CS6923 Machine Learning, Fall 2019
Prof. Linda Sellie, NYU School of Engineering

Homework 6

Submit on NYU Classes by Sat. Oct. 19 at 8:00 p.m. You may work together with one other person on this homework. If you do that, hand in JUST ONE homework for the two of you, with both of your names on it. You may *discuss* this homework with other students but YOU MAY NOT SHARE WRITTEN ANSWERS OR CODE WITH ANYONE BUT YOUR PARTNER.

1. Suppose you have 30 training examples with two binary features x_1 and x_2. 20 of the training examples are of class A and 10 of the training examples are of class B.

   For those 30 training examples:

   - 10 training examples of class A have $x_1 = 1$
   - 10 training examples of class B have $x_2 = 1$
   - 10 training examples of class A have $x_1 = 0$
   - 0 training examples of class B have $x_1 = 0$

   For those 30 training examples:

   - 10 training examples of class A have $x_2 = 1$
   - 5 training examples of class B have $x_2 = 1$
   - 10 training examples of class A have $x_2 = 0$
   - 10 training examples of class B have $x_2 = 0$

   Which feature would be chosen for the root, and what would the gain be if you used the

   - *entropy criterion*
   - *Gini criterion*
   - *misclassification criterion*

   Show your calculations.

2. As discussed in class, the entropy of a dataset S is defined as follows:

$$H(S) = -\sum_{l \in L} \frac{N_l}{N} \log_2 \frac{N_l}{N}$$

   where L is the set of class labels for the examples in S (e.g., + or -), N is the number of examples in S, and N_l is the number of examples in S that have label l.

(In lecture, we focused on the definition for the case where there are 2 classes. The above definition is the generalization to k class problems, for $k \geq 2$. Note that the logarithm is base 2, even with more than 2 classes! It is standard when defining the entropy of a random variable to use the same base for the logarithm, no matter the size of the domain in question. We use base 2, which is one standard choice. The other standard choice is to take the natural log.)

Let x_i be a discrete-valued attribute of the examples in dataset S, and let V be the set of possible values of attribute x_i. For $v \in V$, let S_v be the subset of examples in S satisfying $x_i = v$.

Consider a classification problem with categorical attributes. Applied to such a problem, the decision tree algorithm discussed in class builds a decision tree where each node is labeled by an attribute x_i. When building the tree, at each node, the algorithm chooses the decision (i.e., the x_i) that minimizes

$$\sum_{v \in V} \frac{|S_v|}{|S|} H(S_v)$$

This is equivalent to saying that it chooses the x_i that maximizes the following quantity, which is sometimes called the Information Gain of x_i (with respect to S)

$$\text{Information-Gain}(S) = H(S) - \sum_{v \in V} \frac{|S_v|}{|S|} H(S_v)$$

It can be shown that Information-Gain(S) is always greater than or equal to 0. Note: We will consider $0 \log_2 0$ to be equal to 0.[1]

(a) Answer the following question without using a calculator. Which dataset has higher entropy: A dataset with 4 positive examples and 5 negative examples, or a dataset with 3 positive examples and 6 negative examples?

(b) Two of the examples have identical attribute values. You should treat them as 2 separate examples in all of your calculations.

[1] $\log 0$ is undefined, but we will consider it to have the value 0. https://www.cs.princeton.edu/courses/archive/spr07/cos424/papers/mitchell-dectrees.pdf

	x_1	x_2	x_3	x_4	y
$\mathbf{x}^{(1)}$	1	1	0	1	+
$\mathbf{x}^{(2)}$	1	1	1	0	+
$\mathbf{x}^{(3)}$	0	0	1	0	+
$\mathbf{x}^{(4)}$	0	1	0	1	-
$\mathbf{x}^{(5)}$	1	0	1	1	+
$\mathbf{x}^{(6)}$	0	1	1	1	+
$\mathbf{x}^{(7)}$	0	1	0	1	-
$\mathbf{x}^{(8)}$	1	1	0	0	-
$\mathbf{x}^{(9)}$	1	0	0	0	-

(Note: Sometimes datasets will have 2 examples with the same attribute values, but different labels. This is sometimes due to an error in the labeling, but can also just reflect the fact that for a given vector of attribute values x, there isn't only one correct label. Consider a dataset where each example corresponds to a customer, and the label indicates whether the customer bought a certain product. One customer might buy the product, and another customer with the same attribute values might not.

As we described it in class, a decision tree algorithm with post-pruning will first grow the tree until it has 0% training error, and then prune it to be smaller. But if two examples have the same attribute values but with different labels, this isn't possible.)

Run the decision tree algorithm described in class (without any pruning) to produce a tree on the dataset above having the smallest error possible. Form a leaf of the tree whenever all examples reaching the leaf have the same label, OR all examples reaching the leaf have the same values for all the attributes. For any leaf containing examples that all have the same attribute values, predict the majority label, or + if there is a tie. [2]

(c) In information theory, the entropy of a discrete random variable taking values in $\{1, \ldots, n\}$ is $-\sum_{i=1}^{n} P[X = i] \log_2 P[X = i]$. This is the expected number of bits needed to encode a randomly drawn value of X, under the "most efficient"code. It is standard to use $H(X)$ to denote the entropy of X.

Given two discrete random variable, X and Y, both taking values in a finite set, the conditional entropy of Y given X, written $H(Y|X)$ is defined to be

$$H(Y|X) = \sum_x P[X = x] * \left(\sum_y -P[Y = y|X = x] * \log_2 P[Y = y|X = x] \right)$$

[2] Note: There are better ways to handle having examples with the same attribute values but different labels! Usually you would do at least some limited pre-pruning, and form a leaf when a node either has few examples reaching it, or is almost pure (even if you'd then apply post-pruning). We're keeping it simple here for the homework exercise.

Here $\sum_x$ and $\sum_y$ denote, respectively, the sum over all possible values x of X and y of Y. The quantity

$$H(Y) - H(Y|X)$$

is called the *mutual information* between X and Y. Interestingly, this quantity is symmetric with respect to X and Y, and is also equal to

$$H(X) - H(X|Y)$$

We can relate this quantity to our definition of Information Gain. Consider an example drawn uniformly at random from from dataset S. Let Y be the label of the example, and let X be the value of attribute x_i in the example. Using the definition of $H(Y|X)$, we see that the information gain of x_i in dataset S is $H(Y) - H(Y|X)$. This is why the Information Gain of x_i is sometimes called the mutual information between attribute x_i and the label.

Consider the dataset in part (b) again. Let Y be the label of an example drawn uniformly at random from the dataset. let X be the value of x_1 in the example. What are the values of $H(Y)$ and $H(Y|X)$? What is the value of $H(Y) - H(Y|X)$?

(d) What is the entropy of a labeled dataset S, with z possible labels, if each label appears equally often among the examples in the dataset?