Name:_____ Roll Number:_____

Quiz-4

Max points: 20 **Max Time: 20 mins**

Q.1. [6 points]

Mark True (T) or False (F), fill in the blanks, or choose the correct choice for the statements below.

1. Entropy of a dataset having three classes and equal number of instances of each class is:

(A) 0 (B) 1 (C) 0.82 (D) 1.58

2. Approximate inductive bias of a decision tree is:

(A) All the attributes in a dataset must be selected (B) Deeper trees are preferred

(C) Shallower trees are preferred (D) None of the given options

3. To incorporate continuous attributes in a decision tree algorithm, we can discretize them so that the information gain is maximized.

(A) True (B) False

4. Bootstrapping in random forests means ~~that~~:

(A) Creating multiple boosted datasets from the original one sampled without replacement

(B) Creating multiple boosted datasets from the original one sampled with replacement

5. All features of a classification problem must be used in random forests.

(A) True (B) False

6. Decision trees have lower variance than the random forests.

(A) True (B) False

Q.2. [3+2+5+4 points]

a) Gain ratio should be used instead of information gain in decision trees. Why?

Attributes with many values bias selection in their favor if information gain alone is used. To undo this effect, we may use gain ratio, which penalizes the information gain by the splitting tendency of such attributes. Splitting tendency is measured the same way as the information gain but with respect to the attribute values instead of the target classification.

Name:_____ Roll Number:_____

b) Give a problem associated with the gain ratio metric and a solution to address it.

Gain ratio, however, has its own problem when the fraction of a certain attribute value dominates the dataset and drives it to infinity or a very large number. To remedy this, we may opt to use the gain ratio metric only for those attributes with above average gain.

c) Why do we use the rule post-pruning in decision trees? How it's done?

Rule post-pruning is a method that tries to reduce the variance in a decision tree. It is done as follows:
i) Induce the decision tree and let it overfit the data
ii) Represent each path from the root to the leaves as a decision rule
iii) Prune, i.e. remove, any antecedents by which estimated accuracy (e.g. over a validation set) improves
iv) Use the rules by the order of their estimated accuracy when classifying unknown instances

Name:_____ Roll Number:_____

d) How does a random forest outputs a decision for classification and regression?

For classification, the mode of the predicted labels is considered. For regression, we may average the value over all the decision trees.