← Go Back to Advanced Machine Learning

Course Content

FAQ - Bagging and Random Forest

1. In Bagging, we can employ various models as long as the ensemble contains one particular type of algorithm(task)?

In bagging, we use the same algorithms and give different data points to obtain different models.
Bagging is homogenous therefore all the algorithms used have to be the same. So suppose there is a bagging model with n_estimators as 10 and if we use the Decision tree then all 10 algorithms have to be decision tree. if any other algorithm supposes Logistic Regression then all 10 has to be Logistic Regression.

All the algorithms used in one bagging model have to be the same. 5 Decision Tree and 5 Logistic Regression algorithms won't work, all 10 have to be Decision Tree or all 10 have to be Logistic Regression.

2. How much difference between a model's train and test performance is considered as overfitting or underfitting?

There cannot be a universal threshold to decide the fit of a model. Underfit models are easy to identify but overfit models are relatively more difficult to identify as we can be sure of the model's fitness only when it is used in production. In practice, what matters is the actual test error, i.e. how often the predictions are wrong on new data. The limits on that are strictly a matter of business risk.

Theoretically, you can get any level of closeness you want between training and testing metrics, provided you have large enough sample space.
For example, if you are using a complex model like a deep neural network and have hundreds of thousands (or more) of observations, then you are entitled to see an average test error to be very close to an average training error, say well within 1%. But if you are using a simple model like linear regression or pruned decision tree and/or you only have a few observations,

say in the range of hundreds, then average test error exceeding average training error by even 10% may not be unreasonable.

So it boils down to how much difference is acceptable to that domain or industry, then you have to increase the data size until the train-test difference is smaller than that.

3. To build each individual tree in Random forest, whether a subset of only rows is taken or that of columns is also taken?

To prepare data for an individual tree, a subset of rows is taken but the entire set of features are available from the original data. While splitting at the nodes, a random subset of features is selected. Hence both, row and column sampling is done in Random Forest.

4. I am getting the below error with Bagging Classifier with Logistic Regression as base_estimator:

```
AttributeError: 'str' object has no attribute 'decode'
```

How to solve it?

The LogisticRegression function has a parameter named solver, which was earlier by default set to 'liblinear', but after an update in sklearn, the default solver was changed to 'lbfgs'. Kindly try setting the solver as liblinear, and try the code again, as shown below:

```
bagging_lr=BaggingClassifier(base_estimator=LogisticRegression(solver='liblinear', random_s
bagging_lr.fit(X_train,y_train)
```

Previous Next