CS-233 Theoretical Exercise

Mar 2025

1 Understanding Decision Tree

Question 1: Are the following statements true?

1. Decision trees are prone to overfitting, especially when they are too deep.
2. Decision trees intrinsically perform feature selection during the training process.
3. A decision tree can only be used for classification problems.
4. Decision trees have linear decision boundaries, similar to logistic regression.
5. Decision trees make predictions based on the majority class or average value in leaf nodes, while KNN makes predictions based on the majority class or average value of nearest neighbors.

2 Building Decision Tree

Question 2: Considering the data in Table 2, a dataset of 8 students about whether they like the famous movie Gladiator. Our goal is to build a decision tree classifier using *Gender, Major* as features and whether or not the students like the movie.

Gender	**Major**	**Like**
Male	Math	Yes
Female	History	No
Male	CS	Yes
Female	Math	No
Female	Math	No
Male	CS	Yes
Male	History	No
Female	Math	Yes

Throughout this question, we will use Entropy as the splitting criterion. You may use entropy with base 2: $Q(\mathcal{S}) = -\sum_{k=1}^{K} p^k \log_2 p^k$. Some useful values: $\log(\frac{1}{2}) = -1$, $\log(\frac{1}{4}) = -2$, $\log(\frac{3}{4}) = -0.41$.

Question 2.1: Initial Entropy. To start with, we have all 8 samples on the root node. Then what is the initial entropy of this dataset?

Question 2.2: Information Gain. Now, to split the samples (i.e. grow the tree), we need to compare the information gain for the two features (i.e. Gender and Major). Which feature is the best to split the data as per your results?
Question 2.3: Purity Measures. Apart from the Gini Index and Entropy we have seen in the class, one might argue to use a misclassification error, i.e. $1 - \max(p_k)$, to build the tree. Can you think of the pros and cons of this measure?

Hint: think in terms of its computational cost and its effectiveness as a split criterion.

3 Random Forests

Question 3: You are designing a decision forest for a **multi-class classification problem** with t decision trees trained using bagging. Each tree uses simple weak learners that split the data using a single feature at a time. At each node, the tree randomly selects a subset of k features (out of d total features) and chooses the one that gives the best split according to an information gain criterion. The full dataset has n samples and d real-valued features. For tree i, a new training set $X^{(i)}$ is created using bagging.

Question 3.1: Bagging. Describe how the training set $X^{(i)}$ is constructed using bagging. Why is using *sampling with replacement* important? How should we handle duplicate data points?

Question 3.2: Training time complexity. Fill in the blanks to derive the overall running time to construct a random forest using both bagging and random feature selection.

Let h be the depth (or height) of the deepest tree in the forest. You must use the tightest possible bounds in terms of n, d, t, k, h, and n'.

Consider choosing the best split at a tree node that contains n' sample points. We can choose the best split for these n' points in $O(\_\_\_\_)$ time. Therefore, the time per sample point in that node is $O(\_\_\_\_)$. Each sample point in $X^{(i)}$ participates in at most $O(\_\_\_\_)$ nodes, so it contributes at most $O(\_\_\_\_)$ to the total time.
Therefore, the total time to train a single tree is $O(\_\_\_\_)$.
With t trees, the total time to train the entire forest is $O(\_\_\_\_)$.

Question 3.3: Feature selection effect. Suppose instead of selecting k random features at each split, you always evaluate all d features. How does this affect the diversity of the trees in the forest? What might be a performance consequence?

Question 3.4: Diversity and generalization. Random forests rely on the idea that averaging many uncorrelated models can improve generalization. Why is it important to both **bag the data** and **randomize the feature selection**? What could go wrong if only one of these techniques is used?

Question 4:
You are given a training dataset with 3 Boolean features X_1, X_2, and X_3, where $X_i \in \{0, 1\}$. The label is defined by the rule $Y = X_1 \vee X_2$, that is, $Y = 1$ if $X_1 = 1$ or $X_2 = 1$, and $Y = 0$ otherwise. The dataset contains all 8 possible combinations of these features:

X_1	X_2	X_3	Y
0	0	0	0
1	0	0	1
0	1	0	1
1	1	0	1
0	0	1	0
1	0	1	1
0	1	1	1
1	1	1	1

Question 4.1: Tree accuracy. What is the training error rate of a depth-0 decision tree on this dataset?

Question 4.2: Splitting by error rate. If your splitting criterion is *training error rate*, which feature (or features) would you choose to split on at the root? Briefly explain your answer.

Question 4.3: Splitting by information gain. If your splitting criterion is *information gain*, which feature (or features) would you choose to split on at the root? Briefly explain your answer.