Project One: Part III - Predictive Modeling and Cross Validaton

(You are expected to give a descriptive title)

Contents

This is part III of project one focusing on the applications of cross-validation methods in predictive modeling.

1 Cross-validation for Predictive Modeling

The idea is to use **data-driven approaches** to data splitting and then apply cross-validation methods to select the final model from a pool of candidate models based on **predictive performance metric** such as **MSE** for linear regression models and **accuracy**, **sensitivity**, or **specificity** for logistic regression models.

Suggested Components in the Predictive Analysis

- *random splitting* - using random splitting for all data partitions.
- *Two-way data splitting* - data split into 75% for training and validation and 25% for testing.
- *5-fold cross-validation* - using a 5-fold cross-validation algorithm on the training data

2 Prediction Linear Regression

The primary predictive performance metric for linear regression modeling is the mean square error (the average squared error between predicted and the observed values of the response variable in its original scale).

Other predictive performance metrics that can also be used are R^2 or R^2_{adj}.

Likelihood-based metrics such as AIC and SBC can be used if the likelihood functions of all candidate models are at the same scale. These measures are not as intuitive as the MSE since MSE is a squared '**distance**' in the Euclidean space.

If the response variables in all candidate models are at the same scale, the MSE is expected to be used in the cross-validation for model selection.

3 Logistic Predictive Modeling

The primary tool for assessing the global predictive performance of logistic models is ROC curve analysis (this includes the area under the ROC curve - AUC). ROC curve suggested for this assignment.

Other predictive performance measures that can be considered are **accuracy**, **sensitivity**, and **specificity**.

Reporting ROC and AUC is required when comparing candidate models.

After the final model is identified, you need to use the 25% testing data set to report the **actual** performance of the corresponding models. The performance measure is similar the actual performance when the model is implemented new real data.