Introduction

This project involves using support vector machines and dimension reduction to analyze two sets of data, a gene expression data set and a spam emails data set, in order to identify patterns in the data and compare the effectiveness of different models. For the gene expression data, I aimed to use various types of visualizations obtained from applying PCA and sparse PCA to see how the principal components explain the variation in the data and to discover patterns in the cancer types. For the spam emails data, I aimed to compare the classification accuracy between a linear SVM model and a radial SVM model.

Methods and Results

To analyze the gene expressions, I applied PCA to the data and first created a histogram visualization of the first 4 PC loading vectors. These histograms had a fairly wide distribution with a large amount of nonzero entries, suggesting no dominant features. Next, I plotted the PVE and cumulative PVE at each number of principal components, which showed that the first few principal components explained most of the variance in the data. Looking closer at the first 50 principal components, there was an elbow after about the tenth PC, indicating that these first 10 captured the majority of the variance, about 53%. Then, I plotted the score vectors of PC1 and PC2 along with color-coded and shape-coded points to distinguish the classes. A large portion of the points were around the origin. The KIRC cancer type had the most distinct signature, being separate from the other classes. It formed a diagonal cluster, indicating correlation with both PC1 and PC2. PC1 scores appear to be indicative of the COAD class due to its shape in following the PC1 axis. The LUAD class was the least distinct, as it was clustered around 0 and heavily overlapped with the other classes. PRAD and BRCA shared a similar pattern to each other and had more variance than the LUAD class as their points were more spread out.

Then, I applied sparse PCA and created the same visualizations. This time, the number of principal components was set to 10, since it explained the majority of the variance. The loadings histograms were different; there was a very high peak at 0. The adjusted PVE plot had an elbow after the fourth PC. The adjusted CPVE at 4 PCs was 33%. The score vectors plot was very similar to the PCA one, showing the same patterns for the cancer types.

The next task was analyzing the spam emails, specifically, attempting to classify spam and non-spam using SVM models with the predictors in the dataset. First, I checked for highly correlated features using the correlation coefficients. I found that 3 of the features had high correlation of over 0.8, so I omitted those in order to avoid collinearity and create a more accurate model. I split the data into training and testing subsets, then applied PCA to the training data to identify patterns in the scores plot. The plot showed a notable difference between the non-spam and spam clusters. They both clustered around 0, but trailed into distinct directions. Since they both went upward along the PC2 axis about the same amount but differ horizontally along PC1, we can infer that PC1 effectively captures the difference between the two classes. Overall, the plot suggested that there are distinguishable patterns between spam and non-spam emails.

I built an SVM model with linear kernel using the training set, with the optimal cost value of 0.1 obtained by cross-validation. I then built an SVM model with radial kernel with the optimal cost and gamma values: 5 and 0.5, respectively. I applied these models to the test set to make predictions on the classifications, and examined the resulting classification tables to determine accuracy. The SVM with linear kernel misclassified 53 observations in the test set, resulting in a 47% accuracy. The SVM with radial kernel misclassified 44 observations, resulting in a 56% accuracy. Neither of the models had high accuracy percentages, correctly classifying only about half of the test observations. In addition, although the radial SVM had a higher accuracy percentage, it predicted FALSE (non-spam) for every observation in the test set. On the other hand, the linear SVM mostly predicted TRUE, with 97 TRUE predictions out of the 100 test observations. This suggests that these models are not very effective at fitting the data to predict spam vs non-spam emails. Further, they each seem to favor predicting one class.

Discussion

For the first task, the scores plots revealed patterns between the cancer types, such as the KIRC class being clearly separated from the rest and the similarity between BRCA and PRAD. It would be interesting to conduct further investigation to find out why this is the case. Though the scores plots were the same for PCA and sparse PCA, the loadings histograms were noticeably different.

For the second task, the SVM models did not perform very well at classifying spam emails from non-spam emails, even with using cross-validation to determine the optimal parameters. The scores plot showed a very distinct difference between the FALSE (non-spam) and TRUE (spam) clusters, suggesting that patterns exist in the data, so it would be valuable to apply other classifiers to see if they are more accurate. It would also be interesting to investigate why the models predict one class significantly more than the other.