Introduction to Statistical Learning Chapter 5

Cheat-sheet

Introduction

- resampling methods are critical in modern statistics
- involve repeatedly drawing a sample from the training set, and refitting a model on the sample
- resampling methods can be computationally expensive
- two commonly used resampling methods: cross-validation and the boot-strap
- cross-validation can be used to estimate the test error associated with a statistical learning method (can evaluate performance or select appropriate flexibility)
- process of evaluating a model's performance is model assessment
- process of selecting the appropriate amount of flexibility is model selection

Cross-validation

- test error is the average error associated with predicting a new observation with a model
- test error, clearly, is a good metric to determine if ones model is accurate
- however, sometimes a designated test set is not available
- in this section, consider a class of methods that put aside a subset of the training observations in the fitting process, and then applying the model to those held out observations

The Validation Set Approach

- randomly divide the available training observations into a training set and a validation set (also known as a hold-out set)
- model is fit on the training data, then we observe the predictions for the validation set observations and compare to the true response
- the resulting validation set error rate provides an estimate of the test-error rate (typically assessed using MSE)
- can repeat this process of randomly splitting the data into training and validation data
- drawbacks: the validation test error rate can be highly variable depending on which observations are included in the set, and also since only a subset of the training observations are used to fit the model, we note the validation set error rate may overestimate the error rate for a model fit on all the data

Leave-One-Out Cross-validation

- attempts to address the validation set approach draw-backs
- once again, we split up the training observations into two parts
- this time, we set aside one training observation (x_1, y_1) and fit the model on the remaining n-1 observations
- we take the MSE of this predicted value and its true response
- we repeat this procedure n times for the n observations
- we take the average of the n LOOCV approaches $CV_{(n)} = \frac{1}{n} \sum MSE_i$
- this approach has far less bias than the validation set approach
- LOOCV will always yield the same results (every observation is left-out at some point)
- can be expensive to implement (must fit n models)

K-fold Cross-validation

- randomly divide training set into k-folds (k-groups), of approximately the same size

- the first fold is treated as the validation set, and the model is fit on the remaining k-1 folds
- the MSE is then calculated on the one held-out fold
- this procedure is repeated K times, with a different set of observations being treated as the validation set each time
- then, we compute the average of these values to get $CV_{(k)} = \frac{1}{k} \sum MSE_i$
- LOOCV may pose computational concerns, whereas K-fold cross-validation allows us to choose how big the groups should be, to best fit our computational power
- variability from how the observations are divided into the folds exists, but is less than the variability in the validation set approach
- in the real-world, we do not know the true test MSE, hence it is difficult to assess the accuracy of the cross-validation approximation
- sometimes we are truly interested in the test error estimate, but other times we are only interested in the minimum point of the test error estimate

Bias-Variance Trade-off for K-fold Cross Validation

- a potentially more important advantage of k-fold cross validation is that it gives a more accurate estimate of the true test MSE
- this has to do with the bias-variance trade-off
- the validation set approach can result in an overestimate of the true test MSE because the full model will more accurately approximate the true function versus a model with half the observations left out
- the mean of many correlated quantities has higher variance than the mean of many quantities that are not as heavily correlated (LOOCV has similar observations in each set; these estimates are heavily correlated)
- k-fold cross-validation with 5 or 10 for k have been shown empirically to not have either high bias or high variance in the test error rate

Cross-validation on Classification Problems

- in the qualitative case, instead of using test error rate, we use the number of misclassified observations as our error metric i.e: $C_{(v)} = \frac{1}{n} \sum Err_i$ is the form of the metric for LOOCV
- recalling: the training error tends to increase as the complexity of the model (or flexibility) increases, the test error typically is in a U-shape, where the optimal solution is typically in between high variance models and high bias models, and we note that the CV error rate is a good estimate of the test-error rate

The Bootstrap

- the bootstrap is a powerful tool in modern statistics that can be used to quantify the uncertainty associated with an estimator
- the bootstrap is useful because it can quantify variability in methods for which obtaining variability is hard to obtain, and is hard to output with statistical software
- the bootstrap method helps us emulate the process of obtaining new sample sets, so we can determine the variability of a quantity
- instead of generating new samples from the population, we repeatedly sample the observations from the original data set
- how it works: randomly choose n observations from the data-set, with replacement (meaning the same observation can occur twice in the bootstrap data set), we use this bootstrap data set to generate a new estimate for the quantity; this process is repeated B times, for some large quantity B to produce B bootstrap data sets and B estimates for the quantity and finally we can then compute the standard errors of boot-strap estimates using a formula