Weekly report of lessons

Name: Suryam Arnav Kalra
Roll No: 19CS30050
The week: 30th August 2021 to 3rd September 2021

The topics covered:

- Gain Function
- ID3 Algorithm
- Hypothesis Space Search in Decision Trees
- Restriction bias vs Preference bias and Inductive bias in ID3
- Overfitting the data and Avoiding overfitting
- Evaluating subtrees to prune, Reduced error pruning and Rule post pruning
- Extensions of basic algorithm
- Gini Index
- Regression Tree and Pruning of regression tree
- Sequential covering algorithms and General to Specific beam search
- Oblique decision tree and Random decision forest

Summary topic wise:

- Gain Function: It is a measure of reduction of uncertainty.
- ID3 Algorithm: ID3 is a greedy algorithm that grows the tree top-down, at each node selecting the attribute that best classifies the local training examples. This process continues until the tree perfectly classifies the training examples, or until all attributes have been used.
- Hypothesis Space Search in Decision Trees: Conduct a search of the space of decision trees which can represent all possible discrete functions using a greedy heuristic search: hill climbing without backtracking.
- Restriction bias vs Preference bias: Restriction bias => Incomplete hypothesis space and Preference bias => Incomplete search strategy. ID3 has preference bias.
- Inductive bias in ID3: The algorithm has preference for short trees and for those with high information gain attributes near the root.
- Overfitting the Data: A hypothesis h overfits the training data if there exists another hypothesis h' which has greater error than h on the training data but smaller error on the test data than h.
- Avoiding overfitting: There are two basic approaches to avoid overfitting:
  - Prepruning: Stop growing the tree at some point during construction when it is determined that there is not enough data to make reliable choices.
  - Postpruning: Grow the full tree and then remove nodes that seem not to have sufficient evidence.
- Evaluating subtrees to prune: We can use methods such as Cross-Validation, Statistical Testing and Minimum Description Length to evaluate subtrees.
- Reduced Error Pruning: It considers each of the decision nodes in the tree to be candidates for pruning and then permanently prunes the node with the greatest increase in accuracy on the validation set.

- Rule Post Pruning: Convert the best-fit tree to an equivalent set of rules and then prune each rule independently of others and sort the final set of rules.
- Extensions of basic algorithm: Some of the extensions of the basic algorithm are:
    - Continuous valued attributes: Create a discrete attribute from continuous variables and choose appropriate thresholds.
    - Attributes with many values: Use $GainRatio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)}$ where $SplitInformation(S,A) = -\sum \frac{|S_i|}{|S|} \log \left(\frac{|S_i|}{|S|}\right)$
    - Unknown Attribute Values: Assign most common value of attribute among other examples with the same target value.
    - Attributes with cost: Replace $Gain(S,A)$ by $\frac{Gain(S,A)^2}{Cost(A)}$
- Gini Index: It is another sensible measure of impurity : $Gini = 1 - \sum p(i)^2$
- Regression Tree: These are a type of decision trees in which the target variable can take continuous value (real numbers) and they partition the attribute space into a set of rectangular subspaces each with its own predictor.
- Pruning of Regression Trees: We can apply pre-pruning and post-pruning methods in which the tree that minimizes the squared error on VS is chosen.
- Sequential Covering Algorithms: It is based on the strategy of learning one rule and then removing the data it covers and repeating the process until all examples are covered.
- General to Specific Beam Search: It is similar to the ID3 algorithm but preferentially explores the paths of DT covering most examples by a combination of BFS and DFS.
- Oblique Decision Tree: In contrast to the Decision Tree, here in a node a combination of multiple attributes is checked for further division.
- Random Decision Forest: We obtain multiple trees from randomly constructed subspaces and use a voting procedure to obtain the class label of the data being predicted.

Concepts challenging to comprehend:

Pruning methods along with sequential covering algorithms are a little bit challenging to comprehend.

Interesting and exciting concepts:

Regression Trees, Oblique Decision Trees and Random Decision Forest are quite interesting and exciting to learn.

Concepts not understood:

After going through the book and the video lectures the concepts are clearly understood.

Any novel idea of yours out of the lessons:

Decision Tree structure can be particularly used by students for a variety of uses. It can help us determine what subjects to choose based on a variety of attributes such as the score distribution, the average workload and many more. It can also be used by MNCs to make predictions on the requirements of their buyers by analyzing their sales patterns and maximize profits by making better business models.