Cross Validation Method

Matheus Morroni

January 2020

1 Introduction

Validation can be defined as "process of deciding whether the numerical results quantifying hypothesized relationships between variables, are acceptable as descriptions of the data". In several cases, an model error estimation is made after training, this method is called evaluation of residuals. In this method, a numerical estimate of the difference in predicted and original responses is calculated, this process is also called training error.

However, this only gives to the professional an idea about how well his model does on data used to train it. Now its possible that the model is underfitting or overfitting the data. So, the issue with this evaluation technique is that the process does give an indication of how well the learner will generalize to an independent/unseen data set. Getting this idea about her model is called cross validation.

Exist many different processes under the term Cross Validation Method, ahead these processes are presented.

2 Holdout Method

The basic process to solve this problema involves removing a part of the training data and using it to get predictions from the algorithm trained on rest of the data. The error estimation then tells how the model is doing on unseen data or the validation group. This is a simple kind of cross validation method, and it is called the holdout method.

Although this process does not take any overhead to compute and is better than traditional validation, it still suffers from some issues of high variance. The reason is it is not certain which data points will end up in the validation group and the result might be totally different for different groups tested.

3 K-Fold Cross Validation Method

K-Fold Cross Validation is a method that provides ample data for training the model and also has a huge data for validation. This kind of approach is important because if the professional does not have enough data to train her model, removing a part of it for validation can imply a problem of underfitting. By reducing the training data, she risks losing important patterns or trends in data group, which in turn increases error induced by bias. So when he utilises the K-Fold Cross Validation method he reduces these risks.

In K-Fold Cross Validation, the data is separated into k subsets. Now the Holdout method is repeated k times, such that each time, one of the k subsets is used as the validation group and the one of the rest subsets are put together to form a training set. The error estimation is averaged over all k trials to get the total effectiveness of the model. This process implies that every data point gets to be in a validation group exactly once, and gets to be in a training

group k-1 times.

This significantly reduces bias as the professional is using most of the data for fitting, and also significantly reduces variance as most of the data is also being used in validation group. Interchanging the training and validation groups also adds to the effectiveness of this process. As an important rule and empirical evidence, k equals 5 or 10 is generally preferred, but nothing is determined and it can take any value.

4 Stratified K-Fold Cross Validation Method

To avoid a large imbalance in the response variables, likewise, in classification, there might be several times more negative samples than positive samples, it was created a slight variation of K-Fold Cross Validation called Stratified K-Fold Cross Validation where each fold contains approximately the same ratio of samples of each target class as the complete set, or in case of prediction problems, the mean response value is approximately equal in all the folds.

5 Leave-P-Out Method and Exhaustive Methods

The Cross Validations methods explained earlier are also classify as non-exhaustive cross validation methods. These do not compute all ways of splitting the original sample, i.e. the professional has to decide how many subsets need to be prepared. Also, these approximations of method explained below, also called Exhaustive Methods, that computes all possible ways the data can be split into training and validation groups.

The Leave-P-Out approach leaves p data points out of training data,

i.e. if there are N data points in the original group then, n-p groups are used to train the model and p points are used as the validation group. This process is repeated for all combinations in which original group can be separated this way, and then the error is averaged for all trials, to give overall effectiveness.

This method can be classify as exhaustive because it needs to train and validate the model for all combination possibilities, and for a large p, it can become computationally impracticable.

A particular case of this approach is when $p = 1$. This is also called Leave-One-Out Cross Validation. This method is generally used over the previous one because it does not suffer from the over-computation, as number of possible combinations is equal to number of data points in original group.

6 Conclusion about Cross Validation Method

The Cross Validation Method is a very useful approach to verify the effectiveness of the model, mainly in cases where the overfitting mitigate is necessary. It is also of use in checking the hyper model parameters, in the sense that which parameters will result in lowest validation error. This is all the basic the student needs to get started with cross validation. To start work with all kinds validation techniques, you can access this link, in this site you can run this method with just a few lines of code in python.

7 References

- Cross Validation in Machine Learning, Prashant Gupta, 2017
- Scikit Learn Cross Validation Evaluating Estimator Performance (link referred in the text)