

## 0.1 Cross Validation

- Cross-validation is primarily a way of measuring the predictive performance of a statistical model. Every statistician knows that the model fit statistics are not a good guide to how well a model will predict: high  $R^2$  does not necessarily mean a good model. It is easy to over-fit the data by including too many degrees of freedom and so inflate  $R^2$  and other fit statistics. For example, in a simple polynomial regression I can just keep adding higher order terms and so get better and better fits to the data. But the predictions from the model on new data will usually get worse as higher order terms are added.
- One way to measure the predictive ability of a model is to test it on a set of data not used in estimation. Data miners call this a test set and the data used for estimation is the training set. For example, the predictive accuracy of a model can be measured by the mean squared error on the test set. This will generally be larger than the MSE on the training set because the test data were not used for estimation.
- However, there is often not enough data to allow some of it to be kept back for testing. A more sophisticated version of training/test sets is **leave-one-out cross-validation (LOOCV)** in which the accuracy measures are obtained as follows. Suppose there are  $n$  independent observations,  $y_1, \dots, y_n$ .
- Let observation  $i$  form the test set, and fit the model using the remaining data. Then compute the error ( $e_i^* = y_i - \hat{y}_i$ ) for the omitted observation. This is sometimes called a predicted residual to distinguish it from an ordinary residual. Repeat step 1 for  $i = 1, \dots, n$ . Compute the MSE from  $e_1^*, \dots, e_n^*$ . We shall call this the CV.
- This is a much more efficient use of the available data, as you only omit one observation at each step. However, it can be very time consuming to implement (except for linear models see below).
- Other statistics (e.g., the MAE) can be computed similarly. A related measure is the PRESS statistic (predicted residual sum of squares) equal to  $n \times MSE$ .

### 0.1.1 variations

Minimizing a CV statistic is a useful way to do model selection such as choosing variables in a regression or choosing the degrees of freedom of a nonparametric smoother. It is certainly far better than procedures based on statistical tests and provides a nearly unbiased measure of the true MSE on new observations.

However, as with any variable selection procedure, it can be misused. Beware of looking at statistical tests after selecting variables using cross-validation the tests do not take account of the variable selection that has taken place and so the p-values can mislead.

It is also important to realise that it doesn't always work. For example, if there are exact duplicate observations (i.e., two or more observations with equal values for all covariates and for the  $y$  variable) then leaving one observation out will not be effective.

Another problem is that a small change in the data can cause a large change in the model selected. Many authors have found that  $k$ -fold cross-validation works better in this respect.

In a famous paper, Shao (1993) showed that leave-one-out cross validation does not lead to a consistent estimate of the model. That is, if there is a true model, then LOOCV will not always find it, even with very large sample sizes. In contrast, certain kinds of leave- $k$ -out cross-validation, where  $k$  increases with  $n$ , will be consistent. Frankly, I don't consider this is a very important result as there is never a true model. In reality, every model is wrong, so consistency is not really an interesting property.

### 0.1.2 Cross Validation

Cross-validation, sometimes called rotation estimation, is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set.

- It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. It is worth highlighting that in a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (testing dataset).
- The goal of cross validation is to define a dataset to "test" the model in the training phase (i.e., the validation dataset), in order to limit problems like overfitting, give an insight on how the model will generalize to an independent data set (i.e., an unknown dataset, for instance from a real problem), etc.

### 0.1.3 Cross Validation

The confusion table is a table in which the rows are the observed categories of the dependent and the columns are the predicted categories. When prediction is perfect all cases will lie on the diagonal. The percentage of cases on the diagonal is the percentage of correct classifications. The cross validated set of data is a more honest presentation of the power of the discriminant function than that provided by the original classifications and often produces a poorer outcome. The cross validation is often termed a jack-knife classification, in that it successively classifies **all cases but one** to develop a discriminant function and then categorizes the case that was left out. This process is repeated with each case left out in turn. This is known as leave-1-out cross validation.

This cross validation produces a more reliable function. The argument behind it is that one should not use the case you are trying to predict as part of the categorization process.

### 0.1.4 Error Rates

We can evaluate error rates by means of a training sample (to construct the discrimination rule) and a test sample.

An optimistic error rate is obtained by reclassifying the training data. (In the *training data* sets, how many cases were misclassified). This is known as the **apparent error rate**.

The apparent error rate is obtained by using in the training set to estimate the error rates. It can be severely optimistically biased, particularly for complex classifiers, and in the presence of over-fitted models.

If an independent test sample is used for classifying, we arrive at the **true error rate**. The true error rate (or conditional error rate) of a classifier is the expected probability of misclassifying a randomly selected pattern. It is the error rate of an infinitely large test set drawn from the same distribution as the training data.

### 0.1.5 Misclassification Cost

As in all statistical procedures it is helpful to use diagnostic procedures to assess the efficacy of the discriminant analysis. We use **cross-validation** to assess the classification probability. Typically you are going to have some prior rule as to what is an **acceptable misclassification rate**.

Those rules might involve things like, "what is the cost of misclassification?" Consider a medical study where you might be able to diagnose cancer.

There are really two alternative costs. The cost of misclassifying someone as having cancer when they don't. This could cause a certain amount of emotional grief. Additionally there would be the substantial cost of unnecessary treatment.

There is also the alternative cost of misclassifying someone as not having cancer when in fact they do have it.

A good classification procedure should

- result in few misclassifications
- take *prior probabilities of occurrence* into account
- consider the cost of misclassification

For example, suppose there tend to be more financially sound firms than bankrupt firm. If we really believe that the prior probability of a financially distressed and ultimately bankrupted firm is very small, then one should classify a randomly selected firm as non-bankrupt unless the data overwhelmingly favor bankruptcy.

There are two costs associated with discriminant analysis classification: The true misclassification cost per class, and the expected misclassification cost (ECM) per observation.

Suppose there we have a binary classification system, with two classes: class 1 and class 2. Suppose that classifying a class 1 object as belonging to class 2 represents a more serious error than classifying a class 2 object as belonging to class 1. There would an assignable cost to each error.  $c(i|j)$  is the cost of classifying an observation into class  $j$  if its true class is  $i$ . The costs of misclassification can be defined by a cost matrix.

	Predicted Class 1	Predicted Class 2
Class 1	0	$c(2 1)$
Class 2	$c(1 2)$	0

### 0.1.6 Expected cost of misclassification (ECM)

Let  $p_1$  and  $p_2$  be the prior probability of class 1 and class 2 respectively. Necessarily  $p_1 + p_2 = 1$ .

The conditional probability of classifying an object as class 1 when it is in fact from class 2 is denoted  $p(1|2)$ . Similarly the conditional probability of classifying an object as class 2 when it is in fact from class 1 is denoted  $p(2|1)$ .

$$ECM = c(2|1)p(2|1)p_1 + c(1|2)p(1|2)p_2$$

(In other words: the sum of the cost of misclassification times the (joint) probability of that misclassification.

A reasonable classification rule should have ECM as small as possible.

## 0.2 Training data sets

A training set is a set of data used in various areas of information science to discover potentially predictive relationships. Training sets are used in artificial intelligence, machine learning, genetic programming, intelligent systems, and statistics. In all these fields, a training set has much the same role and is often used in conjunction with a test set.

### 0.3 Cross Validation

Cross validation techniques for linear regression employ the use ‘leave one out’ re-calculations. In such procedures the regression coefficients are estimated for  $n - 1$  covariates, with the  $Q^{th}$  observation omitted.

Let  $\hat{\beta}$  denote the least square estimate of  $\beta$  based upon the full set of observations, and let  $\hat{\beta}^{-Q}$  denoted the estimate with the  $Q^{th}$  case excluded.

In leave-one-out cross validation, each observation is omitted in turn, and a regression model is fitted on the rest of the data. Cross validation is used to estimate the generalization error of a given model. alternatively it can be used for model selection by determining the candidate model that has the smallest generalization error.

Evidently leave-one-out cross validation has similarities with ‘jackknifing’, a well known statistical technique. However cross validation is used to estimate generalization error, whereas the jackknife technique is used to estimate bias.

#### 0.3.1 Cross Validation: Updating standard deviation

The variance of a data set can be calculated using the following formula.

$$S^2 = \frac{\sum_{i=1}^n (x_i^2) - \frac{(\sum_{i=1}^n x_i)^2}{n}}{n - 1} \quad (1)$$

While using bivariate data, the notation  $Sxx$  and  $Syy$  shall apply to the variance of  $x$  and of  $y$  respectively. The covariance term  $Sxy$  is given by

$$Sxy = \frac{\sum_{i=1}^n (x_i y_i) - \frac{(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n}}{n - 1} \quad (2)$$

Let the observation  $j$  be omitted from the data set. The estimates for the variance identities can be updating using minor adjustments to the full sample estimates. Where  $(j)$  denotes that the  $j$ th has been omitted, these identities are

$$Sxx^{(j)} = \frac{\sum_{i=1}^n (x_i^2) - (x_j)^2 - \frac{((\sum_{i=1}^n x_i) - x_j)^2}{n-1}}{n - 2} \quad (3)$$

$$Syy^{(j)} = \frac{\sum_{i=1}^n (y_i^2) - (y_j)^2 - \frac{((\sum_{i=1}^n y_i) - y_j)^2}{n-1}}{n - 2} \quad (4)$$

$$Sxy^{(j)} = \frac{\sum_{i=1}^n (x_i y_i) - (y_j x_j) - \frac{((\sum_{i=1}^n x_i) - x_j)((\sum_{i=1}^n y_i) - y_j)}{n-1}}{n - 2} \quad (5)$$

The updated estimate for the slope is therefore

$$\hat{\beta}_1^{(j)} = \frac{Sxy^{(j)}}{Sxx^{(j)}} \quad (6)$$

It is necessary to determine the mean for  $x$  and  $y$  of the remaining  $n - 1$  terms

$$\bar{x}^{(j)} = \frac{(\sum_{i=1}^n x_i) - (x_j)}{n - 1}, \quad (7)$$

$$\bar{y}^{(j)} = \frac{(\sum_{i=1}^n y_i) - (y_j)}{n - 1}. \quad (8)$$

The updated intercept estimate is therefore

$$\hat{\beta}_0^{(j)} = \bar{y}^{(j)} - \hat{\beta}_1^{(j)} \bar{x}^{(j)}. \quad (9)$$

## 0.4 Updating Estimates

### 0.4.1 Updating of Regression Estimates

Updating techniques are used in regression analysis to add or delete rows from a model, allowing the analyst the effect of the observation associated with that row. In time series problems, there will be scientific interest in the changing relationship between variables. In cases where there a single row is to be added or deleted, the procedure used is equivalent to a geometric rotation of a plane.

Updating techniques are used in regression analysis to add or delete rows from a model, allowing the analyst the effect of the observation associated with that row.

### 0.4.2 Updating Standard deviation

A simple, but useful, example of updating is the updating of the standard deviation when an observation is omitted, as practised in statistical process control analyzes. From first principles, the variance of a data set can be calculated using the following formula.

$$S^2 = \frac{\sum_{i=1}^n (x_i^2) - \frac{(\sum_{i=1}^n x_i)^2}{n}}{n-1} \quad (10)$$

While using bivariate data, the notation  $Sxx$  and  $Syy$  shall apply hither to the variance of  $x$  and of  $y$  respectively. The covariance term  $Sxy$  is given by

$$Sxy = \frac{\sum_{i=1}^n (x_i y_i) - \frac{(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n}}{n-1}. \quad (11)$$

### 0.4.3 Updating of Regression Estimates

Updating techniques are used in regression analysis to add or delete rows from a model, allowing the analyst the effect of the observation associated with that row. In time series problems, there will be scientific interest in the changing relationship between variables. In cases where there a single row is to be added or deleted, the procedure used is equivalent to a geometric rotation of a plane.

$$(X^T X \pm x_i x_i^T)^{-1} = (X^T X)^{-1} \mp \frac{(X^T X)^{-1} (x_i x_i^T (X^T X)^{-1})}{1 - x_i^T (X^T X)^{-1} x_i} \quad (12)$$

### 0.4.4 Updating Regression Estimates

Let the observation  $j$  be omitted from the data set. The estimates for the variance identities can be updating using minor adjustments to the full sample estimates. Where  $(j)$  denotes that the  $j$ th has been omitted, these identities are

$$Sxx^{(j)} = \frac{\sum_{i=1}^n (x_i^2) - (x_j)^2 - \frac{((\sum_{i=1}^n x_i) - x_j)^2}{n-1}}{n-2} \quad (13)$$

$$Syy^{(j)} = \frac{\sum_{i=1}^n (y_i^2) - (y_j)^2 - \frac{((\sum_{i=1}^n y_i) - y_j)^2}{n-1}}{n-2} \quad (14)$$

$$Sxy^{(j)} = \frac{\sum_{i=1}^n (x_i y_i) - (y_j x_j) - \frac{((\sum_{i=1}^n x_i) - x_j)(\sum_{i=1}^n y_i) - y_k}{n-1}}{n-2} \quad (15)$$

The updated estimate for the slope is therefore

$$\hat{\beta}_1^{(j)} = \frac{Sxy^{(j)}}{Sxx^{(j)}} \quad (16)$$

It is necessary to determine the mean for  $x$  and  $y$  of the remaining  $n-1$  terms

$$\bar{x}^{(j)} = \frac{(\sum_{i=1}^n x_i) - (x_j)}{n-1}, \quad (17)$$

$$\bar{y}^{(j)} = \frac{(\sum_{i=1}^n y_i) - (y_j)}{n-1}. \quad (18)$$

The updated intercept estimate is therefore

$$\hat{\beta}_0^{(j)} = \bar{y}^{(j)} - \hat{\beta}_1^{(j)} \bar{x}^{(j)}. \quad (19)$$

#### 0.4.5 Inference on intercept and slope

$$\hat{\beta}_1 \pm t_{(\alpha, n-2)} \sqrt{\frac{S^2}{(n-1)S_x^2}} \quad (20)$$

$$\frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} \quad (21)$$

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_0)} \quad (22)$$

#### Inference on correlation coefficient

This test of the slope is coincidentally the equivalent of a test of the correlation of the  $n$  observations of  $X$  and  $Y$ .

$$\begin{aligned} H_0 : \rho_{XY} &= 0 \\ H_A : \rho_{XY} &\neq 0 \end{aligned} \quad (23)$$