**Assignment 3**

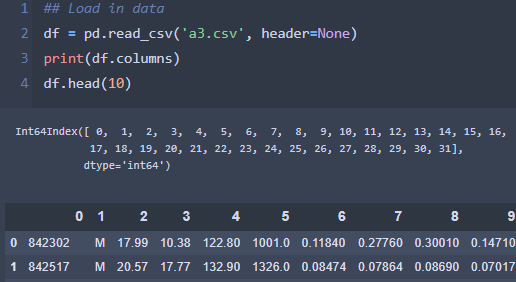
**BDA500 – Machine Learning**

**Alex Natis**

I declare this is my own work in accordance with Seneca College’s Academic Policies. No part has been copied manually or electronically from any other source (including web sites) or distributed to other students.

**Task 1: Study the dataset. Properly prepare the data for model training**

I first loaded in the data, printed the head, and checked the names of the columns.



I knew I would be building multiple models and requested to compare their time and accuracy so I created a dataframe to hold the results.

Text

Description automatically generated

I then split the data into training and testing data respectively and scaled the features to prepare for modelling.

Text

Description automatically generated

**Tasks 2, 3, and 4: Train three different Classifiers on the dataset and record the time to train each model and evaluate them using two different methods**

The 3 classifiers I chose to model the data were Logistic Regression, Linear Kernel SVC, and RBF Kernel SVC. I trained each of the models, timed them, and then evaluated them on both the train and test data using f1 score and accuracy. Then I stored the results in the result dataframe.

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

**Task 5 and 6: Use PCA to reduce the dataset's dimensionality, with an explained variance ratio of 95%. Check/ display the number of Principal Component (dimensions), and the explained variance ration for each Principal Component.**

I used PCA with an argument of .95 to reduce the dimensionality to the point where only 95% of the variance was explained. I then printed the overall explained variance and the explained variance for each principal component.

Text

Description automatically generated

**Task 7 and 8: Train the three models again based on the reduced dataset and record the time to train each model.**

Once again, I trained each of the models, timed them, and then evaluated them on both the train and test data using f1 score and accuracy. Then I stored the results in the result dataframe.

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

**Task 9 and 10: Use PCA to reduce the dataset's dimensionality to two and display the explained variance ration for each Principal Component.**

I used PCA to reduce the number of features to 2 principal components by pass the argument n\_components = 2. Then I printed the total explained variance as well as the explained variance for each principal component.

Text

Description automatically generated

**Task 11 and 12: Train the three models again based on the reduced dataset and record the time to train each model.**

Once again, I trained each of the models, timed them, and then evaluated them on both the train and test data using f1 score and accuracy. Then I stored the results in the result dataframe.

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

**Task 13: Visualize the dataset from task 9)**

I used plotly express to plot the dataset.

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

**Tasks 14, 15, 16, 17, 19, and 20: Compare the models from task 2) and find the best fit. Predict using the model in task 14. Evaluate the performance from task 15. Using the same algorithm chosen from task 14, but the trained model on the reduced dataset from task 5) and task 9), predict and evaluate the performance. Show your observation, discussion, thoughts, and your understanding. Discuss PCA from your experiment, such as whether PCA will speed up the training, or always speed up, or on all ML algorithms.**

Results of the models’ accuracy on the training set as well as predictions on the test set have all been computed and stored throughout the PCA testing process. The result dataframe can be seen below.

Graphical user interface, text, application

Description automatically generated

I found these results quite interesting. Although the RBF kernel SVC model had the lowest accuracy and f1 score on the full dimensionality training data, it had the highest accuracy on the full dimensionality test data. As dimensionality was reduced, the RBF kernel SVC model’s accuracy suffered the most from the reduced dimensionality out of the 3 models.

The results show that using PCA to reduce dimensionality does not always speed up training. The linear kernel SVC model was slower when it was trained with 2 components compared to training on full dimensionality or PCA of 95% variance. On the other hand, logistic regression’s training time consistently benefited each time the data’s dimensionality was reduced. The linear kernel SVC benefited from the first reduction of dimensionality, but this improvement was reversed when dimensionality was reduced again to 2 principal components.