Sampling Methods for Supervised Learning

TIET, PATIALA

Training a Supervised Model

- Supervised Learning occurs when an algorithm learns from examples data and associated target responses (which can be numeric values or string labels).
- Supervised learning is where we have input variables (x) and an output variable (Y) and use an algorithm to learn the mapping function from the input to the output **Y = f(x)** .
- A supervised Learning model is trained in following three ways:
  - Holdout Method
  - K-fold Cross Validation
  - Bootstrap Sampling

Holdout Method

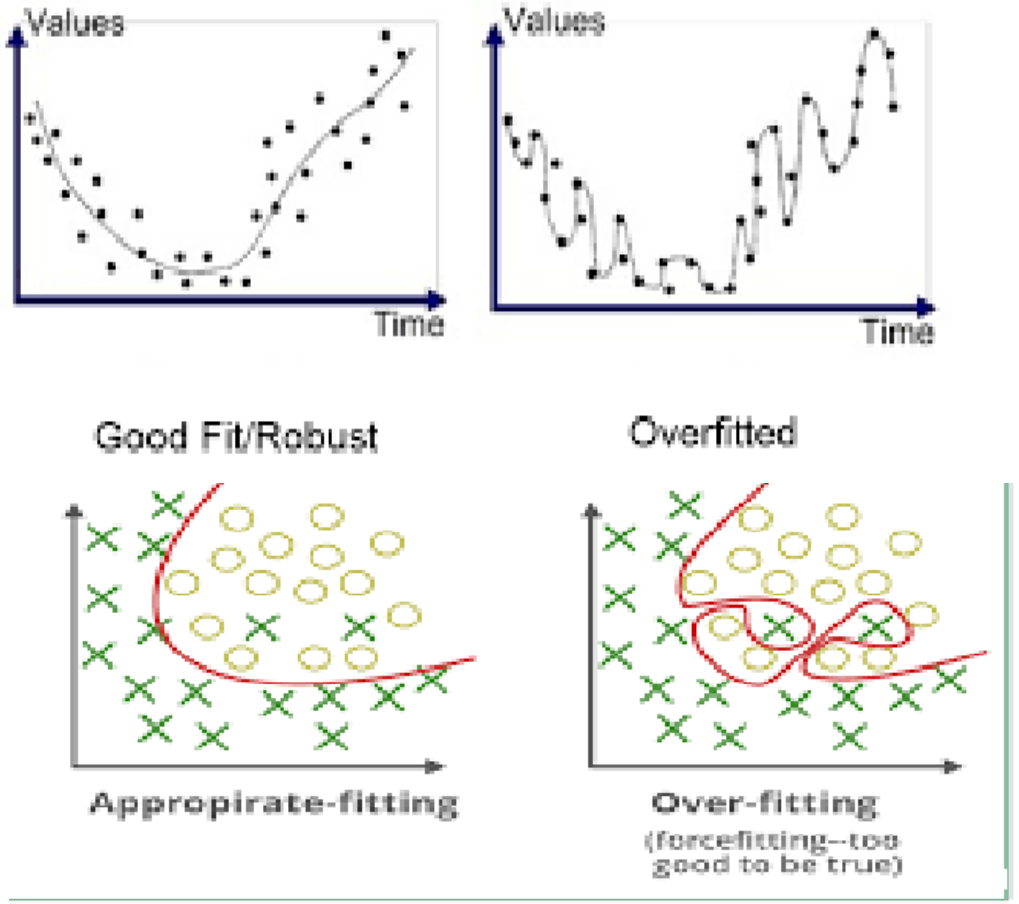
- In this method, a part of the input data is held back (that is how the name holdout originates) for evaluation of the model.
- Hold-out is when we split up your dataset into a ‘train’ and ‘test’ set.
- The **training set** is what the model is trained on, and **the test set** is used to see how well that model performs on unseen data.
- A common split when using the hold-out method is using 70-80% of data for training and the remaining 20-30% of the data for testing.

Overfitting

- Overfitting is a modeling error that occurs when a function is too closely fit to a limited set of data points.
- Overfitting the model generally takes the form of making an overly complex model to explain behavior in the data under study.
- In reality, the data often studied has some degree of error or random noise within it.
- Thus, attempting to make the model conform too closely to slightly inaccurate data can infect the model with substantial errors and reduce its predictive power.
- Overfitting results in good performance with training data se but poor performance with test set.

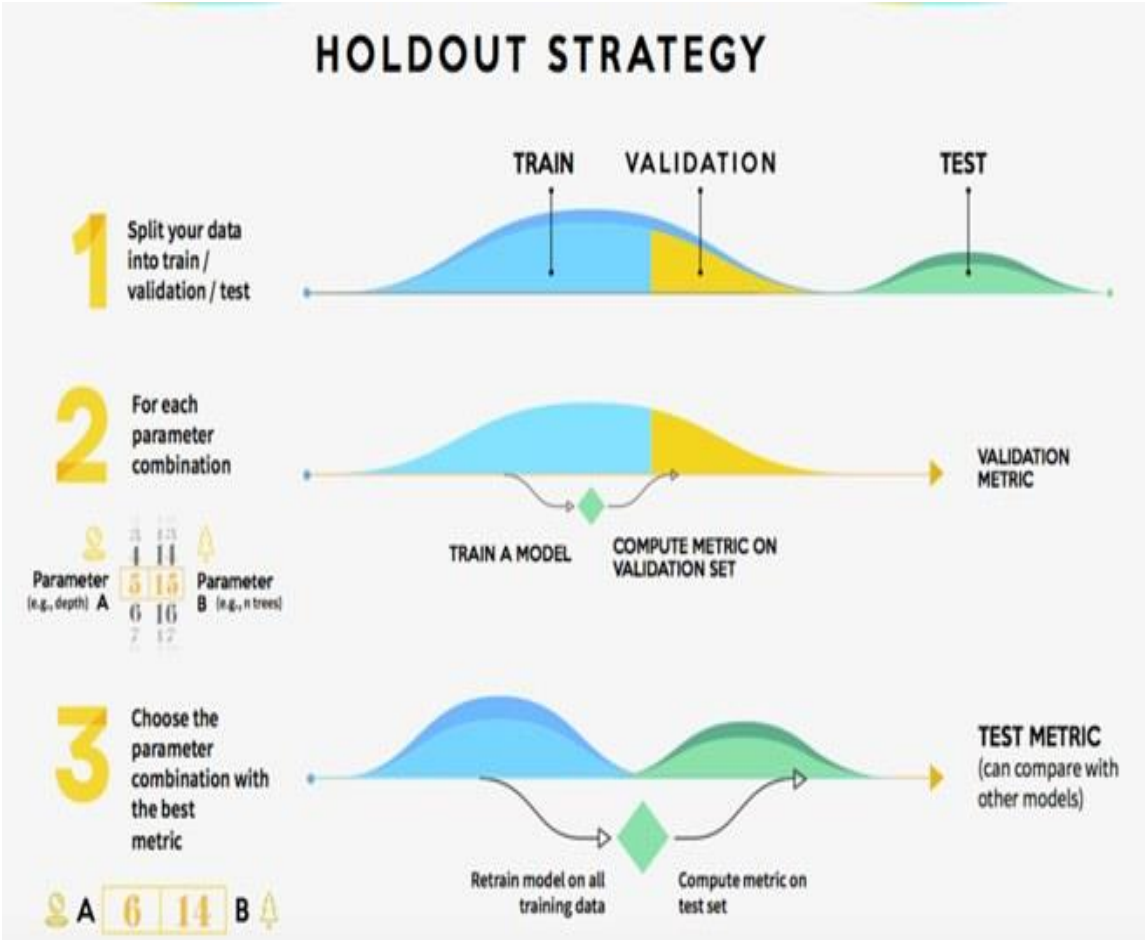
Overfitting

- Overfitting can be avoided by:
  - hold back of a validation data set.
  - using resampling techniques like k-fold cross validation.
  - remove the data points which have little or no-predictive power (feature selection)



Holdout of a Validation Dataset

- When evaluating machine learning models, the validation step helps us find **the best parameters** for our model while also preventing it from becoming **overfitted.**
- In this method, we have training set, test set and validation set.
- Validation set actually is a part of training set, because it is used to build your model. It is usually used for parameter selection and to avoid overfitting.



Holdout Method

- **Pros of the hold-out strategy:** Fully independent data; only needs to be run once so has lower computational costs.
- **Cons of the hold-out strategy:** Performance evaluation is subject to higher variance given the smaller size of the data.

K-Fold Cross Validation

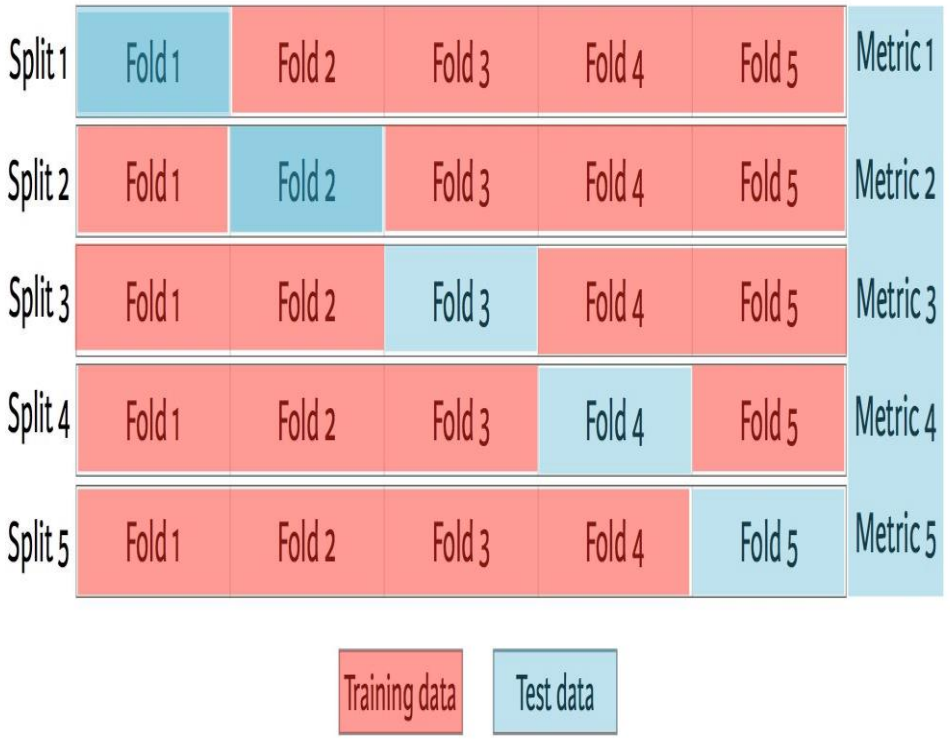
- Cross-validation or ‘k-fold cross-validation’ is when the dataset is randomly split up into ‘k’ groups.
- One of the groups is used as the test set and the rest are used as the training set.
- The model is trained on the training set and scored on the test set.
- Then the process is repeated until each unique group as been used as the test set.
- For example, for 5-fold cross validation, the dataset would be split into 5 groups, and the model would be trained and tested 5 separate times so each group would get a chance to be the test set.

K-Fold Cross Validation

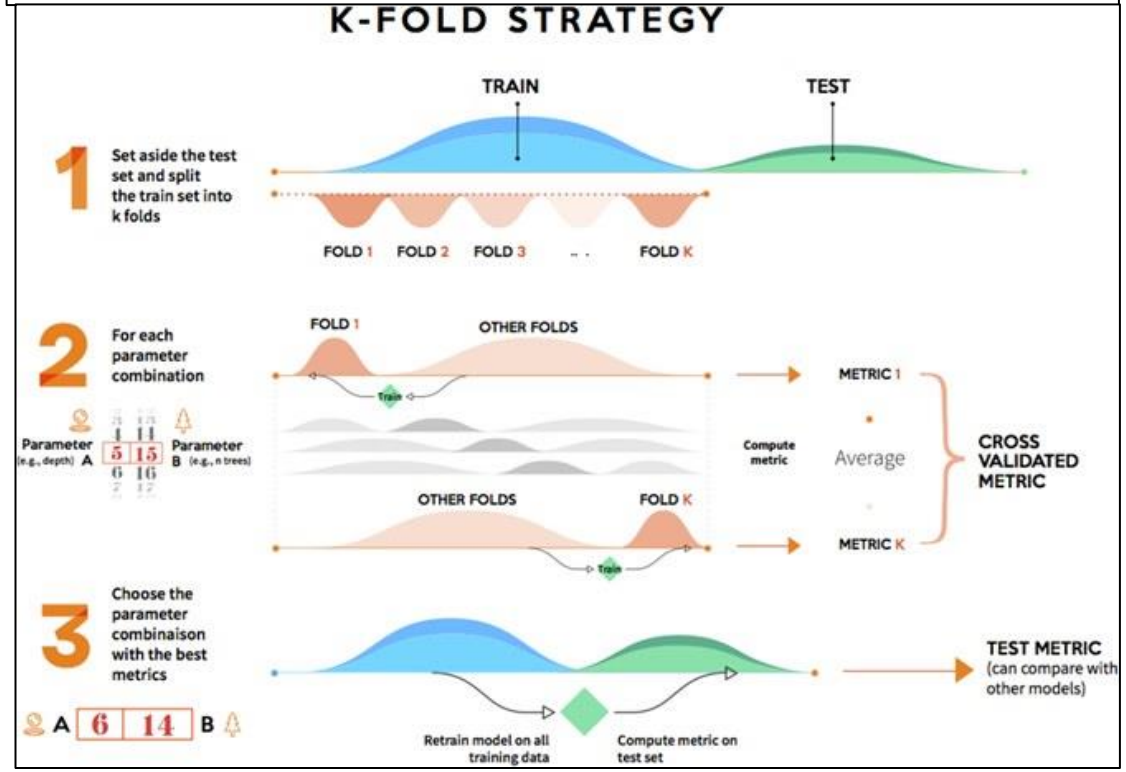
- Cross-validation or ‘k-fold cross-validation’ is when the dataset is randomly split up into ‘k’ groups.
- One of the groups is used as the test set and the rest are used as the training set.
- The model is trained on the training set and scored on the test set.
- Then the process is repeated until each unique group as been used as the test set.
- For example, for 5-fold cross validation, the dataset would be split into 5 groups, and the model would be trained and tested 5 separate times so each group would get a chance to be the test set.

K-Fold Cross Validation

5-FOLD CROSS VALIDATION PROCESS



CROSS VALIDATION



K-Fold Cross Validation

- A variant of cross validation is Leave one cross validation
- Leave one cross validation : We leave one point out (validation) , train for n-1 points . This is done for all n points at the end of the process the errors are averaged out.
- **Pros of the K-fold strategy:** Prone to less variation because it uses the entire training set.
- **Cons of the K-fold strategy:** Higher computational costs; the model needs to be trained K times at the validation step (plus one more at the test step).

Bootstrap Sampling

- It uses the technique of Simple Random Sampling With Replacement (SRSWR).
- Bootstrapping randomly picks data instances from the input dataset , with the possibility of the same data instance picked multiple times.
- Bootstrap sampling is used in a machine learning ensemble algorithm called bootstrap aggregating (also called bagging).
- It helps in avoiding overfitting and improves the stability of machine learning algorithms.
- In bagging, a certain number of equally sized subsets of a dataset are extracted with replacement. Then, a machine learning algorithm is applied to each of these subsets and the outputs are ensembled

Bootstrap Sampling