Machine Learning in Public Health

Lecture 5: CV and Boostrap

Dr. Yang Feng

Today's agenda

- Cross-Validation
- Bootstrap

Resampling Methods

- Resampling methods: repeatedly draw samples from a training set and refit a model of interest (or compute certain estimates) on each sample in order to obtain additional information about the fitted model (or those estimates)
- Common resampling methods include: cross-validation and bootstrap

Cross-validation (CV)

- Cross-validation (CV) can be used to estimate the test error associated with a given statistical learning method
  - *to evaluate its performance (model assessment)*
  - *or to select the best model (model selection)*
- When to use CV to estimate test error?
  - *when you don't have a designated test set*
- CV can be used for both classification and regression
- CV has a few variants; we only discuss the canonical version
- A precursor of CV is the validation set approach

The validation set approach

- Validation set approach: randomly divide the available set of observations 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.

FIGURE 5.1. *A schematic display of the validation set approach. A set of n observations are randomly split into a training set (shown in blue, containing observations 7, 22, and 13, among others) and a validation set (shown in beige, and containing observation 91, among others). The statistical learning method is fit on the training set, and its performance is evaluated on the validation set.*

Drawbacks of the validation set approach

- the validation estimate of the test error rate can be highly variable
- only about a half of the observations are used to train the model (inefficient use of data).
- Cross-validation: a refinement of the validation set approach that addresses these two issues.

Leave-One-Out Cross-Validation (LOOCV)

- Suppose the training data contains $\{(x_1, y_1), \cdots, (x_n, y_n)\}$
- First, use $n-1$ observations $\{(x_2, y_2), \cdots, (x_n, y_n)\}$ to train and use the remaining observation (x_1, y_1) to evaluate the performance: $MSE_1 = (y_1 - \hat{y}_1)^2$
- Repeat this procedure by using (x_2, y_2) as the validation data, training on the $n-1$ observations $(x_1, y_1), (x_3, y_3), \cdots (x_n, y_n)$, and compute $MSE_2 = (y_2 - \hat{y}_2)^2$
- Repeat this approach in total n times, which produces n squared errors $MSE_1, \cdots, MSE_n$
- The LOOCV estimate for the test MSE is the average of these n estimates:

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} MSE_i \, .$$

- There is no need to randomly shuffle the training data before implementing LOOCV. Why?

LOOCV

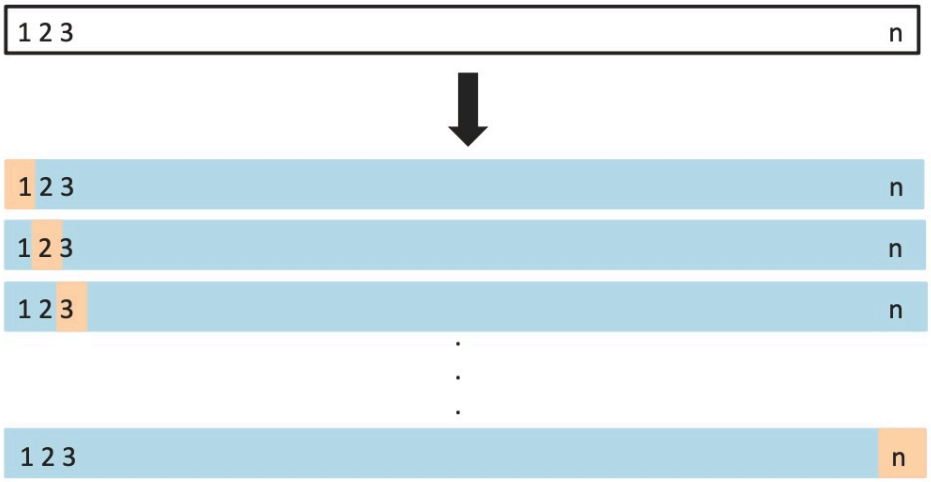


FIGURE 5.3. *A schematic display of LOOCV. A set of n data points is repeatedly split into a training set (shown in blue) containing all but one observation, and a validation set that contains only that observation (shown in beige). The test error is then estimated by averaging the n resulting MSE's. The first training set contains all but observation 1, the second training set contains all but observation 2, and so forth.*

Potential problems for LOOCV

- Computationally, LOOCV has the potential to be expensive to implement, since the model has to be fit n times
  - *However, with least squares linear or polynomial regression, 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* $h_i = \mathbf{x}_i{}^T(X^TX)^{-1}\mathbf{x_i}$ *is the leverage of the* i*th observation.*
- The n fitted models are each trained on an almost identical set of observations, and so averaging does not reduce variance much
  - *Think about letting* $X_i \sim N(0,1)$ *but insist* $X_1 = \cdots = X_n$
  - *What is the variance of* $\bar{X}$*?*

k-fold CV

- k-fold CV: randomly divide the set of observations into k groups, or folds, of approximately equal size.
- The first fold is treated as a validation set, and the method is fit on the remaining $k-1$ folds. The mean squared error, MSE_1, is then computed on the observations in the held-out fold.
- This procedure is repeated k times; each time, a different group of observations is treated as a validation set.
- This process results in k estimates of the test error, $MSE_1, MSE_2, \cdots, MSE_k$.
- The k-fold CV estimate is computed by averaging these values,

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

- We commonly use $CV_{(k)}$ to estimate the test error (model evaluation).

5-fold CV

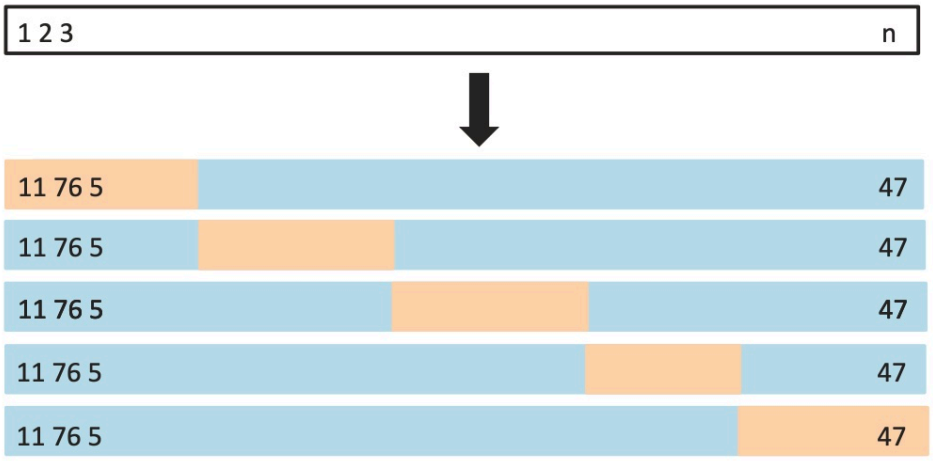


FIGURE 5.5. *A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.*

- LOOCV is a special case of k-fold CV in which $k = n$

Bias-variance trade-off for k-fold CV

- The larger the k, the smaller bias for estimating the **test error**
- However, large k (e.g., $k = n - 1$) means higher variance
- So for prediction accuracy, the better choices are usually not the extreme choices for k.
- The usual choices in practice for k in k-fold CV are .

k-fold CV for model selection

- Suppose in the `Default` dataset, we are deciding between two models
- model 1: use `rm` to predict `medv`
- model 2: use `rm` and `ptratio` to predict `medv`
- model 3: use `rm`, `ptratio`, and `zn` to predict `medv`
- To choose between these two models using 5-fold CV, we calculate $CV_{(5)}$ for models 1, 2, 3 and choose the model with the smaller $CV_{(5)}$.
- The same idea extends to the case where we need to choose among many models.
- After you have chosen the model, which of the 5 linear prediction equations do you actually use?
- Answer: none. We refit the model with all available data.

Use k-fold CV for classification

- The basic idea of CV for classification is the same as that for regression, except the way to calculate $CV_{(k)}$
  - *For example, the LOOCV error rate is*

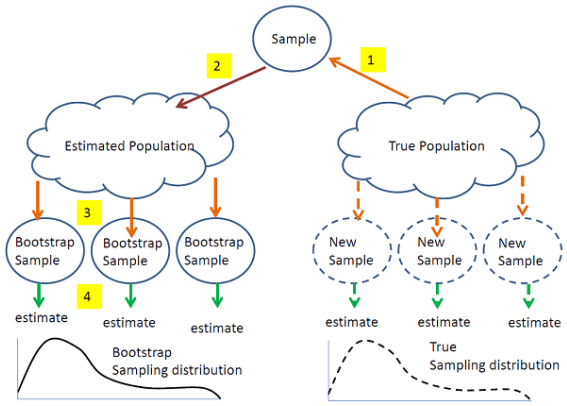
$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} Err_i \,,$$

  - *where* $Err_i = I(y_i \neq \hat{y}_i)$

Bootstrap

Bootstrap in Statistics

- Bootstrap: in statistics, bootstrap is any procedure that relies on random sampling with replacement (from the original sample)
- Bootstrap allows assigning **measures of accuracy** (defined in terms of standard error, variance, confidence intervals, prediction error or some other such measure) to *sample estimates*.
- This is a rather strange idea when we first look at it



Importance of Bootstrap

- On the other hand, it is one of the most influential ideas in Statistics invented in the the second half of the 20th century
- As a verb, bootstrap means "get oneself out of a situation using existing resources (without extra help)" (quote: pull oneself over a fence by one's bootstraps)
- Bootstrap is computationally heavy
- It is a class of widely used procedures. But its theory is beyond this class.
- Bootstrap has different versions (e.g., moving block bootstrap for time series data); we only discuss the simplest kind
- We illustrate the basic bootstrap idea with two toy examples

What is "sample with replacement" (the first example)?

- Suppose there is a box that contains a black ball and a red ball
- Draw one ball from the box, put it back, and then draw another ball from the box; repeat this process multiple times (see different outcomes?)

```
set.seed(1)
sample(c("red","black"), size = 2, replace = TRUE)
```

```
## [1] "red"   "black"
```

```
sample(c("red","black"), size = 2, replace = TRUE)
```

```
## [1] "red" "red"
```

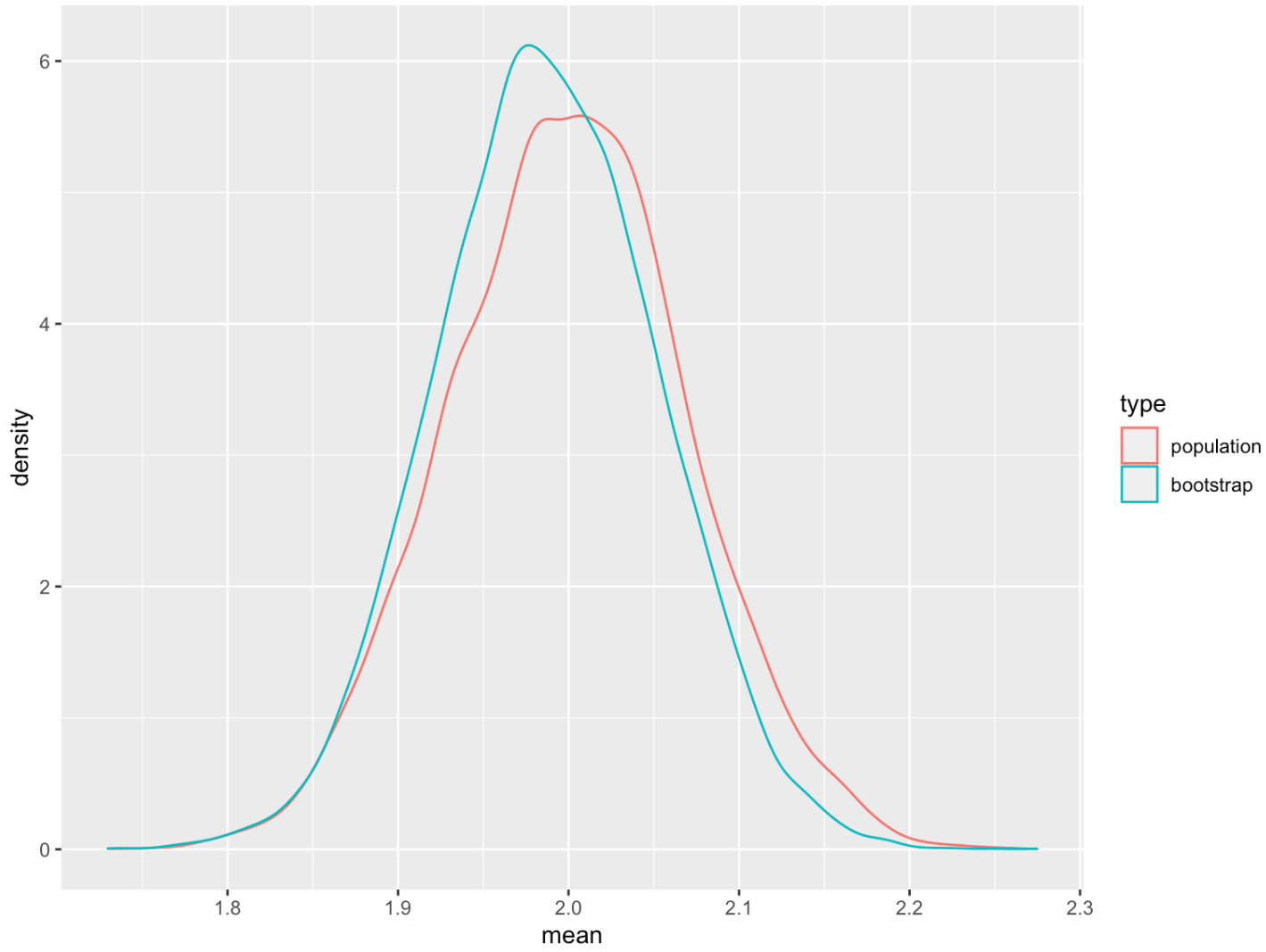
```
sample(c("red","black"), size = 2, replace = TRUE)
```

```
## [1] "black" "red"
```

- When we do sample with replacement, it is possible to get one element twice

- What if we sample two elements without replacement from the above box?

Estimating the distribution of sample mean using a single sample



Next Class

Linear Model Selection and Regularization

- Subset selection
- Ridge regression
- Lasso