Homework 4 (Written)

Problem 1

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:

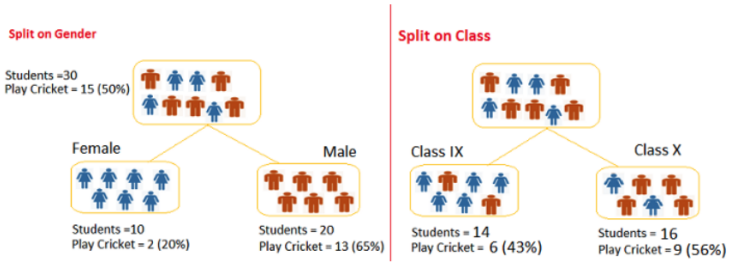

Figure 1: Split on Class and Gender

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

(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.

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

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

Problem 2 (MATH 5027 ONLY)

Show that the parameters α_m in the AdaBoost algorithm is updated by Eqn. (5) Lecture 8 (with $\eta = 1$) by differentiating the cost function below with respect

to α_m and set it 0:

$$
\begin{aligned}
E &= e^{-\alpha_m/2} \sum_{n \in \mathcal{T}_m} w_n^{(m)} + e^{\alpha_m/2} \sum_{n \in \mathcal{M}_m} w_n^{(m)} \\
&= \left(e^{\alpha_m/2} - e^{-\alpha_m/2}\right) \sum_{n=1}^{N} w_n^{(m)} I\left(y_m\left(\mathbf{x}_n\right) \neq t_n\right) + e^{-\alpha_m/2} \sum_{n=1}^{N} w_n^{(m)}
\end{aligned}
$$

where $\mathcal{T}_m$ represents the set of data points that are correctly classified by $y_m(\boldsymbol{x})$, and $\mathcal{M}_m$ represents the remaining misclassified points.