Cross Validation

Re-sampling technique

Ranvirsing Shailendrasing Sisodiya

Department of Statistics
K.B.C. North Maharashtra University, Jalgaon

April 12, 2022

Overview

1. Bias-Variance, Overfitting, Underfitting
2. What is cross validation?
3. Hold Out method
4. Leave out one cross validation (LOOCV)
5. K-Fold Cross-Validation
6. Time Series Cross-Validation
7. Conclusion
8. References

Basics

Bias: Bias is the measurement of how accurately a model can capture a training dataset or In the simplest terms, Bias is the difference between the Predicted Value and the Expected Value.
Variance: Variance is the measurement of how accurately a model can do prediction on a test dataset.
Underfitting: A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data.
In a nutshell, Underfitting – High bias and low variance.
Overfitting: A statistical model fits exactly against its training data. When this happens, the algorithm unfortunately cannot perform accurately against unseen data.
In a nutshell, Overfitting – High variance and low bias

Why do we need Cross-Validation?

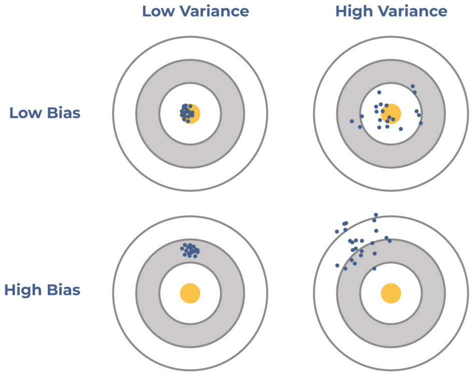

Image Source: ActiveWizards

What is cross validation?

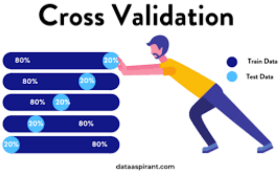

Image Source: `https://dataaspirant.com/cross-validation/`

Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments, one used to learn or train a model and the other used to validate the model.

Hold Out method

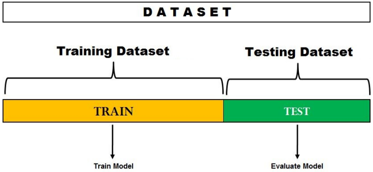

Image Source: DataVedas

A natural approach is to split the available data into two non-overlapped parts, one for training and the other for testing. The data can be divided into 70-30 or 60-40, 75-25 or 80-20, or even 50-50 depending on the use case.
As a rule, the proportion of training data has to be larger than the test data.

Hold Out method

Approach:

1. Randomly divide the available set of observations into two parts, a training set and a test set or hold-out set.
2. Fit the model on the training set.
3. Use the resulting fitted model to predict the responses for the observations in the test set.
4. The resulting test set error rate is typically assessed using the MSE in the case of a quantitative response. This provides an estimate of the test error rate.

Hold Out method

Drawbacks:

- In the Hold out method, the test error rates are highly variable (high variance) and it totally depends on which observations end up in the training set and test set.
- Only a part of the data is used to train the model (high bias) which is not a very good idea when data is not huge and this will lead to overestimation of test error.

One of the major advantages of this method is that it is computationally inexpensive compared to other cross-validation techniques.

Leave One Out Cross-Validation

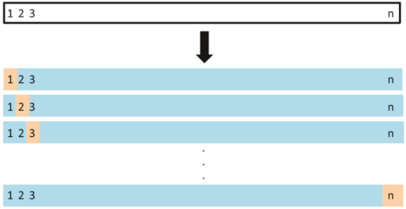

Image Source: An Introduction to Statistical Learning (ISLR)

In this method, we select a single observation (x_1, y_1) as test data, and everything else $(x_2, y_2) \cdots (x_n, y_n)$ is labeled as training data and the model is trained. Now the 2^{nd} observation (x_2, y_2) is selected as test data and the model is trained on the remaining data. This process continues 'n' times and the average of all these iterations is calculated and estimated as the test set error.

Leave out one cross validation (LOOCV)

Approach:

1. The above process continues ‘n’ times and the average of all these iterations is calculated and estimated as the test set error.
2. i.e. we produce n squared errors $MSE_1, MSE_2, \cdots, MSE_n$.
3. The LOOCV estimate for the test MSE is the average of these n test error estimates:

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

Leave out one cross validation (LOOCV)

LOOCV has a couple of major advantages over the Hold Out approach.

- LOOCV has far less bias.
- LOOCV approach tends not to overestimate the test error rate as much as the Hold Out approach does.
- Performing LOOCV multiple times produces similar results. This is not typically true for the Hold Out method.

One of the major disadvantages of this method is that it has the potential to be expensive to implement, since the model has to be fit n times. This can be very time consuming if n is large, and if each individual model is slow to fit.

K-Fold Cross-Validation

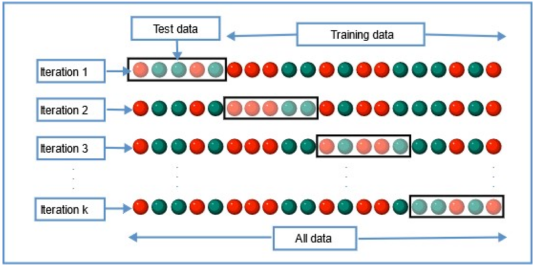

Image Source: Wikipedia

In this resampling technique, the whole data is divided into k sets of almost equal sizes. The first set is selected as the test set and the model is trained on the remaining k-1 sets. The test error rate is then calculated after fitting the model to the test data.

K-Fold Cross-Validation

Approach:

1. The above process continues ‘k’ times and the average of all these iterations is calculated and estimated as the test set error.
2. i.e. we produce n squared errors $MSE_1, MSE_2, \cdots, MSE_k$.
3. The LOOCV estimate for the test MSE is the average of these k test error estimates:

$$CV_n = \tfrac{1}{k} \textstyle\sum_{i=i}^{k} MSE_i$$

Notice that LOOCV is a special case of k-fold CV when k = n.
But usually in K-Fold CV, the no of folds k is less than the number of observations in the data $(k < n)$.

K-Fold Cross-Validation

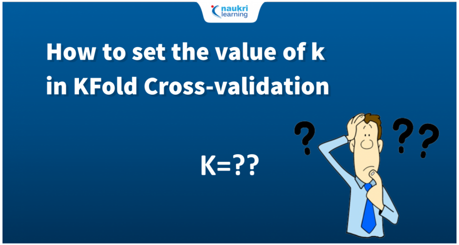

Image Source: Naukri.com

The key configuration parameter for k-fold cross-validation is k that defines the number folds in which to split a given dataset. Common values are k=3, k=5, and k=10, and by far the most popular value used in applied machine learning to evaluate models is k=10.

K-Fold Cross-Validation

Advantages:

- Computation time is reduced as we repeated the process only 10 times when the value of k is 10.
- Every data points get to be tested exactly once and is used in training k-1 times.
- The variance of the resulting estimate is reduced as k increases and it has reduced bias.

One of the major disadvantages of this method is that the training algorithm is computationally intensive as the algorithm has to be rerun from scratch k times.

Time Series Cross-Validation

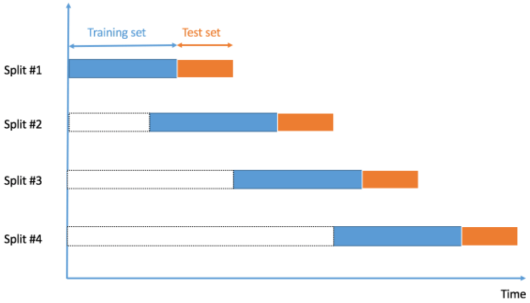

Image Source: Analytics Vidhya

Time series data is data that is collected at different points in time. As the data points are collected at adjacent time periods there is potential for correlation between observations. This is one of the features that distinguishes time-series data from cross-sectional data.

Time Series Cross-Validation

Approach:

1. As the order of the data is very important for time series related problems, so we split the data into training and validation set according to time, also called as “Forward chaining” method or rolling cross-validation.
2. We start with a small subset of data as the training set. Based on that set we predict later data points and then check the accuracy.
3. The Predicted samples are then included as part of the next training dataset and subsequent samples are forecasted.

Conclusion

Cross-Validation is a very powerful tool. It helps us better use our data, and it gives us much more information about our algorithm performance. In complex machine learning models, it's sometimes easy not pay enough attention and use the same data in different steps of the pipeline. This may lead to good but not real performance in most cases, or, introduce strange side effects in others. We have to pay attention that we're confident in our models. Cross-Validation helps us when we're dealing with non-trivial challenges in our Data Science projects.

References

Springer (ISLR)
An Introduction to Statistical Learning with Applications in R

Geeks for Geeks
ML — Underfitting and Overfitting
Link

Dima Shugla (2018)
5 Reasons why you should use Cross-Validation in your Data Science Projects
Link

Thank You !