Foundations of Machine Learning (CS 725)

FALL 2024

Lecture 11:

- Midsem Prep

Instructor: Preethi Jyothi

Question 1

In the limit, as the number of training examples increase, MAP and MLE estimates become the same.

True

The MAP estimate approaches the MLE estimate when N approaches infinity

False

Question 2

During decision tree construction, if you reach a node where the maximum information gain over all splits is zero, then all training examples at that node have the same label.

True

False

Counterexample is XOR.

Question 3

The sum of squares error on training data using the weights obtained after minimizing the ridge regression objective is at least as large as the sum of squares error on training data using the weights obtained after minimizing the unregularized least squares objective.

True

$$||(y - \mathbf{w}_r^T \mathbf{x})||_2^2 \geq ||(y - \mathbf{w}_o^T \mathbf{x})||_2^2$$

Since $\mathbf{w}_o = \arg\min_{\mathbf{w}} ||(y - \mathbf{w}^T \mathbf{x})||_2^2$, by definition the inequality above holds true.

False

Question 4

Let X have a uniform distribution over integers in an interval $[0, \theta]$:

$$p(X = x; \theta) = \begin{cases} \frac{1}{\theta} & \text{if } \ 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

Suppose n samples $x_1, \ldots, x_n$ are drawn i.i.d based on $p(x; \theta)$. What is the MLE estimate of θ?

Solution: The likelihood function is:

$$p(x_1, \ldots, x_n; \theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n & \text{if all } \ 0 \leq x_i \leq \theta \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

This likelihood is maximized by making θ as small as possible, while restricting all of the data to come from the interval $[0, \theta]$. MLE for $\theta = \max(x_1, \ldots, x_n)$.

Question 5

Recall that in the original perceptron algorithm for binary classification, given a training instance ($\mathbf{x}$,y), where $\mathbf{x}$ is a d-dimensional real vector and $y \in \{+1, 1\}$ is the label, the weight vector $\mathbf{w}$ is updated as follows:

$$\text{if } y\mathbf{w} \cdot \mathbf{x} \leq 0 \text{ then } \mathbf{w} \leftarrow \mathbf{w} + y\mathbf{x}$$

Now, consider a variant with two weight vectors $\mathbf{w}^+, \mathbf{w}^-$. The update rule, given a training instance ($\mathbf{x}$,y) is as follows:

$$\text{if } y\mathbf{w}^+ \cdot \mathbf{x} \leq y\mathbf{w}^- \cdot \mathbf{x} \text{ then}$$
$$\mathbf{w}^+ \leftarrow \mathbf{w}^+ + y\mathbf{x} \text{ and } \mathbf{w}^- \leftarrow \mathbf{w}^- - y\mathbf{x}$$

Suppose we initialize $\mathbf{w}^+ = \mathbf{w}^- = 0$ and apply the update rule on n training instances. In parallel, suppose you run the original perceptron algorithm on the same sequence of inputs, with initialization $\mathbf{w} = 0$. At the end of this iteration, given $\mathbf{w}$, compute the values of $\mathbf{w}^+$ and $\mathbf{w}^-$. Justify your answer.

Solution: The following invariant is maintained:

$$\mathbf{w}^+ + \mathbf{w}^- = 0$$

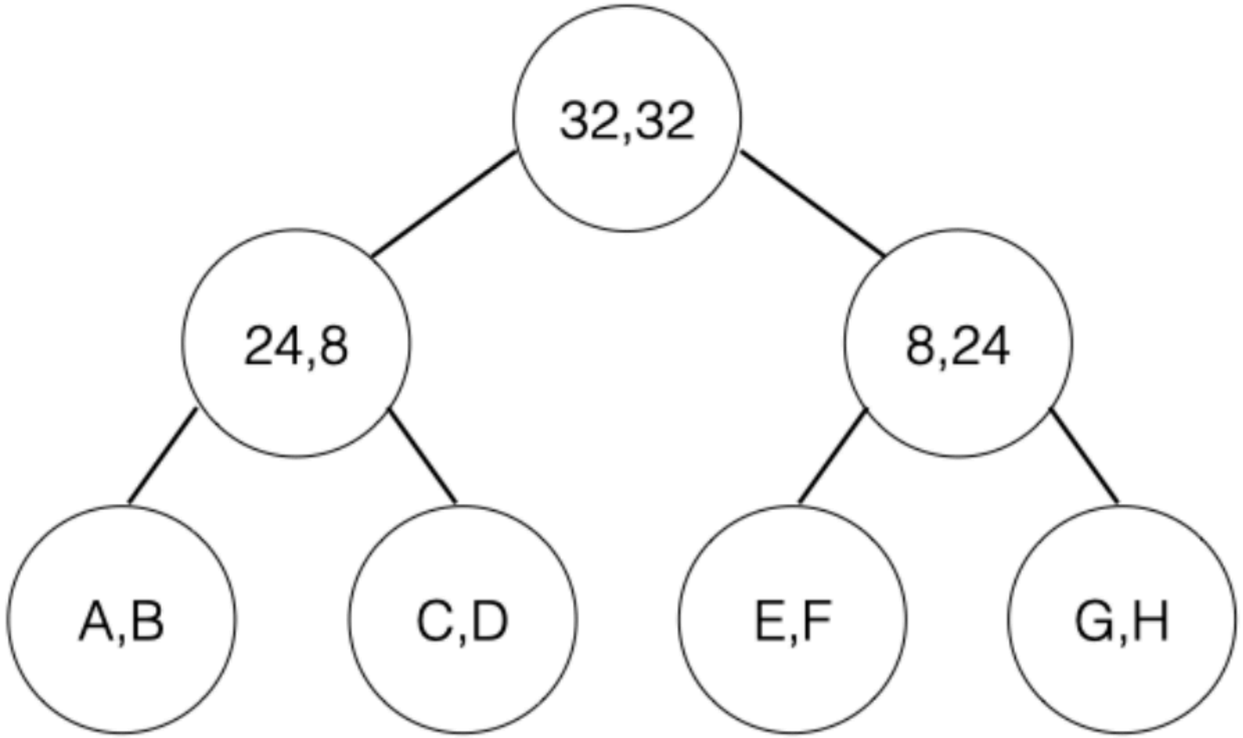
$$\text{if } y\mathbf{w}^+ \cdot \mathbf{x} \leq y\mathbf{w}^- \cdot \mathbf{x}$$

becomes

$$\text{if } 2y\mathbf{w}^+ \cdot \mathbf{x} \leq 0$$

Thus, $\mathbf{w}^+ = \mathbf{w}$ and $\mathbf{w}^- = -\mathbf{w}$

Question 6

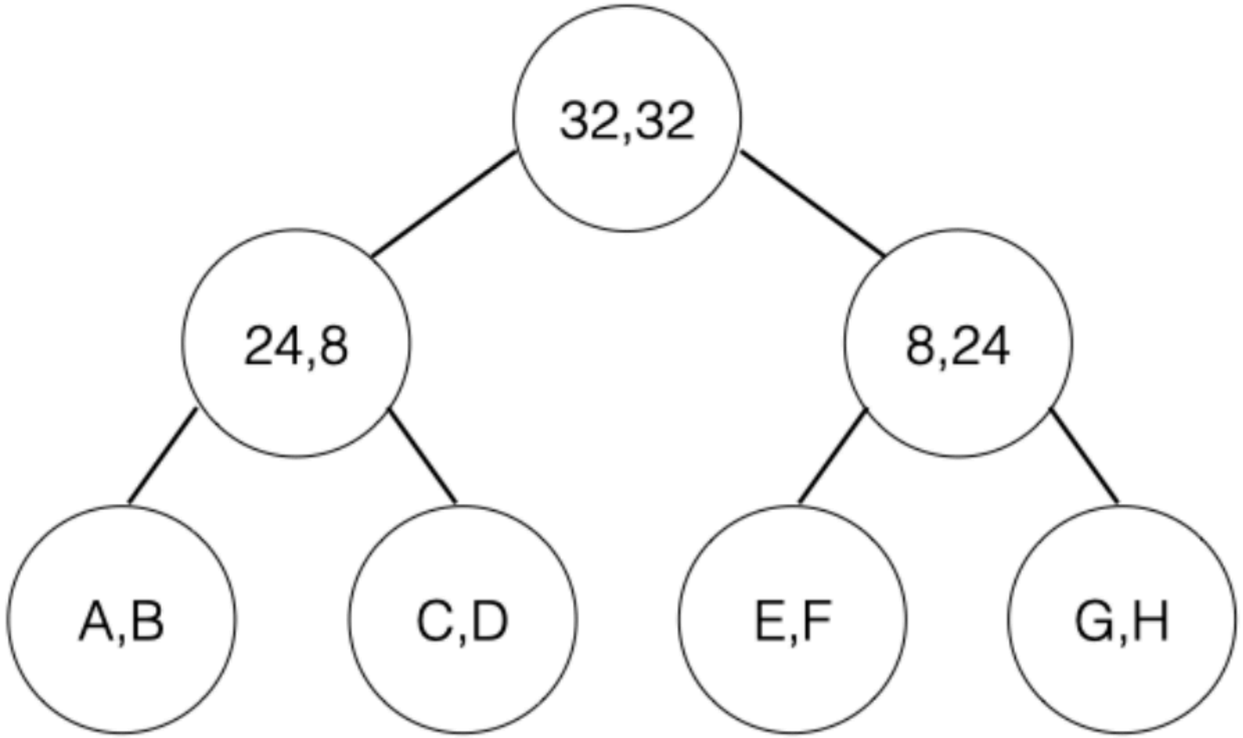


Consider the following decision tree for a binary classification task. The numbers within each node represent the number of training instances in each label. For example, 24,8 means there are 24 examples in class 0 and 8 examples in class 1.

(A) For what A = _____ and B = _____ will we get the smallest possible value of information gain at the node marked 24,8?

12,4 OR 0,0 OR 24,8 OR 3,1 OR 21,7 OR any other A:B that maintains the 3:1 ratio

Question 6

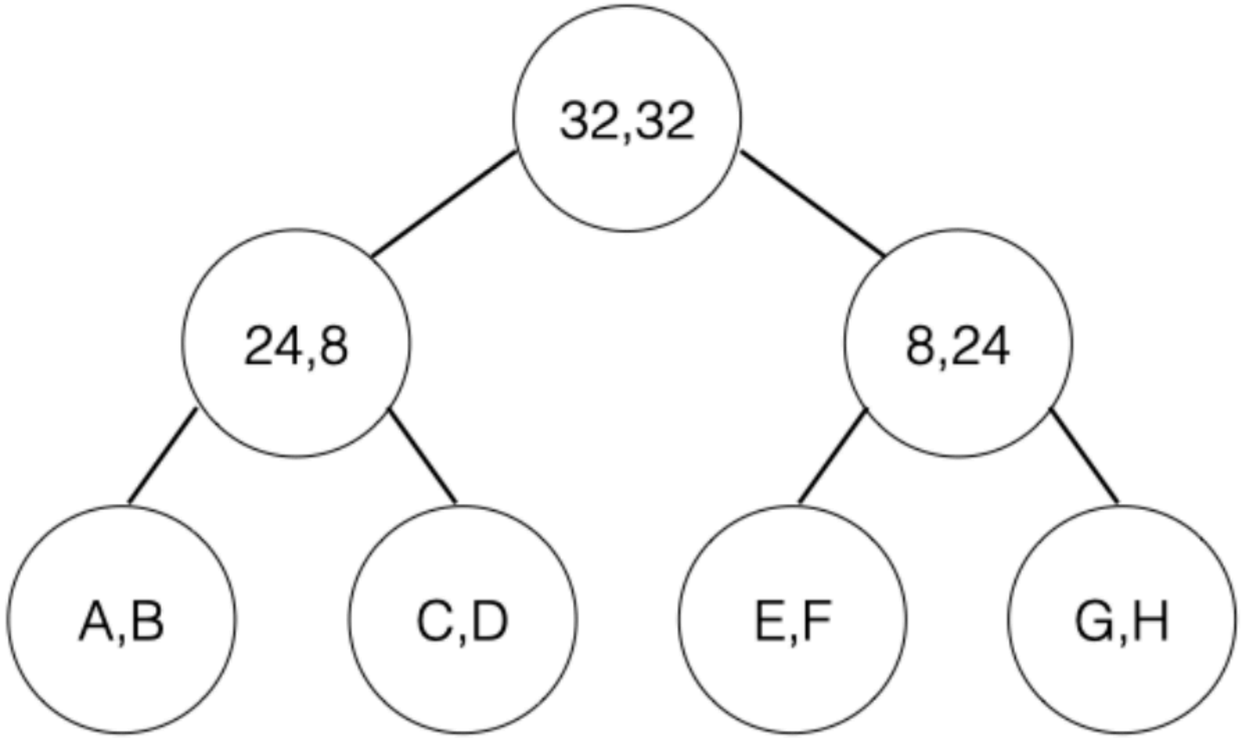


Consider the following decision tree for a binary classification task. The numbers within each node represent the number of training instances in each label. For example, 24,8 means there are 24 examples in class 0 and 8 examples in class 1.

(B) For what G = _____ and H = _____ will we get the largest possible value of information gain at the node marked 8,24?

0,24 OR 8,0

Question 6



Consider the following decision tree for a binary classification task. The numbers within each node represent the number of training instances in each label. For example, 24,8 means there are 24 examples in class 0 and 8 examples in class 1.

(C) What is the information gain at the root node for the given split? For all log values, consider log to the base 2. Assume $\log_2(3) = 1.6$.

$$H(Y) - H(Y|X) = 1 - \left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} \right) = 0.2$$