ITPC-308
Machine learning Concepts
6th Semester (Assignment - 1)

18124009
Bindheshwari Singh

Set 2

Ques 1 :- Explain unbiased learner with an example.

Ans 1 :- ⇒ The obvious solution to the problem of assuring that the target concept is in the hypothesis space H is to provide a hypothesis space capable of representing every teachable concept.

- Every possible subset of the instances (the power set of X)

⇒ What is the size of the hypothesis space H (the power set of X)?

- In EnjoySport, the size of the instance space X is 96.
- The size of the power set of X is $2^{|X|}$ ⇒ The Size of H is 2^{96}
- Our Conjunctive hypothesis space is able to represent only 973 of these hypotheses. ⇒ a very biased hypothesis space.

⇒ Let the hypothesis space H to be the power set of X.

- A hypothesis space H to be the power set of X.
- A hypothesis can be represented with disjunctions, Conjunctions and negations of our earlier hypotheses.
- The target concept "Sky = Sunny or Sky = Cloudy" could then be described as

<Sunny, ?, ?, ?, ?, ?> V <Cloudy, ?, ?, ?, ?, ?>

Problem : Our Concept learning algorithm is now Completely unable to generalize beyond the observed examples.

- three positive examples (x1, x2, x3) and two negative examples (x4, x5) to the learner.

// S: {x1 V x2 V x3} and G: {¬(x4 V x5)} → NO GENERALIZATION

// Therefore, the only examples that will be unambiguously classified by S and G are the observed training examples themselves

Ques 2: — How to decide or select the attribute that is most useful for classifying examples for decision tree learning?

Ans: — Information gain measures how well a given attributes separates the training examples according to their target classification.

// In order to define information gain precisely, we use a measure commonly used in information theory, called entropy.

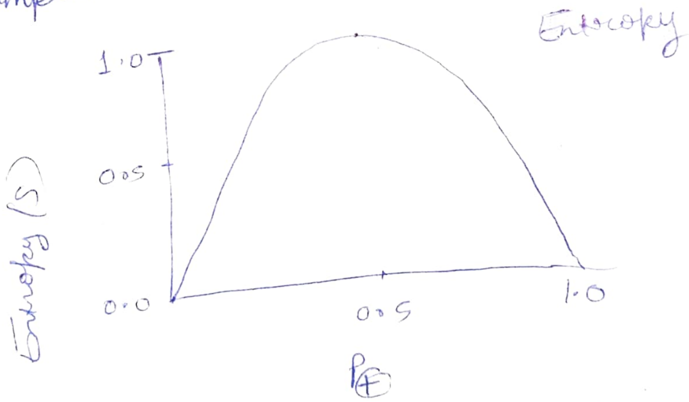
// Entropy characterizes the (im) purity of an arbitrary collection of examples.

$$\text{Entropy}(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

Entropy (S) → Sample of training Examples

p_+ → Proportion of positive examples

p_- → proportion of negative examples.



// Entropy - Non Boolean Target classification:

If the target attribute can take on c different values, then the entropy of S relative to this c-wise classification is defined as

$$\text{Entropy}(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

p_i → proportion of S belonging to class i.

⇒ entropy is a measure of the impurity in a collection of training examples.

⇒ Information gain is a measure of the effectiveness of an attribute in classifying the training data.

⇒ Information gain measures the expected reduction in entropy by partitioning the examples according to an attribute

$$Gain(S, A) = Entropy(S) - \sum_{\boxed{v \in Values(A)}} \left(|S_v| / |S| \right) Entropy(S_v)$$

A → Attribute

S → a Collection of examples

$Values(A)$ → Possible values of attribute A

S_v → the Subset of S for which attribute A has value v

⇒ Which attribute is the best classifier?

eg S: [29+, 35-]

Attributes : A and B

possible values for A : a, b

possible values for B : c, d

$$Entropy([29+, 35-]) = -29/64 \log_2 29/64 - 35/64 \log_2 35/64 = 0.99$$

A ?

a: [21+, 5-]

b: [8+, 30-]

$E([29+, 35-]) = 0.99$

B ?

c: [18+, 33-]

d: [11+, 2-]

$E([29+, 35-]) = 0.99$

A ?

a: [21+, 5-]

b: [8+, 30-]

$E([21+, 5-]) = 0.71$ $E([8+, 30-]) = 0.74$

$$Gain(S, A) = Entropy(S) - 26/64 * Entropy([21+, 5-]) - 38/64 * Entropy([8+, 30-])$$

$= 0.27$

B ?

c: [18+, 33-]

d: [11+, 2-]

$E([18+, 33-]) = 0.94$

$E([11+, 2-]) = 0.62$

$$Gain(S, B) = Entropy(S) - 51/64 * Entropy([18+, 33-]) - 13/64 * Entropy([11+, 2-])$$

$= 0.12$

A provides greater information gain than B

A is a better classifier than B.

Ques 3. How to avoid overfitting the data in the decision tree learning?

Ans- Overfitting is a significant practical difficulty for decision tree models and many other predictive models.
Overfitting happens when the learning algorithm continues to develop hypothesis that reduce training set error at the cost of an increased test set error. There are several approaches to avoiding overfitting in building decision trees.

1) Pre-pruning that stop growing the tree earlier, before it perfectly classifies the training set.
2) Post-pruning that allows the tree to perfectly classify the training set, and then post prune the tree.

Practically, the second approach of post pruning overfit trees is more successful because it is not easy to precisely estimate when to stop growing the tree.
The important step of tree pruning is to define a criterian be used to determine the correct final tree size using one of the following methods:

1. use a distinct dataset from the training set (called validation set), to evaluate the effect of post-pruning nodes from the tree.
2. Build the tree by using the training set, then apply a statistical test to estimate whether pruning or expanding a particular node is likely to produce an improvement beyond the training set.
   - Error estimation
   - Significance testing (e.g., chi-square test)

3. Minimum Description Length Principle: Use an explicit measure of the Complexity for encoding the training set and the decision tree, stopping growth of the tree when this encoding size (Size(tree) + size(misclassifications(tree)) is minimized.

The first method is the most common approach. In this approach the available data are separated into two sets of examples: a training set, which is used to build the decision tree, and a validation set, which is used to evaluate the impact of pruning the tree. The second method is also a common approach. Here, we explain the error estimation and chi^2 test.

Post-pruning Using Error estimation :-

Error estimate for a sub-tree is weighted sum of error estimates for all its leaves. The error estimate (e) for a node is:

$$e = \left(f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}\right) \Big/ \left(1 + \frac{z^2}{N}\right)$$

where:

→ f is the error on the training data.
→ N is the number of instances covered by the leaf
→ z from normal distribution.

Post-pruning Using chi^2 test

In chi^2 test we construct the corresponding frequency table and calculate the chi^2 value and its probability.

	Bronze	Silver	Gold
Bad	4	1	4
Good	2	1	2

chi^2 = 0.21 Probability = 0.90 degree of freedom = 2

If we require that the probability has to be less than limit (e.g., 0.05), therefore we decide not to split the node.

5)

Ques 4. How will you justify that Shorter trees are preferred over larger trees?

Ans: A decision tree that uses Information Gain to decide on the branches tends to overfit with increasing depth. A fully grown decision tree will have an entropy of 0. While this might sound great, most probably the tree is overfitting the training data and will perform bad on test data. A shorter tree most of the time generalizes better.

Hope this.