Machine Learning In Python

Subject : Split Dataset to training and testing set

Lecturer : Reza Akbari Movahed

Hamedan University of Technology

Winter 2020

Split Dataset to training and testing set

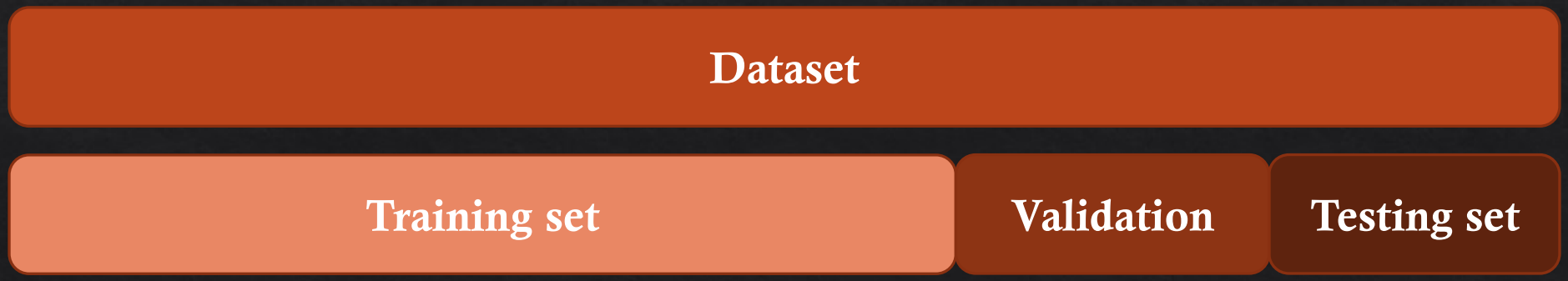
Random Splitting to training and testing sets

- Split dataset (Feature Matrix and Labels) into training and testing subsets, randomly.
- Training Dataset: The sample of data used to fit the model.
- Testing Dataset: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.
- Generally, this method uses a proportion to divide dataset to the training and testing sets.

Split Dataset to training and testing set

Random Splitting to training, testing and validation sets

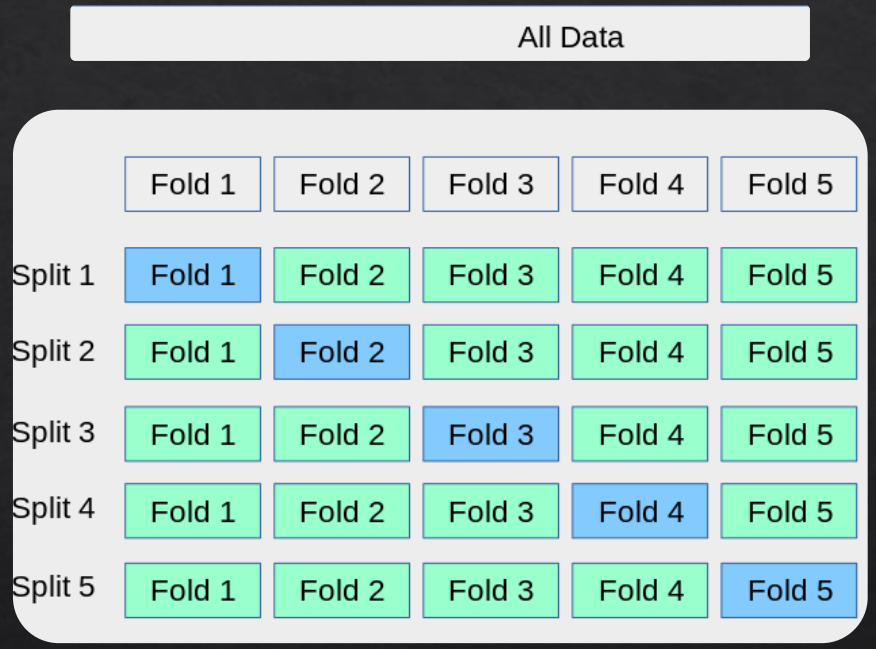
- Split dataset (Feature Matrix and Labels) into random training, testing, and validation subsets, randomly.
- Training Dataset: The sample of data used to fit the model.
- Testing Dataset: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.
- Validation Dataset: The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters.
- Generally, this method uses a proportion to divide dataset to the training, testing, and validation sets.



Split Dataset to training and testing set

Cross Validation Technique (K-Fold)

1. Split the entire data randomly into k folds.
2. Then train the model using the k – 1 folds and test the model using the remaining Kth fold.
3. Repeat this process until every K-fold serve as the test set.
4. Then take the average of your recorded scores. That will be the performance metric for the model.



Split Dataset to training and testing set

Cross Validation Technique (K-Fold)

Three common tactics for choosing a value for k are as follows :

- **Representative**: The value for k is chosen such that each train/test group of data samples is large enough to be statistically representative of the broader dataset.
- **k=10**: The value for k is fixed to 10, a value that has been found through experimentation to generally result in a model skill estimate with low bias a modest variance.
- **k=n**: The value for k is fixed to n, where n is the size of the dataset to give each test sample an opportunity to be used in the hold out dataset. This approach is called leave-one-out cross-validation.