Laboratory Exercise for week 9

IST-717 Big Data Analysis

Daniel Pearl

**Goal:**

This lab assignment is a classification problem to identify various types of clothing using MNIST fashion data. A model will be built using 10 different types of clothing: t-shirts, trousers, pullovers, dresses, coats, sandals, shirts, sneakers, bags, and ankle boots. Three different models will be utilized for this project: Random Forest, Naïve Bayes, and Neural Networks. Each of these models will be compared in speed and accuracy in their ability to predict new inputs of clothing data. The trade-offs of each will be discussed.

**Business Questions:**

* What is the accuracy of each classification method?
* What are the trade-offs of each approach?
* What is the compute performance of each approach?

**Data Sources:**

All data will come from the MNIST fashion database. MNIST is the Modified National Institute of Standards and Technology. Data in this database comes from Zalando’s article images:

<https://github.com/zalandoresearch/fashion-mnist>

**Cleaning the data:**

The images are saved as arrays using the fetch\_openml function. Variable ‘X’ contains an array of numbers for each image representing the RGB code (0-255). Each RGB number is the color of a pixel. Variable ‘y’ contains the classification number for each item of clothing. The arrays are divided into training and testing subsets using the sklearn train\_test\_split package. The training and testing subsets are divided by 255 in order to convert the array values to a 0-1 value. Because of how these numbers are stored in memory, this will slightly speed up the models.

**Data Analysis Methods:**

The first analytic method is the Random Forest model from the sklearn ensemble package. The second analytic method is the naïve Bayes model also from the sklearn package. The last analytic is a neural network. The accuracy of each model is measured and printed along with a confusion matrix to determine the strength of each clothing category.

Random Forest parameters:

For the RandomForestClassifier function, the parameter ‘n\_estimators’ is the number of decision trees in the forest.

Neural Network (MLP Classifier):

For the MLPClassifier function, the solver parameter uses the ‘Ibfgs’ optimizer. The number of hidden layers is set to 15. The random state is set to 1 which is the number of generations for weights.

**Visualizations**

Random Forest Confusion Matrix:

[[1839 0 34 68 7 0 173 0 12 0]

[ 0 2077 5 38 2 0 10 0 0 0]

[ 23 1 1748 27 211 1 102 0 17 0]

[ 50 6 15 1966 64 0 45 0 2 0]

[ 1 2 166 81 1748 0 113 0 4 0]

[ 0 0 0 1 0 1942 1 63 7 26]

[ 316 1 238 49 174 0 1251 0 30 0]

[ 0 0 0 0 0 23 0 2017 3 69]

[ 6 1 5 8 8 8 23 7 2002 0]

[ 0 0 0 1 0 35 3 76 2 1946]]

Random Forest Image Mismatches:

A screenshot of a cell phone

Description automatically generated

Naïve Bayes Confusion Matrix:

Naive Bayes Score: 0.6047619047619047

[[1275 57 70 461 185 0 40 0 45 0]

[ 5 1991 36 85 5 0 10 0 0 0]

[ 9 10 767 185 1081 0 34 0 44 0]

[ 17 380 17 1662 60 0 6 0 6 0]

[ 1 44 107 367 1573 0 4 0 19 0]

[ 1 1 7 2 0 545 9 1362 24 89]

[ 209 37 280 486 869 0 70 0 108 0]

[ 0 0 0 0 0 9 0 2073 5 25]

[ 4 4 66 200 293 7 27 9 1458 0]

[ 2 0 3 3 1 37 10 705 16 1286]]

Naïve Bayes Image Mismatches:

A screenshot of a cell phone

Description automatically generated

**Data Analysis Results**

|  |  |  |
| --- | --- | --- |
| Model | Model Accuracy | Model Speed |
| Random Forest | 0.8812 | 96.125s |
| Naïve Bayes | 0.600 | 1.603s |
| Neural Network | 0.8810 | 51.937s |

**Discussion and Conclusion**

What is the accuracy of each classification method?

Both the Random Forest and the Neural Network had an accuracy of 88.1 percent which massively outperformed the Naïve Bayes model which had an accuracy of 60 percent.

What are the trade-offs of each approach?

Between Random Forest and naïve Bayes, Random Forest is better for image classification problems. It is however, much slower to train than naïve Bayes. Because Random Forest is an ensemble method that iterates through different decision trees, it is usually pretty accurate. The neural network model is just as accurate if not more than the Random Forest model and doesn’t take as long to train. The Random Forest model however requires less pre-processing and only requires the number of iterations for each decision tree. For this set of fashion images, training a Neural Network model would be the best decision.

What is the compute performance of each approach?

The naïve Bayes model performed the fastest with a speed of 1.603 second. The Random Forest model was computed in 96.126 seconds which was the longest of the three models. The Neural Network took 51.937 seconds to train. Given the faster training speed and similar accuracy, the Neural Network model had the greatest performance.