Discriminant Analysis & Cross Validation

Linear Discriminant Analysis (LDA) is a supervised learning classification model which uses a linear combination of continuous features to predict one categorical outcome. LDA may also be used as a dimension-reduction technique since it picks out the most salient features and uses those values to model the dependent variable. Discriminant Analysis makes the assumption that the data follows a multivariate normal distribution with category mean μ and variance Σ. Therefore, one should test the data for this normality assumption before proceeding with a model.

Linear Discriminant Analysis also makes the assumption that each class has equal covariance matrices, or that for every class k, $\Sigma_k = \Sigma$. If this assumption is not made, or Σ_k is unique for each class K, then we have Quadratic Discriminant Analysis (QDA). While QDA is more flexible than LDA, it also requires more estimation parameters. Therefore, we expect the accuracy of a QDA model to be inversely proportional to the dimensionality of the data.

We use these models in combination with a k-fold cross validation technique. While this procedure can be used on any combination of data set and model, it is extremely useful on small datasets. Rather than train-and-test splitting the data into two dataframes, the data is randomly assigned into k groups, or folds, with 10 being the standard value. This standard value had been experimentally proven to optimize the bias-variance tradeoff. Then follows k iterations, with each group treated as the test dataset, and the remaining $k-1$ groups combined into the training dataset that fits the model. The technique scores each iteration, and takes the mean score of the k values. Additionally, this cross validation technique may also be called multiple times (said n repeats) for better accuracy. Notably, this changes the entries in the k groups at the start of each cross-validation repeat, which will remain unchanged for the k iterations.

The scores from cross-validation can be used to determine the most appropriate and accurate model. Likewise, if none of the models tested yield accurate results, it indicates that further hyperparameter tuning is necessary. These changes must be made within the Cross Validation loop, or else the data may be biased before fitting the model.