

# Statistics - Week 6

JIMMY TSZ MING YUE\*

University of Sydney  
jyue6728@uni.sydney.edu.au

Semester 2 Statistics

2018

## Contents

<b>1</b>	<b>Training error Vs Test error</b>	<b>1</b>
1.1	Validation Set Approach . . . . .	1
1.2	Drawbacks of validation set approach . . . . .	2
1.3	Overall Classification accuracy rate . . . . .	2
1.4	Confusion Matrix . . . . .	3
1.5	Sensitivity and Specificity . . . . .	3
<b>2</b>	<b>The Bootstrap</b>	<b>3</b>
2.1	Bootstrap . . . . .	3
2.2	A simple example . . . . .	4
2.3	The real world situation . . . . .	4
2.4	Generic Bootstrap Procedure . . . . .	5
<b>3</b>	<b>Tutorial</b>	<b>5</b>

## 1 Training error Vs Test error

**Definition. 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.

**Definition. Training Error** is calculated by applying statistical learning to the observations used in its training

Often, training error rate and test error rates are quite different, in particular the training error can dramatically underestimate the test error rate.

### 1.1 Validation Set Approach

To minimise these errors, the best solution is to take a large designated test set. But this is not always possible, as such we need alternative methods such as BIC, which makes mathematical adjustments to the training error rate in order to estimate the test error rate, which is the subject of later discussion. However there are other methods,

**Validation Set Approach** which comprises of considering a class of methods that estimate the test error by holding out a subset of the training observations from the fitted process, and

---

\*440159151

then applying the statistical learning method to those held out observations. We first randomly divide the available set of samples into two parts: a training set and a validation or hold-out set. The model is then 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 in the case of a qualitative (discrete) response.

## 1.2 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 (the ones included in the training 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 data set. As such the widely used approach for estimating test error is:

**K - Fold cross validation** 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. 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.

We provide the formulation:

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  $n_k = n/K$ . Then compute:

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

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)

We can extend this for classification problems:

**Cross - Validation for classification** We divide the data into  $K$  roughly equal-sized parts  $C_1, \dots, C_K$ .  $C_k$  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  $n_k = n/K$ . Then let us compute:

$$CV(K) = \sum_{k=1}^K \frac{n_k}{n} \text{Err}_k \quad (2)$$

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

Then the estimated standard deviation of  $CV_k$  is:

$$\hat{\text{SE}}(CV_K) = \sqrt{\sum_{k=1}^K (\text{Err}_k - \bar{\text{Err}})^2 / (K - 1)} \quad (3)$$

## 1.3 Overall Classification accuracy rate

$$\text{ACC} = 1 - \frac{1}{N} \sum_{i=1}^N I(y_i \neq \hat{y}_i) \quad (4)$$

There are several disadvantages:

1. Makes no distinctions about the type of errors being made. In spam filtering, the cost of erroneously deleting an important email is likely to be higher than incorrectly allowing a spam email past a filter.
2. Does not consider the natural frequencies of each class.

## 1.4 Confusion Matrix

## 1.5 Sensitivity and Specificity

**Definition. Accuracy**

$$ACC = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (5)$$

**Definition. Sensitivity**

$$Sen = \frac{TP}{(TP + FN)} \quad (6)$$

**Definition. Specificity**

$$Spe = \frac{TN}{(TN + FP)} \quad (7)$$

**Definition.  $F_1$**

$$F_1 = \frac{2TP}{2TP + FP + FN} \quad (8)$$

which is the harmonic mean

**Definition. GM**

$$GM = \sqrt{\frac{TP}{TP + FN} \frac{TN}{TN + FP}} \quad (9)$$

## 2 The Bootstrap

### 2.1 Bootstrap

**Definition. 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.

For example it can provide an estimate of the standard error of a coefficient, or a confidence interval for that coefficient.

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad (10)$$

## 2.2 A simple example

Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of  $X$  and  $Y$ , respectively where  $X$  and  $Y$  are random quantities. We will invest a fraction  $\alpha$  of our money in  $X$ , and will invest the remaining  $1 - \alpha$  in  $Y$ . We wish to choose  $\alpha$  to minimise the total risk or variance, of our investment. In other words, we want to minimise  $\text{Var}(\alpha X + (1 - \alpha)Y)$ . We can show that the value that minimises this risk is subsequently given by;

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}} \quad (11)$$

where  $\sigma_X^2 = \text{Var}(X)$ ,  $\sigma_Y^2 = \text{Var}(Y)$  and  $\sigma_{XY} = \text{Cov}(X, Y)$ .

But the values of  $\sigma_X^2$ ,  $\sigma_Y^2$  and  $\sigma_{XY}$  are unknown. We can compute estimates for these quantities;  $\hat{\sigma}_X^2$ ,  $\hat{\sigma}_Y^2$  and  $\hat{\sigma}_{XY}$ , using a data set that contains measurements for  $X$  and  $Y$ . We can then estimate the value of  $\alpha$  that minimises the variance of our investment using:

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}} \quad (12)$$

To estimate the standard deviation of  $\hat{\alpha}$ , we repeated the process of simulating 100 paired observations of  $X$  and  $Y$ , and estimating  $\alpha$  1000 times. We thereby obtained 1000 estimates for  $\alpha$ , which we can call  $\hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_{1000}$ . The mean over all 1000 estimates for  $\alpha$  is:

$$\bar{\alpha} = \frac{1}{1000} \sum_{r=1}^{1000} \hat{\alpha}_r = 0.5996 \quad (13)$$

very close to  $\alpha = 0.6$ , and the standard deviation of the estimates is:

$$\sqrt{\frac{1}{1000 - 1} \sum_{r=1}^{1000} (\hat{\alpha}_r - \bar{\alpha})^2} = 0.083 \quad (14)$$

This gives us a very good idea of the accuracy of  $\hat{\alpha}$ :  $\text{SE}(\hat{\alpha}) \approx 0.083$ . So roughly speaking, for a random sample from the population, we would expect  $\hat{\alpha}$  to differ from  $\alpha$  by approximately 0.08, on average.

## 2.3 The real world situation

The procedure outlined above cannot be applied, because for real data we cannot generate new samples from the original population.

However, the bootstrap approach allows us to use a computer to mimic the 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.

## 2.4 Generic Bootstrap Procedure

Denoting the first bootstrap data set by  $Z^{*1}$  we use  $Z^{*1}$  to produce a new bootstrap estimate for  $\alpha$ , which we call  $\hat{\alpha}^{*1}$ . This procedure is repeated  $B$  times for some large value of  $B$  (say 100 or 1000), in order to produce  $B$  different bootstrap data sets  $Z^{*1}, Z^{*2}, \dots, Z^{*B}$  and  $B$  different corresponding  $\alpha$  estimates;  $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \dots, \hat{\alpha}^{*B}$ . We estimate the standard error of these bootstrap estimates using the formula,

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1} \sum_{r=1}^B (\alpha^{*r} - \bar{\alpha}^*)^2} \quad (15)$$

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

## 3 Tutorial

The “breast.txt” dataset contains benign and malignant breast tumour samples. Each sample is measured by various factors including:

1. Clump Thickness
2. Uniformity of Cell Size
3. Uniformity of Cell Shape
4. Marginal Adhesion
5. Single Epithelial Cell Size
6. Bare Nuclei
7. Bland Chromatin
8. Normal Nucleoli
9. Mitoses

The last column contains class levels with “1” being malignant and “0” as benign:

1. Download the breast dataset file from the course webpage

**Solution.** We first import the dataset

```
> breast <- read.table("breast.txt", sep = "\t", header = TRUE)
> df <- data.frame(breast)
> Column.names <- c("Clump_Thickness", "Uniformity_Size", "Uniformity_Shape",
+ "Marginal_Adhesion", "Single_Ep_Size", "Bare_Nuclei", "Bland_Chromatin",
+ "Normal_Nucleoli", "Mitoses")
> colnames(df) <- Column.names
>
```

□

2. Design a 10-fold cross validation procedure to evaluate logistic regression and  $k$ -NN  $k = 3$  classification accuracy.

**| Solution.**

□

3. Calculate TP, TN, FP, FN and compute sensitivity and specificity for each classifier
4. Compute F1 score and compare the two classifiers.

>