STA 529 2.0 Data Mining

Dr Thiyanga S. Talagala

Resampling methods

Lecture 4

Introduction

Drawing samples from a training set repeatedly and refitting a model of interest on each sample to learn more about the fitted model.

- most commonly used resampling methods
  - cross-validation
  - bootstrap

Cross validation

- Cross-validation is a technique for evaluating a machine learning model and testing its performance
- Useful, especially if the amount of data available is limited
- estimate the test error rate by holding out a subset of the training observations from the model training process, and then applying the model to those held out observations.

In-class visualization

The Validation Set Approach

The Validation Set Approach

- Randomly dividing the available data into two parts, a training set and a validation set or hold-out set.
- The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the validation set and compute error
- Simple and is easy to implement.

The Validation Set Approach

- in-class

Leave-One-Out Cross-Validation

- LOOCV involves splitting the set of observations into two parts.
- Here, a single observation (x_1, y_1) is used for the validation set, and the remaining observations $\{(x_2, y_2), ..., (x_n, y_n)\}$ make up the training set.

LOOCV advantages over the validation set approach

- Less bias. Why?

k-Fold Cross-Validation

- alternative to LOOCV
- randomly k-fold CV dividing the set of observations into k groups, or folds, of approximately equal size
- the first fold is treated as a validation set, and the model is fit on the remaining $k1$ folds

In-class visualization

Acknowledgement

Content is based on

Gareth James ▪ Daniela Witten ▪ Trevor Hastie ▪ Robert Tibshirani

An Introduction to Statistical Learning with Applications in R