# Machine Learning

**Assignment 5.4**

Submitted By: Ranji Raj

November 24, 2020

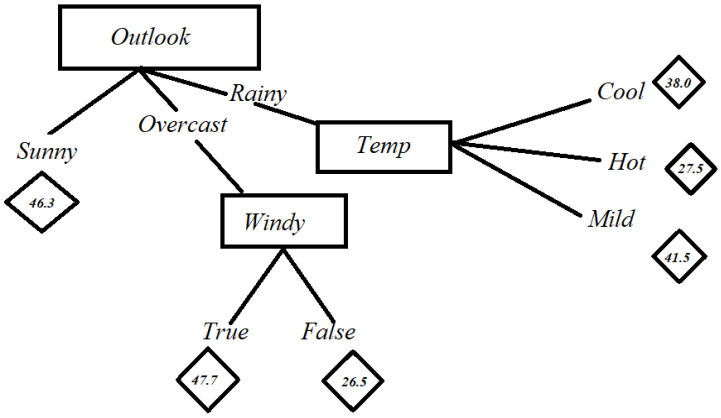


Figure 1: Regression Trees

## a) Regression Tree vs. Classification Tree

| Parameter | Classification Trees | Regression Trees |
|---|---|---|
| Target column | Categorical | Continuous |
| Split criteria | Information Gain | Variance Reduction |
| Application | Spam classification | Predicting pandemic cases |

Table 1: Basic Differences

## b) Role of SSE in deciding the split points

- To calculate the homogeneity of a numerical sample, *SSE=variance* (or equivalently standard deviation) is used.

$$\sigma = \sqrt{\frac{\sum_{i=i}^{n}(x_i - \mu)^2}{N}}$$

  where, $\mu$ is the mean given as, $\frac{\sum x}{N}$, N = Number of samples.

- Interpretation of the standard deviation value when equal to zero indicates the sample is pure.

- Constructing a regression tree is all about finding an attribute that returns the **highest** standard deviation reduction.

- The process of recursive split using SSE is done in the following way:
  1. The standard deviation of the tree is calculated for the first time.
  2. The dataset is then split on different features.
  3. The standard deviation for each branch is calculated.
  4. The resulting standard deviation is subtracted from the standard deviation before the split. This leads to reduction is standard deviation.
  5. The attribute with the largest standard deviation reduction is chosen for the decision node.
  6. The dataset is divided based on the values of the selected attribute. This process is run recursively on the non-leaf branches, until all data is processed.

## c) Stopping criteria

- The **Coefficient of Variation (CV)**(or synonymously, Coefficient of Deviation) can be used to terminate tree growth when smaller than a threshold.

$$CV = \frac{\sigma}{x}$$

- Other natural way of tree halting is, when all the training examples are reached or when too few instances remain in the branch.