Analysis Report

Lab Assignment 2: Classification

Garnett Grant

COMP 247 – W2024

Mohammed Saiful Islam

January 20th, 2024

**8. “Use imshow method to plot the values of the three variables you defined in the above point. Note the values in your Analysis report (written response).”**

When attempting to use the “imshow” method to plot the raw values of the three variables defined from question 7, the system throws an error stating:



Based on this error, the raw data has to be reshaped first and then plotted.

After reshaping the raw data into a 28x28 image, a legible image is displayed.

This can be further fit for analysis by including the cmap attribute and changing the color map display type to binary (i.e., black and white or grayscale)

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**11. Print the frequencies of each of the three target classes and note it in your written report in addition provide a screenshot showing a bar chart.**

The frequencies of each of the three target classes are vastly different when comparing the count of the 0 classes vs the other two classes, 1 & 9 respectively.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

**14. Use 3-fold cross validation to validate the training process, and note the results in your written response.**

Upon using 3-fold cross validation to validate the training process, the results are conclusively inefficient. The average score for the model is 56%. The score also observingly fluctuates through lower percentages after each cross-validation execution.

A screenshot of a computer program

Description automatically generated

**15. Use the model to score the accuracy against the test data, note the result in your written response.**

The accuracy against the test data is calculated at 53% (54% rounded). This is an unfavourable outcome.

A screenshot of a computer program

Description automatically generated

**17. Use the classifier to predict the three variables you defined in point 7 above. Note the results in your written response and compare against the actual results**.

Upon using the classifier to predict the three variables defined in point 7 (some\_digit1, some\_digit2, some\_digit3), the classifier predicted these classes as all 0s. 2/3 were correct.

some\_digit1 is 3, some\_digit2 is 2, and some\_digit3 is 5.

Considering the target variables have been transformed into 3 classes (0 made up of digits in between 0 and 3 inclusive, 1 made up of digits in between 4 and 6 inclusive, & 9 made up of digits in between 7 and 9 inclusive), only 2/3 of the predictions are correct.

A screenshot of a computer program

Description automatically generated

**18. Train a Logistic regression classifier using the same training data. Name the classifier LR\_clf\_firstname.**

**(Note this is a multi-class problem make sure to check all the parameters and set multi\_class='multinomial').**

**Try training the classifier using two solvers first “lbfgs” then “Saga”. Set max\_iter to 1200 and tolerance to 0.1 in both cases.**

A screen shot of a computer code

Description automatically generated

**Make sure you note the results in both cases in your written response, and note the main differences in your #written response.**

**Carryout a quick research on the difference between the “lbfgs” and “Saga” solvers and see #how this applies to the results, note that size and dimensions of the dataset. Don’t worry if one doesn’t converge your research should explain why. Note the results of your research in your analysis report.**

**lbfgs**

lbfgs stands for (Limited-memory Broyden-Fletcher-Goldfarb-Shanno). Lbfgs is a solver that has 3 notable features. Firstly, it’s another form of a known solver coined as Newton’s Method (newton-cg). Secondly, it utilizes limited memory i.e., it retains some vectors that demonstrate estimations non-explicitly. Thirdly, it can perform at a more desired rate on smaller datasets which in turn will save more memory. Some notable flaws of lbfgs require more attention to detail as this solver can eventually converge to nothing.

When using lbfgs solver for my classification model, I observed that the time it took to fit the training data was quite long. It took a staggering 3 mins (+ -). It took relatively the same amount of time just to cross validate the score. Although the time of execution was longer, it has been proven to be quite successful with an accuracy of 85%.

**Saga**

Saga stands for (Stochastic Average Gradient Descent). Saga is a solver that supports all penalties that the other solvers do. While it does have a memory cost of O(n), it is still computationally faster than other solvers for large datasets. It is regularly denoted as one of the better choices for a solver.

When utilizing saga, I observed that the time it took to fit the training data was amazingly swift. This solver completed the task in approximately 3.2s! The time it took to calculate the cross-validation score was also amazing, only taking 6.8s! What’s even more impressive is it’s accuracy of 85%. While the accuracy is similar when utilizing the lbfgs solver, the time of execution offers a more desired overall performance.

**Comparison**

In conclusion, while lbfgs and saga both have considerably high accuracy, both have their own flaws that need to be considered. Nevertheless, when it comes to large datasets like the mnist\_784 dataset, saga is the better solver for computational time and accuracy.

A screenshot of a computer

Description automatically generated

**19. Use 3-fold cross validation on the training data and note the results in your written response.**

As mentioned above, the 3-fold cross validation on the Logistic Regression Classifier using the lbfgs solver is very accurate at an acceptable 85%. When contrasting this performance to the initial Naïve Bayes Classifier used in Question 13, it is remarkable. It did however, take 2m and 50s just to execute.

The 3-fold cross validation on the Logistic Regression Classifier using the saga solver is also desirably accurate at a decent 85.3%.

A screenshot of a computer

Description automatically generated

**20. Use the model to score the accuracy against the test data, note the result in your written response.**

A screenshot of a computer

Description automatically generated

**21. Generate the Generate the accuracy matrix precision and recall of the model and note them in your written response.**

The accuracy, precision, and recall of the model using both solvers are fairly similar, however, the Lbfgs solver performs a bitter better precision wise.

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**22. Use the classifier that worked from the above point to predict the three variables you defined in point 7 above. Note the results in your written response and compare against the actual results**

Using lbfgs solver: A screenshot of a computer

Description automatically generated

Actual Result:

A screenshot of a computer

Description automatically generated

**Finally, in your analysis report (written response) Compare the results from both models, investigate why a model performed better than the other, and write your conclusions. (Be thorough)**

When it comes to performance, the saga solver Logistic Regression model outperformed every other model.

When it comes to precision, the lbfgs solver Logistic Regression model, outperformed every other model.

When it comes to accuracy, both Logistic Regression Models utilizing the saga and lbfgs solvers, outperformed the Naïve Bayse Model.

When it comes to prediction, all Models perform poorly.

In conclusion, my models are correctly processing the data it is given, in optimal timing when fine-tuned appropriately. There may be some other inconsistencies that need to be addressed in order for the Model to operate at it’s best capacity to provide the most thorough results which would further allow the model to be used at it’s greatest capacity in a real world situation.