Rupali Vohra (rv5rr)
CS 4710: Classifier (Homework 4)
Tuesday, April 28, 2015

Summary

The implementation I chose for my classifier involved use of the Naïve Bayes method. This allowed me to easily account for both discrete and continuous variables; a simple proportion was used for discrete features, while the assumption of a Gaussian distribution for continuous variables allowed me to predict outcomes based on means and variances found from the training set.

Approach

Though a bit messy, the best way I could think of to keep track of all the various features, whether it was continuous or discrete, and what the outcome was with that given feature was to have multiple maps that kept track of that information. Since I knew I would not be deleting anything from the map, I decided that HashMaps would be the most efficient way to solve the problem.

I ended up creating 9 HashMaps. One of these (featureOptions) keeps track of each feature, and maps it to an array of all the options available for that feature. This was used to populate other HashMaps that are used to keep track of how many times each of these options appears. Two such HashMaps (frequencyMapGreater and frequencyMapLess) were used to keep track of the number of times each option occurred, depending on whether the outcome was ">50K" or "<=50K," respectively. Similarly, I created two HashMaps (valueMapGreater and valueMapLess) to keep track of the values that appeared for each continuous variable, depending on whether the outcome was ">50K" or "<=50K," respectively. The maps for discrete variables were used to calculate proportions for probabilities, and the maps for continuous variables were used to calculate the means and variances for each feature, which was then used to calculate probabilities. The probabilities calculated from the discrete variables were stored in two HashMaps (featureProbGreater and featureProbLess), which mapped probabilities for which the classification was ">50K" in the former and outcomes where the classification was "<=50K" in the latter. Similarly, probabilities for continuous variables were tracked in two HashMaps: numericValsGreater, which stored data for when the outcome was ">50K," and numericValsLess, which stored data for when the outcome was "<=50K."

All of the HashMaps were populated with data from the training set. The section of the training set I actually used while testing varied (more information given below). This variation did not, obviously, affect my general approach to the problem of making predictions. I did so by following the formula: $Y \leftarrow argmax_{y_k} P(Y = y_k) \prod_i P(X_i | Y = y_k)$. In other words, I calculated the probability that each outcome of each feature in the given test case would occur if the classification were ">50K," and compared that value to the probability that that each feature's outcome would occur if the classification were "<=50K," and selected the larger probability as my classification. In the event that these two probabilities tied, I decided to return an outcome based on the decision of a random number generator.

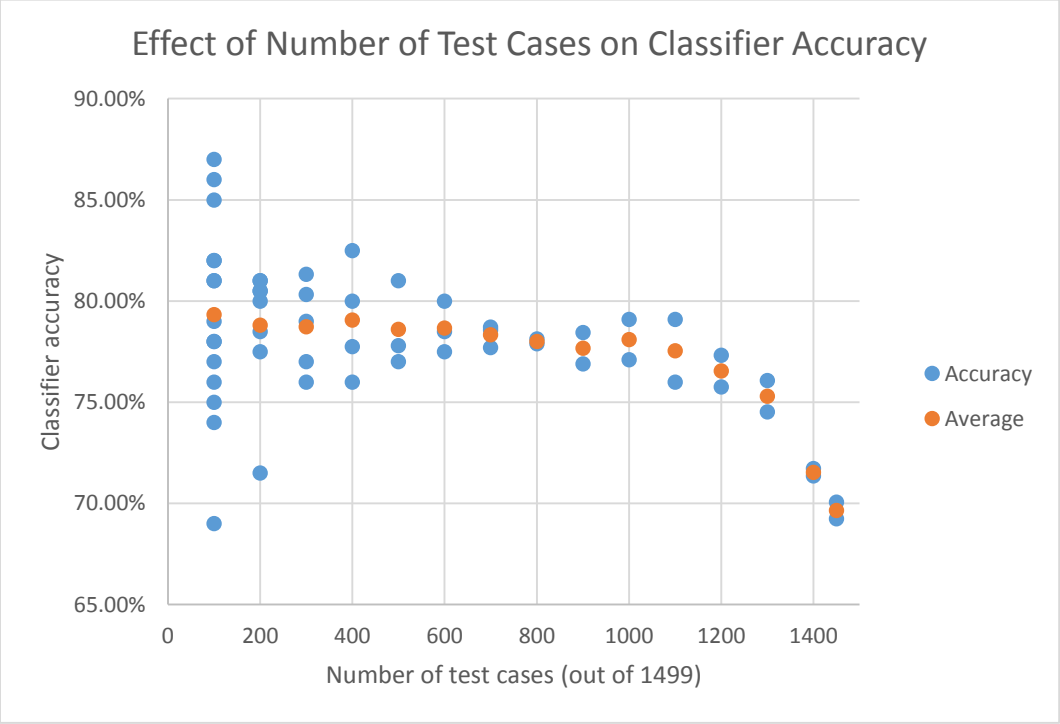
There is a stipulation in the Naïve Bayes approach for classification that mentions the conditional independence for all features in the data set. This is, as stated in the slides, a fairly large assumption, and one that does not hold for the given training set. Specifically, the education_num field is clearly highly correlated with the education_level field, and probably derived from that field. Since I created my Classifier to work without accounting for this at first, I decided it would be interesting to see the difference such an account would make in estimations. The results of this experiment are detailed below.

Testing and Interpretations

In order to speed up the testing process, I created a testing file called TestCaseGenerator.java. This class asks the user to provide how many of the 1499 training set cases the Classifier should use as part of the test cases, and uses the rest as the training set. Then, in order to vary the location of the test data within the training set, the code loops through various positions in the list to create buckets of test cases, and uses the rest of the training set to train the Classifier. This class then creates the test and training files from the original training set (census.train), creates a solution file from what would have been the last column in the test data, runs the classifier's prediction, and then compares the output of the classifier to that of the solution. The code prints out the success rate of the classifier.[1]

As previously stated, I was curious to see how including features that essentially measured the same thing factored into the prediction, so I decided to keep the education-num field for my first few tests. I then adjusted my classifier to ignore education-num in its predictions, and compared the results from that version of the classifier.

The following chart summarizes my results with the classifier that did include education-num in its predictions. There are multiple data points per training set size because I rotated where in the data my test set was located. The various data points depict the variation in results as that test set location changed. If you are curious about which specific sections led to which results, please see the appendix at the end of this write up.

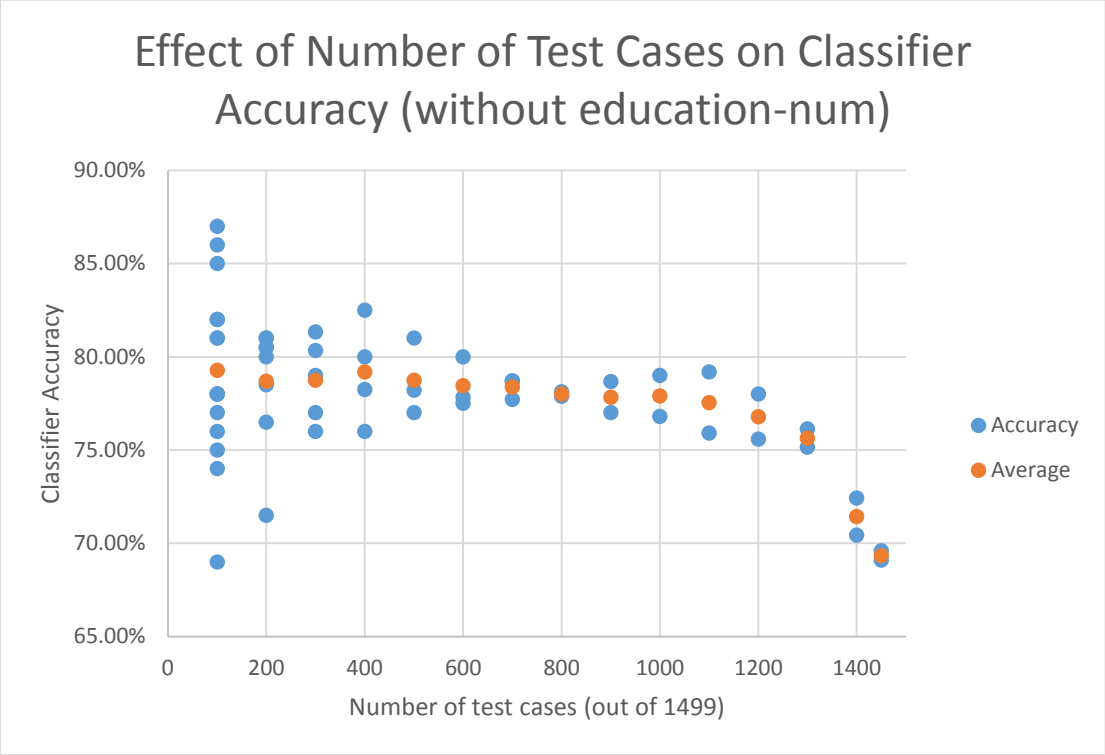


From the graph, we can tell that the number of test cases does not really matter until more than 1000 members of the training set are used as test cases, leaving only 499 cases to train on. An

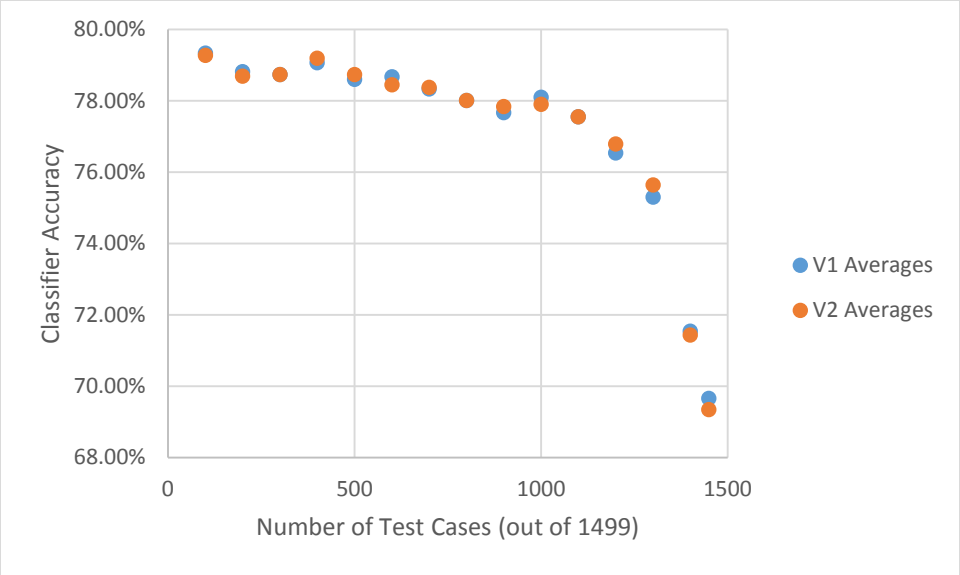
[1] I have included this testing file with my source code, but since the assignment instructions specifically said not to include any other output, the testing file will not work. It relies on the Classifier saving output to a file called "mysol," the code for which has been commented out for submission.

interesting effect is seen especially in the lower test case buckets; there is a lot of variance in the success depending on which bucket the classifier uses for the tests. This makes me wonder if there are certain portions of the training set that are more representative of the general population, and are therefore better suited to predicting outcomes. If this is true, it makes sense that there is less variance in the cases where the test set is large. Because there are only 1499 elements in all, there is a lot of overlap in the two test sets, especially at the greatest sizes of test cases.

I adjusted my classifier to ignore the education-num values in the training set, and ran all of the same tests as before again. The results of these tests are depicted by the graph below. More in depth detail is included in the appendix.



From these two graphs, it is clear that the general trend of the classifier's accuracy is not affected whether education-num is factored in or not. The graph depicting both versions' average accuracies for each bucket size indicate that there is not much of a difference in accuracy by value, either. In this graph, V1 refers to the classifier that includes education-num, while V2 refers to the classifier that ignores education-num.



Appendix

For the following data, the first bolded line describes the test numbers relevant to the given information. This information includes the cardinality of the training set and the cardinality of the test set. The second line describes the average accuracy of the classifier for the given set sizes across all buckets. Finally, each set of two lines describes the information for a given test. The first line of each test describes the section in the training set from which the test set was taken. For example, if the test set was of size 100 and the test section is 0, the first 100 elements of the training set were taken as test cases, and the remaining 1399 elements would be assigned to be the training set. Because 1499 is not evenly divisible by the increments used for testing, the last test section overlaps some data with the test section before it. In other words, the size of the test data is always maintained in any given set of test cases.

Tests 1-2: 49 training set, 1450 test set
Average classifier accuracy: 69.66%
Test 1: test section 0
Classifier accuracy: 69.24%
Test 2: test section 1
Classifier accuracy: 70.07%

Tests 3-4: 99 training set, 1400 test set
Average classifier accuracy: 71.54%
Test 3: test section 0
Classifier accuracy: 71.36%
Test 4: test section 1
Classifier accuracy: 71.72%

Tests 5-6: 199 training set, 1300 test set
Average classifier accuracy: 75.31%
Test 3: test section 0
Classifier accuracy: 74.53%
Test 4: test section 1
Classifier accuracy: 76.07%

Tests 7-8: 299 training set, 1200 test set
Average classifier accuracy: 76.54%
Test 3: test section 0
Classifier accuracy: 77.33%
Test 4: test section 1
Classifier accuracy: 75.75%

Tests 9-10: 399 training set, 1100 test set
Average classifier accuracy: 77.55%
Test 9: test section 0
Classifier accuracy: 76.00%
Test 10: test section 1
Classifier accuracy: 79.09%

Tests 11-12: 499 training set, 1000 test set
Average classifier accuracy: 78.05%
Test 11: test section 0
Classifier accuracy: 77.10%
Test 12: test section 1
Classifier accuracy: 79.10%

Tests 13-14: 599 training set, 900 test set
Average classifier accuracy: 77.67%
Test 13: test section 0
Classifier accuracy: 76.89%
Test 14: test section 1
Classifier accuracy: 78.44%

Tests 15-16: 699 training set, 800 test set
Average classifier accuracy: 78.00%
Test 6: test section 0
Classifier accuracy: 78.13%
Test 7: test section 1
Classifier accuracy: 77.89%

Tests 17-19: 799 training set, 700 test set
Average classifier accuracy: 78.33%

Test 17: test section 0
Classifier accuracy: 78.57%
Test 18: test section 1
Classifier accuracy: 77.71%
Test 19: test section 2
Classifier accuracy: 78.71%

Tests 20-22: 899 training set, 600 test set
Average classifier accuracy: 78.67%
Test 20: test section 0
Classifier accuracy: 78.50%
Test 21: test section 1
Classifier accuracy: 77.50%
Test 22: test section 2
Classifier accuracy: 80.00%

Tests 23-25: 999 training set, 500 test set
Average classifier accuracy: 78.60%
Test 23: test section 0
Classifier accuracy: 77.80%
Test 24: test section 1
Classifier accuracy: 77.00%
Test 25: test section 2
Classifier accuracy: 81.00%

Tests 26-29: 1099 training set, 400 test set
Average classifier accuracy: 79.06%
Test 26: test section 0
Classifier accuracy: 77.75%
Test 27: test section 1
Classifier accuracy: 80.00%
Test 28: test section 2
Classifier accuracy: 76.00%
Test 29: test section 3
Classifier accuracy: 82.50%

Tests 30-34: 1199 training set, 300 test set
Average classifier accuracy: 78.73%
Test 30: test section 0
Classifier accuracy: 77.00%
Test 31: test section 1
Classifier accuracy: 80.33%
Test 32: test section is 2
Classifier accuracy: 76.00%

Test 33: test section 3
Classifier accuracy: 79.00%
Test 34: test section 4
Classifier accuracy: 81.33%

Tests 35-42: 1299 training set, 200 test set
Average classifier accuracy: 78.81%
Test 35: test section 0
Classifier accuracy: 77.50%
Test 36: test section 1
Classifier accuracy: 80.00%
Test 37: test section 2
Classifier accuracy: 80.50%
Test 38: test section 3
Classifier accuracy: 81.00%
Test 39: test section 4
Classifier accuracy: 71.50%
Test 40: test section 5
Classifier accuracy: 80.50%
Test 41: test section 6
Classifier accuracy: 78.50%
Test 42: test section 7
Classifier accuracy: 81.00%

Tests 43-57: 1399 training set, 100 test set
Average classifier accuracy: 79.33%
Test 43: test section 0
Classifier accuracy: 79.00%
Test 44: test section 1
Classifier accuracy: 78.00%
Test 45: test section 2
Classifier accuracy: 77.00%
Test 46: test section 3
Classifier accuracy: 82.00%
Test 47: test section 4
Classifier accuracy: 81.00%
Test 48: test section 5
Classifier accuracy: 81.00%
Test 49: test section 6
Classifier accuracy: 85.00%
Test 50: test section 7
Classifier accuracy: 78.00%
Test 51: test section 8
Classifier accuracy: 69.00%

Test 52: test section 9
Classifier accuracy: 74.00%
Test 53: test section 10
Classifier accuracy: 76.00%
Test 54: test section 11
Classifier accuracy: 86.00%
Test 55: test section 12
Classifier accuracy: 82.00%
Test 56: test section 13
Classifier accuracy: 75.00%
Test 57: test section 14
Classifier accuracy: 87.00%

Without education-num results:

Tests 1-2: 49 training set, 1450 test set
Average classifier accuracy: 69.34%
Test 1: test section 0
Classifier accuracy: 69.10%
Test 2: test section 1
Classifier accuracy: 69.59%

Tests 3-4: 99 training set, 1400 test set
Average classifier accuracy: 71.43%
Test 3: test section 0
Classifier accuracy: 70.43%
Test 4: test section 1
Classifier accuracy: 72.43%

Tests 5-6: 199 training set, 1300 test set
Average classifier accuracy: 75.73%
Test 3: test section 0
Classifier accuracy: 75.15%
Test 4: test section 1
Classifier accuracy: 76.31%

Tests 7-8: 299 training set, 1200 test set
Average classifier accuracy: 76.79%
Test 3: test section 0
Classifier accuracy: 75.58%
Test 4: test section 1
Classifier accuracy: 78.00%

Tests 9-10: 399 training set, 1100 test set

Average classifier accuracy: 77.55%
Test 9: test section 0
Classifier accuracy: 75.91%
Test 10: test section 1
Classifier accuracy: 79.18%

Tests 11-12: 499 training set, 1000 test set
Average classifier accuracy: 77.90%
Test 11: test section 0
Classifier accuracy: 76.80%
Test 12: test section 1
Classifier accuracy: 79.00%

Tests 13-14: 599 training set, 900 test set
Average classifier accuracy: 77.83%
Test 13: test section 0
Classifier accuracy: 77.00%
Test 14: test section 1
Classifier accuracy: 78.67%

Tests 15-16: 699 training set, 800 test set
Average classifier accuracy: 78.00%
Test 6: test section 0
Classifier accuracy: 78.13%
Test 7: test section 1
Classifier accuracy: 77.88%

Tests 17-19: 799 training set, 700 test set
Average classifier accuracy: 78.38%
Test 17: test section 0
Classifier accuracy: 78.71%
Test 18: test section 1
Classifier accuracy: 77.71%
Test 19: test section 2
Classifier accuracy: 78.71%

Tests 20-22: 899 training set, 600 test set
Average classifier accuracy: 78.44%
Test 20: test section 0
Classifier accuracy: 77.83%
Test 21: test section 1
Classifier accuracy: 77.50%
Test 22: test section 2
Classifier accuracy: 80.00%

Tests 23-25: 999 training set, 500 test set
Average classifier accuracy: 78.73%
Test 23: test section 0
Classifier accuracy: 78.20%
Test 24: test section 1
Classifier accuracy: 77.00%
Test 25: test section 2
Classifier accuracy: 81.00%

Tests 26-29: 1099 training set, 400 test set
Average classifier accuracy: 79.19%
Test 26: test section 0
Classifier accuracy: 78.25%
Test 27: test section 1
Classifier accuracy: 80.00%
Test 28: test section 2
Classifier accuracy: 76.00%
Test 29: test section 3
Classifier accuracy: 82.50%

Tests 30-34: 1199 training set, 300 test set
Average classifier accuracy: 78.73%
Test 30: test section 0
Classifier accuracy: 77.00%
Test 31: test section 1
Classifier accuracy: 80.33%
Test 32: test section is 2
Classifier accuracy: 76.00%
Test 33: test section 3
Classifier accuracy: 79.00%
Test 34: test section 4
Classifier accuracy: 81.33%

Tests 35-42: 1299 training set, 200 test set
Average classifier accuracy: 78.69%
Test 35: test section 0
Classifier accuracy: 76.50%
Test 36: test section 1
Classifier accuracy: 80.00%
Test 37: test section 2
Classifier accuracy: 80.50%
Test 38: test section 3
Classifier accuracy: 81.00%

Test 39: test section 4
Classifier accuracy: 71.50%
Test 40: test section 5
Classifier accuracy: 80.50%
Test 41: test section 6
Classifier accuracy: 78.50%
Test 42: test section 7
Classifier accuracy: 81.00%

Tests 43-57: 1399 training set, 100 test set
Average classifier accuracy: 79.27%
Test 43: test section 0
Classifier accuracy: 78.00%
Test 44: test section 1
Classifier accuracy: 78.00%
Test 45: test section 2
Classifier accuracy: 77.00%
Test 46: test section 3
Classifier accuracy: 82.00%
Test 47: test section 4
Classifier accuracy: 81.00%
Test 48: test section 5
Classifier accuracy: 81.00%
Test 49: test section 6
Classifier accuracy: 85.00%
Test 50: test section 7
Classifier accuracy: 78.00%
Test 51: test section 8
Classifier accuracy: 69.00%
Test 52: test section 9
Classifier accuracy: 74.00%
Test 53: test section 10
Classifier accuracy: 76.00%
Test 54: test section 11
Classifier accuracy: 86.00%
Test 55: test section 12
Classifier accuracy: 82.00%
Test 56: test section 13
Classifier accuracy: 75.00%
Test 57: test section 14
Classifier accuracy: 87.00%