CS57800 Statistical Machine Learning

Homework 1

I-Ta Lee
Department of Computer Science
lee2226@purdue.edu

September 15, 2015

1 Foundations

1.

$$x_1 + x_2 + 3x_3 = 4 \tag{1}$$

$$x_1 + 2x_2 + 4x_3 = 5 \tag{2}$$

Subtracting (1) from (2), we have

$$x_2 + x_3 = 1$$

This is equal to

$$x_2 = 1 - x_3$$

Let $x_3 = t$, we have

$$x_2 = 1 - t$$
$$x_3 = t$$

Put this back to equation (1), we get

$$x_1 + (1 - t) + 3t = 4$$
$$x_1 = 3 - 2t$$

So the line of intersection of the two planes is

$$\begin{cases} x_1 = 3 - 2t \\ x_2 = 1 - t \\ x_3 = t \end{cases} \quad \blacksquare$$

2. First we determine the two vectors $\overrightarrow{PQ}$ and $\overrightarrow{PR}$:

$$\overrightarrow{PQ} = (1-0)\hat{i} + (-1-0)\hat{j} + (1-0)\hat{k} = \hat{i} - \hat{j} + \hat{k}$$
$$\overrightarrow{PR} = (4-0)\hat{i} + (3-0)\hat{j} + (7-0)\hat{k} = 4\hat{i} + 3\hat{j} + 7\hat{k}$$

Then we get the normal vector

$$\overrightarrow{PQ} \times \overrightarrow{PR} = \begin{bmatrix} \hat{i} & \hat{j} & \hat{k} \\ 1 & -1 & 1 \\ 4 & 3 & 7 \end{bmatrix}$$
$$\begin{aligned} \overrightarrow{PQ} \times \overrightarrow{PR} &= -7\hat{i} + 4\hat{j} + 3\hat{k} + 4\hat{k} - 7\hat{j} - 3\hat{i} \\ &= -10\hat{i} - 3\hat{j} + 7\hat{k} \end{aligned}$$

The equation of the plane is

$$-10x - 3y + 7z = d$$

Plug in point $P(0,0,0)$ to find the value of d

$$-10x - 3y + 7z = 0$$

So the vector orthogonal to the plane is $(-10, -3, 7)$ ■

3. (a)

$$\begin{aligned} f'(x) &= (6x)(x^{\frac{1}{2}}) + (3x^2)(\frac{1}{2}x^{-\frac{1}{2}}) \\ &= 6x^{\frac{3}{2}} + \frac{3}{2}x^{\frac{3}{2}} \\ &= \frac{15}{2}x^{\frac{3}{2}} \end{aligned}$$

(b)

$$\begin{aligned} f'(x) &= \frac{1}{2}(e^{2x} + e)^{-\frac{1}{2}}(2e^{2x}) \\ &= e^{2x}(e^{2x} + e)^{-\frac{1}{2}} \end{aligned}$$

(c)

$$\begin{aligned} f'(x) &= 3[\ln(5x^2 + 9)]^2(5x^2 + 9)^{-1}(10x) \\ &= 30x(5x^2 + 9)^{-1}[\ln(5x^2 + 9)]^2 \end{aligned}$$

4. (a)

$$\frac{\partial f}{\partial x} = y^3 + 2xy^2$$
$$\frac{\partial f}{\partial y} = 3xy^2 + 2x^2y$$

(b)

$$\frac{\partial f}{\partial x} = e^{2x+3y} + 2xe^{2x+3y} = (1+2x)e^{2x+3y}$$
$$\frac{\partial f}{\partial y} = 3xe^{2x+3y}$$

5. The order from the lowest to the highest is

$$10^8 \prec \log^4 \sqrt{n} \prec 2^{\log_2 n} \prec \sqrt{n^3}\log^2 n \prec 2^{3\log_2 n} \prec 2^n \prec (\frac{5}{3})^{2n}$$

Proof each comparison:

(a) $10^8 \prec \log^4 \sqrt{n}$ By using L'Hospital's Rule, we have

$$\lim_{n\to\infty} \frac{10^8}{log^4\sqrt{n}} = 0$$

(b) $\log^4 \sqrt{n} \prec 2^{log_2 n}$

$$\log^4 \sqrt{n} = (\frac{1}{2}\log n)^4$$
$$2^{log_2 n} = n = (n^{\frac{1}{4}})^4$$

By using L'Hospital's Rule, we have

$$\lim_{n\to\infty} \frac{\log\sqrt{n}}{n^{\frac{1}{4}}} = \lim_{n\to\infty} \frac{\frac{1}{2}\frac{1}{\ln 10}\frac{1}{n}}{\frac{1}{4}n^{\frac{-3}{4}}} = \lim_{n\to\infty} \frac{2}{n^{\frac{1}{4}}\ln 10} = 0$$

(c) $2^{log_2 n} \prec \sqrt{n^3}\log^2 n$

$$2^{log_2 n} = n$$
$$\sqrt{n^3}\log^2 n = n^{\frac{3}{2}}\log^2 n$$

It is obvious that $n < n^{\frac{3}{2}}\log^2 n$

(d) $\sqrt{n^3}\log^2 n \prec 2^{3\log_2 n}$

$$\sqrt{n^3}\log^2 n = n^{\frac{3}{2}}\log^2 n$$
$$2^{3\log_2 n} = n^3$$

By using L'Hospital's Rule, we have

$$\lim_{n\to\infty} \frac{n^{\frac{3}{2}}\log^2 n}{n^3} = \lim_{n\to\infty} \frac{\log^2 n}{n^{\frac{3}{2}}} = \lim_{n\to\infty} \frac{2\log n\frac{1}{n\ln 10}}{\frac{3}{2}n^{\frac{1}{2}}} = \lim_{n\to\infty} \frac{2\log n}{\frac{3}{2}n^{\frac{3}{2}}\ln 10}$$
$$= \lim_{n\to\infty} \frac{2\frac{1}{n\ln 10}}{\frac{9}{4}n^{\frac{1}{2}}\ln 10} = \lim_{n\to\infty} \frac{8}{9n^{\frac{3}{2}}(\ln 10)^2} = 0$$

(e) $2^{3\log_2 n} \prec 2^n$

We simply compare $3\log_2 n$ and n by using L'Hospital's Rule

$$\lim_{n\to\infty} \frac{3\log_2 n}{n} = \lim_{n\to\infty} \frac{\frac{3}{n\ln 2}}{1} = \lim_{n\to\infty} \frac{3}{n\ln 2} = 0$$

(e) $2^n \prec (\frac{5}{3})^{2n}$

$$2^n = (2^{\frac{1}{2}})^{2n}$$

It is obvious that $2^{\frac{1}{2}} < \frac{5}{3}$, so $(2^{\frac{1}{2}})^{2n} \prec (\frac{5}{3})^{2n}$ ■

6. (a) Let x_1, x_2, x_3 be the values of three rolls respectively. Since they are independent, we have

$$\begin{aligned} E[x_1 + x_2 + x_3] &= E[x_1] + E[x_2] + E[x_3] = 3E[x_1] \\ &= 3[\frac{1}{6}(1+2+3+4+5+6)] = 3\frac{7}{2} = \frac{21}{2} \end{aligned}$$

(b) Let x_1, x_2, x_3 be the values of three rolls respectively. Since they are independent, we have

$$\begin{aligned} E[x_1x_2x_3] &= E[x_1]E[x_2]E[x_3] = E[x_1]^3 \\ &= [\frac{1}{6}(1+2+3+4+5+6)]^3 = (\frac{7}{2})^3 = \frac{343}{8} \end{aligned}$$

(c) Let x_1, x_2, x_3 be the values of three rolls respectively and let $X = x_1 + x_2 + x_3$.

$$\begin{aligned} Var[X] &= E[(X - E[X])^2] \\ &= (\frac{1}{6})^3[(3-\frac{21}{2})^2 + 3(4-\frac{21}{2})^2 + 6(5-\frac{21}{2})^2 + 10(6-\frac{21}{2})^2 + 15(7-\frac{21}{2})^2 \\ &\quad + 21(8-\frac{21}{2})^2 + 25(9-\frac{21}{2})^2 + 27(10-\frac{21}{2})^2 + 27(11-\frac{21}{2})^2 + 25(12-\frac{21}{2})^2 \\ &\quad + 21(13-\frac{21}{2})^2 + 15(14-\frac{21}{2})^2 + 10(15-\frac{21}{2})^2 + 6(16-\frac{21}{2})^2 + 3(17-\frac{21}{2})^2 + (18-\frac{21}{2})^2] \\ &= \frac{1}{216}[(\frac{-15}{2})^2 + 3(\frac{-13}{2})^2 + 6(\frac{-11}{2})^2 + 10(\frac{-9}{2})^2 + 15(\frac{-7}{2})^2 + 21(\frac{-5}{2})^2 + 25(\frac{-3}{2})^2 + 27(\frac{-1}{2})^2 \\ &\quad + 27(\frac{1}{2})^2 + 25(\frac{3}{2})^2 + 21(\frac{5}{2})^2 + 15(\frac{7}{2})^2 + 10(\frac{9}{2})^2 + 6(\frac{11}{2})^2 + 3(\frac{13}{2})^2 + (\frac{15}{2})^2] \\ &= \frac{1}{216}[\frac{225}{4} + \frac{507}{4} + \frac{726}{4} + \frac{810}{4} + \frac{735}{4} + \frac{525}{4} + \frac{225}{4} + \frac{27}{4} + \frac{27}{4} + \frac{225}{4} + \frac{525}{4} + \frac{735}{4} + \frac{810}{4} \\ &\quad + \frac{726}{4} + \frac{507}{4} + \frac{225}{4}] \\ &= \frac{7560}{864} = \frac{35}{4} \end{aligned}$$

2 Programming Report

2.1 Experiment Settings

In this project, I experimented the ID3 algorithm with 4 different parameters, including:

- Maximum depth of the tree: which limits the height of the tree
- Reduced error pruning: which makes post-pruning to the tree
- Threshold splitting: which splits training samples by using a real-value threshold. In order to do this, we also have an input parameter to indicate the value range of attributes
- Default label: used when we use majority votes but the numbers of votes for two labels are the same

We use two basic methods to construct the tree. The first one splits the training samples in each tree node by using discrete attribute values, which we call it "value splitting". For example, we use integers from 1 to 10 as the branching conditions and each tree node, except leaf nodes, has ten branches. The second one uses a real-value threshold to split the samples into two groups: one is smaller than or equal to the threshold and the other is larger than the threshold. In addition to these two methods, we apply reduced error pruning as the post-pruning method, which help us to generalize our model. We also experiment with the maximum depth limitation as the generalization method, though it does not provide much improvement.

Moreover, I have added three recursion terminating conditions to the ID3 algorithm introduced in the course slide:

- when the depth is equal to the parameter of maximum depth if specified
- when there is no attribute to be selected for a node, i.e., all the attributes have been selected by the upper-layer nodes
- when the partitioned training samples do not contain some values of a selected attribute in a node, we add a leaf node for that value branch by using majority vote

These conditions are discovered during my implementation and are required to make the algorithm complete.

2.2 Results

In the first experiment, we run the basic ID3 algorithm and tune the parameter of maximum depth. Figure 1 shows the impact of tuning the maximum depth. We can see that the training accuracy increases and the validating accuracy decrease as the maximum depth becomes larger. This meets our expectation that higher depth limitation will result in training overfit. We also expect that the testing accuracy will have similar trend to the validating accuracy. However, the testing accuracy does not change with this parameter. I think this is because the number of testing

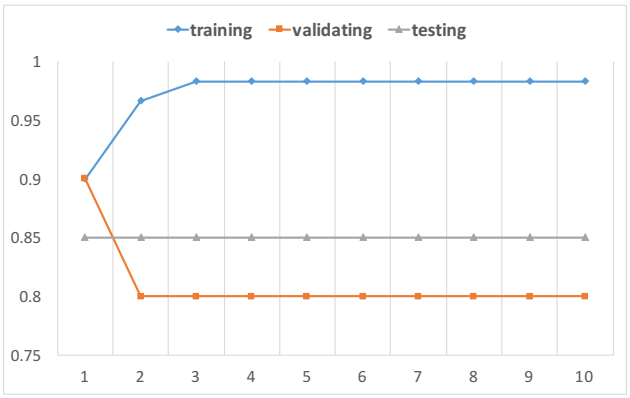

Figure 1: The Impact of Maximum Depth to Accuracy

data is only twenty, which is not enough to show the decreasing trend of testing accuracy. Moreover, all of the accuracies stay the same when the maximum depth is larger than 2.

In the second experiment, we test the impact of using value splitting and threshold splitting, where value splitting uses discrete value of our attributes, i.e., integers of 1 to 10, to split trees in to multiple groups in each node and threshold splitting uses a real-value threshold to split the tree into two groups in each node. Figure 2 shows the results, where in the first six bars we do not apply the reduced error pruning and in the last six bars we do so. Before we prune the tree, we can see that the value splitting has better performance in both validating and testing. In my opinion I think this is because the value splitting method gives more specific classification on each attribute, i.e., 10 branches for each layer, which results in better model, while the threshold splitting only has two branches in each layer. However, after pruning the tree, both two scenarios achieve great performance which are 90% in validating and 85% in testing.

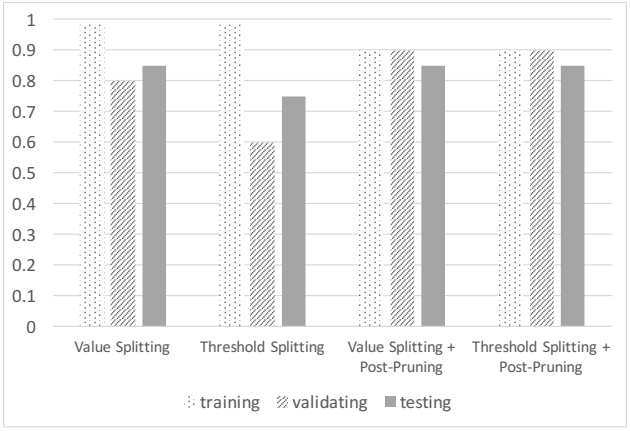

Figure 2: Value splitting and threshold splitting, and their post-pruning results

2.3 Error Analysis

Table 1 lists the testing samples that our model failed to predict. I analyzed the possible reasons resulting in these errors. The first reason would be the noisy training data. For the error of the

Index	Data	Label
6	5 3 4 1 4 1 3 1 1	0
8	10 6 3 6 4 10 7 8 4	1
16	3 1 1 1 2 1 3 1 1	1

Table 1: Testing samples that our model failed to predict

testing sample of index 6, our model started splitting on attribute index 2, because of its highest information gain. This attribute value of the testing sample of index 6 is 4, and in the training samples there are only two samples having this attribute value. These two training samples have one positive and one negative label, which the influence of noisy data. The second reason is that the number of training examples is not enough. For instance, the testing sample of index 8 and three training samples have the same attribute value in index 2 (which is 3) but bear completely different labels. The three training samples has negative labels, while the testing sample has the positive label. For another example, the testing sample of index 16 which has the attribute value beginning with [3, ?, 1, 1]. The three attributes (index 0, 2, 3) have the highest information gain in their layer. However, the training samples with these three attribute values have the opposite label to our testing sample of index 16. These two examples show that errors can be avoided If we have more samples belonging to this subtree.

2.4 Discussions

I have implemented nearly all the variations of decision trees introduced in the class but did not implement the dynamic strategy, the strategy that uses the prediction accuracy of validation set to help making the node creation decision. I think if we use the recursive method to construct the tree, it is infeasible to build the tree using dynamic validation. Since in the recursive method we builds tree in bottom-up order, the "deeper" leaves and internal nodes are first connected and then the root does so. When making the decision about adding a new node, we cannot use the validation set to measure the accuracy, since the root is not connected to other nodes. In addition, if we validate the tree after the tree construction, there is no different from the post-pruning strategy. Therefore, I think the only way to implement such strategy is to use non-recursive methods, which is inappropriate to merge into my current code base.