Bochra CHEMAM

Data science

MOHAMMED VI POLYTECHNIC UNIVERSITY

"The best way to learn data science is to apply data science."

Activity 8 Cross-Validation

28/04/2021

For this Activity session we will:

- Apply Cross-validation methods and learn how they can improve the scores

Remind :

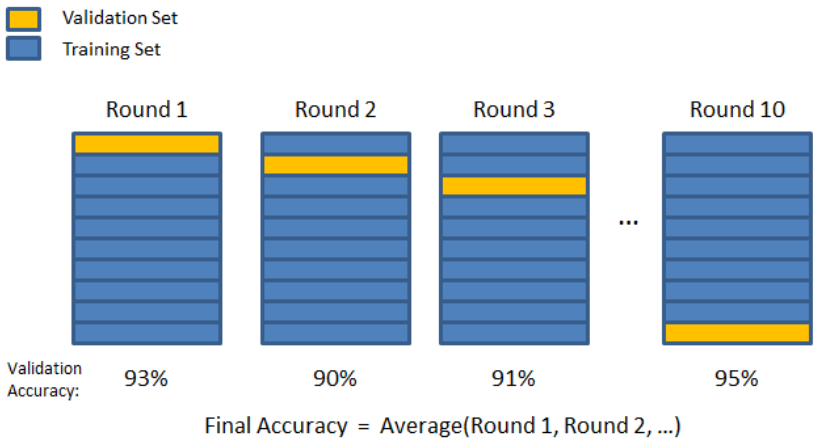
Cross-Validation is a technique that involves reserving a particular sample of a dataset on which you do not train the model. Later, you test your model on this sample before finalizing it.

Here are the steps involved in cross-validation:

1. You *reserve* a sample data set
2. Train the model using the remaining part of the dataset
3. Use the reserve sample of the test (validation) set. This will help you in gauging the effectiveness of your model's performance.

- **k-fold cross-validation:**

  Below is the visualization of a k-fold validation when k=10.



For example:

- **Model1**: Trained on Fold1 + Fold2, Tested on Fold3
- **Model2**: Trained on Fold2 + Fold3, Tested on Fold1
- **Model3**: Trained on Fold1 + Fold3, Tested on Fold2

Benefits of Cross-Validation

- It helps evaluate the quality of your model.
- It helps to reduce/avoid problems of overfitting and underfitting.
- It lets you select the model that will deliver the best performance on unseen data.

Data science

"The best way to learn data science is to apply data science."

What are Overfitting and Underfitting?

- **Overfitting** refers to the condition when a model becomes too data-sensitive and ends up capturing a lot of noise and random patterns that do not generalize well to unseen data. While such a model usually performs well on the training set, its performance suffers on the test set.
- **Underfitting** refers to the problem when the model fails to capture enough patterns in the dataset, thereby delivering a poor performance for both the training as well as the test set.

Excercice

Using the Iris dataset. It only has 5 attributes and 150 rows, meaning it is small and easily fits into memory. Plus all of the numeric attributes are in the same units and the same scale not requiring any special scaling or transforms to get started.

Load the dataset

Analyze your dataset: use Describe method, shape, dtypes, corr()

Investigate the class distribution of the dataset

Data visualization:

1- Use Univariate plots to better understand each attribute (box and whisker plots, Histograms)
2- Use Multivariate plots to understand the relationships between the attributes(scatter plots)
3- Check the correlation

Validation dataset for evaluation

1- To evaluate the model using 80% of the dataset for modeling and hold back 20% for the Test.
2- Use 2 fold cross-validation or plus
3- Evaluate the algorithms using accuracy

Models

Let's test different classification techniques for our problem

- k-Nearest Neighbors (KNN).
- CART (tree decision)

1. Calculate **testing accuracy**
2. Use the **average testing accuracy** as the estimate of out-of-sample accuracy

References

1. Iris dataset: https://www.kaggle.com/uciml/iris
2. Theory: http://www.scholarpedia.org/article/K-nearest_neighbor
3. https://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/

Data science

MOHAMMED VI POLYTECHNIC UNIVERSITY

"The best way to learn data science is to apply data science."