Diploma in Computer Science and Engineering

5TH Semester

Artificial Intelligence and Machine Learning

Organization of week 8 of Session 6

❑ Explore and list the Ensemble Algorithms

We shall examine them in depth in this section. The algorithms on which we shall concentrate are as follows:

- Bagging algorithms:
  - Bagging meta-estimator
  - Random forest
- Boosting algorithms:
  - AdaBoost
  - GBM
  - XGBM
  - Light GBM
  - CatBoost

❑ Random Forest

Another ensemble machine learning algorithm that uses the bagging method is Random Forest. This approach is an expansion of the bagging estimator. Decision trees serve as the foundation estimators in random forests. In contrast to the bagging meta estimator, random forest chooses a set of features at random, using those characteristics to determine the optimum split at each decision tree node.

Looking at it step-by-step, this is what a random forest model does:

- Random subsets are created from the original dataset (bootstrapping).
- At each node in the decision tree, only a random set of features are considered to decide the best split.
- A decision tree model is fitted on each of the subsets.
- The final prediction is calculated by averaging the predictions from all decision trees.

To summarize, a Random forest **randomly** selects data points and features and builds **multiple trees (Forest).**

❑ Hyper parameters

- **n_estimators:**
  - It defines the number of decision trees to be created in a random forest.
  - Generally, a higher number makes the predictions stronger and more stable, but a very large number can result in higher training time.
- **criterion**:
  - It defines the function that is to be used for splitting.
  - The function measures the quality of a split for each feature and chooses the best split.
- **max_features**:
  - It defines the maximum number of features allowed for the split in each decision tree.
  - Increasing max features usually improve performance but a very high number can decrease the diversity of each tree.
- **max_depth**:
  - Random forest has multiple decision trees. This parameter defines the maximum depth of the trees.
- **min_samples_split:**
  - Used to define the minimum number of samples required in a leaf node before a split is attempted.
  - If the number of samples is less than the required number, the node is not split.
- **min_samples_leaf:**
  - This defines the minimum number of samples required to be at a leaf node.
  - Smaller leaf size makes the model more prone to capturing noise in train data.
- **max_leaf_nodes:**
  - This parameter specifies the maximum number of leaf nodes for each tree.
  - The tree stops splitting when the number of leaf nodes becomes equal to the max leaf node.
- **n_jobs**:
  - This indicates the number of jobs to run in parallel.
  - Set value to -1 if you want it to run on all cores in the system.
- **random_state**:
  - This parameter is used to define the random selection.
  - It is used for comparison between various models.

❑ Applications

There are mainly four sectors where Random forest is mostly used:

- **Banking:** The banking sector primarily uses this algorithm for the identification of loan risk.
- **Medicine:** With the help of this algorithm, disease trends and risks of the disease can be identified.
- **Land Use:** We can identify the areas of similar land use by this algorithm.
- **Marketing:** Marketing trends can be identified using this algorithm.

Advantage of Random Forest

- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the overfitting issue.

Disadvantages of Random Forest

- Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.