CSE 151: Introduction to Machine Learning Winter 2017

Problem Set 3

Instructor: Kamalika Chaudhuri **Due on:** Feb 17, 2016

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, but **please submit your answer to Problem 4 along with your homework.**

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 $\mathcal{Y} = \{1, 2, \ldots, k\}$. 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:

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

2. Let X be a discrete random variable which takes values $x_1, \ldots, x_m$ and let Y be a discrete random variable which takes values $x_{m+1}, \ldots, x_{m+n}$. (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 whether a client is likely to default on their credit card payment based on their past behaviour and other characteristics. We will use a decision tree for this purpose.

Download the files `hw3train.txt`, `hw3validation.txt` and `hw3test.txt` from the class website. These are your training, validation 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 22 coordinates; they are named Feature 1, Feature 2, ..., Feature 22, respectively. The coordinates are separated by spaces. The last (23rd) coordinate represents the label of an example, that is, whether the card-holder defaults on their credit card bill in October, where 1 means yes, and 0 means no.

1. First, build an ID3 Decision Tree classifier based on the data in `hw3train.txt`. **Do not use pruning.** Draw the first three levels decision tree that you obtain. For each node that you draw, 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. (Hint: If your code is correct, the root node will involve the rule Feature $5 < 0.5$.)

2. What is the training and test error of your classifier in part (1), where test error is measured on the data in `hw3test.txt`?

3. Now, prune the decision tree developed in part (1) using the data in `hw3validation.txt`. While selecting nodes to prune, select them in Breadth-First order, going from left to right (aka, from the Yes branches to the No branches). Write down the validation and test error after 1 and 2 rounds of pruning (that is, after you have pruned 1 and 2 nodes from the tree.)

4. Download the file `hw3features.txt` from the class website. This file provides a description in order of each of the features – that is, it tells you what each coordinate means. Based on the feature descriptions, what do you think is the most salient or prominent feature that predicts credit card default? (Hint: More salient features should occur higher up in the ID3 Decision tree.)