MP4 Report

Rafal Stapinski

November 28 2018

1 a.2

Given the large number of features, using the multiplicative representation of joint probability would yield a massive amount of results, taking up an exponentially increasing amount of memory/time to compute and access

2 a.5

10% yields accuracy of 0.730
20% yields accuracy of 0.789
30% yields accuracy of 0.801
40% yields accuracy of 0.804
50% yields accuracy of 0.810
60% yields accuracy of 0.810
70% yields accuracy of 0.811
80% yields accuracy of 0.824
90% yields accuracy of 0.823
100% yields accuracy of 0.824

The basic features see a great improvement until up to around 80% of the training data, at which point they most likely see some overfitting to the training data and accuracy does not significantly improve.

3 a.6

After a few tests, starting with k = 1, it appeared that the smallest shift (likely due to the nature of the numbers being smoothed being so small) a very small k was selected of 0.0000001 due to it not shifting the data too much and still being able to approximate a value of 0 fairly reasonably. After testing with many values, this had consistently given the highest accuracy for any slice of the testing data.

4 b.1

10% yields accuracy of 0.744
20% yields accuracy of 0.784
30% yields accuracy of 0.810
40% yields accuracy of 0.816
50% yields accuracy of 0.821
60% yields accuracy of 0.824
70% yields accuracy of 0.829
80% yields accuracy of 0.832
90% yields accuracy of 0.828
100% yields accuracy of 0.832

The first new feature set is derived by convolution a gaussian kernel over the image of the character in order to extract some more abstract (larger scale) features from it. Similarly to how this would be used in a CNN to identify images, the intent was to create a new image of higher class features in it that could be more easily acted upon by the classifier.

The next feature set is done by comparing which sides (horizontal and vertical) of the character have more value in them. This is done to help separate out different types of characters. For example, an 8 and 0 would be roughly equally heavy on the top/bottom and left/right whereas a 7 might be equal on the left/right but would have more ink on the top than the bottom.

Lastly, a set of features was extracted to determine the amount of blank space, edge space, and inside space taken up by the character. This is done to differentiate numbers like 8 which have a high inside and edge count, from numbers like 1 which have a low inside and edge count, to numbers like 0 which have a higher proportion of edge counts over inside counts.

The advances features outperform the basic features by a few points, almost across the board. It seems as if after reaching approximately 80% of the training data, the classifier runs into some overfitting. This could be helped by running multiple passes of a kernel for different layers resulting in higher order features, however for the sake of performance, these were left out. The kernel used in the submission was chosen as it optimized accuracy with performance.

Calculating the kernel over the image takes some time. With 100% of the data, the training process taken just under 2 minutes on a single core of a 2.5GHz commercial CPU, and prediction on the 1000 data points takes around 20 seconds.

5 b.2

10% yields accuracy of 0.752
20% yields accuracy of 0.792
30% yields accuracy of 0.808
40% yields accuracy of 0.818

50% yields accuracy of 0.827
60% yields accuracy of 0.826
70% yields accuracy of 0.827
80% yields accuracy of 0.831
90% yields accuracy of 0.833
100% yields accuracy of 0.835

The combination of the feature sets outperforms the individual sets as they target different aspects of the characters leading to less overlap in the features, and therefore can reach a higher accuracy much quicker with a lot less training data. In this case, it appears that the combination of the features reaches close to its peak accuracy (which is very close to that of the advanced features alone, and better than the basic features alone) after seeing only around 50% of the test data. The combination of features also seems to be less affected by overfitting even after seeing all of the data, does not reach as significant of a plateau in accuracy.