Assignment #5 (PCA)
John Moderwell
MSDS 422

Summary

The objective of this assignment was to use the MNIST data in Python to determine the effect of performing a Principal Component Analysis on a Random Forest model. This would help management determine if PCA should be used to transform the data before creating an RF model. Two projects were created with one that implemented PCA and another that did not. Performance was measured by comparing F1 scores and timing how long each method took to build, fit and test the models.

Research Design

The MNIST dataset consists of 70,000 handwritten numbers. For both methods, the data was sorted into a training set (60,000) and a test set (10,000). The data was shuffled to ensure that there was a roughly equal representation of each digit. The first method, without PCA, builds off the tree model work from last week. The logic behind the Random Forest model involves creating many decision trees and then combining them to produce a more accurate prediction. For this assignment, the RandomForestRegressor in the Sci-kit Lean package from Python was used to build the model. The same model was used for both projects, the only difference being that one used PCA-transformed data. After creating the Random Forest model, two metrics were used to measure performance in both projects. The first metric was total time to build, fit and test the model. It is likely that management values efficiency and so the ability for a model to handle large datasets with ease is a very desirable attribute. A model that cannot handle large datasets may not be the best choice for management. The second metric that was used to measure performance was an F1 score. An F1 score is the harmonic mean of precision and recall.

Precision measures the number of true positives found relative to all positives found. Recall measures the number of true positives found out of all positives actually in the set, therefore measuring the ratio of positive instances that are accurately identified by the classifier. The F1 score is a combination of these two metrics and a values closer to 1 are desired.

The second project applied Principal Component Analysis (PCA) to the dataset. PCA is a dimensionality reduction algorithm. It works by identifying the axis that accounts for the most variance in the training set. After, the algorithm identifies a perpendicular axis that has the largest amount of remaining variance. This process continues until the components required for the problem are all identified. For this assignment, the objective was to keep 95% of the variance in the MNIST dataset. The algorithm identified 154 required components out of 784. Once these components were identified, training and test sets were created from the transformed data. The Random Forest model was then evaluated on this data.

Programming Overview

For this assignment, the programming language Python was used and executed in the Jupyter Notebook environment. Several packages were utilized throughout the project including: Pandas and Numpy for data handling/manipulation. The machine learning portion of the assignment was done entirely using the Sci-Kit Learn package. Relevant features included the RandomForestRegressor, F1 Score, PCA and FactorAnalysis. The package Time was also used to measure how well models handled a large dataset like MNIST.

Results and Recommendations

Assignment #5 (PCA)
John Moderwell
MSDS 422

The Random Forest model without PCA resulted in an F1 score of .9722. The time it took to build, fit and evaluate the model was 315.51 seconds. On the other hand, the Random Forest model with PCA resulted in an F1 score of .8981. The time it took to build, fit and evaluate this model was 8.53 seconds. It is apparent that PCA has a significant effect on model efficiency, but it also resulted in a less accurate score. My recommendation for management values precision over decreased efficiency because the main objective of a computer vision algorithm is accurate classification. With this in mind, it seems that the Random Forest model without PCA is a more appropriate model to use in this situation. PCA has many applications but resulting in a worse F1 score makes it less than ideal for this assignment.