Kyle Elbaum and Nicholas Lum
Machine Problem 4 Report
November 28, 2018

Part A

2) The major drawback using the standard multiplicative representation of joint probability is the inclusion of all the whitespace in a 28 x 28 image. It would count beyond the edge of a digit in the image.

5) Evaluation of the basic Naïve Bayes Classifier:
10: 0.733
20: 0.772
30: 0.792
40: 0.789
50: 0.8
60: 0.801
70: 0.805
80: 0.815
90: 0.813
100: 0.817

6) We chose k = 1. If choosing k > 1 or k > 1, the percentage decreases. Also, we wanted to avoid receiving a domain error when performing natural log on the probability. The result is reasonably good because the evaluation is on basic implementation of the classifier without extra features in Part B.

Part B

1)

a) Using '#' as the only True variable in a binary feature set.
Tests if using only the interior provides better results
Checks if using less 'data' than basic will yield better results

b) Using '+' as the only True variable in a binary feature set.
Tests if using only the exterior provides better results
Checks if using even less 'data' than the above makes a positive difference

c) Each index contains data about the pixel and the pixel below
- Indexes with no pixel below treat the index as a blank
- Feature can take one of four values
  1. Empty, empty
  2. Empty, filled

3. Filled, empty
4. Filled, filled

- Gives extra power to each statistic index as they contain data for two pixels
- Checks if having more ‘data’ provides more reliable results.

2) Extract advanced features using different percent of testing data:
10: 0.742
20: 0.775
30: 0.799
40: 0.804
50: 0.815
60: 0.816
70: 0.823
80: 0.83
90: 0.829
100: 0.827

Using only the basic feature set predicts with 81.7% accuracy
Using only the advanced feature set predicts with 82.7% accuracy

3) Using all the training data, extract_final_features() predicts with 83.1% accuracy using only the third new feature set.