Machine Learning

Assignment 5.1

Submitted By: Ranji Raj

November 27, 2020

Bias-Variance Trade-Off

- The ideology of any ML model is to best estimate the **target** function f which best aims at approximating the output variable Y given the input data X.
- When a model gives a prediction it is often always associated with a prediction error with an error margin which is decomposed into:
  1. **Irreducible Error:** It is that portion of the error which is always present regardless which kind of algorithm is used. Factors causing could be poor attribute selected that has no significance to model prediction.
  2. **Bias Error:** Bias are the simplifying assumptions made by a model to make the target function easier to learn. Mathematically, it is the difference between the *mean of the estimates* and the *actual* value. Linear algorithms, in general have a **high bias** which makes them fast learners and easier to understand but, poor in flexibility. Due to which they have lower predictive power on unseen problems that are complex. *Decision Trees are an example of **low-bias** algorithm.*
  3. **Variance Error:** Variance is the marginal estimate of the target function that varies when different training data was used. In particular, non-linear algorithms that have a remarkable adaptive capability have **high variance**. *Decision Trees are an example of **high-variance** algorithm.*

Overfitting in Decision Trees

- If the number of levels are too high or higher amount of splits i.e. a sophisticated tree, then it tends to overfit.
- Intuitively, when there are too many decision nodes to go through before arriving at the result. i.e number of nodes to traverse before reaching the leaf nodes are

high, the conditions that we are checking against becomes multiplicative. Meaning, the computation becomes *condition1 AND condition2 AND condition3 AND condition4 AND condition5.*

- Only if all the conditions are satisfied, a decision is reached. This works very well for the training set as you are continuously narrowing down the data.
- But, as soon a new test data is fed, even if one of the parameters deviates slightly, the condition will not be met and it will take the wrong branch.
- Hence, Decision Trees are called as **unstable** classifiers. In other words, the bias-variance trade-off does depend on the tree depth and also sensitive as to where and how it splits the data.
- In the context of decision trees, careful supervision of the input features considering their type and clever decision of performing splits can prevent the potential problems of over-fitting.

TWO alternate methods for feature splits

- Node-splitting is simply the process of dividing a node into multiple sub-nodes to create relatively homogeneous nodes.
- Depending upon the type of target variable we can broadly split for:
  1. **Continuous** target variable: By using **Reduction in Variance**, when dealing with regression problems. It is so-called because it uses variance as the underlying measure for deciding the feature on which node is split into child nodes.
  2. Mathematically written as,
  $$\frac{\sum(X-\mu)^2}{N}$$
  where X = input features, μ = mean, N = Number of examples.
  3. Variance, is used for calculating the homogeneity of a node. If a node is pure, then the variance is *zero*.
  4. **Categorical** target variable: By using **Information Gain**, based on the concept of entropy.
  5. Mathematically written as,
  $$Gain = 1 - Entropy$$
  and
  $$Entropy = -\sum_{i=1}^{C} p_i \log_C p_i$$
  6. Lower the value of entropy, higher the purity of the node. The entropy of a homogeneous node is *zero*.