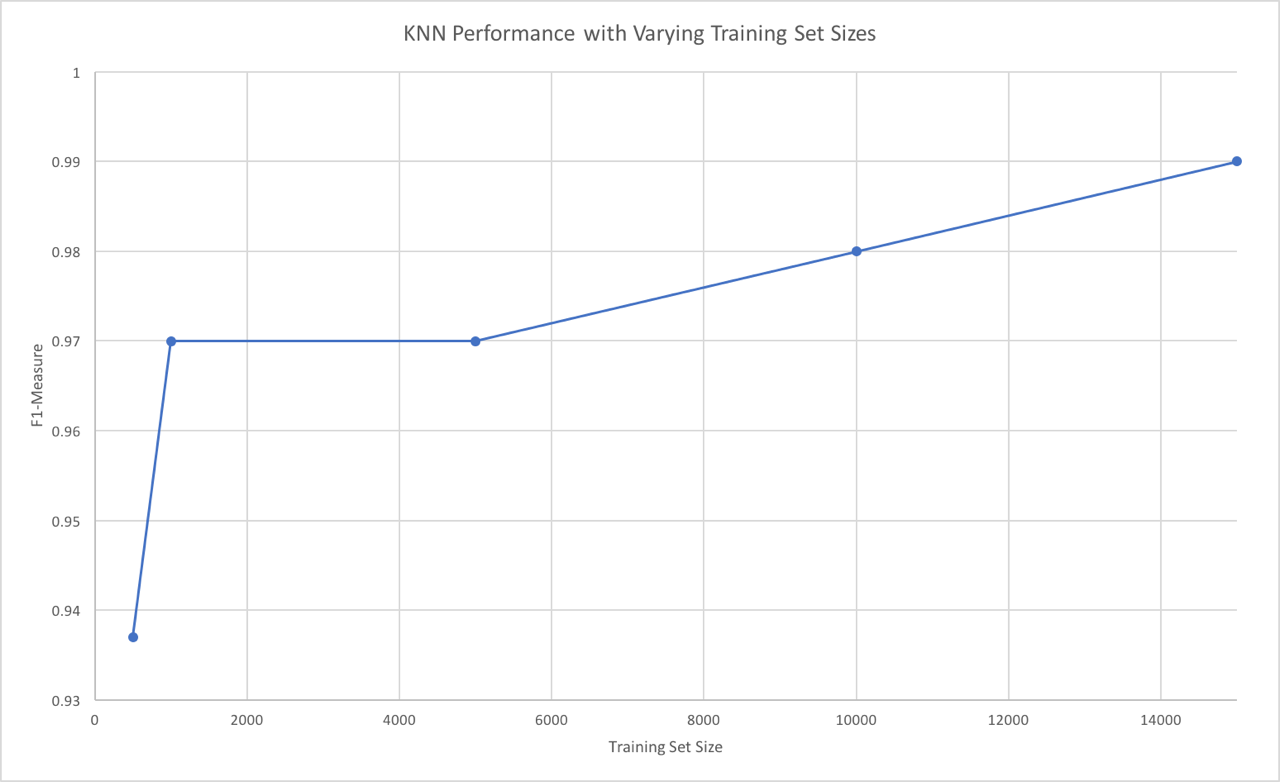
In order to determine the best hyperparamters for KNN, we conducted two experiments. First, we looked to determine which training set size produced the best performance on the testing data. As such, we developed 5 different classifiers, using varying training set sizes: 500, 1000, 5000, 10000, and 15000. Each classifier was tested on the same testing data, so that the only variation was the training set size. In both the testing and training data sets, there is an average distribution of digits, where each digit is represented by 1/10th of the data. With these classifiers, we found the F1-measure, and printed out the corresponding confusion matrix. Examining the F1-measures, there is a clear trend of performance increasing with training data size.

Our second experiment, which looked to determine the optimal number of neighbors, used a training set size of 15,000 and a testing set size of 200. In both the testing and training data sets, there is an average distribution of digits, where each digit is represented by 1/10th of the data. We trained 5 different classifiers, each utilizing a different number of neighbors within the KNN classifier: 1, 5, 10, 20, and 50. The training data and testing data were kept constant for each classifier, such that the only difference was the number of neighbors. With these classifiers, we found the F1-measure, and printed out the corresponding confusion matrix. These results are shown below.

We chose F1-measure as a means to determine performance of a classifier because it provides a holistic measure of accuracy. F1-measure takes both precision and recall into account, and in turn is often used to gauge classification performance.

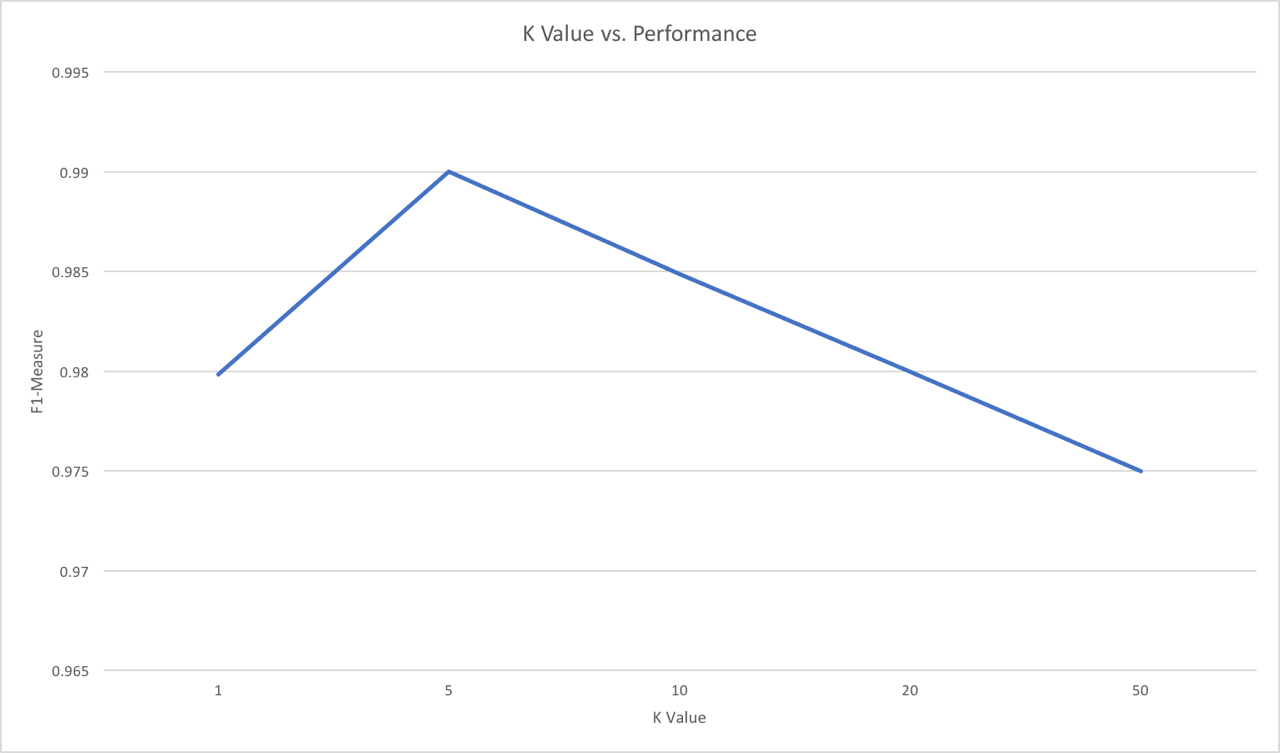
**1) KNN Training Set Size vs. Performance on Test Data**

|  |  |
| --- | --- |
| **Testing Size** | **F1-measure** |
| 500 | 0.937035376904 |
| 1000 | 0.969971810389 |
| 5000 | 0.969968683435 |
| 10000 | |  | | --- | |  | | 0.979968683435 | |  | |
| 15000 | 0.989993746091 |

****

**2) KNN Parameters**

|  |  |
| --- | --- |
| **Number of Neighbors (K)** | **F1-measure** |
| 1 | 0.97985566637 |
| 5 | 0.989993746091 |
| 10 | 0.98485566637 |
| 20 | 0.979968683435 |
| 50 | 0.974968683435 |



As shown in both the results table and graph above, k=5 produced the best performing classifier. Analyzing the numbers, it appears that increasing k improves performance to a point, but then performance dramatically dips once it reaches a certain threshold. We believe this is due to the inherent nature of clustering. Having too small of a k is suscpetible to noise, resulting in lower performance. That being said, once k is too large, the algorithm takes data into account that is beyond the immediate cluster when classifiying a given inputted data point. In turn, classification performance goes down significantly because data points beyond scope have an influence during classification. This behavior is clearly demonstrated in the results above.

**3)** Below is a confusion matrix for the best performing KNN classifier, with an F1-measure of 0.989993746091.

**KNN Confusion Matrix (k=5, training set = 15000)**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Predicted** | | | | | | | | | | | |
| **Actuals** |  | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| **0** | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **1** | 0 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2** | 0 | 0 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **3** | 0 | 0 | 0 | 20 | 0 | 0 | 0 | 0 | 0 | 0 |
| **4** | 0 | 0 | 0 | 0 | 19 | 0 | 0 | 1 | 0 | 0 |
| **5** | 0 | 0 | 0 | 1 | 0 | 19 | 0 | 0 | 0 | 0 |
| **6** | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 | 0 |
| **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 |
| **8** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 |
| **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 |

**Problem 4.**

**A.** The code used for this problem can be found in classifier\_2.py. It can be run using the following command:

*python classifier\_2.py*

**B.** In order to determine the best hyperparameters for SVM, we conducted two experiments. The first experiment that looked to determine training set size is the same as the first experiment for KNN. The same description can be found below.

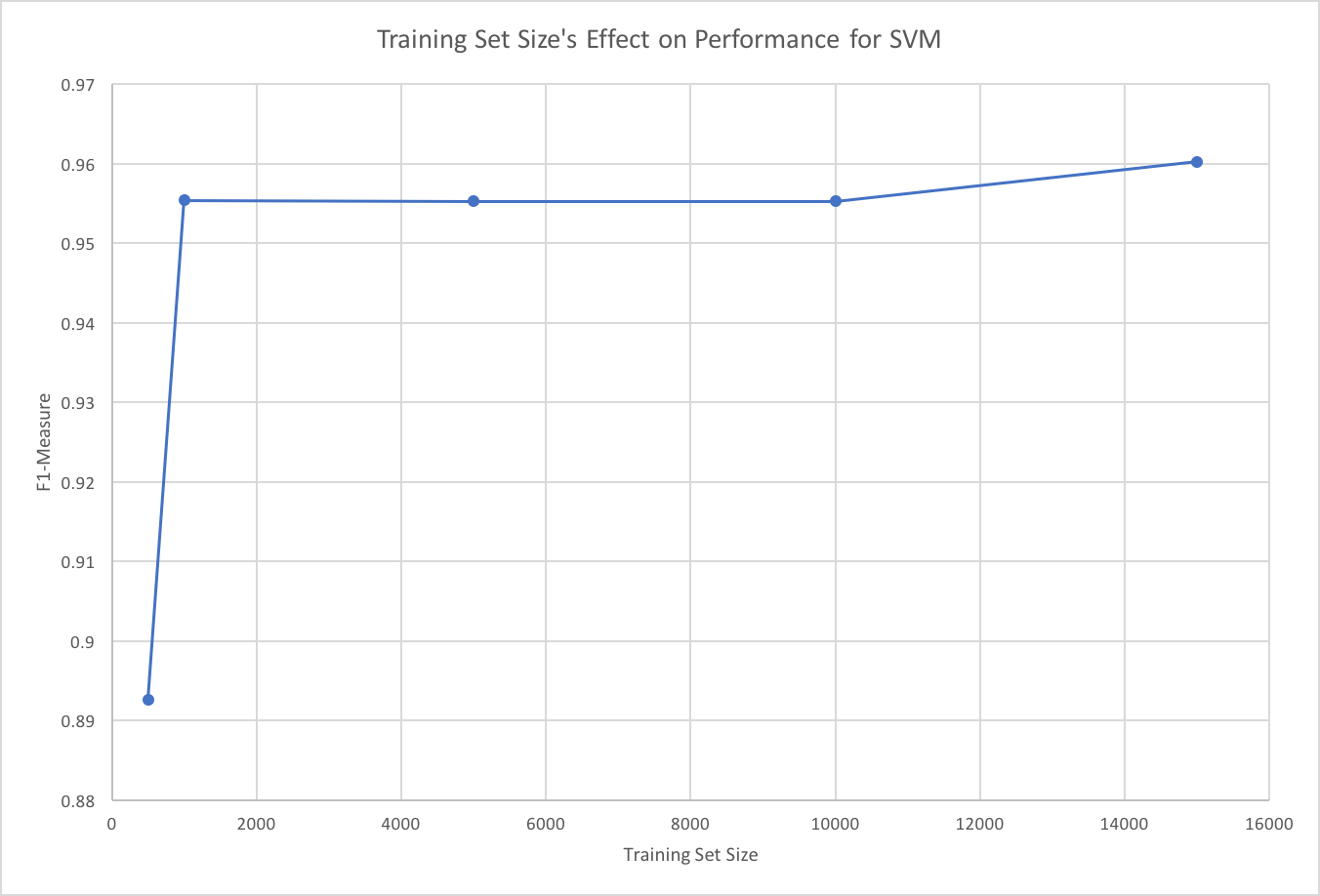
First, we looked to determine which training set size produced the best performance on the testing data. As such, we developed 5 different classifiers, using varying training set sizes: 500, 1000, 5000, 10000, and 15000. Each classifier was tested on the same testing data, so that the only variation was the training set size. In both the testing and training data sets, there is an average distribution of digits, where each digit is represented by 1/10th of the data. With these classifiers, we found the F1-measure, and printed out the corresponding confusion matrix. Examining the F1-measures, there is a clear trend of performance increasing with training data size.

Our second experiment, which looked to determine the optimal C value (the cost of misclassification) and the best kernel, used a training set size of 15,000 and a testing set size of 200. In both the testing and training data sets, there is an average distribution of digits, where each digit is represented by 1/10th of the data. We trained 16 different classifiers, with a grid search over the parameter C of {1, 5, 10, 100} and the kernels of {“linear”, “poly”, “rbf”, and “sigmoid”}. The training data and testing data were kept constant for each classifier, such that the only difference was the combination of C and kernel. With these classifiers, we found the F1-measure. These results are shown below.

We chose F1-measure as a means to determine performance of a classifier because it provides a holistic measure of accuracy. F1-measure takes both precision and recall into account, and in turn is often used to gauge classification performance.

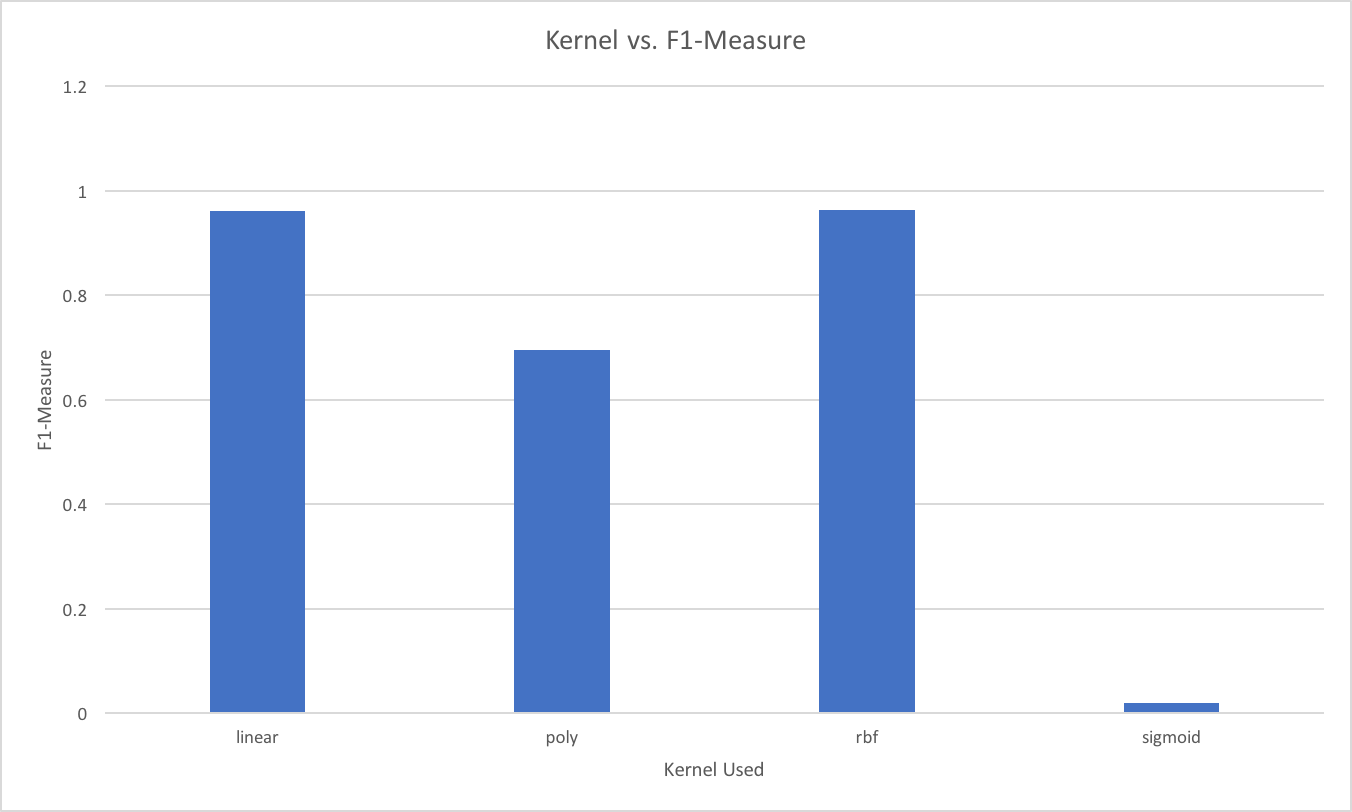
**1) SVM Training Set Size vs. Performance on Test Data**

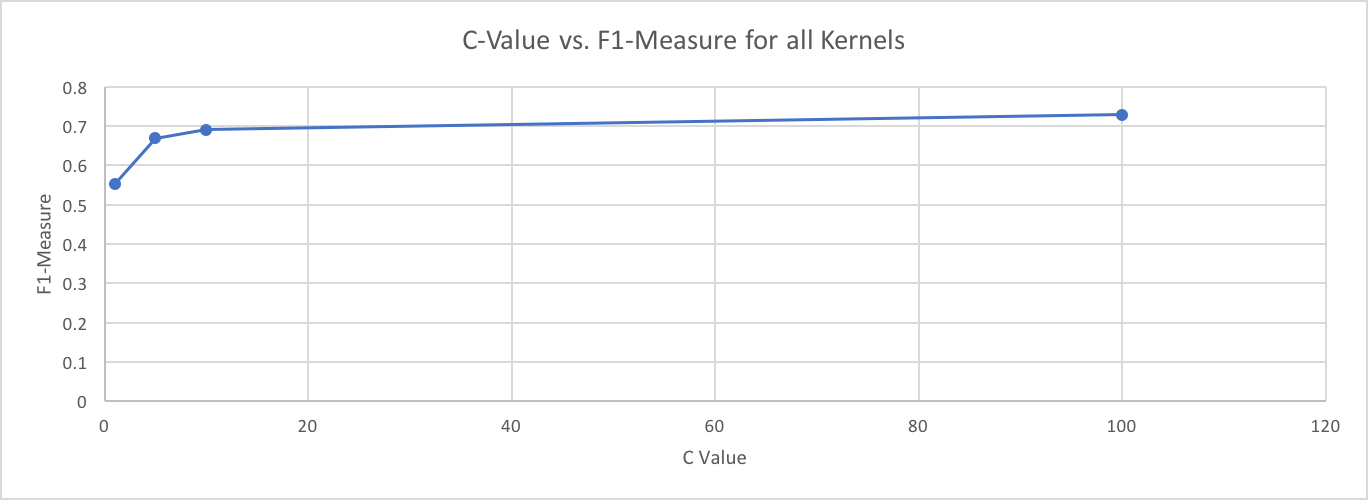
|  |  |
| --- | --- |
| **Testing Size** | **F1-measure** |
| 500 | 0.892654623674 |
| 1000 | 0.955399647074 |
| 5000 | 0.955280822808 |
| 10000 | |  | | --- | |  | | 0.955280822808 | |  | |
| 15000 | 0.960280822808 |

****

**2) SVM Parameters**

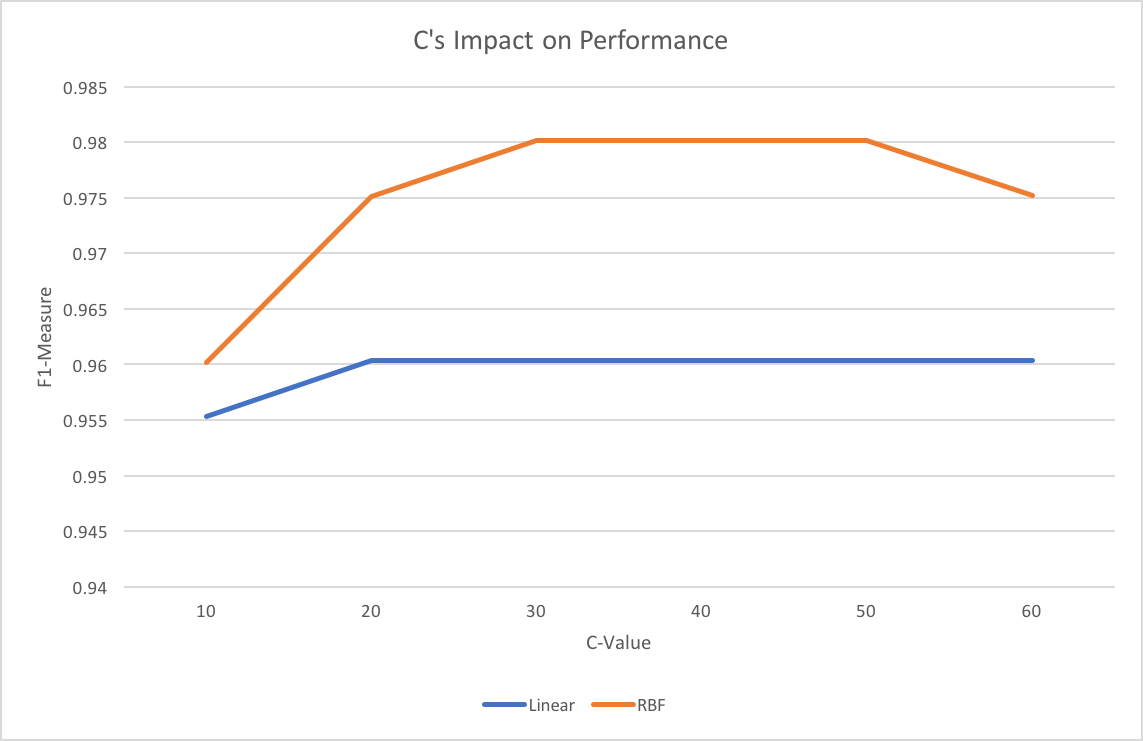
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Kernel** | **C Value** | | | |  |
| **1** | **5** | **10** | **100** | **Averages** |
| **Linear** | 0.965216653 | 0.964984365 | 0.955352578 | 0.960311693 | 0.961466322 |
| **Polynomial** | 0.263257587 | 0.731320374 | 0.826653224 | 0.960405901 | 0.695409271 |
| **RBF** | 0.960280823 | 0.960167806 | 0.960167806 | 0.975225587 | 0.963960505 |
| **Sigmoid** | 0.018181818 | 0.018181818 | 0.018181818 | 0.018181818 | 0.018181818 |
| **Averages** | 0.55173422 | 0.668663591 | 0.690088856 | 0.72853125 |  |

****

****

For kernels, linear and rbf performed comparably, with rdf performing slightly better. Sigmoid performed horribly. Our results also showed that a higher cost to misclassification yields a better classifier across all kernels tested. Combined, the best classifier used rbf with a cost of 100.

However, because C=100 did only marginally better than C=10, and because there was little performance improvement between linear and rbf kernels, we chose to run a second experiment between comparing linear and rbf kernels across C-values of {10, 20, 30, 40, 50, 60}. The intuition to further test the space between 10 and 100 turned out to be correct—a C-value of 30-50 is the correct optimum (with identical performance). We select 40 as the average of the 3.



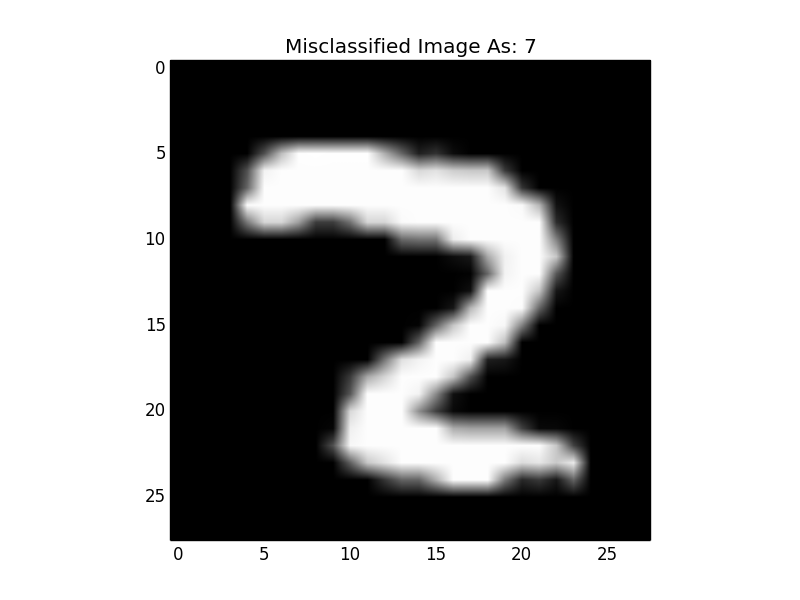
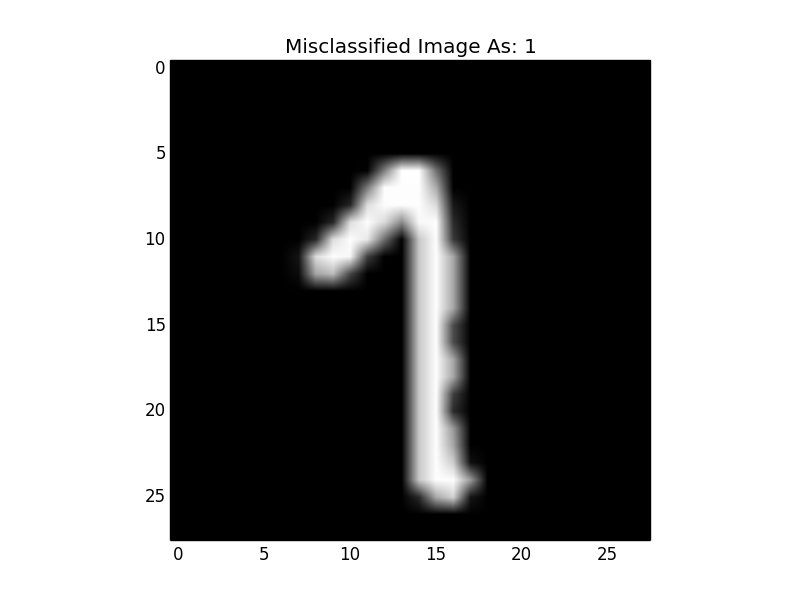
**3)** Below is a confusion matrix for the best performing SVM classifier, with an F1-measure of 0.980115697311.

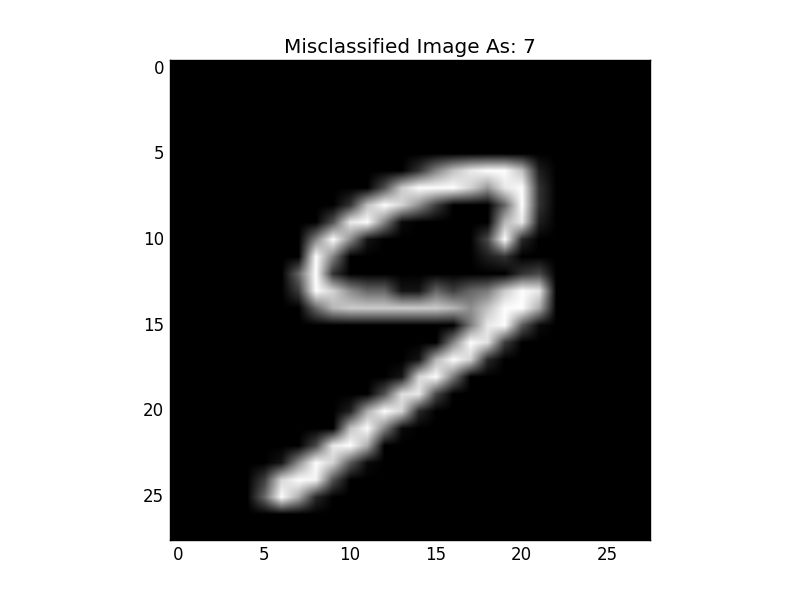
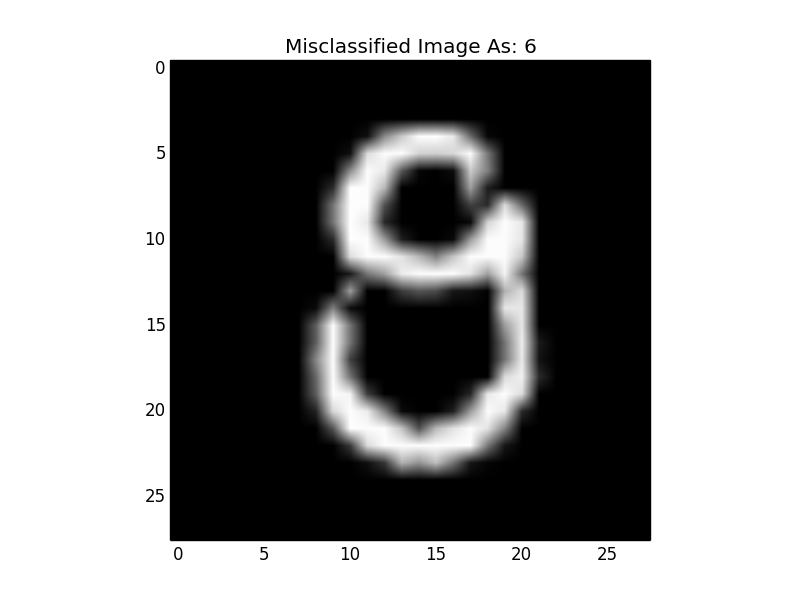
**SVM Confusion Matrix (C=40, training set = 15000, kernel=rbf)**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Predicted** | | | | | | | | | | | |
| **Actuals** |  | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| **0** | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **1** | 0 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2** | 0 | 0 | 19 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **3** | 0 | 0 | 0 | 19 | 0 | 0 | 1 | 0 | 0 | 0 |
| **4** | 0 | 0 | 0 | 0 | 19 | 1 | 0 | 0 | 0 | 0 |
| **5** | 0 | 0 | 0 | 0 | 1 | 19 | 0 | 0 | 0 | 0 |
| **6** | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 | 0 |
| **7** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 |
| **8** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 |
| **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 |

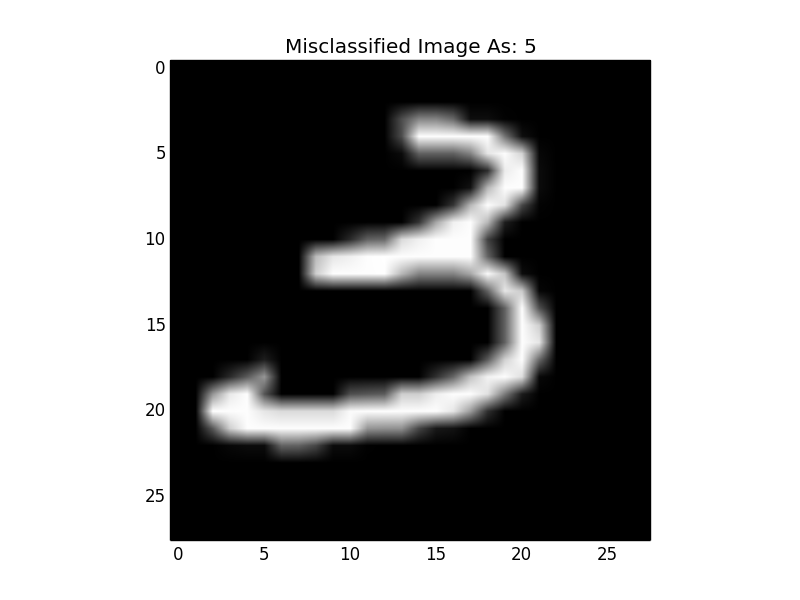
**Problem 5.**

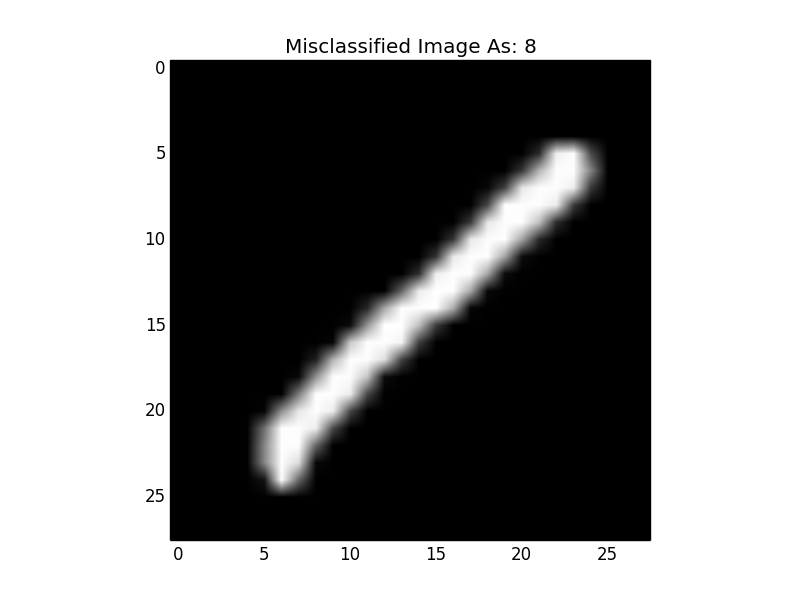
Misclassified images for KNN

****



Misclassified images for SVM



****

**Problem 6**

Our SVM classifier with the optimal hyper-parameters had an F1-measure of 0.9801. The KNN classifier with the optimal hyper-parameters had an F1-measure of 0.9810. The KNN classifier performed better than the SVM by 0.0918%. This is hardly a practical improvement—the classifiers were similar in performance. However, what little performance improvement the KNN classifier had was likely since there was a very high probability of any given handwritten digit being similar to several other handwritten digits and a somewhat low chance of a “8” for instance would be similar to a “1” from the training set. This type of problem is ideal for a KNN classifier.

**Problem 7**