Validation

Prof. Eric A. Suess

May 6, 2020

Introduction

Today we return the the **supervised learning** setting, **classification** and **prediction**.

Today we will review the second part of Chapter 10 Evaluating Model Performance.

We will discuss:

- holdout method: training, validation, test data
- cross-validation
- bootstrapping and the 0.632 bootstrap
- the R package caret

Estimating future performance

Split the data further. We have been using **training** and **test** datasets. We should add a third dataset, **validation**.

Develop the model(s) on the **training** dataset and then validate on the **validation** dataset. Finally, after choosing the final model(s) use the **test** dataset to see how the model(s) perform on **unseen** data.

Sampling

We have been using **random sampling**.

The **caret** package can be used to perform **stratified sampling**, which may blance the datasets better.

```
library(caret)
in_train <- createDataPartition(credit$default,
     p = 0.75)
```

Final Model

The author suggests:

Since the models trained on larger datasets generally perform better, a common practice is to retain the model on the **full set** of data after the final model has been selected and evaluated, allowing the model maximum use of available data.

Repeated holdout

To further evaluate the model, one can repeatly sample the training data and fit the model.

The final model would result from "averaging" over all of the models fit.

This process is referred to at **repeated holdout**.

Cross-validation

A formalization of the **repeated holdout** method is **k-fold cross-validation**.

Here **k folds** are randomly selected and the model is trained on each **k-1** subsets and validated on the remaining fold.

The final model would again result from “averaging” over all of the models fit.

Aside: this is similar to the **leave-one-out method** or **jackknifing**.

Bootstrapping

An alternative to **k-fold cross-validation** is **bootstrap sampling**.

Here the *training* and *test* datasets are created by sampling **with replacement**. The nonselected examples make up the **test** datasets.

When using **bootstrapping** the process is repeated many times and the results “averaged” at the end.