# CSE 151A - Homework 07

Due: Wednesday, May 20, 2020

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. Unless otherwise noted by the problem's instructions, show your work or provide some justification for your answer. Homeworks are due via Gradescope on Wednesday at 11:59 p.m.

**Essential Problem 1.**

In class, we saw the entropy and Gini coefficient as measures of uncertainty in the data. Suppose that a given set of data consists of 20 people at high risk for heart disease and 10 people at low risk. Calculate the entropy and the Gini coefficient. Use base 2 for any logarithms.

**Essential Problem 2.**

Suppose you are training a decision tree classifier to predict whether it will rain in the next hour or not. To do so, you will use two features: the current temperature (in Fahrenheit) and the current pressure (in millibars). Here is some training data that you have collected:

| Temperature | Pressure | Rain? |
|---|---|---|
| 65 | 1001 | Yes |
| 72 | 1003 | Yes |
| 79 | 1030 | No |
| 55 | 1022 | Yes |
| 62 | 1025 | No |
| 71 | 1010 | Yes |
| 73 | 1011 | No |

Train the decision tree to a depth of two. In other words, decide on a root question, splitting the data into two groups, then decide on another question for each group. Your resulting decision tree should have three interior nodes, and four leaf nodes. Use the Gini coefficient to measure uncertainty.

**Correction (5/19/2020)**: it turns out that only two interior nodes (questions) are necessary to achieve zero training error, so your tree should have two interior nodes and three leaf nodes.

**Essential Problem 3.**

Suppose that boosting has been used to train the following four decision stumps:

$$H_1(\vec{x}) = \begin{cases} 1 & x_1 \geq 3 \\ -1 & \text{otherwise} \end{cases}$$

$$H_2(\vec{x}) = \begin{cases} -1 & x_2 \geq 1 \\ 1 & \text{otherwise} \end{cases}$$

$$H_3(\vec{x}) = \begin{cases} -1 & x_2 \geq -2 \\ 1 & \text{otherwise} \end{cases}$$

$$H_4(\vec{x}) = \begin{cases} 1 & x_1 \geq 4 \\ -1 & \text{otherwise} \end{cases}$$

Assume that the "performance" of each decision stump was $\alpha_1 = 2$, $\alpha_2 = 5$, $\alpha_3 = 1$, $\alpha_4 = 10$.

Suppose $\vec{x} = (2, 3)$. What does the overall boosting classifier $H(\vec{x})$ predict for this point? Show your work.

**Plus Problem 1.** (6 plus points)

Draw the decision boundary of the boosting classifier described in Essential Problem 03. Mark the regions where the classifier predicts class +1, and where the classifier predicts class -1.

**Plus Problem 2.** (6 plus points)

**a)** In class, we saw that the entropy of a distribution of a variable that takes two possible values, the first with probability $p$ and the second with probability $1 - p$, is given by

$$p \log \frac{1}{p} + (1 - p) \log \frac{1}{1 - p}$$

In general, if a random variable $X$ takes $K$ possible values, each with probability $p_1, \ldots, p_K$, then the entropy of the distribution of $X$ is defined to be

$$-\sum_{i=1}^{K} p_i \log p_i$$

Show that this general equation is equal to the first when $K = 2$.

**b)** Just like we have defined conditional probability, we also have a notion of conditional entropy. Intuitively, this is a measure of uncertainty in one random variable given another random variable is known. More formally, given a random variable $Z$, the conditional entropy of a random variable $X$, $H(X|Z)$, is defined as:

$$H(X|Z) = \sum_{z} P(Z = z) H(X|Z = z)$$

Let $X \in 1, 2, 3$ and $Z \in a, b$ and let the joint distribution of $X$ and $Z$ be given as follows:

| | | $Z$ | |
|---|---|---|---|
| | | a | b |
| | 1 | 1/4 | 1/8 |
| $X$ | 2 | 1/8 | 3/8 |
| | 3 | 1/16 | 1/16 |

Find H(X|Z). Use base 2 for any logarithms.