Instructions

• This is a 40 point homework.

• For Problem 4, you are free to use any programming language you wish. Please submit a printout of your code 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.

<table>
<thead>
<tr>
<th></th>
<th>Cancer</th>
<th>No Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene A</td>
<td>$\frac{7}{17}$</td>
<td>$\frac{10}{17}$</td>
</tr>
<tr>
<td>No Gene A</td>
<td>$\frac{5}{17}$</td>
<td>$\frac{12}{17}$</td>
</tr>
<tr>
<td>Gene B</td>
<td>$\frac{2}{20}$</td>
<td>$\frac{18}{20}$</td>
</tr>
<tr>
<td>No Gene B</td>
<td>$\frac{3}{20}$</td>
<td>$\frac{17}{20}$</td>
</tr>
</tbody>
</table>

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 $X|Y$ for $y = 0, 1$ and $X|Z$ 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 do you think are more informative about the cancer?

Solutions

1. First, we can compute the marginal distributions of $Y$ and $Z$ as follows,

<table>
<thead>
<tr>
<th>$y$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(Y = y)$</td>
<td>$\frac{7}{17}$</td>
<td>$\frac{10}{17}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$z$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(Z = z)$</td>
<td>$\frac{2}{20}$</td>
<td>$\frac{18}{20}$</td>
</tr>
</tbody>
</table>

Then, by definition of conditional probability, i.e. $P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$, we can get the conditional distributions of $X|Y$ as follows.

<table>
<thead>
<tr>
<th>$x$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(X = x</td>
<td>Y = 0)$</td>
<td>$\frac{1}{7}$</td>
</tr>
<tr>
<td>$P(X = x</td>
<td>Y = 1)$</td>
<td>$\frac{1}{5}$</td>
</tr>
</tbody>
</table>

Similarly we have the conditional distributions of $X|Z$ as follows,

<table>
<thead>
<tr>
<th>$x$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(X = x</td>
<td>Z = 0)$</td>
<td>$\frac{1}{3}$</td>
</tr>
<tr>
<td>$P(X = x</td>
<td>Z = 1)$</td>
<td>$\frac{1}{7}$</td>
</tr>
</tbody>
</table>
2. By the definition of conditional entropy, \( H(X|Y) = P(Y = 0)H(X|Y = 0) + P(Y = 1)H(X|Y = 1) \).

\[
H(X|Y = 0) = -P(X = 0|Y = 0) \log P(X = 0|Y = 0) - P(X = 1|Y = 0) \log P(X = 1|Y = 0)
= -\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2}
= \log 2
\]

Similarly we have

\[
H(X|Y = 1) = -P(X = 0|Y = 1) \log P(X = 0|Y = 1) - P(X = 1|Y = 1) \log P(X = 1|Y = 1)
= -\frac{1}{6} \log \frac{1}{6} - \frac{5}{6} \log \frac{5}{6}
= \log 6 - \frac{5}{6} \log 5
\]

Thus

\[
H(X|Y) = P(Y = 0)H(X|Y = 0) + P(Y = 1)H(X|Y = 1)
= 2 \log 2 + 3 \left( \log 6 - \frac{5}{6} \log 5 \right)
= 2 \log 2 + 3 \log 6 - \frac{1}{2} \log 5
\]

For \( H(X|Z) \), we can get

\[
H(X|Z = 0) = -P(X = 0|Z = 0) \log P(X = 0|Z = 0) - P(X = 1|Z = 0) \log P(X = 1|Z = 0)
= -\frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log \frac{2}{3}
= \log 3 - \frac{2}{3} \log 2
\]

Similarly we have

\[
H(X|Z = 1) = -P(X = 0|Z = 1) \log P(X = 0|Z = 1) - P(X = 1|Z = 1) \log P(X = 1|Z = 1)
= -\frac{3}{11} \log \frac{3}{11} - \frac{8}{11} \log \frac{8}{11}
= \log 11 - \frac{3}{11} \log 3 - \frac{8}{11} \log 8
\]

Thus

\[
H(X|Z) = P(Z = 0)H(X|Z = 0) + P(Z = 1)H(X|Z = 1)
= \frac{9}{20} \left( \log 3 - \frac{2}{3} \log 2 \right) + \frac{11}{20} \left( \log 11 - \frac{3}{11} \log 3 - \frac{8}{11} \log 8 \right)
= -\frac{3}{2} \log 2 + \frac{3}{10} \log 3 + \frac{11}{20} \log 11
\]

Using natural logarithm, the numerical values are shown as follows.

| \( H(X|Y = 0) \) | 0.693147180560 |
| \( H(X|Y = 1) \) | 0.450561208866 |
| \( H(X|Y) \) | 0.547595597544 |
| \( H(X|Z = 0) \) | 0.63651416829 |
| \( H(X|Z = 1) \) | 0.5859526183 |
| \( H(X|Z) \) | 0.6087053158 |

3. From the table above, \( H(X|Y) < H(X|Z) \). This suggests that there is less uncertainty in \( X \) when given \( Y \) than when given \( Z \). Therefore gene A is more informative about the 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.

Solutions

1. False.  
Counterexample: Consider a classifier for data which uses one feature (called Feature1).

![Figure 1: Two Decision Trees which are equal (see definition in question) but have different structures](image1)

2. False. 
If \( T \) and \( T' \) produce zero error on the same training set \( S \subseteq \mathcal{X} \), then, \( \forall x \in S \), \( T(x) = T'(x) \). However, the training set typically does not include all elements in feature space \( \mathcal{X} \). Thus, there exist such \( x_0 \in \mathcal{X} - S \) that \( T(x_0) \neq T'(x_0) \). For example, consider the following training set:

<table>
<thead>
<tr>
<th>Feature 1</th>
<th>Feature 2</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

For training set above, the two decision trees shown in Figure 2 both produce zero error. However, for the point \( x_1 = (0, 1) \) or the point \( x_2 = (1, 0) \), these two trees would give different predictions. Hence they are not equal.

![Figure 2: Two Decision Trees with Zero Error on S](image2)
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:

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

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:

$$
Z = X \text{ with probability } \alpha \\
   = Y \text{ with probability } 1 - \alpha
$$

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

Solutions

1. Observe that $X$ is a random variable which takes values $k = 1, 2, 3, \ldots$. For a fixed integer $k$, we need $k$ flips to get the first head if the first $k - 1$ tosses come up tails, and the $k$-th toss comes up a head. Therefore,

$$
p_k = \Pr(X = k) = \frac{1}{2^{k-1}} \cdot \frac{1}{2} = \frac{1}{2^k}
$$

Therefore,

$$
H(X) = - \sum_{k=1}^{\infty} p_k \log p_k = - \sum_{k=1}^{\infty} \frac{1}{2^k} \log \frac{1}{2^k} = \sum_{k=1}^{\infty} \log 2 \cdot \frac{k}{2^k}
$$

The last step follows because $\log \frac{1}{2^k} = -k \log 2$. From the expressions given above, the sum is:

$$
\sum_{k=1}^{\infty} \frac{k}{2^k} = \sum_{k=0}^{\infty} \frac{k}{2^k} = \frac{\frac{1}{2}}{(1 - \frac{1}{2})^2} = 2
$$

Thus, $H(X) = 2 \log 2$.

2. Let $p_i = \Pr(X = x_i)$ and let $q_j = \Pr(Y = x_{m+j})$. Then, $H(X) = - \sum_{i=1}^{m} p_i \log p_i$ and $H(Y) = - \sum_{j=1}^{n} q_j \log q_j$. By definition of $Z$, $Z$ takes values $x_i$, $1 \leq i \leq m$ with probability $\alpha p_i$, and values $x_{m+j}$, $1 \leq j \leq n$ with probability $(1 - \alpha)q_j$. Therefore,

$$
H(Z) = - \sum_{i=1}^{m} \alpha p_i \log \alpha p_i - \sum_{j=1}^{n} (1 - \alpha)q_j \log (1 - \alpha)q_j
$$

$$
= - \sum_{i=1}^{m} \alpha p_i \log \alpha - \sum_{i=1}^{m} \alpha p_i \log p_i - \sum_{j=1}^{n} (1 - \alpha)q_j \log (1 - \alpha) - \sum_{j=1}^{n} (1 - \alpha)q_j \log q_j
$$

$$
= \alpha H(X) + (1 - \alpha) H(Y) - \alpha \log \alpha - (1 - \alpha) \log (1 - \alpha)
$$

Here the last step follows from the observation that $\sum_{i=1}^{m} p_i = 1$ and $\sum_{j=1}^{n} q_j = 1$. 
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 folks default on their credit card bill in October, 1 means yes, 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.)

Solutions

1. Figure 3 shows the first three levels of the decision tree. There are multiple possible decision trees even if we ignore the effect of spacing in the training examples. Nevertheless, all of them share the same top 3 layers. In this solution, if multiple splittings achieves maximum information gain, we pick one uniformly at random and recurse.

When splitting each node, we employ Information Gain as the criterion for selecting a (feature, threshold) pair. The set of thresholds for a particular feature to be considered at a node are chosen according to the following approach. First, we sort the training samples $S$ on the feature $f$ being considered. There are only a finite number of these values, so let us denote them in sorted order by $v_1 < v_2 < \cdots < v_n$. Any threshold value lying between $v_i$ and $v_{i+1}$ will have the same effect of dividing the data points associated with the node into those whose value of the feature $f$ lies in $\{v_k : k \leq i\}$ and those whose value is in $\{v_k : k > i\}$. There are thus only $n - 1$ possible splits on $f$. We choose the midpoint of each interval, i.e. $\frac{v_i + v_{i+1}}{2}$, as the representative threshold.
For the next two problems, depending on the randomness in tie breaking in the decision tree training procedure, we might end up with different trees. The results on the table below is from one of the trees generated by ID3. In general the results might slightly deviate from the result below.

For demonstration purposes, for each node, we also present (1) "#validation error examples", the number of validation examples that passes through this node such that the subtree rooted here classifies them incorrectly, and (2) "#validation error examples if this is leaf", the number of validation examples that passes through this node such that their labels does not agree with the label of this node. If the latter is less than the former, the subtree rooted at this node can be pruned to a leaf.

2. Table 1 shows the errors:

<table>
<thead>
<tr>
<th>#Prunings</th>
<th>Training Error</th>
<th>Validation Error</th>
<th>Test Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>0</td>
<td>0.167</td>
<td>0.175</td>
</tr>
</tbody>
</table>

Table 1: Training, validation and test errors of decision tree without pruning.

The decision trees generated is shown (again) in Figure 4, with the feature names incorporated.

3. The decision tree after pruning 1 node:
Figure 5: The top 3 layers of the trained decision tree, with 1 node pruned.

The decision tree after pruning 2 nodes:

Figure 6: The top 3 layers of the trained decision tree, with 2 nodes pruned.

Table 2 shows the errors.

<table>
<thead>
<tr>
<th>#Prunings</th>
<th>Training Error</th>
<th>Validation Error</th>
<th>Test Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0845</td>
<td>0.119</td>
<td>0.118</td>
</tr>
<tr>
<td>2</td>
<td>0.105</td>
<td>0.107</td>
<td>0.103</td>
</tr>
</tbody>
</table>

Table 2: Training, validation and test errors of decision tree with pruning.

4. As shown in Figure 4, the most salient feature is PAYMENT_DELAY_SEPTEMBER. Features BILL_AMT1, LIMIT_BAL, PAY_AMT4, PAY_AMT5 are also relevant.