Cross-Validation Using K-Fold With Scikit-Learn

Isheunesu Tembo · Follow

3 min read · Mar 14, 2022

Imagine that you are a medical research scientist and your task is to find a cure for a particular disease , you perform experiments and test the drug on animals before you administer it to humans .

The same happens with machine learning we train the models , test them , then deploy them for real world use.

Learning the performance of a prediction function and testing it on the same data is a methodological mistake. A model would just repeat the labels of the samples that it has just seen would have have a perfect score but would fail to predict anything useful on yet unseen data , this situation is what we call overfitting. To avoid it , it is common to hold out part of available data as test set.

When evaluating different hyparameters for estimators such as the C setting that must be manually set for an SVM classifier , there is still a high risk of overfitting on

the test set because parameters can be tweaked until the estimator performs optimally. This way , knowledge about the test set can “leak” into the model and evaluation metrics will no longer report on the generalization performance.

To solve this problem , yet another part of the dataset can be held as a “validation set”.

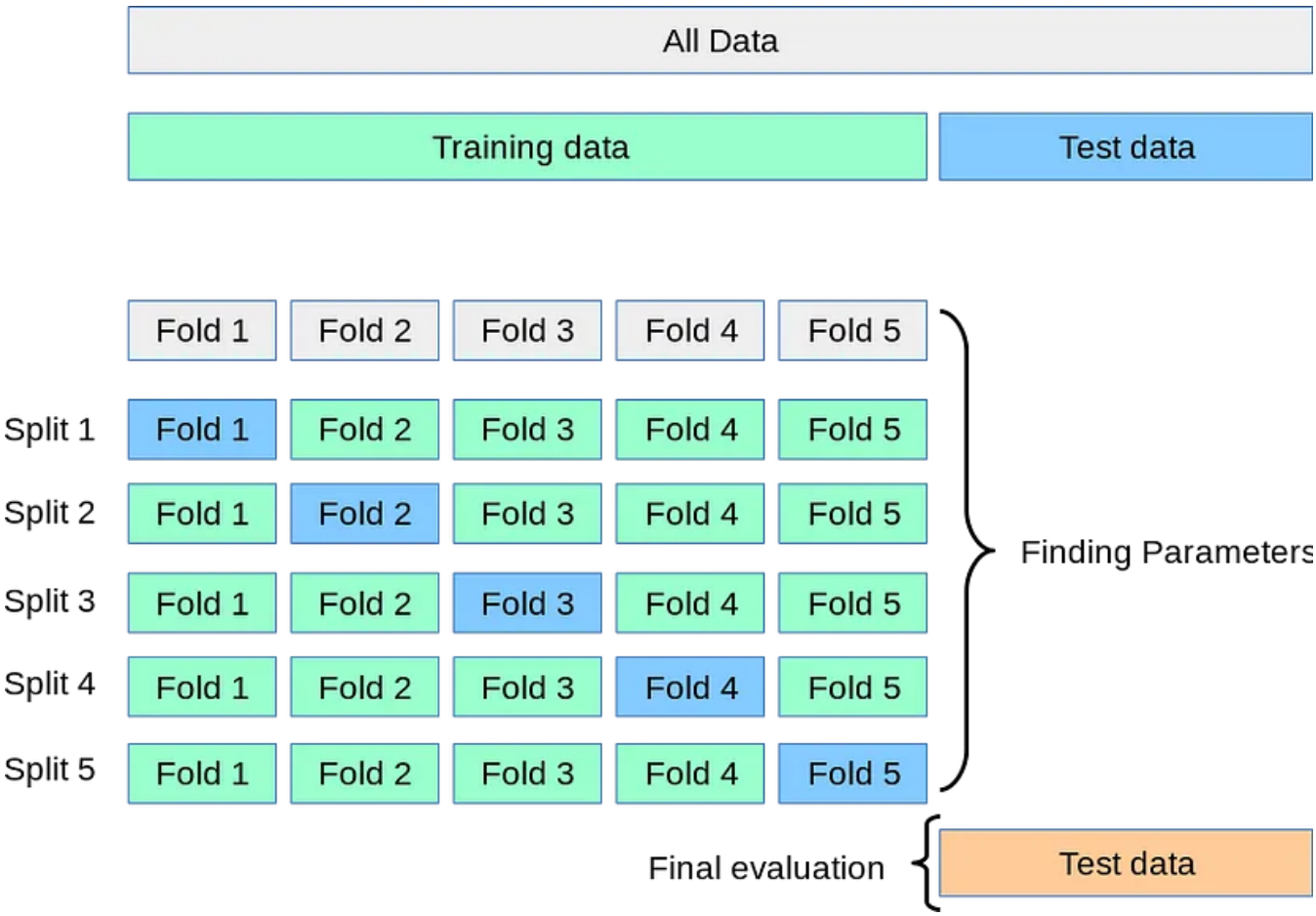
By partitioning data another problem arises which is we drastically reduce the number of samples which can be used for learning by the model , and the results can depend on a particular random choice for the pair (training , validation) sets.

A solution to this problem is a procedure called cross-validation , but the validation set is no longer needed when doing CV. Using an approach called K-fold , the training set is split into k smaller sets.

The following procedure is followed for each of the K-fold :

1 .A model is trained using K-1 of the folds as training data

2.The resulting model is validated on the remaining part of the data.



Let see a code example :

```
import numpy as np
import pandas as pd
import os

from sklearn.datasets import load_iris
iris=load_iris()

X=iris.data
y=iris.target

X=(X-np.min(X))/(np.max(X)-np.min(X))
```

In [5]:

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test=train_test_split(X,y,test_size=0.3)
```

In [6]:

```
from sklearn.neighbors  import KNeighborsClassifier
knn= KNeighborsClassifier(n_neighbors=3)
```

In [7]:

```
from sklearn.model_selection import cross_val_score
accuracies=cross_val_score(estimator=knn,X=x_train,y=y_train,cv=10)
accuracies
```

```
array([0.83333333, 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 1.        ])
```

```
print("average accuracy :",np.mean(accuracies))
print("average std :",np.std(accuracies))
```

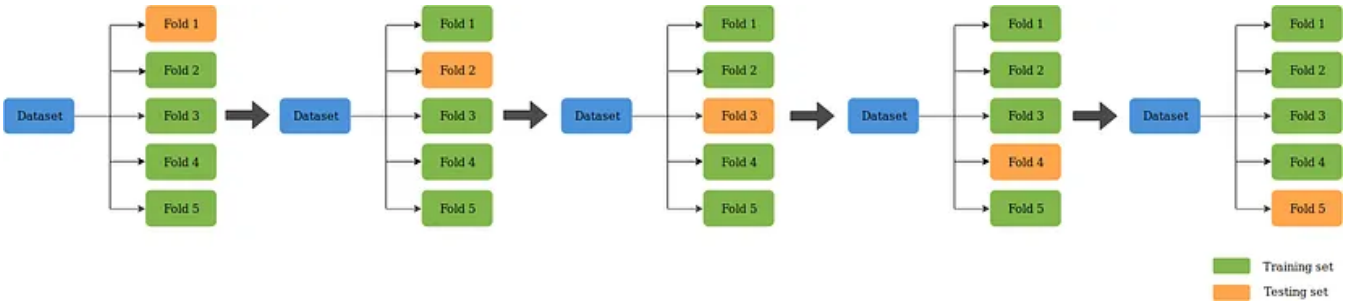
```
average accuracy : 0.9833333333333334
average std : 0.04999999999999999
```

```
knn.fit(x_train,y_train)
print("test accuracy :",knn.score(x_test,y_test))
```

```
test accuracy : 0.9555555555555556
```

GridSearch

```
from sklearn.model_selection import GridSearchCV

grid ={"n_neighbors":np.arange(1,50)}
knn= KNeighborsClassifier()
knn_cv=GridSearchCV(knn,grid,cv=10) #GridSearchCV
knn_cv.fit(X,y)

rint("tuned hyperparameter K:",knn_cv.best_params_)
print("tuned parametreye göre en iyi accuracy (best
score):",knn_cv.best_score_)

tuned hyperparameter K: {'n_neighbors': 13}
tuned parametreye göre en iyi accuracy (best score): 0.98
```

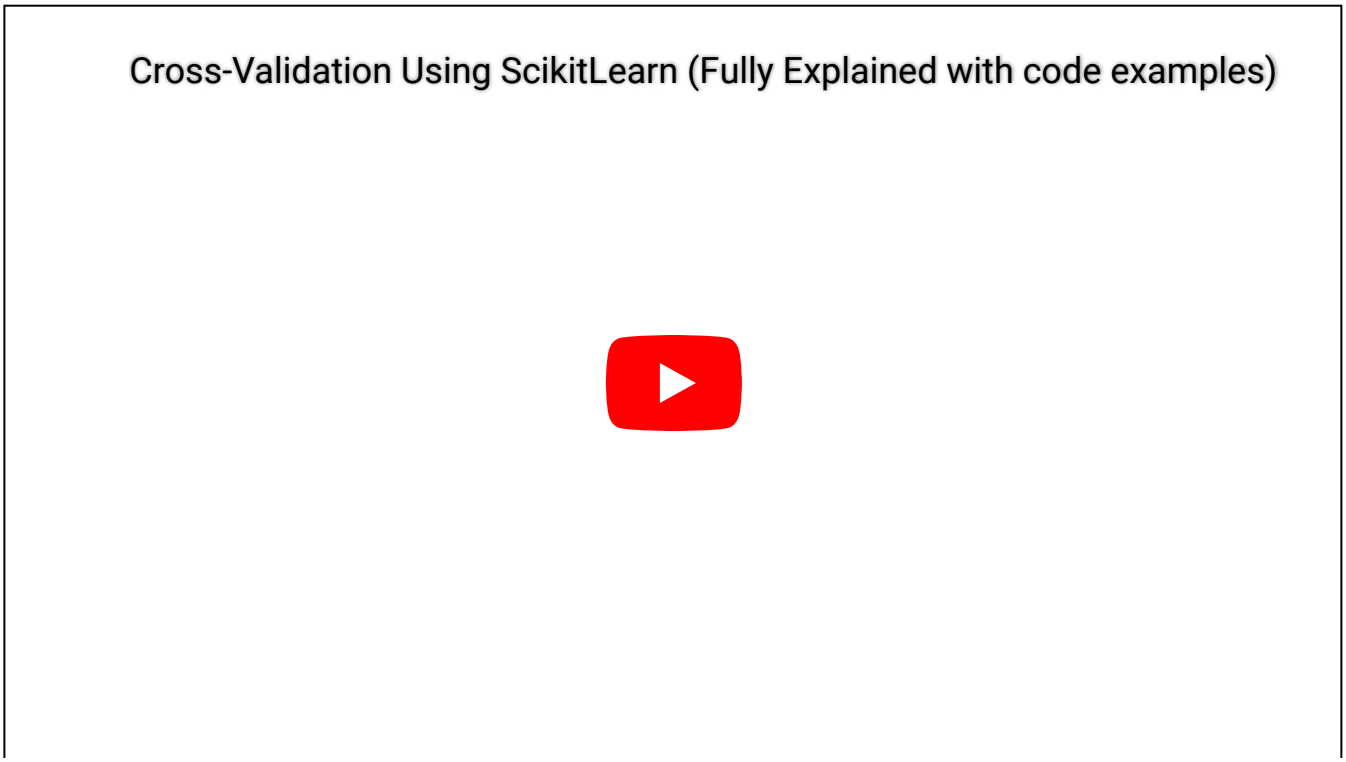


What is K-Fold Cross Validation

K-Fold CV is where a given data set is split into a K number of sections/folds where each fold is used as a testing set at some point. Lets take the scenario of 5-Fold cross validation(K=5). Here, the data set is split into 5 folds. In the first iteration, the first fold is used to test the model and the rest are used to train the model. In the second iteration, 2nd fold is used as the testing set while the rest serve as the training set. This process is repeated until each fold of the 5 folds have been used as the testing set.

```
from sklearn.model_selection import KFold
scores=[]
kFold=KFold(n_splits=10,random_state=42,shuffle=False)
for train_index,test_index in kFold.split(X):
    print("Train Index: ", train_index, "\n")
    print("Test Index: ", test_index)

    X_train, X_test, y_train, y_test = X[train_index],
X[test_index], y[train_index], y[test_index]
    knn.fit(X_train, y_train)
    scores.append(knn.score(X_test, y_test))
```

```
knn.fit(X_train,y_train)
scores.append(knn.score(X_test,y_test))

print(np.mean(scores))

0.9393939393939394

cross_val_score(knn, X, y, cv=10)
```



Machine Learning | K Fold Cross Validation | K Nearest Neighbours | Scikit Learn

Follow

Written by Isheunesu Tembo

32 Followers

Machine Learning Enthuasiast

More from Isheunesu Tembo

Isheunesu Tembo in Analytics Vidhya

Scikit-Learn Pipeline

Let's say you are a machine learning engineer and you are hired to create a machine learning algorithm by the bank to determine fraudulent...

4 min read · Feb 5, 2020

65

Isheunesu Tembo

Java-Object Oriented Programming(Abstraction) Explanation

What is abstraction?

3 min read · Sep 28, 2018

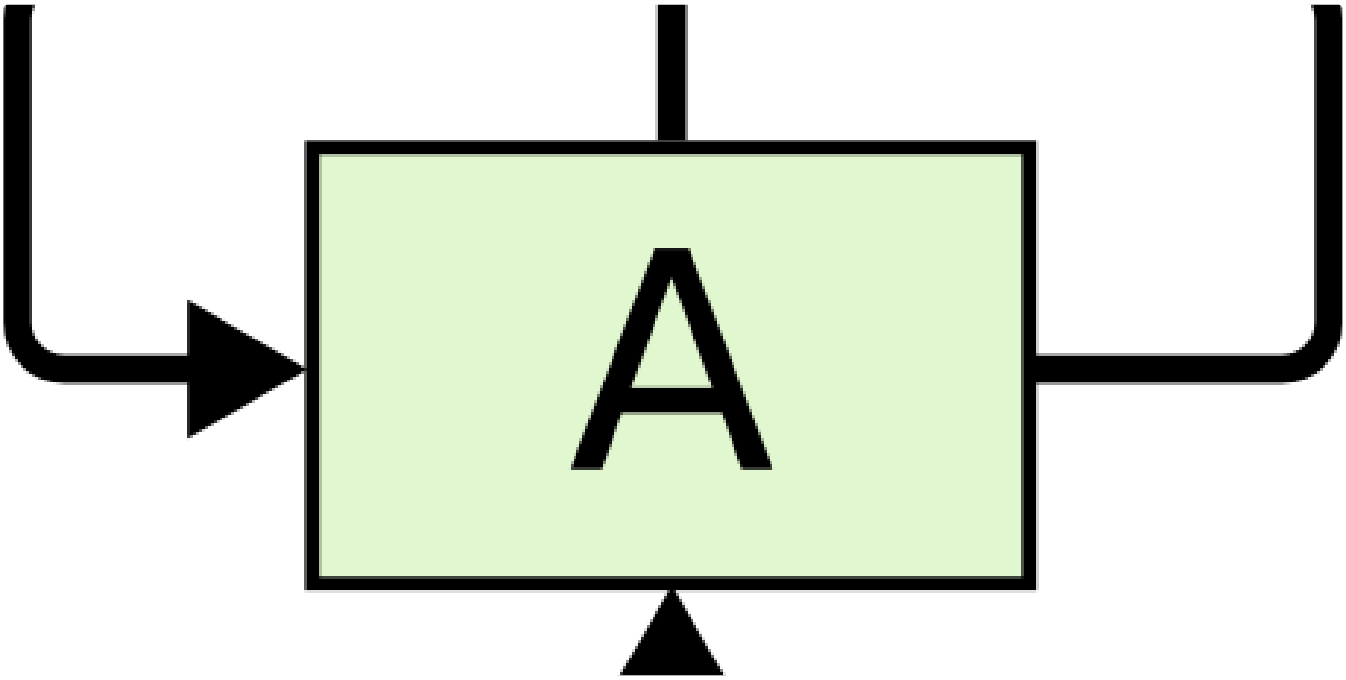
Isheunesu Tembo

Natural Language Processing , what is it ? How can we represent text using the Tensorflow Tokenizer

One thing that makes us great as humans is our ability to learn new languages and communicate with other humans ,but can computers match...

4 min read · Mar 14, 2022

16



Isheunesu Tembo

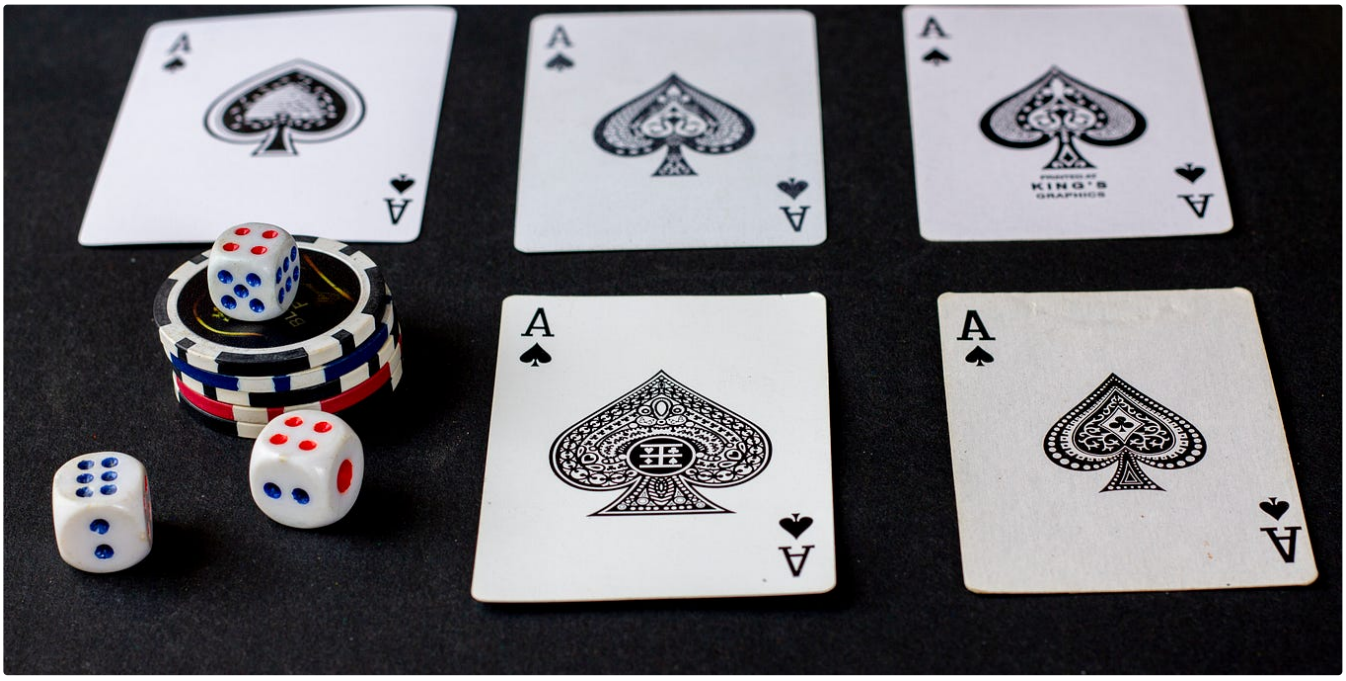
Recurrent Neural Networks

Humans don’t start their thinking from scratch every second.Lets say you are thinking about something most of the time your thoughts are...

7 min read · Feb 13, 2019

See all from Isheunesu Tembo

Recommended from Medium



Egor Howell in Towards Data Science

How To Correctly Perform Cross-Validation For Time Series

Avoid the common pitfalls in applying cross-validation to time series and forecasting models.

· 5 min read · Jan 10

171

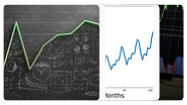
Bee Guan Teo in DS Notes

Multiple Linear Regression with Scikit-Learn—A Quickstart Guide

A step by step tutorial to perform multiple linear regression in Python

· 5 min read · Dec 31, 2022

Lists

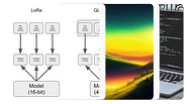


Predictive Modeling w/ Python

18 stories · 44 saves

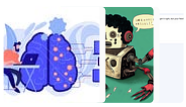
Practical Guides to Machine Learning

10 stories · 62 saves



Natural Language Processing

369 stories · 19 saves



The New Chatbots: ChatGPT, Bard, and Beyond

13 stories · 25 saves

Dr. Soumen Atta, Ph.D.

Regression models: a concise tutorial of real-life examples with Python implementations (Part I)

In this tutorial, we will discuss seven regression models with real-life examples and Python implementations. Before reading this tutorial...

· 10 min read · Feb 15

Amado de Jesús Vázquez Acuña

Unbalanced Dataset? do this trick..

In this article you will learn the most powerful techniques to deal with this annoying problem. Believe me that after knowing this, you...

✦ · 4 min read · Mar 24

12

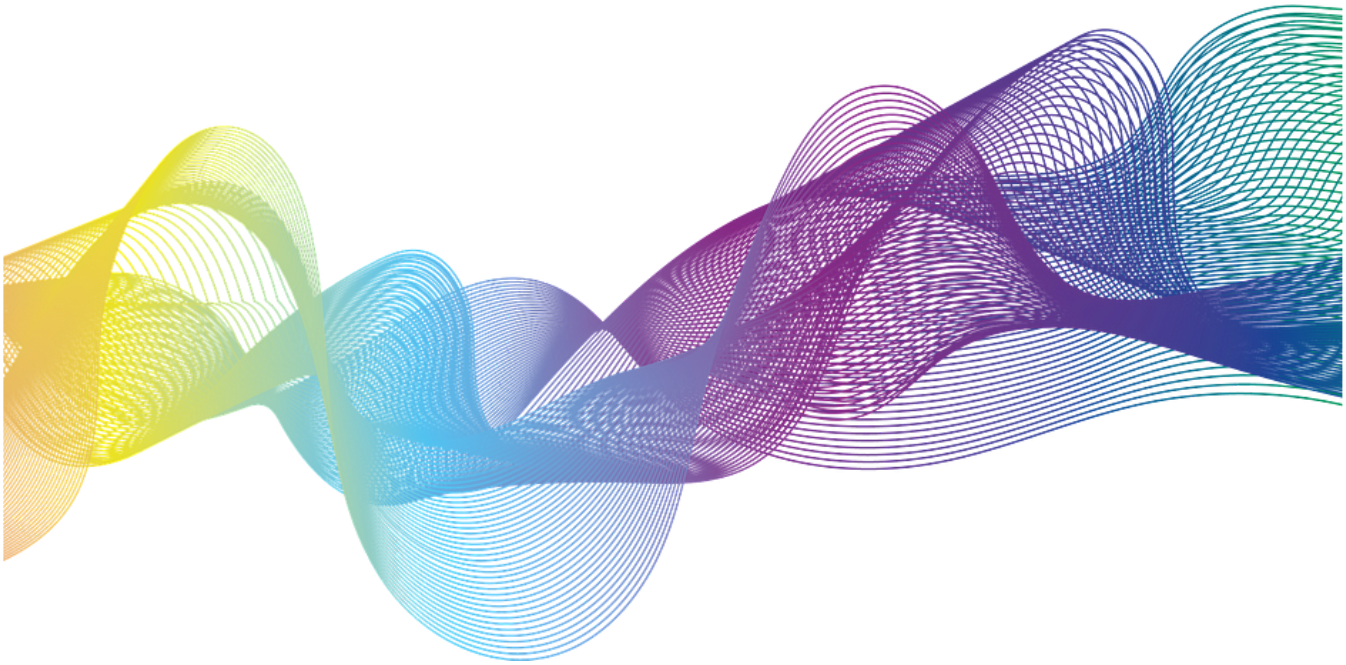
Sadrach Pierre, Ph.D. in Towards Data Science

Mastering P-values in Machine Learning

Understanding P-values and ML use cases

✦ · 7 min read · Jan 6

172



Federico Trotta in MLearning.ai

How To Easily Validate Your ML Models With Learning Curves

Discover the power of learning curves to validate your ML models

· 8 min read · Feb 4