Jonathan Pedoeem REU UNCW 2018

06/04/2018 Professor Chen

**Project 6 & 7**

1. **Introduction**

Support Vector Machines is an alternative classification and regression method to the ones used previously. It offers a novel approach to creating either linear, polynomial, or radial boundaries in a multidimensional space.

Principal Component Analysis is a method of creating basis vectors that align in the direction of highest variance. Picking out the most significant basis vectors can be used as a dimension reduction technique.

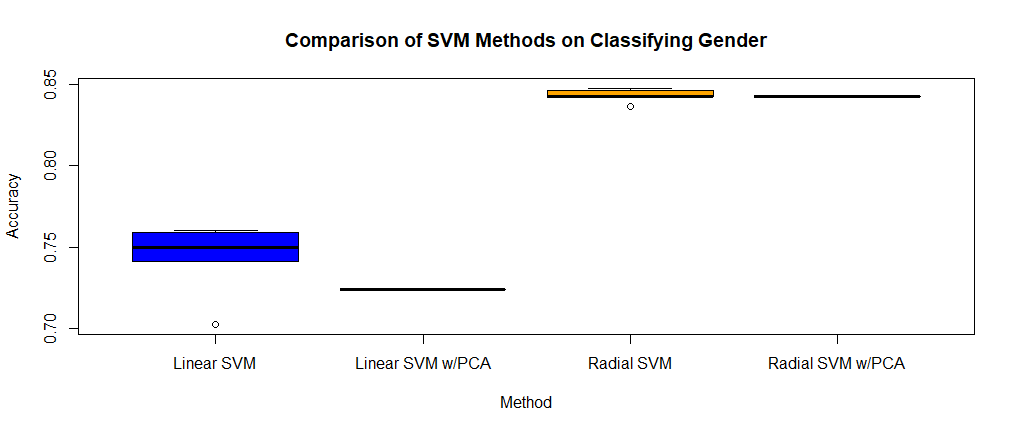
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Comparison of SVM methods** | | | | |
| **Method** | | **Accuracy** | **Standard Dev** | **Hyperparameters** |
| **Linear SVM** | | 74.25% | 2.4% | Cost = 0.001 |
| **Linear SVM with PCA** | 72.34% | 0.0% | Cost = 1 (No Significant Influence) | |
| **Radial SVM** | 84.3% | 0.44% | Cost = 1, gamma = 0.24 (No Significant Influence | |
| **Radial SVM with PCA** | 84.3% | 0.00% | Cost = 1 gamma = 2 (No Significant Influence) | |

1. **Results**

While literature promises high results for Support Vector Machines, the results achieved using both Linear and Radial SVM with and without PCA were not better than any of the methods previously attempted. Linear SVM performed the worst out of any classification method tested. While Radial SVM is towards the middle of the pack. Using Pricipal Component analysis to pick the top 150 principal components did not significantly affect the accuracy.

The hyperparameters were selected using grid search, in most cases the choice of a certain hyperparameter value did not have an influence on the accuracy which can indicated that the data is really spread out.

Table 1 Comparison of Support Vector Machine Methods on Gender Classification

 Figure 1 Boxplot Comparing Different SVM methods

Dimension Reduction using Principal Component Analysis was also tested on a selection of Classification and Regression techniques. Using PCA to select the 150 most significant directions in the BIF data instead of selecting the first 150 improved LDA on gender by 5.4% and improved the MAE for Linear Regression on Age by 2.506 years. However, Random Forest and K-NN did not see any improvements when suing PCA. By its nature, it is understandable why Random Forest did not show any improvement. Random forest works better with more predictors, therefore selecting only a few highly Principal Components will affect its performance

|  |  |  |
| --- | --- | --- |
| **Gender Classification with and without PCA** | | |
|  | **Without PCA** | **With PCA** |
| **LDA** | 86.1% | 91.5% |
| **Random Forest** | 85.3% | 84.3% |

Table 2 Comparison of Classification Techniques with and without PCA

|  |  |  |
| --- | --- | --- |
| **Age Regression with and without PCA** | | |
|  | **Without PCA** | **With PCA** |
| **Linear Regression on Age** | 8.789 (MAE) | 6.283 (MAE) |
| **K-NN on Age** | 17.895 (MAE) | 17.791(MAE) |

Table 3 Comparison of Regression Techniques with and without PCA