Lecture 101

Date: 27 March 2020
Scribes: Akash Gupta
Topics: CV and Bootstrap

0.1 Resampling : Cross Validation

Resampling is a method by which we repeatedly take samples from a training dataset and with each sample we fit a model. With this method we will be able to extract new information was hitherto unavailable. We can now get an idea about the variability of the model fit since the model fit would be different with each different sample. **Cross validation** helps us estimate the test error rate associated with a model so that we can evaluate its performance or select the correct amount of flexibility. Evaluating the performance of a model is called **model assessment** and selecting the correct amount of flexibility (degree of closeness of fitting) is called **model selection**.

Remember that test error rate tells about the average error in predicting a new observation through our model. In these concepts we are interested in figuring out the test error rate, which is hard to find since we usually don't have enough observations to get an accurate estimate of test error rate. So we apply various procedures like :- removing a subset of observations from the training set to train the model and then using the model on that subset to get an idea about the test error rate. Other methods apply mathematical corrections to the training error rate to obtain the test error rate. Typically a model is fit on the training set and evaluated on the validation set and both subsets of data belong to a set of decided observations. MSE is used as a measure of validation set error. Then the split of data into training and validation sets is done many times using different observations and consequently MSE is computed. When we do this, we can obtain the variability in test MSE.

0.1.1 LOOCV

In this method the set of observations is divided into two subsets of a single observation (x_1, y_1) and a subset of $n-1$ observations $(x_1, y_1), (x_2, y_2), \cdots, (x_{n-1}, y_{n-1})$. We then fit our model on the $n-1$ observations and use that to make a prediction on the left out observation using x_1. We can then obtain MSE as : $(y_i - \hat{y}_i)^2$. Then we repeat this process many times and each time leaving out a different observation and in a similar manner compute the corresponding MSE. After this we can obtain the LOOCV estimate for the test MSE is the average of all MSEs obtained :-

$$CV_n = \frac{1}{n}\sum_{i=1}^{n} MSE_i \tag{1}$$

0.1.2 K-fold cross validation

In this method, we divide our observations into k groups of folds of approximately equal size. The first fold is often considered as the validation set and the model is fit on the remaining $k-1$ folds. After this the model is tested on the left out fold to obtain the first MSE. In this manner, the splitting of the data into folds is repeated k times where a different fold is treated as the validation fold and hence we obtain k estimates of MSE :- $MSE_1, MSE_2, \cdots, MSE_k$. After this we estimate the CV estimate of test error rate by averaging these MSE values :-

$$CV_k = \frac{1}{k}\sum_{i=1}^{k} MSE_i \tag{2}$$

Note that we often interested in the minimum point of the MSE curve in the MSE vs. Flexibility graph. This lowest point of the MSE curve is a good point of telling us the desired level of flexibility for our model. Typically k is chosen as 5 or 10 such that an optimal level of bias variance trade off is maintained.

0.1.3 Cross validation in classification

In this situation instead of looking at MSE to quantify the test error rate, we use the number of misclassified observations. As an example, the LOOCV error rate looks like :-

$$CV_n = \frac{1}{n}\sum_{i=1}^{n} Err_i \tag{3}$$

$$Err_i = I(y_i \neq \hat{y}_i) \tag{4}$$

Note that in equation 4, I is an indicator function which assigns 1 to Err if $y_i \neq \hat{y}_i$ and assigns 0 if $y_i = \hat{y}_i$. The latter condition implies correction classification and the former suggests incorrect classification and when we divide the sum of these vales by n we get the proportion of incorrectly classified values. We notice that if the actual Bayes decision boundary is given we can check the accuracy of our model by comparing the true error rate with CV error rate.

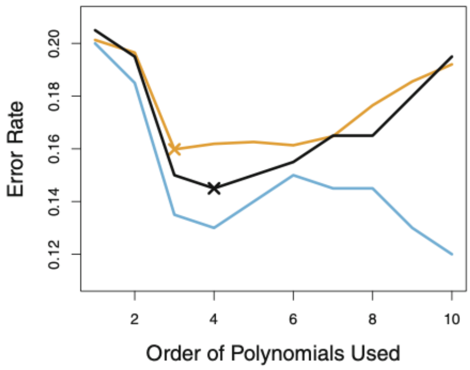


Figure 1: cross validation error rate

We can ultimately see that the CV error (Black line) rate is not too far behind the true error (Orange line) rate. In case we are computing the test MSE using k-folds approach, we will apply the CV operator with the indicator variable to the left out fold and compare it with the true error rate if given. Additionally if we want to compare models of different flexibilities we can see the model for which the estimated MSE (CV error rate) is minimum. Here we can see that the model with 4 order nonlinear equation (degree of flexibility) has the least estimated error rate.

0.2 Bootstrap

Bootstrap is a method used to estimate the uncertainty associated with our model or estimator. For example it can be used to estimate the standard errors of the coefficients of regression line. This general process can be applied to any model to determine standard errors associated with its parameters. Before we jump into bootstrap technique, we would show a toy example of how accuracy of an estimator is arrived at using an approach that is somewhat similar to what bootstrap does.

- Suppose we have some money that we can invest in two assets that give returns X or Y. Note that X and Y are random variables.

- Invests α of your money in X and $1-\alpha$ in Y.

- In essence because returns on both assets are uncertain and subject to volatility or variability, we, as investors would naturally like to minimize that variability in our portfolio returns. So we would carefully choose α such that we minimize the expression :-

$$Var(\alpha X + (1-\alpha)Y) \tag{5}$$

- In general the formula for the minimum value of α is given by :-

$$\alpha = \frac{\sigma_y^2 - \sigma_{xy}}{\sigma_x^2 + \sigma_y^2 - 2\sigma_{xy}} \tag{6}$$

- Since in reality the variances and covariances are unknown we compute estimates of these values using samples and put them in equation 6 to obtain the estimator for α :-

$$\hat{\alpha} = \frac{\hat{\sigma}_y^2 - \hat{\sigma}_{xy}}{\hat{\sigma}_x^2 + \hat{\sigma}_y^2 - 2\hat{\sigma}_{xy}} \tag{7}$$

- We can easily simulate 100 pairs of X and Y returns, compute the sample variances and get an estimate of α.

- If we simulate data like this 1000 times and get 1000 estimates of α we would see that our values of α range from 0.53 to 0.65.

- If we want to know the accuracy of our estimate for α we need to find out the standard error.

- Mean of our estimate is :-

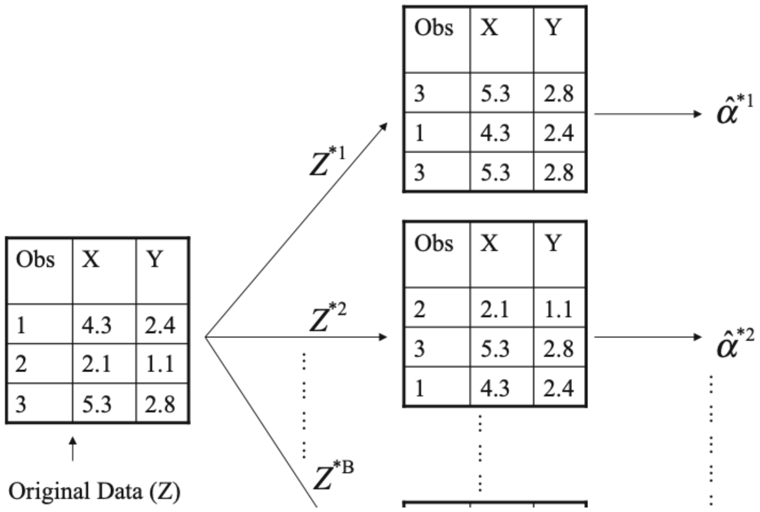
$$\frac{1}{1000}\sum_{i=1}^{1000}\hat{\alpha}_i = \bar{\alpha} = 0.599 \tag{8}$$

- the True value of α was found to be 0.6 which is quite close to our estimate.
- Finally we can compute the standard error as well :-

$$\sqrt{\frac{1}{1000-1}\sum_{i=1}^{1000}(\hat{\alpha}_i - \bar{\alpha})^2} = 0.083 \tag{9}$$

- This gives us a good idea about the accuracy of our estimator since we can say that on average our estimator deviates by the true value by almost 0.08.

This was a hassle free method to find the estimate of model accuracy because we could generate many samples from the original population which were IID. However in reality we may not have the luxury of generating samples. This is where bootstrap comes in and helps us to emulate the process of obtaining new sample sets so that the variability of our estimator can be computed. It tries to obtain distinct datasets or samples by repeatedly sampling observations from the same original data set. We will now explain how exactly this works :- This is typically used when have a small number of observations, in this case only 3. This method works in the following way :-



Obs	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

Obs	X	Y
3	5.3	2.8
1	4.3	2.4
3	5.3	2.8

Obs	X	Y
2	2.1	1.1
3	5.3	2.8
1	4.3	2.4

Figure 2: bootstrap

- Randomly select or sample (with replacement) values from the original dataset Z in order to obtain the first bootstrapped dataset Z^{*1}.
- Notice that in this the third observation is contained twice and the second observation is not there. Such sampling is allowed as long as the observations are consistent for X and Y pairs.

- We can then use this bootstrap data to get an estimate of α which is called $\hat{\alpha}^{*1}$.

- this process is repeated B times where B is a large number so that we can obtain B bootstrapped datasets :- $Z^{*1}, Z^{*2}, \cdots, Z^{*B}$ and their corresponding estimates for α as :- $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \cdots, \hat{\alpha}^{*B}$.

- Finally, standard error of these bootstrap estimates can be obtained as follows :-

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{i=1}^{B}\hat{\alpha}^{*i}\right)^2} \tag{10}$$

- We see that the bootstrap method generates a standard error for our estimate of 0.087 which is very close to the estimate of 0.083 obtained using 1000 simulated datasets of 100 observations.

- Hence this process can be used effectively to estimate the standard error of model parameters in case n is less or even generally.