Machine Learning Homework 4

Hunter Carroll

February 2024

Note. For this homework, I decided to use log base 2 for my calculations as Yaning said this was an acceptable approach thanks.

Problem 1

Problem Statement

There is a sample of 30 students, of whom 15 play cricket in leisure time. Suppose we want to build a decision tree to predict who will play cricket in leisure time, considering three variables: gender (boy/girl), class (IX/X) and height (5 to 6 ft). The following figure shows two cases of the tree construction: the first split is on Gender and the first split is on Class:

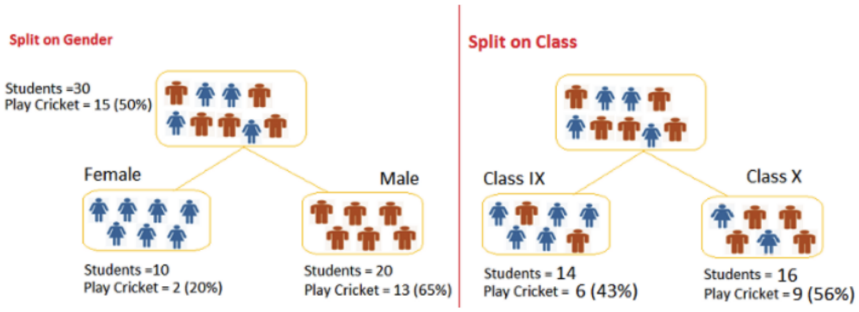

Figure 1: Split on Class and Gender

(a) For "split on gender", calculate the cross-entropy on the parent node (depth=0) and the entropy on EACH leaf node (depth = 1)

Parent Node: The probability that an individual plays cricket is $p_{\text{Play Cricket}} = \frac{15}{30}$ and the probability that an individual does not play cricket is $p_{\text{Doesn't Play}} = \frac{15}{30}$.
Leaf Node Female: The probability that one plays cricket and is female is $p_{\text{Play Female}} = \frac{2}{10}$ and the probability that one does not play cricket and is female is $p_{\text{Doesn't Play Female}} = \frac{8}{10}$.
Leaf Node Male: The probability that one plays cricket and is male is $p_{\text{Play Male}} = \frac{13}{20}$ and the probability that one does not play and is male is $p_{\text{Doesn't Play}} = \frac{7}{20}$
Then we will define cross-entropy,

$$H(X) = -\sum_{x \in X} p(x) \log p(x) = -(p_1 \log(p_1) + p_2 \log(p_2) + ... + p_n \log p_n)$$

This is equivalent to,

$$Q_t(T) = \sum_{k=1}^{K} p_{tk} ln(p_{tk})$$

There is sometimes a negative and sometimes not a negative sign when defining the cross-entropy. We can think of this as the likelihood and negative likelihood either formula will not change the answer we get for the information gain.
Therefore, for the parent node we have,

$$H(\text{Parent}) = -(\frac{15}{30} \log_2 \frac{15}{30} + \frac{15}{30} \log_2 \frac{15}{30}) = 1$$

Max uncertainty in the parent node or referred to as an impure node.
Now for the female leaf node,

$$H(\text{Female}) = -(\frac{2}{10} \log_2 \frac{2}{10} + \frac{8}{10} \log_2 \frac{8}{10}) \approx 0.72$$

Now for the male leaf node,

$$H(\text{Male}) = -(\frac{13}{20} \log_2 \frac{13}{20} + \frac{7}{20} \log_2 \frac{7}{20}) \approx 0.93$$

(b) Define the information gain (IG) as IG = |cross-entropy of parent node - weighted sum of cross-entropy of the children | where in the weighted sum, a weight is the ratio of the number of instances in a child to the total number of instances in the parent. Find the information gain of the splitting.

First, we need to find the weighted sum of cross-entropy of the children nodes,

$$\text{Weighted Sum Children Entropy} = \frac{10}{30} \cdot 0.72 + \frac{20}{30} \cdot 0.93 \approx 0.86$$

Therefore, the information gain of splitting is,

$$IG = |1 - 0.86| = 0.14$$

(c) For "split on class", calculate the cross-entropy on EACH leaf node (depth = 1). Find the information gain.

Leaf Node Class IX: The probability that an individual in Class IX plays cricket is $p_{\text{Class IX Cricket}} = \frac{6}{14}$ and the probability that an individual in Class IX doesn't play cricket is $p_{\text{Class IX No Cricket}} = \frac{8}{14}$.
Leaf Node Class X: The probability that an individual in Class X plats cricket is $p_{\text{Class X Cricket}} = \frac{9}{16}$ and the probability that an individual in Class X doesn't play cricket is $p_{\text{Class X No Cricket}} = \frac{7}{16}$.
Therefore, for leaf node Class IX we have,

$$H(\text{Class IX}) = -(\frac{6}{14}\log_2\frac{6}{14} + \frac{8}{14}\log_2\frac{8}{14}) \approx 0.99$$

Therefore, for leaf node Class X we have,

$$H(\text{Class X}) = -(\frac{9}{16}\log_2\frac{9}{16} + \frac{7}{16}\log_2\frac{7}{16}) \approx 0.99$$

Now the weighted sum of cross-entropy for the children nodes is the following,

$$\text{Weighted Sum Children Entropy} = \frac{14}{30}\cdot 0.99 + \frac{16}{30}\cdot 0.99 \approx 0.99$$

Therefore, the information gain is as follows,

$$IG = |1 - 0.99| = 0.01$$

(d) Which is a better split, split on gender or on class?

The better split is the split on gender. The reason is that the information gain from splitting on gender is greater than the information gain for splitting by class. Therefore, the decision tree will split on gender.
