Homework 1 Solution

Xingyuan Pan

September 14, 2015

1 Decision Trees

1.1 Boolean functions as decision trees

(a) $x_1 \vee (x_2 \wedge x_3)$ **can be represented as:**

if $x_1 = 1$**:** $y = 1$

else if $x_2 = 0$**:** $y = 0$

else if $x_3 = 0$**:** $y = 0$

else: $y = 1$

(b) x_1 **xor** x_2 **can be represented as:**

if $x_1 = 0$**:**

    if $x_2 = 0$**:** $y = 0$

    else: $y = 1$

else:

    if $x_2 = 0$**:** $y = 1$

    else: $y = 0$

(c) The 2-of-3 function can be represented as:

if $x_1 = 1$**:**

  if $x_2 = 1$**:** $y = 1$

  else if $x_3 = 1$**:** $y = 1$

  else: $y = 0$

else:

  if $x_2 = 0$**:** $y = 0$

  else if $x_3 = 1$**:** $y = 1$

  else: $y = 0$

1.2 Restaurant problem

(a)

We have four features: Friday, Hungry, Patrons and Type. The number of values they can have are 2, 2, 3 and 4 respectively. The number of all possible inputs are $2 \times 2 \times 3 \times 4 = 48$. Each input can be assigned as Yes or No, so the total number of functions is 2^{48}. We have already seen 9 inputs in the training data set, which means we can only freely assign labels to the remaining 39 inputs. Therefore the number of functions consistent with the given training data is 2^{39}.

(b)

There are 5 inputs labeled as Yes and 4 inputs labeled as No. The entropy is

$$S = -\frac{5}{9}\log\frac{5}{9} - \frac{4}{9}\log\frac{4}{9} = 0.9911 \tag{1}$$

(c)

Let's consider the feature Friday. When the value of Friday is Yes, there are three examples, one of them is labeled as Yes and two labeled as No. The entropy for these three examples is

$$S(Friday = Yes) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = 0.9183 \tag{2}$$

Similarly when the value of Friday is No, there are six examples, four of them are labeled as Yes and two of them are labeled as No. The entropy for these six example is

$$S(Friday = No) = -\frac{4}{6}\log\frac{4}{6} - \frac{2}{6}\log\frac{2}{6} = 0.9183 \tag{3}$$

So the expectation of the entropy for choosing Friday is

$$S(Friday) = \frac{3}{9}S(Friday = Yes) + \frac{6}{9}S(Friday = No) = 0.9183 \tag{4}$$

And the information gain for choosing Friday is thus 0.9911 - 0.9183 = 0.0728. Similar calculations show the information gain for Hungry, Patrons and Type are 0.2295, 0.6305 and 0.1567.

(d)

Because the feature Patrons has the highest information gain, the root of the tree should use feature Patrons.

(e)

The decision tree for the restaurant problem is shown in Figure 1. The solution is not unique. As long as your tree consistent with training data and the root feature is Patrons, you are correct.

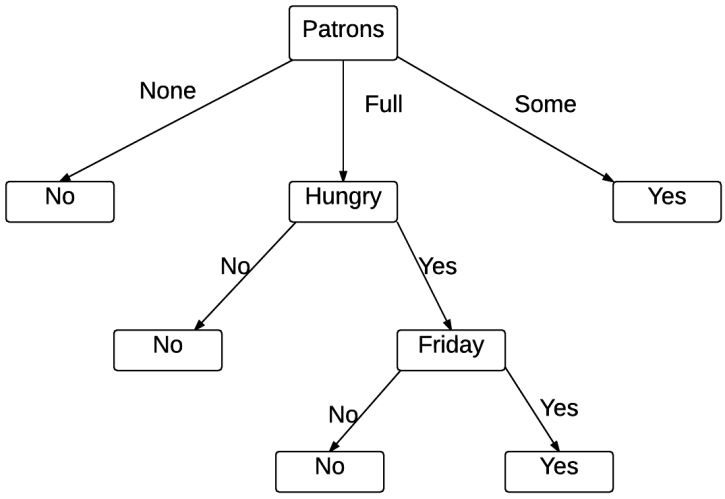

Figure 1: One possible decision tree for the restaurant problem

(f)

According to the tree given in Figure 1, only the first test example is predicted incorrectly. The accuracy on test data is 2/3, or 66.7%.

1.3 Misclassification rate and Gini coefficient

(a)

Using misclassification rate the information gain can be written as

$$\text{Gain}(S, A) = \text{Misclassification}(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} \text{Misclassification}(S_v)$$

The misclassification of the whole training data is $1 - \max(4/9, 5/9) = 4/9$. Let's consider the feature Friday. When the value of Friday is Yes, there are three examples, one of them is labeled as Yes and two labeled as No. The misclassification for these three examples is

$$S(Friday = Yes) = 1 - \max(1/3, 2/3) = 1/3 \tag{5}$$

Similarly when the value of Friday is No, there are six examples, four of them are labeled as Yes and two of them are labeled as No. The misclassification for these six examples is

$$S(Friday = No) = 1 - \max(2/6, 4/6) = 1/3 \tag{6}$$

So the expectation of the misclassification for choosing Friday is

$$S(Friday) = \frac{3}{9} S(Friday = Yes) + \frac{6}{9} S(Friday = No) = 1/3 \tag{7}$$

And the information gain for choosing Friday is thus $4/9 - 1/3 = 1/9$. Similar calculations show the information gain for Hungry, Patrons and Type are 2/9, 1/3 and 1/9. The root attribute should be Patrons.

(b)

The Gini coefficient for the whole training set is

$$\frac{4}{9}(1 - \frac{4}{9}) + \frac{5}{9}(1 - \frac{5}{9}) = 40/81 = 0.4938 \tag{8}$$

Let's consider the feature Friday. When the value of Friday is Yes, there are three examples, one of them is labeled as Yes and two labeled as No. The Gini coefficient for these three examples is

$$S(Friday = Yes) = \frac{1}{3}(1 - \frac{1}{3}) + \frac{2}{3}(1 - \frac{2}{3}) = 4/9 \tag{9}$$

Similarly when the value of Friday is No, there are six examples, four of them are labeled as Yes and two of them are labeled as No. The Gini coefficient for these six example is

$$S(Friday = No) = \frac{2}{6}(1 - \frac{2}{6}) + \frac{4}{6}(1 - \frac{4}{6}) = 4/9 \tag{10}$$

So the expectation of the Gini coefficient for choosing Friday is

$$S(Friday) = \frac{3}{9}S(Friday = Yes) + \frac{6}{9}S(Friday = No) = 4/9 \tag{11}$$

And the information gain for choosing Friday is thus $40/81 - 4/9 = 4/81 = 0.0494$. Similar calculations show the information gain for Hungry, Patrons and Type are 121/810 (0.1494), 53/162 (0.3272) and 7/81 (0.0864). The root attribute should be Patrons.

2 Nearest Neighbors

The Voronoi diagram using Euclidean distance is shown in Figure 2, and The Voronoi diagram using Manhattan distance is shown in Figure 3. Note that in Figure 3, the upper left corner can have either label A or B, since the distances are the same.

In order to get conflict between two distance measures, I only need two data points in my training set: $A(1, 1)$ and $B(1.5, 0)$. My test data example can be $C(0, 0)$. It is easy to see that

$$\text{distance}^{\text{Euclidean}}(AC) = \sqrt{2} \tag{12}$$
$$\text{distance}^{\text{Euclidean}}(BC) = 1.5 \tag{13}$$

and

$$\text{distance}^{\text{Manhattan}}(AC) = 2 \tag{14}$$
$$\text{distance}^{\text{Manhattan}}(BC) = 1.5 \tag{15}$$

The order has been reversed!

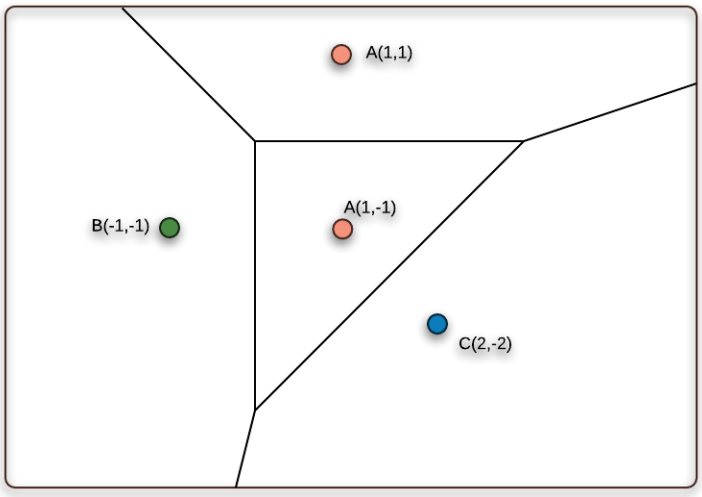

Figure 2: Voronoi diagram using Euclidean distance

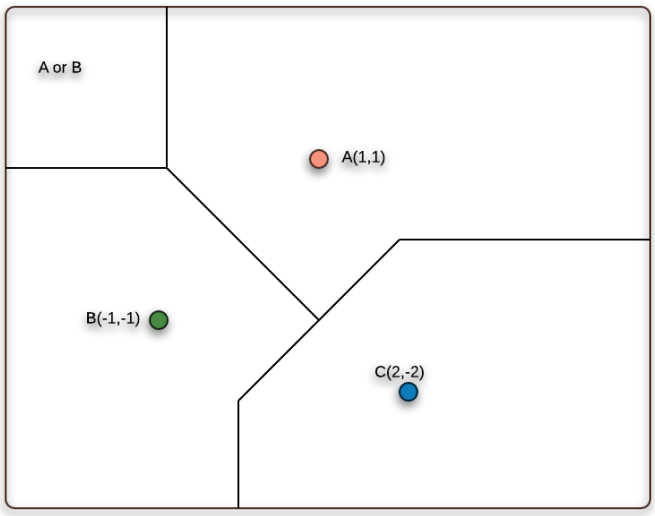

Figure 3: Voronoi diagram using Manhattan distance