# 10-701 Machine Learning, Fall 2012: Homework 1 Solutions

## 1 Decision Trees, [25pt, Martin]

1. **[10 points]**

   (a) For the first split, there are two possibilities to consider – either we split on *Size*="Big" v.s. *Size*="Small", or *Orbit*="Near" v.s. *Orbit*="Far". Let's calculate the corresponding conditional entropies. For brevity, we define

   $$g(p) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}$$

   for $0 \leq p \leq 1$, where we use the convention that $0 \log_2 \frac{1}{0} = 0$. Then we have

   $$\begin{aligned} H(\text{Habitable}|\text{Size}) =& P(\text{Size} = \text{Big}) H(\text{Habitable}|\text{Size} = \text{Big}) \\ &+ P(\text{Size} = \text{Small}) H(\text{Habitable}|\text{Size} = \text{Small}) \\ =& P(\text{Size} = \text{Big}) g[P(\text{Habitable} = \text{Yes}|\text{Size} = \text{Big})] \\ &+ P(\text{Size} = \text{Small}) g[P(\text{Habitable} = \text{Yes}|\text{Size} = \text{Small})] \\ =& \frac{350}{800} g\left[\frac{190}{350}\right] + \frac{450}{800} g\left[\frac{184}{450}\right] \\ =& 0.9841298 \end{aligned}$$

   and

   $$\begin{aligned} H(\text{Habitable}|\text{Orbit}) =& P(\text{Orbit} = \text{Near}) H(\text{Habitable}|\text{Orbit} = \text{Near}) \\ &+ P(\text{Orbit} = \text{Far}) H(\text{Habitable}|\text{Orbit} = \text{Far}) \\ =& P(\text{Orbit} = \text{Near}) g[P(\text{Habitable} = \text{Yes}|\text{Orbit} = \text{Near})] \\ &+ P(\text{Orbit} = \text{Far}) g[P(\text{Habitable} = \text{Yes}|\text{Orbit} = \text{Far})] \\ =& \frac{300}{800} g\left[\frac{159}{300}\right] + \frac{500}{800} g\left[\frac{215}{500}\right] \\ =& 0.99016 \end{aligned}$$

   So we see that the first split must be on *Size*. We do not need to make any further calculations for the two remaining subtrees, since clearly both must be split on *Orbit*. The final tree is given in Figure 1.

   (b) Upon careful examination, the reader may notice that *Temperature* only takes three unique values in the given data set. Recall that we are only considering splits on this feature that can be expressed in the form of a threshold (Temperature $\leq T_0$ v.s.

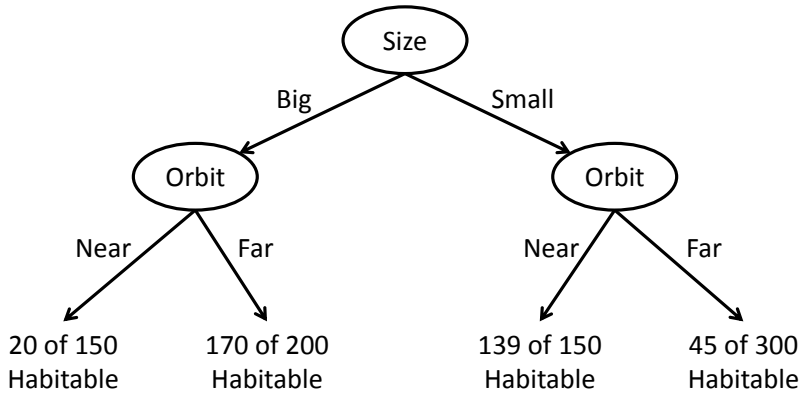


Figure 1: The decision tree for 1.1(a).

Temperature $> T_0$). Hence, along with the binary splits on *Size* and *Orbit*, we also must consider two possible splits on *Temperature*, which can be written as

$$\text{Temperature} \leq 205 \text{ v.s. Temperature} > 205$$

and

$$\text{Temperature} \leq 260 \text{ v.s. Temperature} > 260.$$

We will refer to the resulting binary features as $\text{Temperature}_{205}$ and $\text{Temperature}_{260}$, respectively. (It is important to note that these are not the only splits one might consider based on this data set. For example, we could have defined the first split to be Temperature $\leq$ 240 v.s. Temperature $>$ 240. While this would not affect the overall shape of the tree, it might affect how we later classify data points with values for *Temperature* that were not observed in the training set.)

Thus for the first split in the tree we have

$$\begin{aligned} H(\text{Habitable}|\text{Size}) =& \frac{4}{9} g\left[\frac{2}{4}\right] + \frac{5}{9} g\left[\frac{2}{5}\right] \\ =& 0.9838614, \end{aligned}$$

$$\begin{aligned} H(\text{Habitable}|\text{Orbit}) =& \frac{3}{9} g\left[\frac{1}{3}\right] + \frac{6}{9} g\left[\frac{3}{6}\right] \\ =& 0.9727653, \end{aligned}$$

$$\begin{aligned} H(\text{Habitable}|\text{Temperature}_{205}) =& \frac{3}{9} g\left[\frac{0}{3}\right] + \frac{6}{9} g\left[\frac{4}{6}\right] \\ =& 0.6121972, \end{aligned}$$

and finally

$$\begin{aligned} H(\text{Habitable}|\text{Temperature}_{260}) =& \frac{6}{9} g\left[\frac{3}{6}\right] + \frac{3}{9} g\left[\frac{1}{3}\right] \\ =& 0.9727653 \end{aligned}$$

so the best choice for the first split is clearly $\text{Temperature}_{205}$. Furthermore we see that when $\text{Temperature}_{205} = \text{true}$, the labels in the available data are constant (no habitable planets), so we do not continue splitting on that side of the tree. The remaining data set for $\text{Temperature}_{205} = \text{false}$ is given in Table 1.

Table 1: Planet size, orbit, and temperature vs. habitability.

| Size | Orbit | $\text{Temperature}_{260}$ | Habitable |
|---|---|---|---|
| Big | Near | true | Yes |
| Big | Near | false | Yes |
| Small | Far | true | Yes |
| Small | Near | true | Yes |
| Small | Near | false | No |
| Small | Near | false | No |

The conditional entropies are

$$\begin{aligned} H(\text{Habitable}|\text{Size}, \text{Temperature}_{205} = \text{false}) =& \frac{2}{6} g\left[\frac{2}{2}\right] + \frac{4}{6} g\left[\frac{2}{4}\right] \\ =& 2/3, \end{aligned}$$

$$\begin{aligned} H(\text{Habitable}|\text{Orbit}, \text{Temperature}_{205} = \text{false}) =& \frac{5}{6} g\left[\frac{3}{5}\right] + \frac{1}{6} g\left[\frac{1}{1}\right] \\ =& 0.8091255, \end{aligned}$$

and

$$\begin{aligned} H(\text{Habitable}|\text{Temperature}_{260}, \text{Temperature}_{205} = \text{false}) =& \frac{3}{6} g\left[\frac{3}{3}\right] + \frac{3}{6} g\left[\frac{1}{3}\right] \\ =& 0.4591479 \end{aligned}$$

so the best split is on $\text{Temperature}_{260}$.

The subtree corresponding to $\text{Temperature}_{260} = \text{true}$ is not amenable to further splitting. As for the data in the subtree with $\text{Temperature}_{260} = \text{false}$, we see that the *Orbit* feature cannot be split on further since it only takes on the value "Near". A split on *Size*, however, perfectly classifies the remaining planets. So we can immediately conclude that the final tree looks like Figure 2. According to this tree, the (Big, Near, 280) example would be classified as habitable.

2. **[15 points]**

   (a) This is true simply by virtue of the fact that with each split we can reduce the number of remaining points by a factor of 2, hence at the leaves of the complete tree we can have exactly one data point remaining which can be classified as needed.

   (b) Let $(x_1, y_1)$,...,$(x_n, y_n)$ such that $x_i = i/n$ and $y_i = 0$ for all $i = 1, ..., n$. Let the labels be $+$ for all odd-indexed points, and $-$ for all even-indexed points. A simple inductive argument shows why these points cannot be perfectly classified with less than $n - 1$ splits.

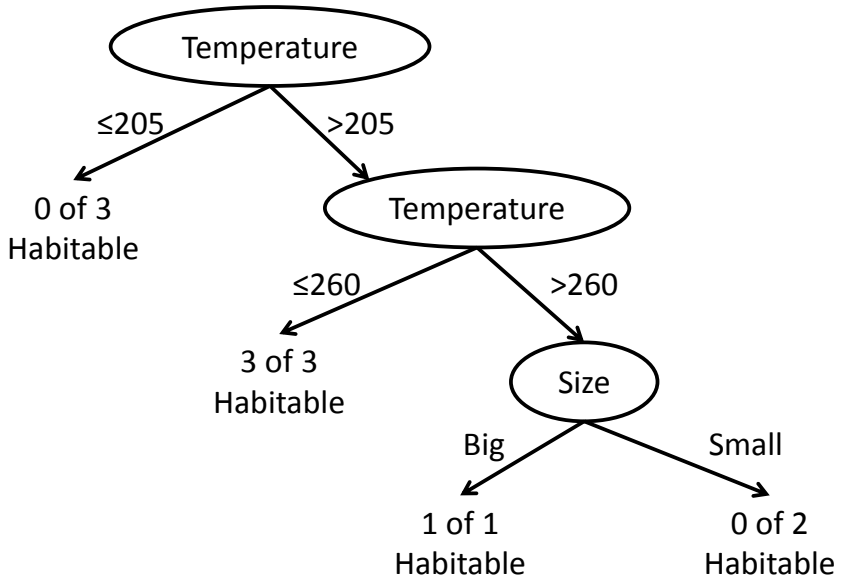


Figure 2: The decision tree for 1.1(b).

(c) Let $x_i = \frac{2\lceil \frac{i}{2} \rceil - 1}{n}$ and $y_i = \frac{2\lfloor \frac{i}{2} \rfloor}{n}$, with odd-indexed points labeled $+$. For $n = 10$, these points are plotted in Figure 3. Clearly these points can be easily separated by a linear decision boundary such as those given by logistic regression or SVM's, but a decision tree splitting only on one coordinate at a time will not be effective.

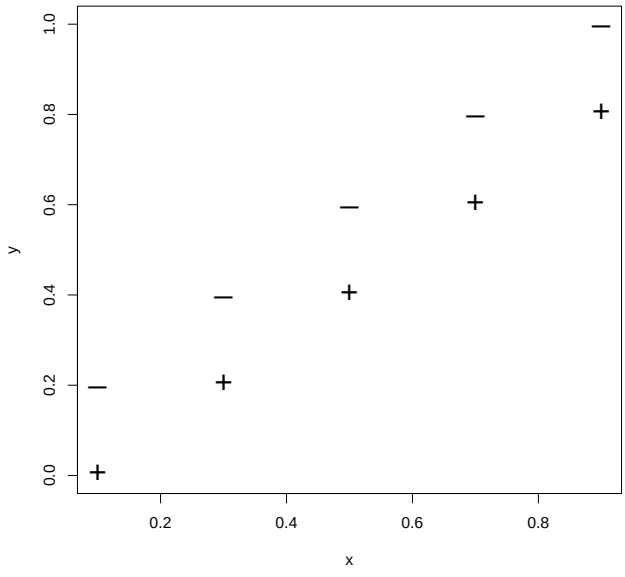


Figure 3: An illustration of the points in 1.2(c).

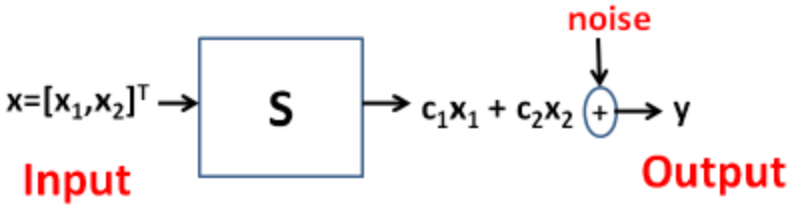


Figure 4: Exercise 2

## 2 Maximum Likelihood Estimation, [25pt, Avi]

Figure 2 shows a system S which takes two inputs $x_1, x_2$ (which are deterministic ) and outputs a linear combination of those two inputs, $c_1x_1+c_2x_2$, introduces an additive error $\epsilon$ which is a random variable following some distribution. Thus the output $y$ that you observe is given by equation 1. Assume that you have $n > 2$ instances $< x_{j1}, x_{j2}, y_j >_{j=1,\ldots,n}$ or equivalently $< x_j, y_j >$, where $x_j = [x_{j1}, x_{j2}]$.

$$y = c_1x_1 + c_2x_2 + \epsilon \tag{1}$$

In other words having $n$ equations in your hand is equivalent to having $n$ equations of the following form: $y_j = c_1x_{j1} + c_2x_{j2} + \epsilon_j,\ j = 1 \ldots n$. The goal is to estimate $c_1, c_2$ from those measurements by maximizing conditional log-likelihood given the input, under different assumptions for the noise. Specifically:

1. **[10 points]** Assume that the $\epsilon_i$ for $i = 1 \ldots n$ are iid Gaussian random variables with zero mean and variance $\sigma^2$.

   (a) Find the conditional distribution of each $y_i$ given the inputs
   **Ans:** $y_i \sim N(c_1x_{j1} + c_2x_{j2}, \sigma^2)$

   (b) Compute the loglikelihood of $y$ given the inputs.
   **Ans:** Since the noise are iid, the likelihood function is given by

   $$L(c_1, c_2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{(y_i - c_1x_{j1} - c_2x_{j2})^2}{2\sigma^2}$$

   Taking the logarithm we get the loglikelihood function which is a function of the two unknown parameters $c_1, c_2$:

   $$l(c_1, c_2) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (yi - c_1x_{j1} - c_2x_{j2})^2$$

   (c) Maximixe the likelihood above to get $c_{ls}$
   **Ans:** Let $y \in \mathcal{R}^n$ be the vector containing the measurements, $X$ the $n \times 2$ matrix with $X_{ij} = x_{ij}$ and $c = [c_1, c_2]^T$ then we are trying to minimize $\|y - Xc\|_2^2$ resulting in a solution $c = (X^TX)^{-1}X^Ty$

2. **[10 points]** Assume that the $\epsilon_i$ for $i = 1 \dots n$ are independent Gaussian random variable with zero mean and variance $Var(\epsilon_i) = \sigma_i$.

   (a) Find the conditional distribution of each $y_i$ given the inputs
   **Ans:** $y_i \sim N(c_1 x_{j1} + c_2 x_{j2}, \sigma_i^2)$

   (b) Compute the loglikelihood of $y$ given the inputs.
   **Ans:** Similar as before

$$l(c_1, c_2) = -\sum_{i=1}^{n} \frac{(y_i - c_1 x_{j1} - c_2 x_{j2})^2}{2\sigma_i^2}$$

   **Ans:** Now we are trying to minimize $\|W(y - Xc)\|^2$ where $W$ is a diagonal matrix with $w_{ii} = \frac{1}{\sigma_i}$ resulting is solution $c = (X^T W^T W X)^{-1} X^T W^T W y$

   (c) Maximixe the likelihood above to get $c_{wls}$

3. **[5 points]** Assume that $\epsilon_i$ for $i = 1 \dots n$ has density $f_{\epsilon_i}(x) = f(x) = \frac{1}{2b} \exp(-\frac{|x|}{b})$. In other words our noise is iid following Laplace distribution with location parameter $\mu = 0$ and scale parameter $b$.

   (a) Find the conditional distribution of each $y_i$ given the inputs

   (b) Compute the loglikelihood of $y$ given the inputs.
   **Ans:**

$$l(c_1, c_2) = -\frac{1}{b} \sum_{i=1}^{n} \|y - Xc\|_1$$

   (c) Comment on why this model leads to more robust solution.
   **Ans:** It is prepared to see higher values of residuals because it has a larger tail. Thus is more robust to noise and outliers