Cross validation

Elie Mulamba
Amisi Fikirini
Salomon Muhirwa

African Master's in Machine Intelligent, AIMS-Senegal

Foundations of Machine Learning

April 13, 2023

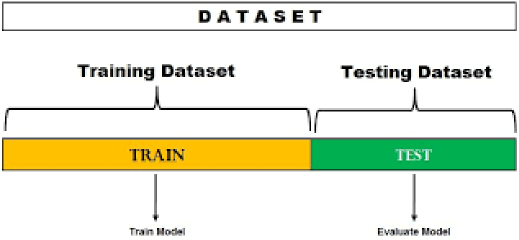
Overview

Overview

Introduction

Statement

Training a machine learning model is very important step as this process allows to confirm or deny if the model is good or not. To do this, the process requires dividing the general dataset into two slices called "Training data" and "Test data".



Overview

Introduction

Definition

Model validation is the process of verifying that a model meets all the requirements that have been set for it.
This includes verifying that the model is accurate and complete, as well as verifying it is consistent with other models.

Overview

Introduction

Importance of Cross Validation

Let M be a set of finite models $M = M_i, ...M_d$ and S a dataset
$S = S_{train} \cup S_{test}$
using empirical risk minimization for model selection:

1. Train each model M_i on S_{train}, to get some hypothesis hi.
2. Test each h_i on S_{test}
3. Pick the hypotheses with the smallest training error. This method will always select a high-variance

Why do we need cross Validation?

- Evaluate model performance
- Avoid overfitting
- Hyperparameter tuning
- Limited data

Overview

Introduction

Working principle

1. Randomly split S into S_{train} and S_{test}
2. Train h_i on S_{train}
3. Measure Generalization Error
4. Retrain the best model on S

Type of Cross validation

- K-Fold Cross-validation
- Leave One Out Cross Validation
- Stratified Cross-Validation

Overview

K-Fold Cross-validation

Algorithm

let's S=$\bigcup_{i=1}^{k} S_i$
K= number of folds
$\forall(i,j)$:S$_i \cap S_j = \phi$
For i =1 to K

- Train h_i on $A = S - S_i$
- Test h_i on S_i and Compute the $Error_i$
- Average $Error_i$

Generalize Error

K-Fold Cross-validation

Advantages & Disadvantages

1. Advantages

   More accurate estimate of model performance: K-fold cross-validation provides a more accurate estimate of the model's performance than other methods like a traintest split because it uses all available data for training and testing.

K-Fold Cross-validation

Advantages & Disadvantages

1. Advantages

More accurate estimate of model performance: K-fold cross-validation provides a more accurate estimate of the model's performance than other methods like a traintest split because it uses all available data for training and testing.

Better utilization of data: By dividing the data into K folds, k-fold cross-validation enables us to use all available data for both training and testing.

K-Fold Cross-validation

Advantages & Disadvantages

1. Advantages

   More accurate estimate of model performance: K-fold cross-validation provides a more accurate estimate of the model's performance than other methods like a traintest split because it uses all available data for training and testing.

   Better utilization of data: By dividing the data into K folds, k-fold cross-validation enables us to use all available data for both training and testing.

   Reduced risk of overfitting: K-fold cross-validation reduces the risk of overfitting, as the model is tested on different subsets of the data during the validation process.

K-Fold Cross-validation

Advantages & Disadvantages

1. Advantages

   More accurate estimate of model performance: K-fold cross-validation provides a more accurate estimate of the model's performance than other methods like a traintest split because it uses all available data for training and testing.

   Better utilization of data: By dividing the data into K folds, k-fold cross-validation enables us to use all available data for both training and testing.

   Reduced risk of overfitting: K-fold cross-validation reduces the risk of overfitting, as the model is tested on different subsets of the data during the validation process.

   Improved model generalization: K-fold cross-validation allows us to estimate how well the model will generalize to new, unseen data.

K-Fold Cross-validation

Advantages & Disadvantages

1. Disadvantages

Increased computational time: K-fold cross-validation requires fitting the model K times.

Sensitivity to data imbalance: If the data set is imbalanced, meaning that some classes have significantly more samples than others, k-fold cross-validation can result in biased estimates of model performance.

Higher variance in performance estimates: The performance estimates obtained from k-fold cross-validation can have higher variance than other methods.

Increased complexity of hyperparameter tuning: K-fold cross-validation can make hyperparameter tuning more complex, as it requires fitting the model multiple times for each combination of hyperparameters.

Overview

Leave One Out Cross-validation

Algorithm

1. Splitting
   Randomly split S_{train} into m disjoint subsets of training examples. where $m =$ the size of S_{train} Let's call these subsets $S_1, ...S_k$
2. Training
   For i= 1 to m
   - Train h_i on $S_1 \cup ... \cup S_{i-1} \cup S_{i+1} \cup ... \cup S_m$
   - Test h_1 on S_i => Compute the $Error_i$
   - Average the $Error_i$

   Generalize the Error

Leave One Out Cross-validation

Advantages & Disadvantages

1. Advantages

Lower Bias: uses almost all of the data for training in each fold

$$bias\_loo = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \qquad (2.1)$$

$$bias\_k - fold = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{m_j}\sum_{i=1}^{j}(y_i - \hat{y}_i)^2 \qquad (2.2)$$

Good for small dataset: because it provides a more accurate estimate of the model's true performance.

Leave One Out Cross-validation

Advantages & Disadvantages

1. Disadvantages

Higher variance

can be expressed as:

$$var\_LOO = \frac{1}{n}\sum_{i=1}^{n}, (y_i - f_i)^2 - bias_{LOO})^2 \tag{2.3}$$

$$var_{k-fold} = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{n_j}\sum_{i=1}^{nj}(y_i - f_i)^2 - bias_{k-fold})^2 \tag{2.4}$$

Where var_{LOO} and var_{k-fold} are the variances of the prediction errors

Computationally expensive: LOO can be computationally expensive,

Overview

Stratified Cross-validation

Advantages & Disadvantages

The algorithm of Stratified k-Fold technique:

$S = \bigcup_{i=1}^{K} S_i$

for i = 1 to K

$S_{train} = S - S_i$

$S_{test} = S_i$

$Calculate\ the\ Error_i$

Generalization Error

Stratified Cross-validation

Advantages & Disadvantages

1. Advantages
   Stratified KFold ensures that the proportion of the feature of interest is the same across the original data, training set and the test set.

Stratified Cross-validation

Advantages & Disadvantages

1. Advantages
   Stratified KFold ensures that the proportion of the feature of interest is the same across the original data, training set and the test set.
   Checking Model Generalization: Cross-validation gives the idea about how the model will generalize to an unknown dataset

AIMS African Institute for Mathematical Sciences SENEGAL

Stratified Cross-validation

Advantages & Disadvantages

1. Advantages
   Stratified KFold ensures that the proportion of the feature of interest is the same across the original data, training set and the test set.
   Checking Model Generalization: Cross-validation gives the idea about how the model will generalize to an unknown dataset
   Checking Model Performance: Cross-validation helps to determine a more accurate estimate of model prediction performance

Stratified Cross-validation

Advantages & Disadvantages

1. Disadvantages
   Higher Training Time: with cross-validation, we need to train the model on multiple training sets.

Stratified Cross-validation

Advantages & Disadvantages

1. Disadvantages

   Higher Training Time: with cross-validation, we need to train the model on multiple training sets.

   Expensive Computation: Cross-validation is computationally very expensive as we need to train on multiple training sets.

Overview

Experiments

Form Scratch

Demonstration on Google Colab

Conclusion

To conclude, we have been able to:

- Experiment with different types of cross-validation
- understanding how cross-validation solves the problem of overfitting
- Assess the effectiveness of each technique in improving accuracy and reducing loss.
- Determine which technique provided the best results in terms of model performance and generalization.
- Gain insights into the relative strengths and weaknesses of each technique, which can be used to inform future model development efforts.

References

- Efron, B., Tibshirani, R. J. (1994). An introduction to the bootstrap. Chapman Hall/CRC.
- Kohavi, R. (1995). Cross-validation and the bootstrap: Estimating the error rate of a prediction rule. Machine Learning, 14(2), 113-143.
- Arlot, S., Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics surveys, 4, 40-79.
- Bengio, Y., Grandvalet, Y. (2004). Cross-validation: what does it estimate and how well does it do it?. Proceedings of the 2004 international conference on artificial intelligence and statistics, 14.