Summary/Review

Classification Problems

The two main types of supervised learning models are:

- Regression models, which predict a continuous outcome
- Classification models, which predict a categorical outcome.

The most common models used in supervised learning are:

- Logistic Regression
- K-Nearest Neighbors
- Support Vector Machines
- Decision Tree
- Neural Networks
- Random Forests
- Boosting
- Ensemble Models

With the exception of logistic regression, these models are commonly used for both regression and classification. Logistic regression is most common for dichotomous and nominal dependent variables.

Logistic Regression

Logistic regression is a type of regression that models the probability of a certain class occurring given other independent variables.It uses a logistic or logit function to model a dependent variable. It is a very common predictive model because of its high interpretability.

Classification Error Metrics

A confusion matrix tabulates true positives, false negatives, false positives and true negatives. Remember that the false positive rate is also known as a type I error. The false negatives are also known as a type II error.

Accuracy is defined as the ratio of true postives and true negatives divided by the total number of observations. It is a measure related to predicting correctly positive and negative instances.

Recall or sensitivity identifies the ratio of true positives divided by the total number of actual positives. It quantifies the percentage of positive instances correctly identified.

Precision is the ratio of true positive divided by total of predicted positives. The closer this value is to 1.0, the better job this model does at identifying only positive instances.

Specificity is the ratio of true negatives divided by the total number of actual negatives. The closer this value is to 1.0, the better job this model does at avoiding false alarms.

The receiver operating characteristic (ROC) plots the true positive rate (sensitivity) of a model vs. its false positive rate (1-sensitivity).

The area under the curve of a ROC plot is a very common method of selecting a classification methods.T

he precision-recall curve measures the trade-off between precision and recall.

The ROC curve generally works better for data with balanced classes, while the precision-recall curve generally works better for data with unbalanced classes.