Course: Machine Learning Techniques (CSE3008) - Module2

Dr.Mohana S D,
Assistant Professor,
(Course In-charge - CSE3008)
School of Computer Science and Engineering & Information Science,
Presidency University Bengaluru.

GAIN MORE KNOWLEDGE
REACH GREATER HEIGHTS
PRESIDENCY
UNIVERSITY

Ensemble Learning

Ensemble Learning

- Ensemble learning is a technique for improving the accuracy of machine learning models by combining the predictions of multiple models.
- The basic idea is to train multiple models on different subsets of the data, and then combine their predictions to make a final prediction.
- Ensemble learning can improve the accuracy and robustness of models, and is widely used in many real-world applications.

Introduction

- Ensemble learning is a powerful technique used in machine learning to improve the accuracy and robustness of models.
- It involves combining multiple individual models to produce a more accurate prediction.
- There are several techniques used in ensemble learning, including bagging, pasting, random patches, random subspaces, and boosting.

- Ensemble learning is a machine learning technique that involves combining multiple models to improve predictive performance.
- One way to implement ensemble learning is by using a subset of instances from the training data.
- The process of using a subset of instances is also known as bagging (short for bootstrap aggregating).

It works based on:

1. Start by randomly selecting a subset of instances from the training data.
2. Train a model on this subset of instances.
3. Repeat steps 1 and 2 multiple times, each time selecting a different subset of instances and training a new model.
4. Combine the predictions of all the models to make a final prediction.

- The idea behind this approach is that by using different subsets of instances, each model will have a slightly different view of the data.
- By combining the predictions of all the models, we can reduce the impact of individual errors and increase the overall accuracy of the ensemble.
- Bagging can be used with many different types of models, including decision trees, neural networks, and support vector machines.
- It is a popular technique in ensemble learning and has been shown to be effective in many applications.

Voting Classifier

- The voting classifier is a simple but effective ensemble learning technique. It involves combining the predictions of multiple individual models to produce a final prediction.
- The individual models can be any type of classifier, such as logistic regression, decision trees, or support vector machines. The voting classifier takes a majority vote of the individual model predictions to produce the final prediction.
- This technique works best when the individual models are diverse and have different strengths and weaknesses.

Random Forest

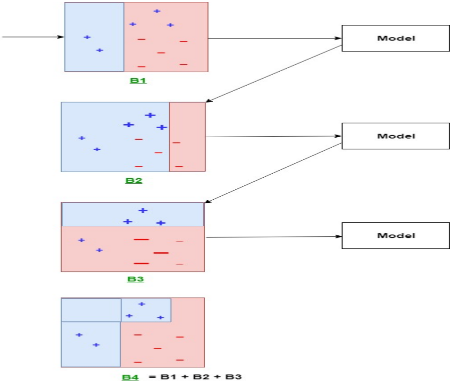
- Random Forest is another popular ensemble learning technique. It is a type of decision tree ensemble that creates a forest of decision trees.
- Each tree is created using a random subset of the data and a random subset of the features.
- This randomness helps to reduce overfitting and improve the accuracy of the model.
- The final prediction is made by taking the majority vote of the predictions of all the trees in the forest.

Boosting

- Boosting is a technique used to improve the accuracy of weak learners by combining them into a strong learner. The basic idea is to create a sequence of models, where each model is trained to correct the errors of the previous model.
- AdaBoost is a popular boosting algorithm that assigns weights to the training data points based on their classification error.
- The final model is a weighted sum of the individual models.
- Gradient boosting is another popular boosting algorithm that uses gradient descent to optimize a loss function.
- It trains a sequence of models, where each model tries to correct the errors of the previous model.
- In summary, ensemble learning is a powerful technique that can be used to improve the accuracy and robustness of machine learning models.
- The techniques discussed in this document, including the voting classifier, random forest, and boosting algorithms like AdaBoost and

AdaBoost

It was the first really successful boosting algorithm developed for the purpose of binary classification. AdaBoost is short for Adaptive Boosting and is a very popular boosting technique that combines multiple “weak classifiers” into a single “strong classifier”.



Gradient Boosting

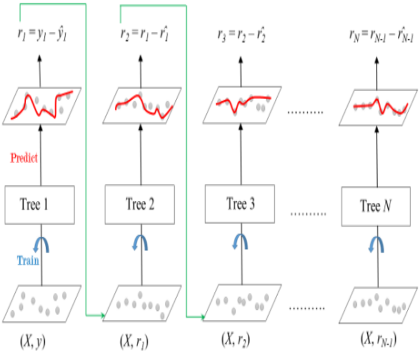
It is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor's error. In contrast to Adaboost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of predecessor as labels. There is a technique called the Gradient Boosted Trees whose base learner is CART (Classification and Regression Trees).

Shrinkage

There is an important parameter used in this technique known as Shrinkage. Shrinkage refers to the fact that the prediction of each tree in the ensemble is shrunk after it is multiplied by the learning rate (eta) which ranges between 0 to 1. There is a trade-off between eta and number of estimators, decreasing learning rate needs to be compensated with increasing estimators in order to reach certain model performance. Each tree predicts a label and final prediction is given by the formula,

$y(pred) = y1 + (eta * r1) + (eta * r2) + ... + (eta * rN)$

Gradient Boosting



Extremly Randomizing and Stacking

Extremely Randomized Trees

- Extremely Randomized Trees, or ExtraTrees for short, is an extension of the Random Forest algorithm.
- In ExtraTrees, the decision trees are built using random splits on random subsets of the features.
- This randomness helps to reduce the variance of the model and improve its accuracy. In addition, ExtraTrees can also be used for regression and classification problems.
- ExtraTrees can be especially useful when dealing with noisy or imbalanced datasets, as it is less prone to overfitting than other ensemble learning techniques.

Stacking

- Stacking is a more complex ensemble learning technique that involves combining the predictions of multiple individual models using another model.
- In stacking, the individual models are first trained on the training data, and their predictions are then used as input features for a second-level model.
- The second-level model is trained on the outputs of the individual models, which can help to improve the accuracy of the predictions.
- Stacking can be used with any type of model, including regression and classification models, and it can be especially effective when the individual models are diverse.
- In summary, Extremely Randomized Trees and Stacking are two powerful techniques that can be used to further improve the accuracy and robustness of machine learning models.
- Extremely Randomized Trees can be especially useful when dealing with noisy or imbalanced datasets, while Stacking can be used to combine the strengths of multiple individual models to produce more accurate predictions.

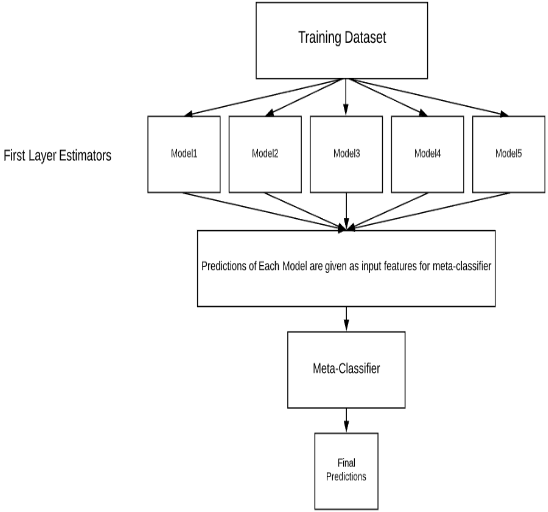
Stacking

- Stacking is a way of ensembling classification or regression models it consists of two-layer estimators.
- The first layer consists of all the baseline models that are used to predict the outputs on the test datasets.
- The second layer consists of Meta-Classifier or Regressor which takes all the predictions of baseline models as an input and generate new predictions.

Stacking

- Then you add a new model which learns from the intermediate predictions the same target.
- This final model is said to be stacked on the top of the others, hence the name.
- Thus, you might improve your overall performance, and often you end up with a model which is better than any individual intermediate model.
- However, that it does not give you any guarantee, as is often the case with any machine learning technique.

Stacking



Training Dataset

First Layer Estimators

Model1

Model2

Model3

Model4

Model5

Predictions of Each Model are given as input features for meta-classifier

Meta-Classifier

Final Predictions