Model Tuning

Session Plan

1. Introduction
2. Discussion Questions on concept
3. Hands-on Case study
4. Extended Discussions and QnA
5. Summary

Discussion Questions

1. What is hyperparameter tuning?
2. Different types of hyperparameter tuning
3. What is data leakage ?
4. How to prevent data leakage using sklearn?

What is hyperparameter tuning?

- Hyperparameters are the parameters that govern the entire training process
- Their value are set before the learning process begins
- They have a significant effect on model’s performance
- The process of finding optimal hyperparameters for a model is known as hyperparameter tuning
- Choosing optimal hyperparameters can lead to improvements in overall model’s performance and can help in reducing both overfitting and underfitting

Different types of hyperparameter tuning

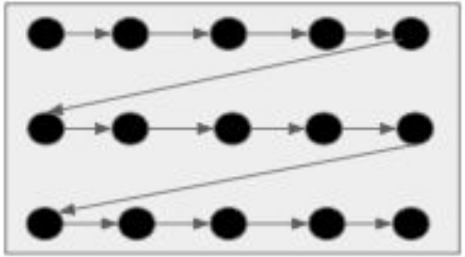
- Some models consists of huge number of hyperparameters, and finding the optimal set of hyperparameters can be a very time consuming process
- To make the process efficient, we'll look at 2 of the most common methods available in sklearn:
  - GridSearchCV
  - RandomizedSearchCV
- Grid search is best used when we have small search space, while Random search is best used when we have large search space
- We can use grid search to get best possible results when we don't have any time constraints, but when we have time constraints, it's better to go with random search
- Randomised search is known to give better results as compared to grid search

Grid Search

Grid search is a technique used to find optimal set of hyperparameters for a model from the provided search space

Let’s understand working of grid search with an example

- Let this grey box be set of all possible hyperparameters
- Let these black circles indicate the search space
- Grid search will iterate over all black circles in a sequence
- And finally gives the best set of hyperparameters based on the best score obtained

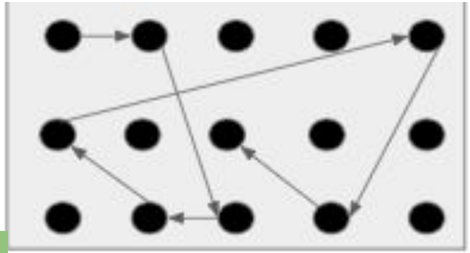


Doesn’t work good on large search spaces

It will find the best set of hyperparameters but with high cost

Randomised Search

- Random Search is another technique to find best set of hyperparameters which takes lesser time than grid search
- Random search is very similar to grid search, the difference is that in random search
  - we define **‘n_iter’ -** not all parameter values are tried out, but rather a fixed number of parameter settings is sampled from the specified distributions. The number of parameter settings that are tried is given by n_iter.
  - the set of hyperparameters is not searched sequentially
  - We can pass a range here instead of just numbers
- Out of entire search space of hyperparameter, only n_iter number of set of hyperparameters will be checked **randomly**



Works good on large search spaces
Gives better results than grid search

It doesn’t guarantee to find the best set of hyperparameters

Data leakage

- Data leakage is when information from outside the training dataset is known to model during the training process
- Sharing of any kind of information between testing and training sets leads to data leakage
- Data leakage can cause you to create overly optimistic, if not completely invalid, predictive models

Using whole dataset to fit standard scaler

will require to know distribution of whole data

That will lead to data leakage

While imputing missing values using KNN

Imputations in train data will also be based on observations in test data

This will lead to data leakage

While doing hyperparameter tuning,

tuning the model parameters on test set

Will lead to data leakage

How to prevent data leakage using sklearn

1. Scale train and test set separately
2. Normalize train and test sets separately
3. Hyperparameter tuning should be done on validation set or using cross validation and not on test set
4. For missing value imputation, fit the imputer on train set and then transform the train and test sets

To do all the above efficiently, we can use pipelines in sklearn

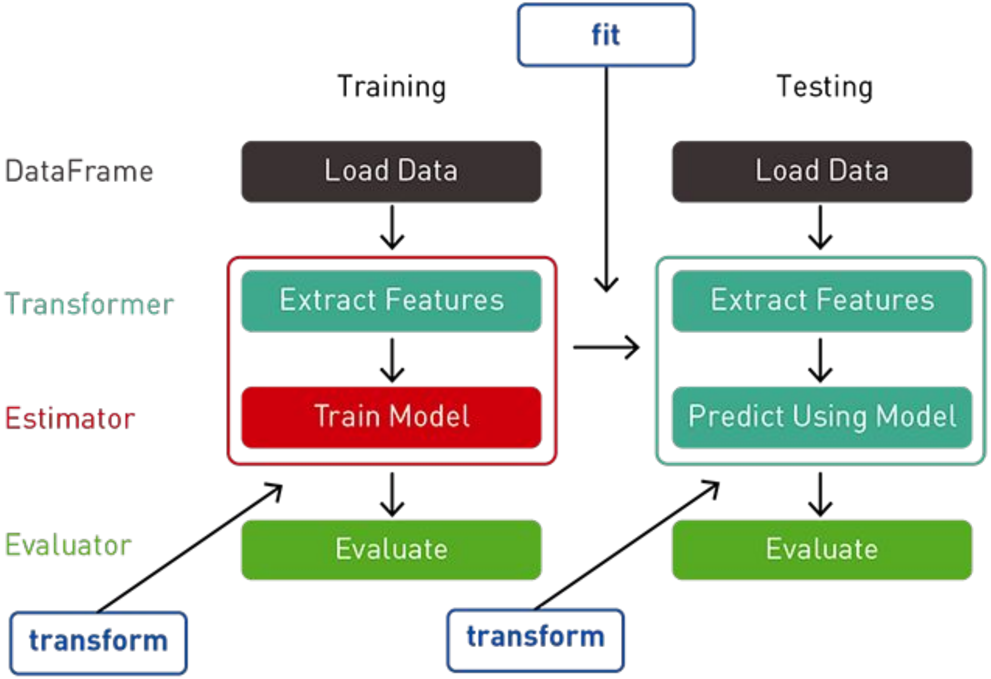


Image source

greatlearning

Power Ahead

Happy Learning !