Machine Learning Fall 2019 Homework 1 Written Solutions

1. (5 points) Show what the recursive decision tree learning algorithm would choose for the first split of the following dataset:

ID	X_1	X_2	X_3	X_4	Y
1	0	0	0	0	0
2	0	0	0	1	0
3	0	0	1	0	0
4	0	0	1	1	0
5	0	1	0	0	0
6	0	1	0	1	1
7	0	1	1	0	1
8	0	1	1	1	1
9	1	0	0	0	1
10	1	0	0	1	1

Assume that the criterion for deciding the best split is entropy reduction (i.e., information gain). If there are any ties, choose the first feature to split on tied for the best score. Show your calculations in your response.

(Hint: this dataset is one of the test cases in the programming assignment, so you should be able to use your answer here to debug your code.)

Solution: First, we calculate the base entropy, which is just the entropy of $P(Y)$:

$$H(Y) = -P(Y=0)\log P(Y=0) - P(Y=1)\log P(Y=1) = 0.693.$$

I'm using the natural log. If you use a different base, that's okay. Your numbers will be different, but you should get the same variable as the best gain.

To calculate the information gain, we first get $P(X_1 = 0) = 0.8$ and $P(X_1 = 1) = 0.2$. Then we get that $P(Y = 0|X_1 = 0) = 5/8$ (and $P(Y = 1|X_1 = 0) = 3/8$) by counting the number of positive examples among the first eight. For the case where $X_1 = 1$, we have that $P(Y = 0|X_1 = 1) = 0.0$ (and $P(Y = 1|X_1 = 0) = 1.0$).

Then we calculate

$$\begin{aligned}G(X_1) = 0.693+&\\ &P(X_1=0)(P(Y=0|X_1=0)\log P(Y=0|X_1=0) + P(Y=1|X_1=0)\log P(Y=1|X_1=0))\\ &+P(X_1=1)(P(Y=0|X_1=1)\log P(Y=0|X_1=1) + P(Y=1|X_1=1)\log P(Y=1|X_1=1))\\ &= 0.693 + 0.8\cdot\left(\frac{5}{8}\log\frac{5}{8} + \frac{3}{8}\log\frac{3}{8}\right) + 0.2\cdot(0\log 0 + 1\log 1)\\ &= 0.693 - 0.529 + 0 = 0.164\end{aligned}$$

Similar calculations for X_2, X_3 and X_4 lead to

$$G(X_2) = 0.693 + 0.6(0.333\log 0.333 + 0.667\log 0.667) + 0.4(0.25\log 0.25 + 0.75\log 0.75) = 0.693 - 0.607 = 0.086.$$

$$G(X_3) = 0.693 + 0.4(0.5\log 0.5 + 0.5\log 0.5) + 0.6(0.5\log 0.5 + 0.5\log 0.5) = 0.693 - 0.693 = 0.$$

$$G(X_4) = 0.693 + 0.5(0.4\log 0.4 + 0.6\log 0.6) + 0.5(0.6\log 0.6 + 0.4\log 0.4) = 0.693 - 0.673 = 0.02.$$

Variable X_1 therefore has the highest information gain of 0.164.

A couple noteworthy cases occur in this exercise. First, when taking the entropy of a distribution with no uncertainty, $p = 1$ or $p = 0$, we get zero entropy because $\log 1 = 0$. This happened in the calculation for X_1. For X_3, the proportions of the labels did not change when we condition on X_3, so we ended up with the same value for the conditional entropy as the base entropy. The information gain is therefore zero, aligning with intuition about what information gain means. No information was gained by considering X_3.

2. A Bernoulli distribution has the following likelihood function for a data set $\mathcal{D}$:

$$p(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0}, \tag{1}$$

where N_1 is the number of instances in data set $\mathcal{D}$ that have value 1 and N_0 is the number in $\mathcal{D}$ that have value 0. The maximum likelihood estimate is

$$\hat{\theta} = \frac{N_1}{N_1 + N_0}. \tag{2}$$

(a) (5 points) Derive the maximum likelihood estimate above by solving for the maximum of the likelihood. I.e., show the mathematics that get from Equation (1) to Equation (2).

Solution: First, we can consider the log likelihood for convenience:

$$\log p(\mathcal{D}|\theta) = N_1 \log\theta + N_0 \log(1-\theta).$$

Since this is a concave function that is differentiable in the domain of interest, we can take its derivative and solve for the zero-derivative point as its maximum.

$$\begin{aligned}
\frac{d\ \log p(\mathcal{D}|\theta)}{d\,\theta} &= \frac{N_1}{\theta} - \frac{N_0}{1-\theta}\\
&= \frac{N_1(1-\theta)}{\theta(1-\theta)} - \frac{N_0\theta}{\theta(1-\theta)}\\
&= \frac{N_1(1-\theta) - N_0\theta}{\theta(1-\theta)}\\
&= \frac{N_1 - N_1\theta - N_0\theta}{\theta(1-\theta)}\\
0 &= \frac{N_1 - (N_1+N_0)\hat{\theta}}{\hat{\theta}(1-\hat{\theta})}\\
0 &= N_1 - (N_1+N_0)\hat{\theta}\\
(N_1+N_0)\hat{\theta} &= N_1\\
\hat{\theta} &= \frac{N_1}{N_1+N_0}
\end{aligned}$$

(b) (5 points) Suppose we now want to maximize a posterior likelihood

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}, \tag{3}$$

where we use the Bernoulli likelihood and a (slight variant[1] of a) symmetric Beta prior over the Bernoulli parameter

$$p(\theta) \propto \theta^{\alpha}(1-\theta)^{\alpha}. \tag{4}$$

Derive the maximum posterior mean estimate.

Solution: The log likelihood is (within an additive constant of)

$$\begin{aligned}
\log p(\theta|\mathcal{D}) &= N_1 \log\theta + N_0\log(1-\theta) + \alpha\log\theta + \alpha\log(1-\theta)\\
&= (N_1+\alpha)\log\theta + (N_0+\alpha)\log(1-\theta).
\end{aligned}$$

[1] For convenience, we are using the exponent of α instead of the standard $\alpha - 1$.

Again, taking the derivative and solving for its zero, we get

$$\frac{d\ \log p(\theta|\mathcal{D})}{d\,\theta} = \frac{N_1 + \alpha}{\theta} - \frac{N_0 + \alpha}{1 - \theta}.$$

Using a change-of-variables, we can see the same algebraic manipulations from part (a) apply to a new counts $N_1' = N_1 + \alpha$ and $N_0' = N_0 + \alpha$, giving

$$\hat{\theta} = \frac{N_1'}{N_1' + N_0'} = \frac{N_1 + \alpha}{N_1 + \alpha + N_0 + \alpha} = \frac{N_1 + \alpha}{N_1 + N_0 + 2\alpha}$$