Homework 1: Support Vector Machines

Analysis

This dataset consists of eight numerical predictors and a categorical label with two classes. There are 1000 examples in the dataset, making its shape 1000, 9. Group A has 405 data points and Group B has 595 data points, meaning that while not perfectly balanced, there isn't an extreme class imbalance. The data also isn't very linear, which can be seen in the nonlinear pairplots and multimodal distributions between the variables. No initial cleaning was necessary for the dataset, as there were enough data points, no null values, and the data itself wasn't too complex. For preproccessing the data, it was split into training and test sets with an 80/20 split. Predictors (all numerical variables) were also z-scored to be scaled to the same distribution.

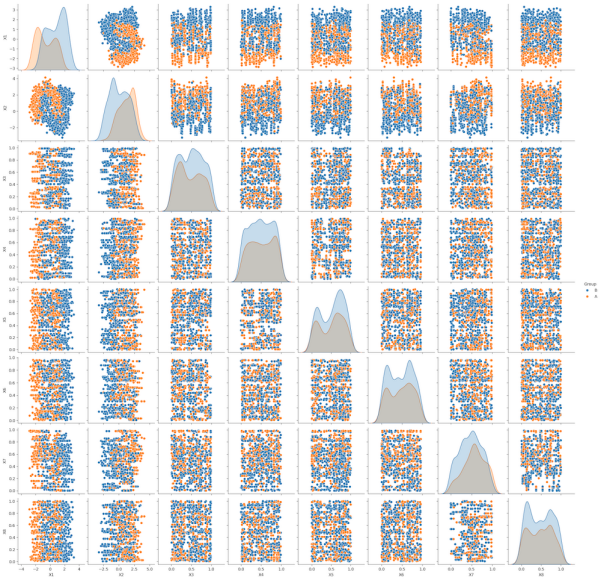

Figure 1. Pairplots of variable relationships and distributions by Group

Methods

Support Vector Machine
A pipeline consisting of the transformed columns and an SVC model was created. SVM models use kernel, gamma, and c parameters. The kernel type determines which function will transform the data points. Gamma is the kernel coefficient and affects the complexity of the model by determining how ridgid the decision boundary is. And C is the slack variable (regularization), which influences the flexibility of the model's classification. Each dataset will require its own variation of these parameters for strong performance. Grid search was applied during the model training process to choose effective values for these parameters. Possible hyperparameter values were organized into a parameter grid, and grid search was performed to evaluate accuracy and identify the best SVM model. Group classification predictions were made with this model for both the train and test set.

Logistic Regression
A pipeline consisting of the transformed columns and a Logistic Regression model was created. The logistic regression model was then fit to the training set and predictions were made with this model for both the train and test set.

K Nearest Neighbors
A pipeline consisting of the transformed columns and an KNN model was created. The number of neighbors (k) is an important parameter the KNN uses to make classifications. More neighbors means more nearby classifications affect the classification of a data point. The optimal number of neighbors can be identified through hyperparameter tuning. Values for k were organized into a parameter grid, and grid search was performed to identify the best k value. The best model was used to make Group classifications on the train and test sets.

Results

Model performance was evaluated with an accuracy metric, ROC AUC score, and confusion matrix. Accuracy measures the proportion of correct classifications to number of examples in the dataset. ROC AUC score assesses the model's ability to differentiate between classes. Confusion matrices show the models true positive, true negative, false positive, and false negative rates.

SVM
The support vector machine model has 0.895 accuracy and a 0.973 ROC AUC score for training set predictions and 0.775 accuracy and 0.832 ROC AUC score on the test set. This means that the model performs well, as high accuracy indicates that the model makes correct classifications most of the time. Furthermore, the ROC AUC score being close to 1 proves that the model can distinguish the classes. However, the model was noticeably better at predicting the training set, as the metrics were higher here, which indicates that this model is slightly overfit. The confusion matrices for the training and test sets also support this behavior. For the training set, there were 260 true negatives, 56 false positives, 456 true positives, and 28 false negatives. For the test set, there were 52 true negatives, 37 false positives, 103 true positives, and 8 false negatives. Thus, the model does a good job at predicting both classes (low false rates, high precision, and high specificity), with slightly better predictions on the training set.

Logistic Regression
The logistic regression model has 0.758 accuracy and 0.839 ROC AUC score for the training set and 0.765 accuracy and 0.863 ROC AUC score for the test set. These strong metrics indicate that the model performs well and is not overfit. The confusion matrices display 216 true negatives, 100 false positives, 390 true positives, and 94 false negatives for the training data and 60 true negatives, 29 false positives, 93 true positives, and 18 false negatives for the test data. This means the model has high precision and decent specificity; the model is good at predicting both classes but slightly better at predicting one class over another.

KNN
The k nearest neighbors model behaved similarly to the logistic regression model. Its accuracy and ROC AUC score for the training set are 0.788 and 0.878, respectively. The test set's accuracy and ROC AUC score are 0.760 and 0.835, respectively. Confusion matrices show training predictions had 212 true negatives, 104 false positives, 418 true positives, and 66 false negatives, and test predictions had 50 true negatives, 39 false positives, 102 true positives, and 9 false negatives. Overall, there was good model performance, minimal overfitting, and a slightly better ability to predict one class over another.

Summary
All of the models had strong performance, but the SVM was the best at making classifications. Even being overfit, the support vector machine had higher accuracy and class distinguishment compared to the logistic regression and knn models. In this scenario, SVM may have performed better because it can handle

complex relationships like nonlinearity, and class imbalances. While logistic regression and KNN can handle these to some extent, they can also be more sensitive to them than support vector machines.

Reflection

This project demonstrated how to evaluate models to identify an optimal one, whether by tuning a specific model or comparing different model types. Hyperparameter tuning is important, especially when many parameters influence a model, because a model configuration for one dataset may not be best for another. The benefits of SVMs were also evident, because it could handle the nonlinear relationships and other complexities in this dataset better than the logistic regression and knn models. Many insights come from evaluating metrics, too. For instance, classification performance, bias and variance (underfit/overfit), or identifying class imbalance. For future projects like this one, I will probably perform more exploratory data analysis at the beginning so that I can have a better understanding of the dataset while debugging.