

Week 5: Resampling methods

Slava Mikhaylov

PUBLG088 Advanced Quantitative Methods

Week 5 Outline

Cross-validation

- Validation-set approach

- K-fold Cross-validation

Bootstrap

- Bootstrap mechanism

- Bootstrap and prediction

- The Bootstrap versus Permutation tests

Additional resources

- ▶ Piazza;
- ▶ R Drop-in Surgeries (Tuesdays 3-4, DMS Watson G15 Public Cluster, back of Science Library, starting today);
- ▶ Homework assignments walk-through session (Mondays, starting in Week 6, after Reading Week, room/time TBA).

Cross-validation

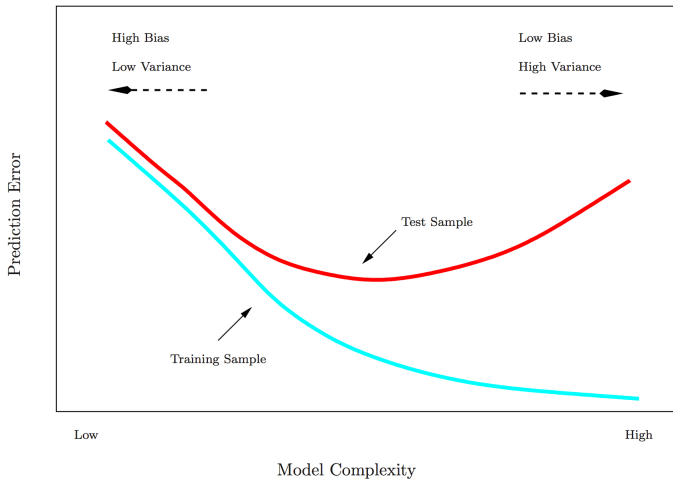
Resampling

- ▶ Today we discuss two resampling methods: cross-validation and the bootstrap.
- ▶ These methods refit a model of interest to samples formed from the training set, in order to obtain additional information about the fitted model.
- ▶ E.g., they provide estimates of test-set prediction error, and the standard deviation and bias of our parameter estimates.

Training Error versus Test error

- ▶ The **test error** is the average error that results from using a statistical learning method to predict the response on a new observation, one that was not used in training the method.
- ▶ In contrast, the **training error** can be easily calculated by applying the statistical learning method to the observations used in its training.
- ▶ Training error rate often is quite different from the test error rate, and in particular the former can **dramatically underestimate** the latter.

Training- versus Test-Set Performance



More on prediction-error estimates

- ▶ Best solution: a large designated test set. Often not available.
- ▶ Some methods make a **mathematical adjustment** to the training error rate in order to estimate the test error rate. These include the **Cp statistic**, **AIC** and **BIC**. They are discussed elsewhere in this course.
- ▶ Here we instead consider a class of methods that estimate the test error by holding out a subset of the training observations from the fitting process, and then applying the statistical learning method to those held out observations.

Validation-set approach

- ▶ We randomly divide the available set of samples into two parts: a **training set** and a **validation** 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.
- ▶ The resulting validation-set error provides an estimate of the test error. This is typically assessed using MSE in the case of a quantitative response and misclassification rate for qualitative response models.

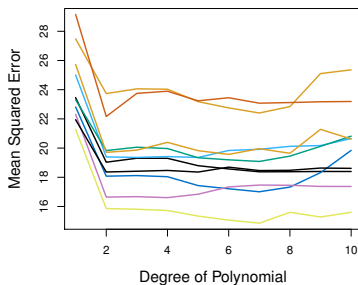
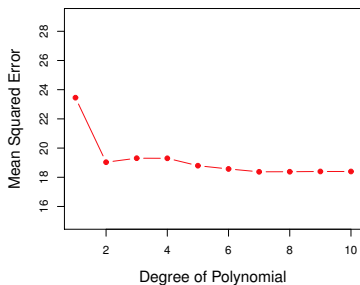
The Validation process



A random splitting into two halves: left part is training set, right part is validation set.

Example

- ▶ We want to compare linear vs higher-order polynomial terms in a linear regression with our Auto dataset.
- ▶ We randomly split the 392 observations into two sets, a training set containing 196 of the data points, and a validation set containing the remaining 196 observations.



Left panel shows single split; right panel shows multiple splits.

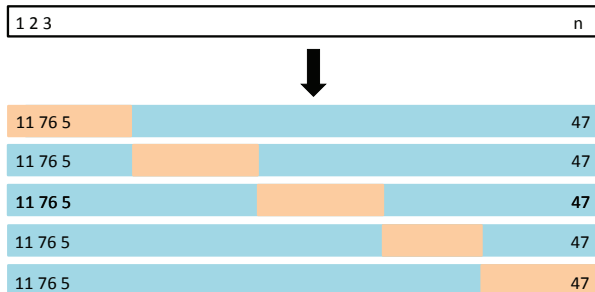
Drawbacks of validation set approach

- ▶ The validation estimate of the test error can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the validation set.
- ▶ In the validation approach, only a subset of the observations – those that are included in the training set rather than in the validation set – are used to fit the model.
- ▶ This suggests that the validation set error may tend to **overestimate** the test error for the model fit on the entire dataset. **Why?**

K-fold Cross-validation

- ▶ Very popular approach for estimating test error.
- ▶ Estimates can be used to select best model, and to give an idea of the test error of the final chosen model.
- ▶ Idea is to randomly divide the data into K equal-sized parts. We leave out part k , fit the model to the other $K - 1$ parts (combined), and then obtain predictions for the left-out k th part.
- ▶ This is done in turn for each part $k = 1, 2, \dots, K$, and then the results are combined.

5-fold CV



Mechanism

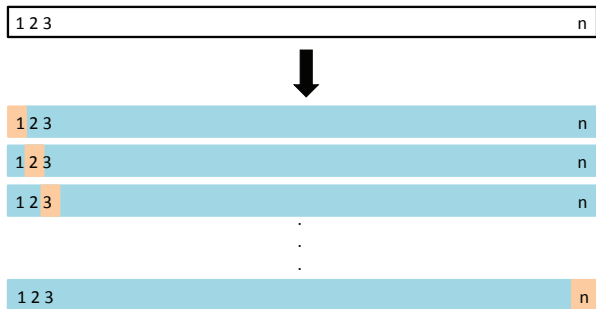
- ▶ Let the K parts be C_1, C_2, \dots, C_K , where C_k denotes the indices of the observations in part k . There are n_k observations in part k : if N is a multiple of K , then $n_k = n/K$.
- ▶ Compute

$$CV_{(K)} = \sum_{k=1}^K \frac{n_k}{n} \text{MSE}_k$$

where $\text{MSE}_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$, and \hat{y}_i is the fit for observation i , obtained from the data with part k removed.

- ▶ Setting $K = n$ yields n -fold or **leave-one out cross-validation** (LOOCV).

LOOCV



Special case of linear regression

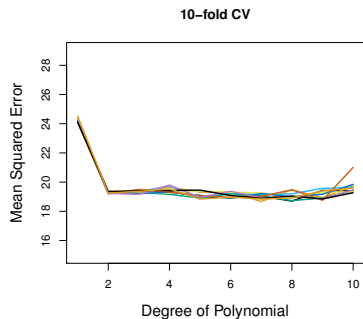
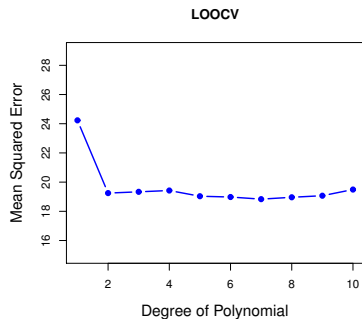
- ▶ With least-squares linear or polynomial regression, there is a shortcut making the cost of LOOCV the same as that of a single model fit.
- ▶ The following formula holds:

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

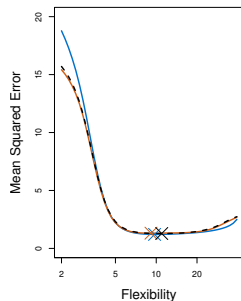
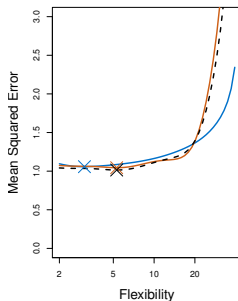
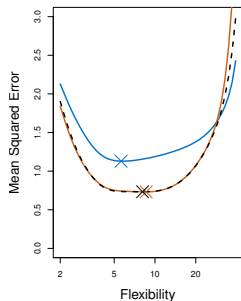
where \hat{y}_i is the i th fitted value from the original least squares fit, and h_i is the leverage (diagonal of the “hat” matrix). This is similar to the ordinary MSE, except the i th residual is divided by $(1 - h_i)$.

- ▶ The estimates from each fold are highly correlated and hence their average can have high variance.
- ▶ A better choice is $K = 5$ or 10 .

Auto data example



True and estimated test MSE for the simulated data



Additional issues with CV

- ▶ Since each training set is only $(K - 1)/K$ as big as the original training set, the estimates of prediction error will typically be biased upward.
- ▶ This bias is minimized when $K = n$ (LOOCV), but this estimate has high variance, as noted earlier.
- ▶ $K = 5$ or 10 provides a good balance for this bias-variance tradeoff.

CV for classification

- ▶ We divide the data into K roughly equal-sized parts C_1, C_2, \dots, C_K . C_k denotes the indices of the observations in part k . There are n_k observations in part k : if n is a multiple of K , then $n_k = n/K$.
- ▶ Compute

$$CV_K = \sum_{k=1}^K \frac{n_k}{n} \text{Err}_k$$

where $\text{Err}_k = \sum_{i \in C_k} I(y_i \neq \hat{y}_i) / n_k$.

- ▶ The estimated standard deviation of CV_K is

$$\widehat{\text{SE}}(CV_K) = \sqrt{\sum_{k=1}^K (\text{Err}_k - \bar{\text{Err}})^2 / (K - 1)}$$

- ▶ This is a useful estimate, but strictly speaking, not quite valid.
Why not?

CV: right and wrong

- ▶ Consider a simple classifier applied to some two-class data:
 1. Starting with 5000 predictors and 50 samples, find the 100 predictors having the largest correlation with the class labels.
 2. We then apply a classifier such as logistic regression, using only these 100 predictors.
- ▶ How do we estimate the test set performance of this classifier?
- ▶ Can we apply cross-validation in step 2, ignoring step 1?

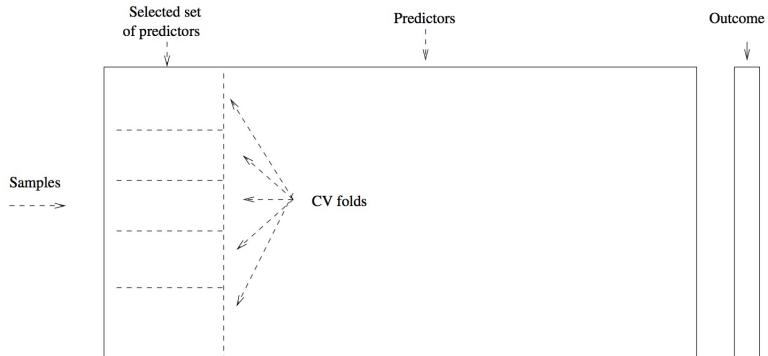
CV: right and wrong

- ▶ This would ignore the fact that in Step 1, the procedure has already seen the labels of the training data, and made use of them. This is a form of training and must be included in the validation process.
- ▶ It is easy to simulate realistic data with the class labels independent of the outcome, so that true test error = 50%, but the CV error estimate that ignores Step 1 is zero. (You can try doing this in class later today.)

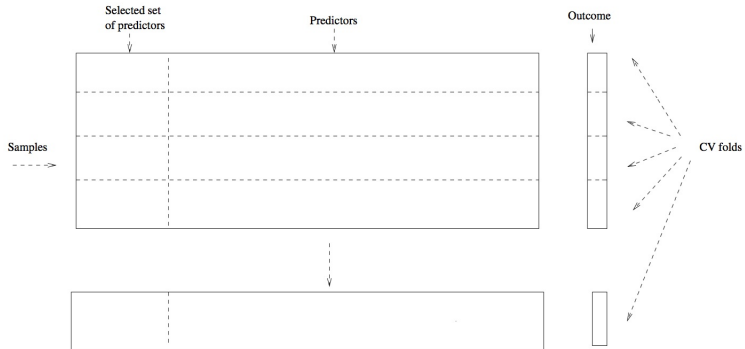
CV: right and wrong

- ▶ **Incorrect:** Apply cross-validation in step 2.
- ▶ **Correct:** Apply cross-validation to steps 1 and 2.

Wrong



Right



Bootstrap

Bootstrap

- ▶ The **bootstrap** is a flexible and powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.
- ▶ E.g., it can provide an estimate of the standard error of a coefficient, or a confidence interval for that coefficient.

Where does the name come from?

- ▶ The use of the term bootstrap derives from the phrase to pull oneself up by one's bootstraps, widely thought to be based on one of the eighteenth century “The Surprising Adventures of Baron Munchausen” by Rudolph Erich Raspe:

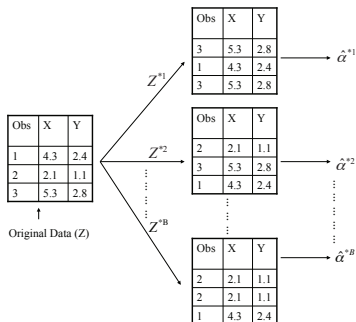
The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps.

- ▶ It is not the same as the term bootstrap used in computer science meaning to “boot” a computer from a set of core instructions, though the derivation is similar.

Intuition

- ▶ We usually care about uncertainty around our estimates from a sample. One way would be to repeatedly sample the population. But we cannot do that in real world.
- ▶ The bootstrap approach allows us to use replicate this process of obtaining new data sets, so that we can estimate the variability of our estimate without generating additional samples.
- ▶ Rather than repeatedly obtaining independent data sets from the population, we instead obtain distinct data sets by repeatedly sampling observations from the original data set with replacement.
- ▶ Each of these “bootstrap data sets” is created by sampling with replacement, and is the same size as our original dataset. As a result some observations may appear more than once in a given bootstrap data set and some not at all.

Example with three observations



- ▶ A graphical illustration of the bootstrap approach on a small sample containing $n = 3$ observations.
- ▶ Each bootstrap dataset contains n observations, sampled with replacement from the original data set.
- ▶ Each bootstrap data set is used to obtain an estimate of α .

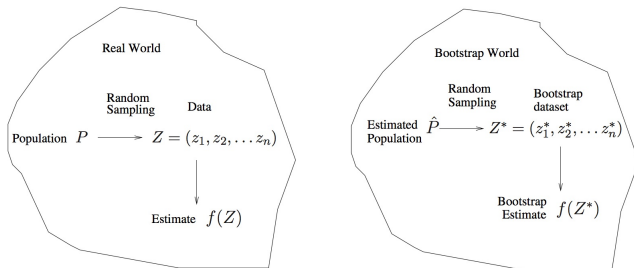
Mechanism

- ▶ Denoting the first bootstrap dataset by Z^{*1} , we use Z^{*1} to produce a new bootstrap estimate for α , which we call $\hat{\alpha}^{*1}$.
- ▶ This procedure is repeated B times for some large value of B (say 1000 or 10,000), in order to produce B different bootstrap datasets, $Z^{*1}, Z^{*2}, \dots, Z^{*B}$, and B corresponding α estimates, $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \dots, \hat{\alpha}^{*B}$.
- ▶ We estimate the standard error of these bootstrap estimates using the formula

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

- ▶ This serves as an estimate of the standard error of $\hat{\alpha}$ estimated from the original data set.

A general picture for the bootstrap



The bootstrap in general

- ▶ In more complex data situations, figuring out the appropriate way to generate bootstrap samples can require some thought.
- ▶ For example, if the data is a time series, we can't simply sample the observations with replacement (**why not?**).
- ▶ We can instead create blocks of consecutive observations, and sample those with replacements. Then we paste together sampled blocks to obtain a bootstrap dataset.

Other uses of the bootstrap

- ▶ Primarily used to obtain standard errors of an estimate (e.g. median).
- ▶ Also provides approximate confidence intervals for a population parameter. Also called a Bootstrap Percentile confidence interval. It is the simplest method (among many approaches) for obtaining a confidence interval from the bootstrap.

Bootstrap and prediction error

- ▶ In cross-validation, each of the K validation folds is distinct from the other $K - 1$ folds used for training: **there is no overlap**. This is crucial for its success.
- ▶ To estimate prediction error using the bootstrap, we could think about using each bootstrap dataset as our training sample, and the original sample as our validation sample.
- ▶ But each bootstrap sample has significant overlap with the original data. About two-thirds of the original data points appear in each bootstrap sample.
- ▶ This will cause the bootstrap to seriously underestimate the true prediction error.
- ▶ The other way around – with original sample = training sample, bootstrap dataset = validation sample – is even worse.

Removing the overlap

- ▶ Can partly fix this problem by only using predictions for those observations that did not (by chance) occur in the current bootstrap sample.
- ▶ But the method gets complicated, and in the end, cross-validation provides a simpler, more attractive approach for estimating prediction error.

The Bootstrap versus Permutation tests

- ▶ The bootstrap samples from the estimated population, and uses the results to estimate standard errors and confidence intervals.
- ▶ Permutation methods sample from an estimated null distribution for the data, and use this to estimate p-values and False Discovery Rates for hypothesis tests.
- ▶ The bootstrap can be used to test a null hypothesis in simple situations. E.g. if $\theta = 0$ is the null hypothesis, we check whether the confidence interval for θ contains zero.
- ▶ Can also adapt the bootstrap to sample from a null distribution (see Efron and Tibshirani book “An Introduction to the Bootstrap” 1993, chapter 16) but there’s no real advantage over permutations.