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.

Solution: Recall the formulas for Entropy and the Gini index as

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

$$Gini = 2p(1-p)$$

where p is the probability of our positive class.

With $p = \frac{20}{20+10} = \frac{2}{3}$ and using $\log_2$, we calculate

$$\begin{aligned} Entropy &= \frac{2}{3} \log \frac{3}{2} + \frac{1}{3} \log 3 \\ &= 0.918 \end{aligned}$$

$$\begin{aligned} Gini &= 2(\frac{2}{3})(\frac{1}{3}) \\ &= \frac{4}{9} = 0.444 \end{aligned}$$

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.

Solution: Recall, at each node we choose a 'question' or *split* that results in the lowest uncertainty. A split on a set S of labeled points will form left- and right-sets, S_L, S_R, as well as the proportion of points in S that fell into each side, p_L, p_R. The uncertainty of a given split is calculated as the weighted sum of uncertainty of each of the resulting sets. Define the following

$$u(S) = 2p(1-p) = 2(1-p)p$$

$$u(\text{split on } S) = p_L u(S_L) + p_R u(S_R)$$

Our root question will examine full data set. For fun, we can calculate this initial uncertainty as

$$u(S) = 2 \cdot \frac{3}{7} \cdot \frac{4}{7} = 0.490$$

We can calculate uncertainty of possible splits, starting with the lowest to highest x_1, then the lowest to highest x_2, where x_1, x_2 are the Temperature and Pressure features. We will calculate the first three below, and the rest using code.

Notation: The symbol : is often used in set notation, you can read it as meaning "such that".

$$\begin{aligned} u(x_1 < 55\ ?) &\Rightarrow \text{No } [3,4],\ \text{Yes } [0,0] \\ &= u(S : x_1 \not< 55) + 0 \\ &= u(S) \\ &= 0.490 \end{aligned}$$

$$\begin{aligned} u(x_1 < 62\ ?) &\Rightarrow \text{No } [3,3],\ \text{Yes } [0,1] \\ &= \frac{6}{7}u(S : x_1 \not< 62) + \frac{1}{7}u(S : x_1 < 62) \\ &= \frac{6}{7}(2 \cdot \frac{1}{2} \cdot \frac{1}{2}) + \frac{1}{7}(2 \cdot 1 \cdot 0) \\ &= 0.429 \end{aligned}$$

$$\begin{aligned} u(x_1 < 65\ ?) &\Rightarrow \text{No } [2,3],\ \text{Yes } [1,1] \\ &= \frac{5}{7}u(S : x_1 \not< 65) + \frac{2}{7}u(S : x_1 < 65) \\ &= \frac{5}{7}(2 \cdot \frac{2}{5} \cdot \frac{3}{5}) + \frac{2}{7}(2 \cdot \frac{1}{2} \cdot \frac{1}{2}) \\ &= 0.486 \end{aligned}$$

$$u(x_1 < 71\ ?) \Rightarrow \dots$$

Once performed over all possible splits on S, we arrive at the minimum uncertainty with the split $x_2 < 1011$, with

$$\begin{aligned} u(x_2 < 1011\ ?) &\Rightarrow \text{No } [3,1],\ \text{Yes } [0,3] \\ &= 0.214 \end{aligned}$$

Note that uncertainty is zero at our node $\{S : x_2 < 1011\}$, thus we consider that node a leaf. However we must continue reducing uncertainty for our node $\{S : x_2 \not< 1011\}$. To do this, we repeat the above process, but examining only this subset of data.

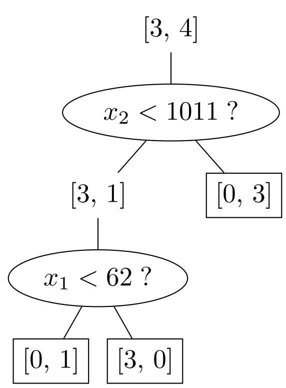
Since the current node was split using x_2, it is wise to examine splits using x_1. The first two are calculated below.

$$\begin{aligned}
u(x_1 < 55\ ?: x_2 \not< 1011) &= u(S : x_2 \not< 1011, x_1 \not< 55) + 0 \\
&= u(S : x_2 \not< 1011) \\
&= 2 \cdot \frac{3}{4} \cdot \frac{1}{4} \\
&= 0.375
\end{aligned}$$

$$\begin{aligned}
u(x_1 < 62\ ?: x_2 \not< 1011) &\Rightarrow \text{No } [0,1],\ \text{Yes } [3,0] \\
&= \frac{1}{4} u(S : x_2 \not< 1011, x_1 \not< 62) + \frac{3}{4} u(S : x_2 \not< 1011, x_1 < 62) \\
&= \frac{1}{4}(2 \cdot 0 \cdot 1) + \frac{3}{4}(2 \cdot 1 \cdot 0) \\
&= 0
\end{aligned}$$

Aha! We've found another split that results in zero uncertainty. We can now consider the nodes $\{S : x_2 \not< 1011, x_1 < 62\}$ and $\{S : x_2 \not< 1011, x_1 \not< 62\}$ to be leaves!

Our final decision tree structure looks like



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.

Solution: Recall the forumla for a boosting classier as

$$\begin{aligned} H(\vec{x}) &= \sum_{t=1}^{T} = \alpha_t H_t(\vec{x}) \\ &= 2(-1) + 5(-1) + 1(-1) + 10(-1) \\ &= -18 \end{aligned}$$

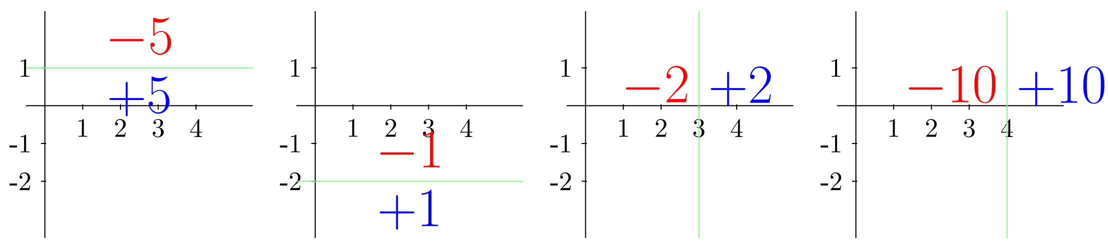
Since $-18 < 0$, we predict the negative class.

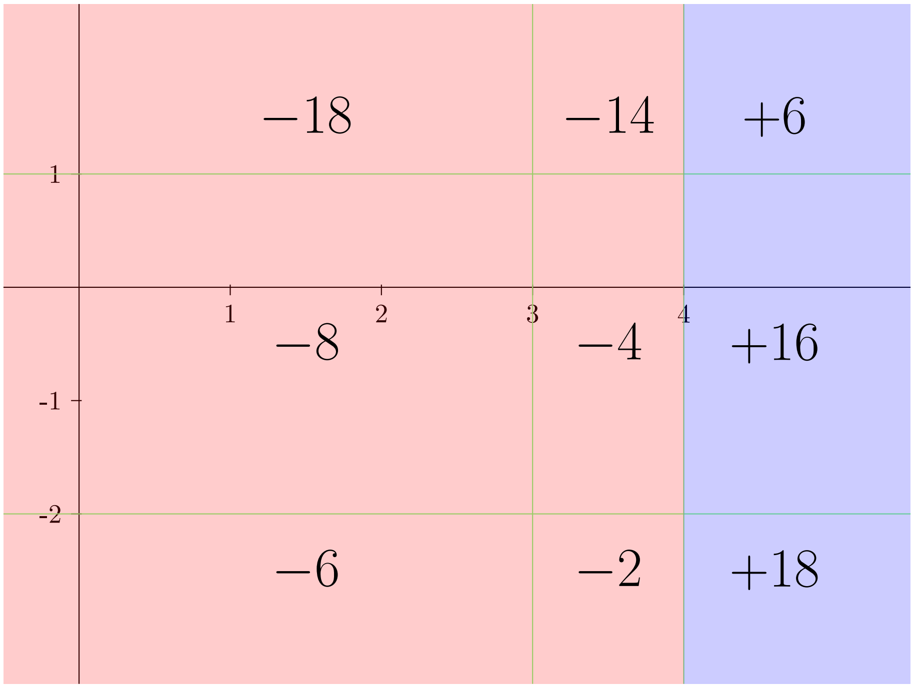
Plus Problem 1. (6 plus points)

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.

Solution: Because it's a bit challenging to properly visualize the addition of multiple weak classifier decision boundaries using weighted transparency, we can go ahead and ditch the transparency in favor of some cold hard numbers.

Simply put, just as we saw in Essential Problem 3, the decision boundary (or regions) for our boosting classifier is just the weighted sum of the decision boundaries (regions) of our weak classifiers.





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 $K = 2$.

Solution: Define $p = p_1$ so that $p_2 = 1 - p$. We have that the entropy of the distribution is

$$-p_1 \log p_1 - p_2 \log p_2 = -p \log p - (1-p) \log(1-p)$$

Recall that $\log 1/x = \log 1 - \log x = 0 - \log x = -\log x$. Therefore:

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

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:

$$
\begin{array}{ccccc}
 & & & Z & \\
 & & \text{a} & & \text{b} \\
 & 1 & 1/4 & & 1/8 \\
X & 2 & 1/8 & & 3/8 \\
 & 3 & 1/16 & & 1/16
\end{array}
$$

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

Solution: To calculate $H(X \mid Z = z)$, we'll need to find all $P(X \mid Z = z)$, since we're calculating the entropy *given that* $Z = z$. Recall

$$P(X \mid Z) = \frac{P(X \cap Z)}{P(Z)}$$

Note that the $P(X \cap Z)$ are the entries of the joint distribution table, and the $P(Z)$ are simply the column sums!

$$P(Z = a) = \frac{1}{4} + \frac{1}{8} + \frac{1}{16} = \frac{7}{16} \qquad P(Z = b) = 1 - P(Z = a) = \frac{9}{16}$$

We can now find all of our conditional $P(X \mid Z)$ by dividing the entire a column by $\frac{7}{16}$ and dividing the entire b column by $\frac{9}{16}$. This results in the following table

$$
\begin{array}{ccccc}
 & & & Z & \\
 & & \text{a} & & \text{b} \\
 & 1 & 4/7 & & 2/9 \\
X \mid Z & 2 & 2/7 & & 6/9 \\
 & 3 & 1/7 & & 1/9
\end{array}
$$

Which enables us to calculate $H(X \mid Z = z)$ using the definition in Part a).

Using $\log_2$, this results in

$$H(X \mid Z = a) = 1.379$$

$$H(X \mid Z = b) = 1.224$$

And we can finally calculate

$$
\begin{aligned}
H(X \mid Z) &= \sum_z P(Z = z) H(X \mid Z = z) \\
&= \frac{7}{16}(1.379) + \frac{9}{16}(1.224) \\
&= 1.292
\end{aligned}
$$