

# Statistical Machine Learning

## Part 6

Horacio Gómez-Acevedo  
Department of Biomedical Informatics

February 18, 2021

# Resampling Methods for Parameter Estimation

Suppose you have applied your state of the art algorithm, but you don't know what is the distribution of a parameter (hyperparameter). The question is: how do I determine the bias and variance?

The **jackknife** and **bootstrap** are *resampling* methodologies that help improving classification.

# Jackknife

It was introduced by Maurice Quenouille around 1950's. Let's start with an example as motivation for the use of the jackknife.

**Example.** Let's suppose that we have  $m$  independent random variables  $X_1, \dots, X_m$  that follow the same distribution. We can define the statistic  $\bar{X}$  defined as  $\frac{X_1 + \dots + X_m}{m}$ . The question is what is the standard deviation of this statistic given a set of observed values  $X_1 = x_1, \dots, X_m = x_m$ ?

Following the definition of variance, we can determine

$$\hat{\sigma}^2(\bar{X}) = \frac{1}{m(m-1)} \sum_{i=1}^m (x_i - \bar{x})^2 \quad (1)$$

That was simple enough, but what about calculating an estimate of the variance for other common statistics as *mode*, or *median* or other statistics?

# Jackknife

Let's define the sample average of the data set deleting the  $j$ th variable as

$$\bar{X}_{(j)} = \frac{1}{m-1} \sum_{k \neq j} X_k$$

We also define the statistic that is the *average* of these averages

$$\bar{X}_{(\bullet)} = \frac{1}{m} \sum_{k=1}^m \bar{X}_{(k)}$$

The **Jackknife** estimate of the standard deviation is

$$\hat{\sigma}_{Jack}^2(\bar{X}) = \frac{m-1}{m} \sum_{i=1}^m (\bar{X}_{(i)} - \bar{X}_{(\bullet)})^2 \quad (2)$$

It can be verified that (1) and (2) coincide; however, this process allows a generalization of this method.

# Jackknife

One of the biggest advantages of the expression (2) is that when we have an estimator  $\hat{\theta}(x_1, \dots, x_m)$  of the statistic  $\theta$ , we can actually estimate the variance of such estimator

$$\hat{\sigma}_{jack}^2 = \frac{m-1}{m} \sum_{i=1}^m (\hat{\theta}_{(i)} - \hat{\theta}_{(\bullet)})^2,$$

where

$$\hat{\theta}_{(i)} = \hat{\theta}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_m)$$

$$\hat{\theta}_{(\bullet)} = \frac{1}{m-1} \sum_{i=1}^m \hat{\theta}_{(i)}$$

# Jackknife bias

It is also possible to obtain the **jackknife bias** estimation  
Recall the definition of bias

$$bias = \theta - E(\hat{\theta})$$

The Jackknife estimate of bias is given by

$$bias_{jack} = (m - 1)(\hat{\theta}_{(\bullet)} - \hat{\theta})$$

# Bootstrap

In a common definition, a *bootstrap* data set is one created by randomly selecting  $m$  points (with replacement) from the training set  $\mathcal{D}$ .

For example if our training data set consists of the points  $\{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}$ , then a bootstrap could be

$$B_1 = \{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}$$

$$B_2 = \{(x_1, y_1), (x_1, y_1), (x_2, y_2)\}$$

$$B_3 = \{(x_2, y_2), (x_3, y_3), (x_2, y_2)\}$$

# Bootstrap

The bootstrap was developed by Bradley Efron in the late 1970s. In the bootstrap setup, the data sets (say  $B_j$ s in our example) are treated as independent sets. The **bootstrap** estimate of a statistic  $\theta$  is defined as

$$\hat{\theta}^{*(\bullet)} = \frac{1}{B} \sum_{b=1}^B \hat{\theta}^{*(b)},$$

where  $\hat{\theta}^{*(b)}$  is the estimate of  $\theta$  for the sample  $b$ .



# Bootstrap bias and variance estimates

The bootstrap estimate of the bias

$$bias_{boot} = \hat{\theta}^{*(\bullet)} - \hat{\theta}$$

Whereas the bootstrap estimate of the variance is

$$\hat{\sigma}^2(\theta) = \frac{1}{B} \sum_{b=1}^B \left( \hat{\theta}^{*(b)} - \hat{\theta}^{*(\bullet)} \right)^2$$

# Subset Selection (Regression)

We have seen that when several predictors are present, it is difficult to determine which ones to keep or discard.

There are some alternatives:

- ▶ Best Subset Selection
- ▶ Stepwise Selection
  - ▶ Forward Selection
  - ▶ Backwards Selection

# Best Subset Selection (Regression)

Let's suppose we have a linear regression model

$$Y = \theta_0 + \theta_1 X_1 + \cdots + \theta_p X_p + \varepsilon$$

In theory, we can make (loads) of models

$$\mathcal{M}_0 : \hat{Y} = \hat{\theta}_0$$

$$\mathcal{M}_1 : \hat{Y} = \hat{\theta}_0 + \hat{\theta}_1 X_1$$

$$\vdots$$

$$\mathcal{M}_p : \hat{Y} = \hat{\theta}_0 + \hat{\theta}_1 X_1$$

$$\mathcal{M}_{p+1} : \hat{Y} = \hat{\theta}_0 + \hat{\theta}_1 X_1 + \hat{\theta}_2 X_2$$

$$\vdots$$

$$\mathcal{M}_{2^p-1} : \hat{Y} = \hat{\theta}_0 + \hat{\theta}_1 X_1 + \cdots + \hat{\theta}_p X_p$$

# Best Subset Selection

## Algorithm for Best Subset Selection

1. For  $k \in \{1, \dots, p\}$ :
  - 1.1 Fit all  $\binom{p}{k}$  models that contain exactly  $k$  predictors.
  - 1.2 Pick the best among these  $\binom{p}{k}$  models, and call it  $\widehat{\mathcal{M}}_k$ . The selection is based either by selecting the smallest RSS, or largest  $R^2$ .
2. Select a single best model from among  $\mathcal{M}_0, \widehat{\mathcal{M}}_1, \dots, \widehat{\mathcal{M}}_p$  using cross-validated prediction error,  $C_p$  (AIC), BIC, or adjusted  $R^2$ .

Notice that in step 3 we have changed our metric. If we were to proceed with the same metric (say largest  $R^2$ ), we will end up with the model including all parameters since  $R^2$  increases monotonically towards 1 as the number of predictors increases.

# Forward Stepwise Selection

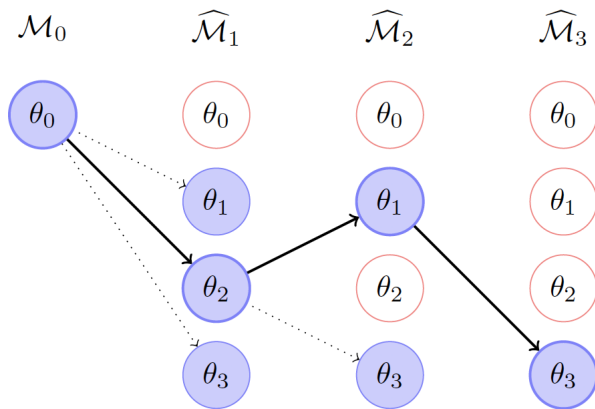
The algorithm for this problem is stated as

1. For  $k = \{0, \dots, p - 1\}$ :
  - 1.1 Consider all  $p - k$  models that augment the predictors in  $\widehat{\mathcal{M}}_k$  with one additional predictor.
  - 1.2 Choose the *best* among these  $p - k$  models and call it  $\widehat{\mathcal{M}}_{k+1}$ .  
The metric is defined as having the smallest RSS or highest  $R^2$ .
2. Select a single best model from among  $\mathcal{M}_0, \widehat{\mathcal{M}}_1, \dots, \widehat{\mathcal{M}}_p$  using cross-validated prediction error  $C_p$  (AIC), BIC, or adjusted  $R^2$ .

So, instead of comparing  $2^p$  models, we will be comparing  $1 + \frac{p(p+1)}{2}$  models.

Note that the forward stepwise tends to do well in practice, it is not guaranteed to find the best possible model!

# Forward Stepwise Selection



# Backwards Stepwise Selection

It requires the same number of steps as the forward selection. The algorithm runs as follows:

1. Let  $\widehat{\mathcal{M}}_p$  denote the full model with all predictors.
2. For  $k = p, p - 1, \dots, 1$ :
  - ▶ Consider all  $k$  models that contain all but one of the predictors in  $\widehat{\mathcal{M}}_k$ , for a total of  $k - 1$  predictors.
  - ▶ Choose the *best* among these  $k$  models, and call it  $\widehat{\mathcal{M}}_{k-1}$ . Again, we consider one of the metrics such as the smallest RSS or largest  $R^2$ .
3. Select a single best model from among  $\mathcal{M}_0, \widehat{\mathcal{M}}_1, \dots, \widehat{\mathcal{M}}_p$  using cross-validated prediction error,  $C_p$  (AIC), BIC or adjusted  $R^2$ .

## Selection criteria

Let's suppose we have fitted a model containing  $d$  predictor, the  $C_p$  estimate of test MSE is computed using the equation

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2),$$

where  $\hat{\sigma}^2$  is an estimate of the variance of the error  $\varepsilon$  associated with each response measurement.  $C_p$  is an unbiased estimate of test MSE. Thus, we choose the model with the lowest  $C_p$  value. The Akaike information criteria (AIC) is defined for a large class of models fit by maximum likelihood.

$$AIC \approx \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

The Bayesian information criteria (BIC) is derived from a Bayesian point of view

$$BIC \approx \frac{1}{n}(RSS + \log(n)d\hat{\sigma}^2)$$



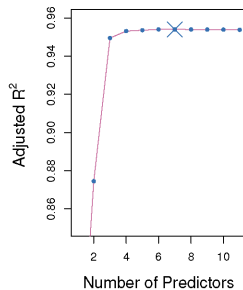
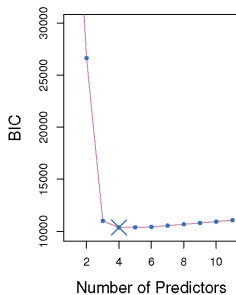
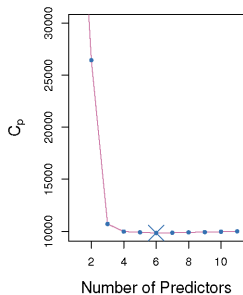
## Adjusted $R^2$

Finally, the so-called **adjusted**  $R^2$  statistic is defined as

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n - d - 1)}{TSS/(n - 1)} = 1 - \frac{n - 1}{n - d - 1} \cdot \frac{RSS}{TSS}$$

Recall that  $R^2 = 1 - RSS/TSS$  where  $TSS = \sum (y_i - \bar{y})^2$  is the total sum of squares. So adding the adjusted  $R^2$  statistic pays a price for the inclusion of unnecessary variables in the model.

# Optimal Selection



# Final Thoughts

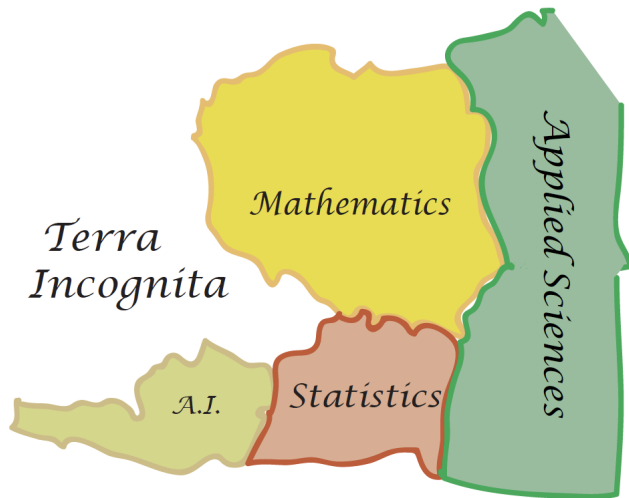


FIGURE 1. The greater world of mathematics and science.

# References

Materials and some of the pictures are from (1),(2), and (3).

1. Gareth James et al. *An Introduction to Statistical Learning with applications in R*. Springer (2015)
2. Richard O. Duda et al. *Pattern Classification* John Wiley (2001).
3. Aurélien Géron. *Hands-on Machine Learning with Scikit-Learn & TensorFlow* O'Reilly (2017)
4. Wiebe R. Pestman *Mathematical Statistics* de Gruyter (1998)
5. Bradley Efron. *The Jackknife, the Bootstrap and other Resampling Plans* SIAM (1982)
6. Bradley Efron *A 250-year argument: Belief, behavior and the bootstrap* Bull. Am. Math. Soc (2012)

I have used some of the graphs by hacking TiKz code from StakExchange, Inkscape for more aesthetic plots and other old tricks of  $\text{\TeX}$