ADS 503: Applied Predictive Modeling

Module 5 Presentation 5.1: Linear Classifiers

Hello class and welcome to the Module 5 Presentation of ADS-503, Applied Predictive Modeling. In this module we are going to learn about various performance measures of classification models. We will also study linear classifier models such as Logistic Regression, Linear Discriminant Analysis, Partial Least Squares Discriminant Analysis, Penalized Models and Nearest Shrunken Centroids.

It is very important to pick the right performance measure for assessing the quality of a classifier, usually this is very dependent on the problem at hand and the costs associated with making a false positive or/and false negative prediction. This slide covers some of the most commonly used performance measures. ROC curve & Lift curve are other measures that are used commonly, Chapter 11 of your textbook discusses more about this topic.

Logistic regression is one of the most common classification algorithms used in practice. This popularity is due to its simplicity and the inferential statements that can be made about model terms. When choosing an algorithm for solving a classification problem usually logistic regression is used as a base model to compare other models against. Logistic regression can also be effective when the goal is solely prediction, but it does require the user to identify effective representations of the predictor data that yield the best performance. Section 12.2 of your textbook discusses more about this topic.

Linear Discriminant Analysis, or LDA, has some similarities with the PCA method we learnt previously. Unlike PCA, which is an unsupervised method, LDA is a supervised method but involves finding linear combinations of predictors that maximize between group variance when compared to within group variance. Practitioners should be particularly rigorous in pre-processing data before using LDA, it is recommended that predictors be centered and scaled and that near-zero variance predictors be removed. Section 12.3 of your textbook discusses more about this topic.

LDA can be extended to cases where multicollinearity is an issue or when the number of predictors is higher than number of observations using partial least squares approach, this approach is called partial least squares discriminant analysis. The main idea behind this approach is to maximize the covariance between the input & output. Section 12.4 of your textbook discusses more about this topic.

This slide details some of the penalties that can be applied to logistic regression models and linear discriminant analysis models. These penalties help to regularize the model by eliminating overfitting and there by providing generalized solutions. Section 12.5 of your textbook discusses more about this topic.

Nearest shrunken centroids is a linear classification algorithm that is well suited for high dimensional problems and when the number of predictors is much larger than number of samples. After taking each class centroids and the overall centroid of the training set, classifying unknown samples can be done in 2 ways. In the first method we can choose the closest centroid to the unknown sample and assign it that class label. In the second method a shrinkage parameter is used to shrink class centroids to the overall centroids and the unknown sample is assigned to the nearest centroid after shrinkage. Section 12.6 of your textbook discusses more about this topic.