Shivangi Mundhra
SUID - 842548148

Homework 6

Shivangi Mundhra | SUID – 842548148 | 10/22/2022

SECTION 1 - INTRODUCTION

In this homework, I am going to try to recognize digits 0 to 9 in handwriting images using a Decision Tree model and a Naïve-Bayes model. I am going to walk through all the aspects of this assignment including what the assignment is and what approaches I am using. In the analysis itself, I will try to explain all reasoning and thought process behind each significant step. The intent is that any person, especially one without technical knowledge and that is not familiar with this problem statement, should be able to read this report and follow along.

About the assignment –

The task at hand, in this assignment, is to be able to recognize the handwritten numbers 0 – 9, using decision tree and Naïve-Bayes algorithms.

About the Decision Tree algorithm –

Decision tree is a machine learning model used for classification purposes. The idea is to create a model that predicts the value of a target variable (in this case, label) by learning simple decision rules inferred from the data features. In this assignment, we want to create a model that predicts the label of handwritten images by analyzing the accompanying data for those labels. Decision Tree algorithm simply creates a set of rules, based on which it bases it decisions for classifying/predicting labels.

About the Naïve-Bayes algorithm –

The Naive Bayes classification algorithm is a probabilistic classifier. It uses Bayes probability for classification. It is called Naïve-Bayes because it treats all the features as independent which in real scenarios is highly unlikely.

Analysis –

I used scikit-learn machine learning libraries to create a decision tree model and a Naïve-Bayes model in python.

1. I start with importing different packages like sklearn, pandas, numpy, etc. Some of these, like sklearn, help us with creation and implementation of the models, some, like matplotlib and graphviz help in creating visualizations and others just help in data manipulation.
2. I store the contents of the data from digit-train.csv into a train dataframe and the data from digit-test.csv into a test dataframe.
3. I split the training and testing dataset in images and labels data. The images data represents the features (columns) that we want to train our data on and use to predict labels.

Shivangi Mundhra
SUID - 842548148

SECTION 2 - DECISION TREE

1. We create a base decision tree model with default settings, using the training data. We use this model to make predictions on the test data set and calculate its accuracy.
2. After testing the decision tree model against the test data set, we receive an accuracy score of 78.63%.
3. This tree might include a chance of overfitting. So, we test accuracies of a few models with various max_depths and min_samples_leaf.
    a. Max_depth - The maximum depth of the tree. If None (default), then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.
    b. Min_samples_leaf – default is 1. The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least min_samples_leaf training samples in each of the left and right branches. This may have the effect of smoothing the model, especially in regression.
4. We set the max_depth (= 10), min_samples_leaf (= 1) variable in the next decision tree model.
5. We test the accuracy of this model on the test data again and find that the accuracy for this model is 78.44%.
6. We also use 5-fold cross validation – a technique to prevent over-fitting and promote model generalization.
7. The 5-fold cross validation score is 75.060 +/- 1.282.

SECTION 3 – NAÏVE-BAYES

In Naïve-Bayes, we apply the Gaussian Naïve-Bayes and the Multinomial Naïve-Bayes to our data and create two separate models.

Gaussian Naïve-Bayes –

1. We create a base Gaussian Naïve-Bayes model with default settings, using the training data. We use this model to make predictions on the test data set and calculate its accuracy.
2. After testing the model against the test data set, we receive an accuracy score of 52.67%.

Shivangi Mundhra
SUID - 842548148

3. We get the confusion matrix for its predictions. The rows represent ground truth, ie, actual values from test labels and the columns represent predictions for the test data.

```
[[385   3   4   1   0   4   8   0   3   6]
 [  1 443   5   6   0   1   6   0   9   7]
 [ 50  23 148  30   3   7  89   1  65   4]
 [107  49  12 116   2   3  20   5  77  55]
 [ 24  10   7   5  40  15  37   0  76 190]
 [101  17   4  12   1  43  18   1 157  34]
 [ 12   3   5   0   1   3 375   0   4   1]
 [  4   6   1   1   3   4   3 132   7 304]
 [ 25  96   2   2   3  12   8   0 172  73]
 [  3   6   1   0   1   1   0   9   8 357]]
```

4. We also print the classification report for this model.

	precision	recall	f1-score	support
0	0.54	0.93	0.68	414
1	0.68	0.93	0.78	478
2	0.78	0.35	0.49	420
3	0.67	0.26	0.37	446
4	0.74	0.10	0.17	404
5	0.46	0.11	0.18	388
6	0.66	0.93	0.77	404
7	0.89	0.28	0.43	465
8	0.30	0.44	0.35	393
9	0.35	0.92	0.50	386
accuracy			0.53	4198
macro avg	0.61	0.53	0.47	4198
weighted avg	0.62	0.53	0.48	4198

   a. Accuracy – Accuracy gives us the percent of predictions that were correct overall.
   b. Precision – Precision explains of the predictions, what percent were correct/actual predictions.
   c. Recall – Recall captures the percent of actuals that were predicted accurately.
5. The macro average of the overall model for accuracy, precision and recall are too low which makes this model unreliable.
6. The 5-fold cross validation score comes to 49.738 +/- 2.497.

Multinomial Naïve-Bayes –

Shivangi Mundhra
SUID - 842548148

1. We create a base Multinomial Naïve-Bayes model with default settings, using the training data. We use this model to make predictions on the test data set and calculate its accuracy.
2. After testing the model against the test data set, we receive an accuracy score of 81.75%.
3. We get the confusion matrix for its predictions. The rows represent ground truth, ie, actual values from test labels and the columns represent predictions for the test data.

```
[[380   0   0   2   0   7   6   0  19   0]
 [  0 445   6   5   0   0   1   0  19   2]
 [  3   7 352   9   6   2  12   3  25   1]
 [  2   8  25 348   2  15   6   3  20  17]
 [  1   1   3   0 290   1   5   0  12  91]
 [ 11   5   2  55   5 268   5   1  25  11]
 [  6   0   4   0   5  14 368   0   7   0]
 [  2   8   1   1  28   0   1 366   8  50]
 [  4  24   9  23   4  16   3   1 297  12]
 [  0   1   1   7  32   1   0   9  17 318]]
```

4. We also print the classification report for this model.

	precision	recall	f1-score	support
0	0.93	0.92	0.92	414
1	0.89	0.93	0.91	478
2	0.87	0.84	0.86	420
3	0.77	0.78	0.78	446
4	0.78	0.72	0.75	404
5	0.83	0.69	0.75	388
6	0.90	0.91	0.91	404
7	0.96	0.79	0.86	465
8	0.66	0.76	0.71	393
9	0.63	0.82	0.72	386
accuracy			0.82	4198
macro avg	0.82	0.82	0.82	4198
weighted avg	0.83	0.82	0.82	4198

   a. Accuracy – Accuracy gives us the percent of predictions that were correct overall.
   b. Precision – Precision explains of the predictions, what percent were correct/actual predictions.
   c. Recall – Recall captures the percent of actuals that were predicted accurately.
5. The macro average of the overall model for accuracy, precision and recall are relatively high as compared to other models.
6. The 5-fold cross validation score should be higher than Gaussian Naïve-Bayes as well.

SECTION 4 – ALGORITHM PERFORMANCE COMPARISON

The Gaussian Naïve-Bayes model's accuracy is too low as compared to the accuracy of the Decision Tree model. The Decision Tree model is also relatively easier to explain since it's just based on feature rule values. Since the Naïve-Bayes model is a probabilistic model, it is relatively harder to explain. Decision tree is a discriminative model, whereas Naive bayes is a generative model. Decision Tree model would generally work better on larger data. Naïve-Bayes also assumes independence in different features and Decision Trees work well with either, however, if not rightly pruned, it might ignore some key features. The Decision Tree model also ran a lot quicker than the Naïve-Bayes models.