Random Forest (RF)

Computational Data Science,
Addis Ababa University

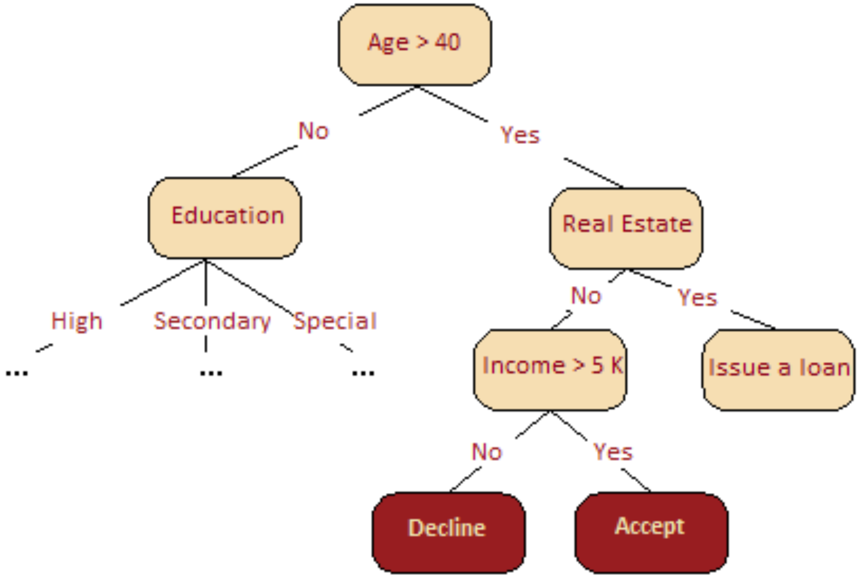
www.aau.edu.et

mesfin.diro@aau.eud.et

+251-912-086156

Decision Tree

- A decision tree is a simple decision making-diagram
- A decision tree is often a generalization of the experts' experience, a means of sharing knowledge of a particular process.
- The decision to grant a loan was made on the basis of some intuitively (or empirically) derived rules that could be represented as a decision tree.



Decision Tree

- Decision trees learn how to best split the dataset into smaller and smaller subsets to predict the target value.
- The condition is presented as the "node" and the possible outcomes as edges(branches)
- This splitting process continues until no further gain can be made or a present rule is met. e.g the maximum depth of the tree is reached.
- Random forests are a large number of decision trees combined (using average or majority rule) at the end of the process
- Gradient boosting machines are also combine decision tree but start the combining process at the beginning, instead of at the end

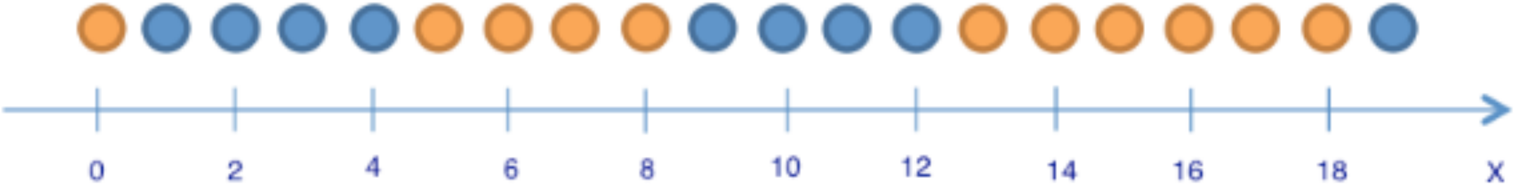
Entropy

- **Entropy can be described as the degree of chaos in the system.**
- **The higher the entropy, the less ordered the system and vice versa. This will help us formalize "effective data splitting”**
- **Entropy is defined for a system with N possible states as follows:**

$$S = -\sum_{i=1}^{N} p_i \log_2 p_i$$

where p_i is the probability of finding the system in the i-th state. This is very important concept used in physics, information theory and other areas

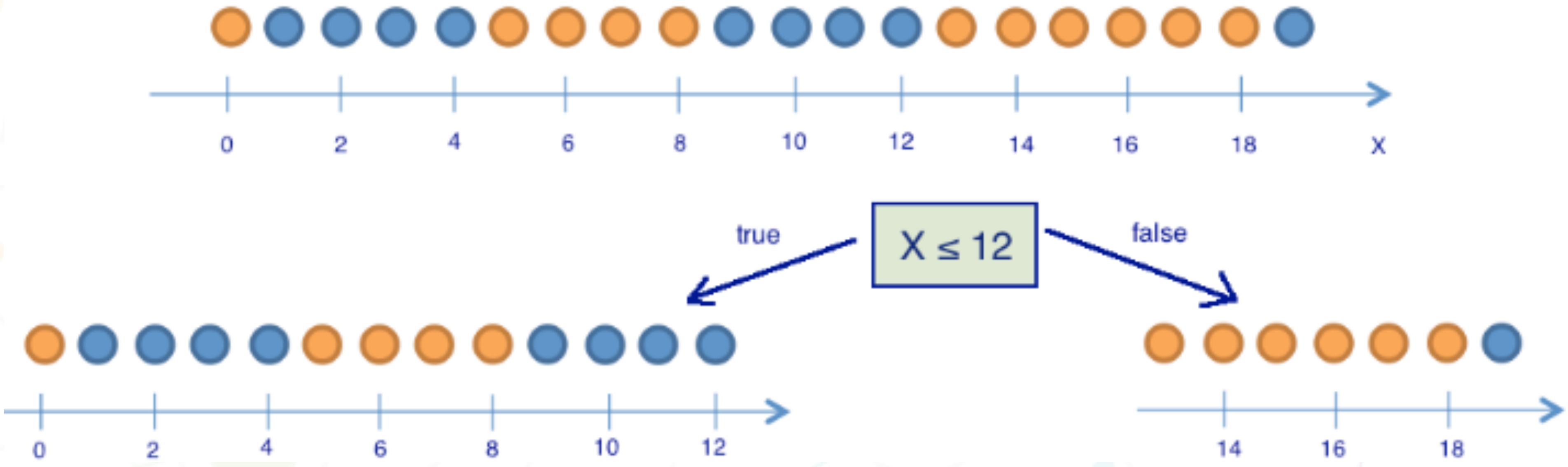
Entropy



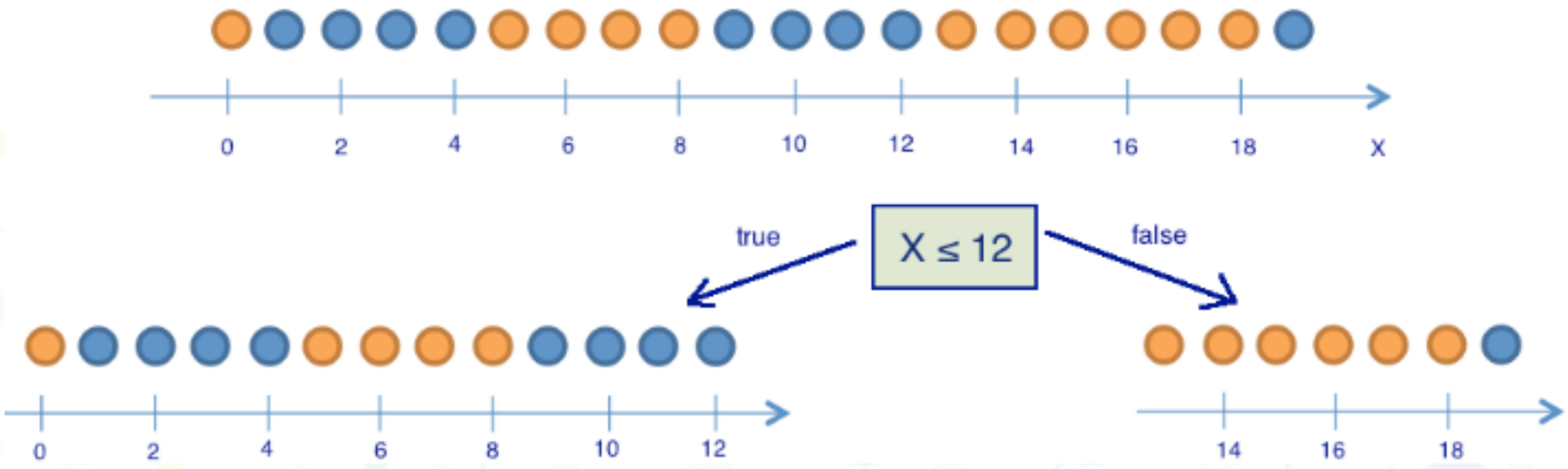
- **There are 9 blue balls and 11 yellow balls. If we randomly pull out a ball, then it will be blue with probability** $p_1 = \frac{9}{20}$ **and yellow with probability** $p_2 = \frac{11}{20}$ **which gives un an entropy** $S_0 = -\frac{9}{20}\log_2\frac{9}{20} - \frac{11}{20}\log_2\frac{11}{20} \approx 1$

Entropy

- The above entropy value by itself may not tell us much, but let's see how the value changes if we were to break the balls into two groups: with the position less than or equal to 12 and greater than 12.



Entropy



- **The left group has 13 balls, 8 blue and 5 yellow. The entropy of this group is**

$$S_1 = -\frac{5}{13}\log_2\frac{5}{13} - \frac{8}{13}\log_2\frac{8}{13} \approx 0.96$$

- **The right group has 7 balls, 1 blue and 6 yellow. The entropy of the right group is** $S_2 = -\frac{1}{7}\log_2\frac{1}{7} - \frac{6}{7}\log_2\frac{6}{7} \approx 0.6$

Information Gain

- Since entropy is, in fact, the degree of chaos (or uncertainty) in the system, the **reduction in entropy** is called **information gain**. Formally, the information gain (IG) for a split based on the variable Q (in this example it's a variable$x \leq 12$ is defined as

$$IG(Q) = S_O - \sum_{i=1}^{q} \frac{N_i}{N} S_i,$$

- Where q is the number of groups after the split, N_i is number of objects from the sample in which variable Q is equal to the i-th value. In our example, our split yielded two groups $(q = 2)$, one with 13 elements $(N_1 = 13)$, the other with 7$(N_2 = 7)$ Therefore, we can compute the information gain as

$$IG(x \leq 12) = S_0 - \frac{13}{20} S_1 - \frac{7}{20} S_2 \approx 0.16.$$

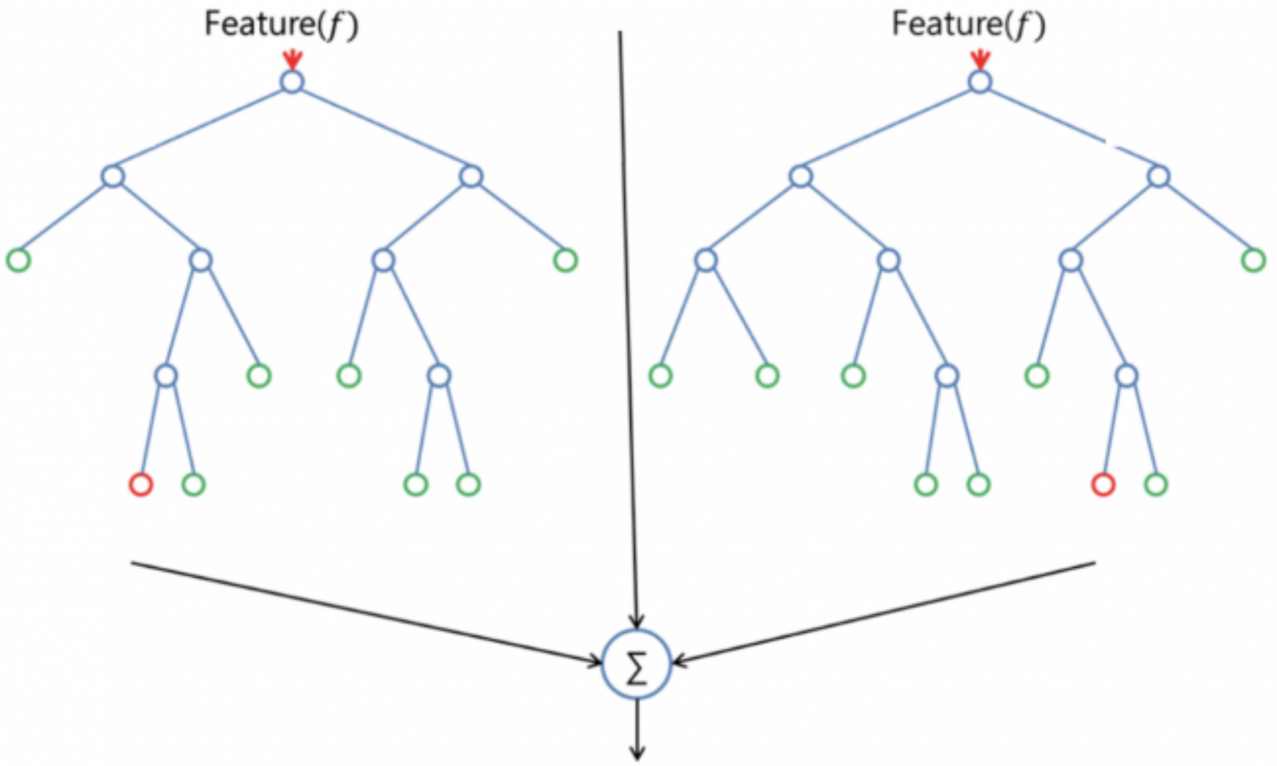
Random Forest(RF)

- Random Forest is a supervised machine learning algorithm which is based on ensemble learning for both classification and regression problems
- It is one of the most flexible and easy to use algorithm. It creates decision trees on the given data samples, gets prediction from each tree and selects the best solution by means of voting. It is also a pretty good indicator of feature importance.
- Random forest algorithm combines multiple decision-trees, resulting in a forest of trees, hence the name Random Forest.
- In the random forest classifier, the higher the number of trees in the forest results in higher accuracy.

Random Forest Algorithm

- **Random forests (RF) construct many individual decision trees at training.**
- **Predictions from all trees are pooled to make the final prediction; the mode of the classes for classification or the mean prediction for regression.**
- **As they use a collection of results to make a final decision, they are referred to as Ensemble techniques:**
  - **Step 1: Start with the selection of Random samples from a given dataset**
  - **Step 2: Construct a decision tree for every feature**
  - **Step 3: Voting for every predicted result**
  - **Step 4 : Select the most voted prediction result as the finial prediction result**

Random Forest Algorithm



Tree Algorithms

ID3

Iterative Dichotomiser 3

- Create a multiway tree for each node for categorical features
- Trees are grown to their maximum
- a pruning step is usually applied to generalize to unseen data

C4.5

Successor to ID3

- Removed the restriction that features must be categorical by dynamically defining a discrete attribute
- It Partitions the continuous attribute value into a discrete set of intervals.

C5.0

Quinlan's latest

- Latest Version release under a propriety license.
- It uses less memory and builds smaller rulesets than C4.5

CART

Classification and Regression Tree

- it is very similar to C4.5
- But it differs in that it supports regression
- does not compute rule set
- it constructs binary tree using the features and threshold

Mathematical Formulation

- Given training vectors x_i and label vector y, a decision tree recursively partitions the feature space such that the samples with the same labels are grouped together
- Let the data at node m be represented by Q_m with N_m samples . For each candidate split $\theta = (j, t_m)$ consisting of a feature j and threshold t_m, partition the data into

$$Q_m^{left}(\theta) = \{(x, y) | x_j \leq t_m\}$$

$$Q_m^{right}(\theta) = Q_m \backslash Q_m^{left}(\theta)$$

Splitting Criteria

- Entropy is the measure of impurity of a node that defined(for a binary class with values) as :

$$\text{entropy(t)} = -\sum_{k} p_{mk} \log(p_{mk})$$

- Gini index is another measures of impurity that measures the divergence between the probability distributions of the target attribute's values. It is defined as:

$$Gini\ Index = \sum_{k} p_{mk}(1 - p_{mk})$$

- Classification Error is computed as:

$$Classification\ Error(t) = 1 - \max[p_{mk}]$$

Where, p_{mk} denote the fraction of records belonging to class m at a given node k

Mathematical Formulation

- **The quality of a candidate split m is then computed using a impurity function or loss function $H()$, the choice of which depends on the task being solved(classification or regression)**

$$G(Q_m, \theta) = \frac{N_m^{left}}{N_m} H(Q_m^{left}(\theta)) + \frac{N_m^{right}}{N_m} H(Q_m^{right}(\theta))$$

- **Select the parameters that minimizes the impurity:**

$$\theta* = \operatorname{argmin}_\theta G(Q_m, \theta)$$

- **Recurse for subsets $Q_m^{left}(\theta*)$ and $Q_m^{right}(\theta*)$ until the maximum allowable depth is reached , $N_m < \min_{samples}$ or $N_m = 1$**

Classification Criteria

- If a target is a classification outcome taking on values $0, 1, \cdots, K-1$, for node m, let

$$p_{mk} = 1/N_m \sum_{y \in Q_m} IG(y = k)$$

- be the proportion of class k observations in node m. If m is a terminal node, $predict\_proba$ for this region is set to p_{mk}. Common measures of impurity are the following:

Gini: $$H(Q_m) = \sum_k p_{mk}(1 - p_{mk})$$

Entropy: $$H(Q_m) = -\sum_k p_{mk} \log(p_{mk})$$

Misclassification: $$H(Q_m) = 1 - \max(p_{mk})$$

Regression Criteria

- **If the target is a continuous value, then for node m, common criteria to minimize as for determining locations for future splits are Mean Squared Error (MSE or L2 error), Poisson deviance as well as Mean Absolute Error (MAE or L1 error).**
- **MSE and Poisson deviance both set the predicted value of terminal nodes to the learned mean value $\bar{y}_m$ of the node whereas the MAE sets the predicted value of terminal nodes to the median $median(y)_m$.**

$$\bar{y}_m = \frac{1}{N_m} \sum_{y \in Q_m} y$$

Mean Squared Error:

$$H(Q_m) = \frac{1}{N_m} \sum_{y \in Q_m} (y - \bar{y}_m)^2$$

Regression Criteria

Half Poisson deviance:

$$H(Q_m) = \frac{1}{N_m} \sum_{y \in Q_m} (y \log \frac{y}{\bar{y}_m} - y + \bar{y}_m)$$

$$median(y)_m = \underset{y \in Q_m}{\mathrm{median}}(y)$$

Mean Absolute Error:

$$H(Q_m) = \frac{1}{N_m} \sum_{y \in Q_m} |y - median(y)_m|$$

THANK YOU!