Weekly report of lessons

Name: Ravi Pratap Singh
Roll No:20CS60R60
The week: 23 sept , 24 sept , 25 sept

The topics covered: Hypothesis space search by ID3,Restriction bias and Preference bias ,Overfitting of data ,pre pruning and post pruning,Attribute with many values,GiniIndex ,Regression Tree, Advantages and disadvantages,Discussion on paper on Oblique decision tree, Conditional Probability, Bayes' Rule

Summary topic wise :

- ID3's hypothesis space of all decision spaces is a *complete* space of finite discrete-valued functions, relative to the available attributes.Because every finite discrete-valued function can be represented by some decision tree, ID3 avoids one of the major risks of methods that search incomplete hypothesis spaces .ID3, in its pure form, performs *no backtracking* in its search.ID3 maintains only a *single current hypothesis* as it searches through the space of decision trees .
- ID3 exhibits purely preference bias. A preference bias is more desirable than restriction bias, because it allows learners to work within a complete hypothesis space that is assured to contain the unknown target function.
- Overfitting a model is a condition where a statistical model begins to describe the random error in the data rather than the relationships between variables. This problem occurs when the model is too complex.
- Pre-pruning that stops growing the tree earlier, before it perfectly classifies the training set. Post-pruning that allows the tree to perfectly classify the training set, and then post prune the tree.
- In case of attribute with many values we can use gainRatio , defined as

$$Gain(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)} \quad , \; SplitInformation(S,A) = -\Sigma \frac{Si}{S} log \frac{Si}{S}$$

- Gini index is another measure of impurity , calculated as

$$Gini(A) = \sum_{v} p(v) \sum_{i \neq j} p(i/v) p(j/v)$$

- Regression Tree is similar to classification . Its uses a set of attributes to predict the value instead of class labels.Impurity of sample is defined in terms of Variance of the output in the learning sample. The best split is one which reduces the most variance.
- Adv- Compared to other algorithms, decision trees require less effort for data preparation during pre-processing.A decision tree does not require normalization of data. Disadv- A small change in the data can cause a large change in the structure of the decision tree causing instability. For a Decision tree sometimes calculation can go far more complex compared to other algorithms.
- Conditional probability : $P(A/B) = P(A) * P(B/A)$

  If A and B are independent : $P(A/B) = P(A)$, $P(B/A) = P(B)$
- Bayes' Rule : $P(B/A) = \frac{P(A/B)\,P(B)}{P(A)}$

Do you find doubt clearing session useful? : Yes,the doubt clearing session was very informative and useful as I was able to think about the subject from others perspective . It helped me to know more about the topics which I was weak at. Some of the discussion of assignment also helped me to go ahead with it.

Should there be more such sessions, and, if so, how frequently? : Yes,These kinds of sessions should be conducted , maybe in a interval of 2 week or so.