CSE 151: Introduction to Machine Learning	Spring 2018

Problem Set 3

Instructor: Kamalika Chaudhuri **Due on:** never

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 do you think are more informative about the cancer?

Solutions

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

y	0	1
$P(Y = y)$	$\frac{2}{5}$	$\frac{3}{5}$

z	0	1
$P(Z = z)$	$\frac{9}{20}$	$\frac{11}{20}$

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

x	0	1
$P(X = x\|Y = 0)$	$\frac{1}{2}$	$\frac{1}{2}$
$P(X = x\|Y = 1)$	$\frac{1}{6}$	$\frac{5}{6}$

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

x	0	1
$P(X = x\|Z = 0)$	$\frac{1}{3}$	$\frac{2}{3}$
$P(X = x\|Z = 1)$	$\frac{3}{11}$	$\frac{8}{11}$

2. By the definition of conditional entropy, $H(X|Y) = P(Y = 0)H(X|Y = 0) + P(Y = 1)H(X|Y = 1)$.

$$\begin{aligned} H(X|Y = 0) &= -P(X = 0|Y = 0)\log P(X = 0|Y = 0) - P(X = 1|Y = 0)\log P(X = 1|Y = 0) \\ &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} \\ &= \log 2 \end{aligned}$$

Similarly we have

$$\begin{aligned} H(X|Y=1) &= -P(X=0|Y=1)\log P(X=0|Y=1) - P(X=1|Y=1)\log P(X=1|Y=1) \\ &= -\frac{1}{6}\log\frac{1}{6} - \frac{5}{6}\log\frac{5}{6} \\ &= \log 6 - \frac{5}{6}\log 5 \end{aligned}$$

Thus

$$\begin{aligned} H(X|Y) &= P(Y=0)H(X|Y=0) + P(Y=1)H(X|Y=1) \\ &= \frac{2}{5}\log 2 + \frac{3}{5}\left(\log 6 - \frac{5}{6}\log 5\right) \\ &= \frac{2}{5}\log 2 + \frac{3}{5}\log 6 - \frac{1}{2}\log 5 \end{aligned}$$

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

$$\begin{aligned} H(X|Z=0) &= -P(X=0|Z=0)\log P(X=0|Z=0) - P(X=1|Z=0)\log P(X=1|Z=0) \\ &= -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} \\ &= \log 3 - \frac{2}{3}\log 2 \end{aligned}$$

Similarly we have

$$\begin{aligned} H(X|Z=1) &= -P(X=0|Z=1)\log P(X=0|Z=1) - P(X=1|Z=1)\log P(X=1|Z=1) \\ &= -\frac{3}{11}\log\frac{3}{11} - \frac{8}{11}\log\frac{8}{11} \\ &= \log 11 - \frac{3}{11}\log 3 - \frac{8}{11}\log 8 \end{aligned}$$

Thus

$$\begin{aligned} H(X|Z) &= P(Z=0)H(X|Z=0) + P(Z=1)H(X|Z=1) \\ &= \frac{9}{20}\left(\log 3 - \frac{2}{3}\log 2\right) + \frac{11}{20}\left(\log 11 - \frac{3}{11}\log 3 - \frac{8}{11}\log 8\right) \\ &= -\frac{3}{2}\log 2 + \frac{3}{10}\log 3 + \frac{11}{20}\log 11 \end{aligned}$$

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

Solutions

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

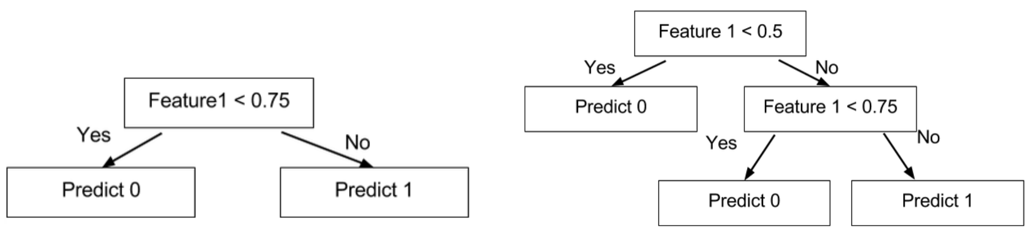

Figure 1: Two Decision Trees which are equal (see definition in question) but have different structures

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:

Feature 1	Feature 2	Label
0	0	0
1	1	1

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.

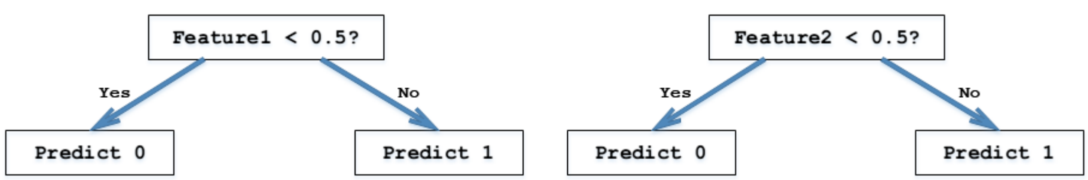

Figure 2: Two Decision Trees with Zero Error on S

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.

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 αp_i, and values x_{m+j}, $1 \leq j \leq n$ with probability $(1-\alpha)q_j$. Therefore,

$$\begin{aligned} H(Z) &= -\sum_{i=1}^{m} \alpha p_i \log \alpha p_i - \sum_{j=1}^{n} (1-\alpha) q_j \log (1-\alpha) q_j \\ &= -\sum_{i=1}^{m} \alpha p_i \log \alpha - \sum_{i=1}^{m} \alpha p_i \log p_i - \sum_{j=1}^{n} (1-\alpha) q_j \log (1-\alpha) - \sum_{j=1}^{n} (1-\alpha) q_j \log q_j \\ &= \alpha H(X) + (1-\alpha) H(Y) - \alpha \log \alpha - (1-\alpha) \log (1-\alpha) \end{aligned}$$

Here the last step follows from the observation that $\sum_{i=1}^{m} p_i = 1$ and $\sum_{j=1}^{n} q_j = 1$.