

Problem Set 3

Instructor: Kamalika Chaudhuri

Due on: May 1, 2015

Instructions

- This is a 40 point homework.
- Problem 4 is a programming assignment. For this problem, you are free to use any programming language you wish.
- Please email your a copy of your code to cse151homeworks@gmail.com.

Problem 1: 8 points

A group of biologists would like to determine which genes are associated with a certain form of liver cancer. After much research, they have narrowed the possibilities down to two genes, let us call them A and B. After analyzing a lot of data, they have also calculated the following joint probabilities.

	Cancer	No Cancer
Gene A	$\frac{1}{2}$	$\frac{1}{10}$
No Gene A	$\frac{1}{5}$	$\frac{1}{5}$

	Cancer	No Cancer
Gene B	$\frac{2}{5}$	$\frac{3}{20}$
No Gene B	$\frac{3}{10}$	$\frac{3}{20}$

1. Let X denote the 0/1 random variable which is 1 when a patient has cancer and 0 otherwise. Let Y denote the 0/1 random variable which is 1 when gene A is present, 0 otherwise, and let Z denote the 0/1 random variable which is 1 when gene B is present and 0 otherwise. Write down the conditional distributions of $X|Y = y$ for $y = 0, 1$ and $X|Z = z$, for $z = 0, 1$.
2. Calculate the conditional entropies $H(X|Y)$ and $H(X|Z)$.
3. Based on these calculations, which of these genes is more informative about cancer?

Problem 2: 8 points

Since a decision tree is a classifier, it can be thought of as a function that maps a feature vector x in some set \mathcal{X} to a label y in some set \mathcal{Y} . We say two decision trees T and T' are *equal* if for all $x \in \mathcal{X}$, $T(x) = T'(x)$.

The following are some statements about decision trees. For these statements, assume that $\mathcal{X} = \mathbb{R}^d$, that is, the set of all d -dimensional feature vectors. Also assume that $\mathcal{Y} = \{1, 2, \dots, k\}$. Write down if each of these statements are correct or not. If they are correct, provide a brief justification or proof; if they are incorrect, provide a counterexample to illustrate a case when they are incorrect.

1. If the decision trees T and T' do not have exactly the same structure, then they can never be equal.
2. If T and T' are any two decision trees that produce zero error on the same training set, then they are equal.

Problem 3: 8 points

1. A fair coin (that is, a coin with equal probability of coming up heads and tails) is flipped until the first head occurs. Let X denote the number of flips required. What is the entropy $H(X)$ of X ? You may find the following expressions useful:

$$\sum_{j=0}^{\infty} r^j = \frac{1}{1-r}, \quad \sum_{j=0}^{\infty} jr^j = \frac{r}{(1-r)^2}$$

2. Let X be a discrete random variable which takes values x_1, \dots, x_m and let Y be a discrete random variable which takes values x_{m+1}, \dots, x_{m+n} . (That is, the values taken by X and the values taken by Y are disjoint.) Let:

$$\begin{aligned} Z &= X \text{ with probability } \alpha \\ &= Y \text{ with probability } 1 - \alpha \end{aligned}$$

Find $H(Z)$ as a function of $H(X)$, $H(Y)$ and α .

Problem 4: Programming Assignment: 16 points

In this problem, we will look at the task of classifying iris flowers into sub-species based on their features using ID3 Decision Trees. We will use four features of the flowers – the petal width, petal length, sepal width, and sepal length. It turns out that these four features will be enough to give us a fairly accurate classification.

Download the files `hw3train.txt` and `hw3test.txt` from the class website. These are your training and test sets respectively. The files are in ASCII text format, and each line of the file contains a feature vector followed by its label. Each feature vector has four coordinates; the coordinates are separated by spaces.

1. Build an ID3 Decision Tree classifier based on the data in `hw3train.txt`. **Do not use pruning.** Draw the decision tree that you obtain. For each node of the decision tree, if it is a leaf node, write down the label that will be predicted for this node, as well as how many of the training data points lie in this node. If it is an internal node, write down the splitting rule for the node, as well as how many of the training data points lie in this node.
2. What is the test error of your classifier in part (1) on the data in `hw3test.txt`?