View on GitHub

stats-learning-notes

Notes from Introduction to Statistical Learning

Previous: Chapter 4 - Classification

Chapter 5 - Resampling Methods

Resampling methods are processes of repeatedly drawing samples from a data set and refitting a given model on each sample with the goal of learning more about the fitted model.

Resampling methods can be expensive since they require repeatedly performing the same statistical methods on N different subsets of the data.

Cross validation is a resampling method that can be used to estimate a given statistical methods test error or to determine the appropriate amount of flexibility.

Model assessment is the process of evaluating a model's performance.

Model selection is the process of selecting the appropriate level of flexibility for a model.

Bootstrap is used in a number of contexts, but most commonly it is used to provide a measure of accuracy of a given statistical learning method or parameter estimate.

Cross Validation

In the absence of a large test set that can be used to determine the test error rate, there are a number of techniques that can be used to estimate the error rate using training data.

The validation set approach involves randomly dividing the available observations into two groups, a training set and a validation or hold-out set. The model is then fit using the training set and then the fitted model is used to predict responses for the observations in the validation set.

The resulting validation set error rate offers an estimate of the test error rate.

Though conceptually simple and easy to implement, the validation set approach has two potential drawbacks.

1. The estimated test error rate can be highly variable depending on which observations fall into the training set and which observations fall into the test/validation set.

2. The estimated error rate tends to be overestimated since the given statistical method was trained with fewer observations than it would have if fewer observations had been set aside for validation.

Cross-validation is a refinement of the validation set approach that mitigates these two issues.

Leave-one-out Cross-Validation

Leave-one-out cross validation is similar to the validation set approach, except instead of splitting the observations evenly, leave-one-out cross-validation withholds only a single observation for the validation set. This process can be repeated n times with each observation being withheld once. This yields n mean squared errors which can be averaged together to yield the leave-one-out cross-validation estimate of the test mean squared error.

$$CV(n) = \frac{1}{n} \sum_{i=1}^{n} MSE_i$$

Leave-one-out cross validation has much less bias than the validation set approach. Leave-one-out cross validation also tends not to overestimate the test mean squared error since many more observations are used for training. In addition, leave-one-out cross validation is much less variable, in fact, it always yields the same result since there's no randomness in the set splits.

Leave-one-out cross validation can be expensive to implement since the model has to be fit n times. This can be especially expensive in situations where n is very large and/or when each individual model is slow to fit.

A shortcut exists for least squares linear or polynomial regression that makes the cost of leave-one-out cross validation the same as a single model fit. Formally stated, the shortcut is

$$CV(n) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - n_i} \right)^2$$

where $\hat{y}_i$ is the ith fitted value from the least squares fit and n_i is the leverage statistic, defined as

$$n_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}$$

for a simple linear regression.

Because the ith residual is divided by $1 - n_i$, each observation is inflated based on the amount the observation influences its own fit which allows the inequality to hold.

Leave-one-out cross validation is a very good general method which can be used with logistic regression, linear discriminant analysis, and many other methods. That said, the shortcutting method doesn't hold in general which means the model generally needs to be refit n times.

K-Fold Cross Validation

K-fold cross validation operates by randomly dividing the set of observations into K groups or folds of roughly equal size. Similar to leave-one-out cross validation, each of the K folds is used as the validation set while the other $K - 1$ folds are used as the test set to generate K estimates of the test error. The K-fold cross validation estimated test error comes from the average of these estimates.

$$CV(k) = \frac{1}{k} \sum_{i=1}^{k} MSE_i$$

It can be shown that leave-one-out cross validation is a special case of K-fold cross validation where $K = n$.

Typical values for K are 5 or 10 since these values require less computation than when K is equal to n.

Cross validation can be used both to estimate how well a given statistical learning procedure might perform on new data and to estimate the minimum point in the estimated test mean squared error curve, which can be useful when comparing statistical learning methods or when comparing different levels of flexibility for a single statistical learning method.

Bias-Variance Trade-Off for K-Fold Cross Validation

There is a bias-variance trade-off inherent to the choice of K in K-fold cross validation. Typically, values of $K = 5$ or $K = 10$ are used as these values have been empirically shown to produce test error rate estimates

that suffer from neither excessively high bias nor very high variance.

In terms of bias, leave-one-out cross validation is preferable to K-fold cross validation and K-fold cross validation is preferable to the validation set approach.

In terms of variance, K-fold cross validation where $K < n$ is preferable to leave-one-out cross validation and leave-one-out cross validation is preferable to the validation set approach.

Cross Validation of Classification Problems

Cross validation can also be useful when Y is qualitative, in which case the number of misclassified observations is used instead of mean squared error.

In the classification setting, the leave-one-out cross validation error rate takes the form

$$CV(n) = \frac{1}{n} \sum_{i=1}^{n} Err_i$$

where $Err_i = I(y \neq \hat{y}_i)$. The K-fold cross validation error rate and the validation set error rate are defined similarly.

The Bootstrap

The bootstrap is a widely applicable tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning approach, including those for which it is difficult to obtain a measure of variability.

The bootstrap generates distinct data sets by repeatedly sampling observations from the original data set. These generated data sets can be used to estimate variability in lieu of sampling independent data sets from the full population.

The sampling employed by the bootstrap involves randomly selecting n observations with replacement, which means some observations can be selected multiple times while other observations are not included at all.

This process is repeated B times to yield B bootstrap data sets, $Z^{*1}, Z^{*2}, \ldots, Z_{*B}$, which can be used to estimate other quantities such as standard error.

For example, the estimated standard error of an estimated quantity $\hat{\alpha}$ can be computed using the bootstrap as follows:

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1} \sum_{r=1}^{B} (\hat{\alpha}^{*r} - \frac{1}{B} \sum_{s=1}^{B} \hat{\alpha}^{*s})^2}$$

Next: Chapter 6 - Linear Model Selection and Regularization

stats-learning-notes maintained by tdg5

Published with GitHub Pages