CS 540: Decision Tree Example:
Packers 2012

1 The Data

We would like to build a tree based on data from the Packers 2012 regular season[1]. The label we would like to predict is Win or Loss ({W,L}). Table 1 shows the data we have available[2]. The attributes we will use in our DT are:

Date: {E,O} Even or Odd day of the month

Home: {H,A} Home or Away game

TempBin: {1,2,3} Cool (Temp < 50), Warm ($50 \leq$ Temp < 75), Hot ($75 \leq$ Temp)

[Aside: Does it really make sense to use Date as we are? Should we always include all of the features that we have? What risks might we run? Since this is an example, we'll just ignore this issue and keep going with our (kind of weird) attributes.]

Game	DayOfMonth	**Date**	**Home**	Temp	**TempBin**	Win
01	9	O	H	72	2	L
02	13	O	H	60	2	W
03	24	E	A	67	2	L
04	30	E	H	64	2	W
05	7	O	A	48	1	L
06	14	E	A	81	3	W
07	21	O	A	70	2	W
08	28	E	H	44	1	W
09	4	E	H	39	1	W
10	18	E	A	70	2	W
11	25	O	A	34	1	L
12	2	E	H	45	1	W
13	9	O	H	33	1	W
14	16	E	A	44	1	W
15	23	O	H	24	1	W
16	30	E	A	70	2	L

Table 1: Packers 2012 Regular Season data

Before we begin, we'll split our dataset into a Training set and Tuning set, so that we'll have some data to use in the pruning phase. For this example, we'll use a 80%/20% split, holding aside 3 examples for a tuning set. We randomly select examples 2, 3 and 8 to be held aside as a tuning set. To be completely explicit, Tables 2 and 3 show the resulting Training and Tuning sets.

[1] All data is taken from pro-football-reference.com: http://www.pro-football-reference.com/teams/gnb/2012.htm

[2] The bye week at game 10 has been ignored

Instance	Date	Home	TempBin	Win
1	O	H	2	L
2	E	H	2	W
3	O	A	1	L
4	E	A	3	W
5	O	A	2	W
6	E	H	1	W
7	E	A	2	W
8	O	A	1	L
9	E	H	1	W
10	O	H	1	W
11	E	A	1	W
12	O	H	1	W
13	E	A	2	L

Table 2: Training Set (n=13)

Instance	Date	Home	TempBin	Win
1	O	H	2	W
2	E	A	2	L
3	E	H	1	W

Table 3: Tuning Set (m=3)

2 Building The Tree

As with any tree, we start by creating the root. Our first job is then to ask “What attribute (or question) should we ask at the root node?”. The question we'll ask is the one with the highest Information Gain (IG). So we'll need to calculate the IG for each of our 3 attributes: I(Win;Date), I(Win;Home) and I(Win;TempBin). We'll start with I(Win;Date).

2.1 Calculating I(Win;Date)

$$
\begin{aligned}
I(\text{Win};\text{Date}) &= H(\text{Win}) - H(\text{Win} \mid \text{Date}) && (1)\\
H(\text{Win}) &= -p_{\text{W}} \log_2 p_{\text{W}} - p_{\text{L}} \log_2 p_{\text{L}} && (2)\\
H(\text{Win} \mid \text{Date}) &= p_{\text{E}} \left[-(p_{\text{W}|\text{E}} \log_2 p_{\text{W}|\text{E}}) - (p_{\text{L}|\text{E}} \log_2 p_{\text{L}|\text{E}})\right] + \dots \\
&\qquad p_{\text{O}} \left[-(p_{\text{W}|\text{O}} \log_2 p_{\text{W}|\text{O}}) - (p_{\text{L}|\text{O}} \log_2 p_{\text{L}|\text{O}})\right] && (3)
\end{aligned}
$$

where we're using the shorthand p_{W} for $\Pr(\text{Win} = \text{W})$ and $p_{\text{W}|\text{E}}$ for $\Pr(\text{Win} = \text{W} \mid \text{Date} = \text{E})$.

We'll estimate all of the probabilities needed using our training set:

$$p_{\text{W}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\text{Win}_i = \text{W}\} = \frac{9}{13} \tag{4}$$

$$p_{\text{L}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\text{Win}_i = \text{W}\} = \frac{4}{13} \tag{5}$$

$$p_{\text{E}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\text{Date}_i = \text{E}\} = \frac{7}{13} \tag{6}$$

$$p_{\text{O}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\text{Date}_i = \text{O}\} = \frac{6}{13} \tag{7}$$

For the conditional probabilities, we look at appropriate subset of the examples. For example, to calculate $p_{\text{W}|\text{E}}$ and $p_{\text{L}|\text{E}}$ we'll look at instances {2,4,6,7,9,11,13}, the 7 instances where Date = E, shown in Table 4.

Instance	Date	Home	TempBin	Win
2	E	H	2	W
4	E	A	3	W
6	E	H	1	W
7	E	A	2	W
9	E	H	1	W
11	E	A	1	W
13	E	A	2	L

Table 4: Training instances where Date = E

Instance	Date	Home	TempBin	Win
1	O	H	2	L
3	O	A	1	L
5	O	A	2	W
8	O	A	1	L
10	O	H	1	W
12	O	H	1	W

Table 5: Training instances where Date = O

$$p_{\mathrm{W|E}} = \frac{1}{7} \sum_{i:\mathrm{Date}_i=\mathrm{E}} \mathbb{1}\{\mathrm{Win}_i = \mathrm{W}\} = \frac{6}{7} \tag{8}$$

$$p_{\mathrm{L|E}} = \frac{1}{7} \sum_{i:\mathrm{Date}_i=\mathrm{E}} \mathbb{1}\{\mathrm{Win}_i = \mathrm{L}\} = \frac{1}{7} \tag{9}$$

To calculate $p_{\mathrm{W|O}}$ and $p_{\mathrm{L|O}}$ we'll use the remaining 6 instances {1,3,5,8,10,12} shown in Table 5 (these are just the items where Date ≠ E, since Date is binary).

$$p_{\mathrm{W|O}} = \frac{1}{6} \sum_{i:\mathrm{Date}_i=\mathrm{O}} \mathbb{1}\{\mathrm{Win}_i = \mathrm{W}\} = \frac{3}{6} \tag{10}$$

$$p_{\mathrm{L|O}} = \frac{1}{6} \sum_{i:\mathrm{Date}_i=\mathrm{O}} \mathbb{1}\{\mathrm{Win}_i = \mathrm{L}\} = \frac{3}{6} \tag{11}$$

p_{W}	9/13
p_{L}	4/13
p_{E}	7/13
p_{O}	6/13
$p_{\mathrm{W\|E}}$	6/7
$p_{\mathrm{L\|E}}$	1/7
$p_{\mathrm{W\|O}}$	3/6
$p_{\mathrm{L\|O}}$	3/6

Table 6: Summary of values calculated

Table 6 shows a summary of all of the values we've found so far. Plugging all of these values into Equations 2, 3 and finally 1 we get:

$$H(\mathrm{Win}) = -\left(\frac{9}{13}\log_2\frac{9}{13}\right) - \left(\frac{4}{13}\log_2\frac{4}{13}\right) \approx 0.89 \tag{12}$$

$$H(\mathrm{Win} \mid \mathrm{Date}) = \frac{7}{13}\left[-\left(\frac{6}{7}\log_2\frac{6}{7}\right) - \left(\frac{1}{7}\log_2\frac{1}{7}\right)\right] + \frac{6}{13}\left[-\left(\frac{3}{6}\log_2\frac{3}{6}\right) - \left(\frac{3}{6}\log_2\frac{3}{6}\right)\right] \approx 0.78 \tag{13}$$

$$\mathbf{I(Win;Date) = 0.89 - 0.78 = 0.11} \tag{14}$$

2.2 Calculating I(Win;Home) and I(Win;TempBin)

Of course we need two more values before we can determine which is the best question to ask at our root node: I(Win;Home) and I(Win;TempBin). The same procedure as was used in the last section is used to calculate both of these values as well. We'll skip to the final equations, but it is a good exercise to double check the values we're using here.

$$\begin{align}
H(\text{Win} \mid \text{Home}) &= p_{\text{H}} \left[-(p_{\text{W|H}} \log_2 p_{\text{W|H}}) - (p_{\text{L|H}} \log_2 p_{\text{L|H}})\right] + \dots \nonumber \\
&\quad\; p_{\text{A}} \left[-(p_{\text{W|A}} \log_2 p_{\text{W|A}}) - (p_{\text{L|A}} \log_2 p_{\text{L|A}})\right] \tag{15} \\
&= \frac{6}{13}\left[-\left(\frac{5}{6}\log_2\frac{5}{6}\right) - \left(\frac{1}{6}\log_2\frac{1}{6}\right)\right] + \frac{7}{13}\left[-\left(\frac{4}{7}\log_2\frac{4}{7}\right) - \left(\frac{3}{7}\log_2\frac{3}{7}\right)\right] \approx 0.83 \tag{16} \\
\mathbf{I(Win; Date)} &= \mathbf{0.89 - 0.83 = 0.06} \tag{17}
\end{align}$$

$$\begin{align}
H(\text{Win} \mid \text{TempBin}) &= p_1 \left[-(p_{\text{W|1}} \log_2 p_{\text{W|1}}) - (p_{\text{L|1}} \log_2 p_{\text{L|1}})\right] + \dots \nonumber \\
&\quad\; p_2 \left[-(p_{\text{W|2}} \log_2 p_{\text{W|2}}) - (p_{\text{L|2}} \log_2 p_{\text{L|2}})\right] + \dots \nonumber \\
&\quad\quad p_3 \left[-(p_{\text{W|3}} \log_2 p_{\text{W|3}}) - (p_{\text{L|3}} \log_2 p_{\text{L|3}})\right] \tag{18} \\
&= \frac{7}{13}\left[-\left(\frac{5}{7}\log_2\frac{5}{7}\right) - \left(\frac{2}{7}\log_2\frac{2}{7}\right)\right] + \dots \nonumber \\
&\quad\; \frac{5}{13}\left[-\left(\frac{3}{7}\log_2\frac{3}{7}\right) - \left(\frac{2}{7}\log_2\frac{2}{7}\right)\right] + \dots \nonumber \\
&\quad\quad \frac{1}{13}\left[-\left(1\log_2 1\right) - \left(0\log_2 0\right)\right] \approx 0.84 \tag{19} \\
\mathbf{I(Win; TempBin)} &= \mathbf{0.89 - 0.84 = 0.05} \tag{20}
\end{align}$$

2.3 Choosing the question at the root

So now we have all the information we need to choose which question we will ask at the root. We choose the attribute which contributes the highest Info Gain. In this case that is **Date**, due to [I(Win;Date) = 0.11] > [I(Win;Home) = 0.06] > [I(Win;TempBin) = 0.05]. Asking this question results in two child nodes: Date = E and Date = O. Our Training Set will now be split between these two nodes. This is represented in Figure 1.

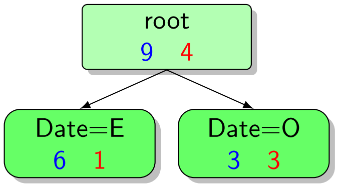

Figure 1: Tree after asking first question. Blue is number of Wins, Red the number of Losses at that node.

2.4 Choosing the next set of questions

We now recurse down into both nodes and continue building the tree. For each node we first must ask (1) is this node pure? (2) have we asked all of the available questions? (2) have we run out of examples in this path? In both cases, the answer to all of these questions is no, so we need to continue and determine what question should be asked at both nodes. Tables 6 and 7 contain the subsets of the data that reach each of the child nodes. We'll speed through this, only showing the math, but again, it is a good exercise to double check the values used here.

Instance	Date	Home	TempBin	Win
2	E	H	2	W
4	E	A	3	W
6	E	H	1	W
7	E	A	2	W
9	E	H	1	W
11	E	A	1	W
13	E	A	2	L

Table 7: Subset of Training Set where Date=E

Instance	Date	Home	TempBin	Win
1	O	H	2	L
3	O	A	1	L
5	O	A	2	W
8	O	A	1	L
10	O	H	1	W
12	O	H	1	W

Table 8: Subset of Training Set where Date=O

To find the question to ask at Date=E:

$$H(\text{Win} \mid \text{Date} = \text{E}) = -\left(\frac{6}{7}\log_2\frac{6}{7}\right) - \left(\frac{1}{7}\log_2\frac{1}{7}\right) \approx 0.59 \tag{21}$$

$$H(\text{Win} \mid \text{Home}, \text{Date} = \text{E}) = \frac{3}{7}\left[-(1\log_2 1) - (0\log_2 0)\right] + \frac{4}{7}\left[-\left(\frac{3}{4}\log_2\frac{3}{4}\right) - \left(\frac{1}{4}\log_2\frac{1}{4}\right)\right] \approx 0.46 \tag{22}$$

$$\mathbf{I(Win; Home \mid Date = D) = 0.59 - 0.46 = 0.13} \tag{23}$$

$$\begin{aligned} H(\text{Win} \mid \text{TempBin}, \text{Date} = \text{E}) = &\frac{3}{7}\left[-(1\log_2 1) - (0\log_2 0)\right] + \dots \\ &\frac{3}{7}\left[-\left(\frac{2}{3}\log_2\frac{2}{3}\right) - \left(\frac{1}{3}\log_2\frac{1}{3}\right)\right] + \dots \\ &\frac{1}{7}\left[-(1\log_2 1) - (0\log_2 0)\right] \approx .39 \end{aligned} \tag{24}$$

$$\mathbf{I(Win; TempBin \mid Date = D) = 0.59 - 0.39 = 0.20} \tag{25}$$

$$\tag{26}$$

So at Date=E we'll choose to ask [TempBin?].

To find the question to ask at Date=O:

$$H(\text{Win} \mid \text{Date} = \text{O}) = -\left(\frac{3}{6}\log_2\frac{3}{6}\right) - \left(\frac{3}{6}\log_2\frac{3}{6}\right) = 1 \tag{27}$$

$$H(\text{Win} \mid \text{Home}, \text{Date} = \text{O}) = \frac{3}{6}\left[-\left(\frac{2}{3}\log_2\frac{2}{3}\right) - \left(\frac{1}{3}\log_2\frac{1}{3}\right)\right] + \frac{3}{6}\left[-\left(\frac{1}{3}\log_2\frac{1}{3}\right) - \left(\frac{2}{3}\log_2\frac{2}{3}\right)\right] \approx 0.92 \tag{28}$$

$$\mathbf{I(Win; Home \mid Date = O) = 1 - .92 = 0.08} \tag{29}$$

$$\begin{aligned} H(\text{Win} \mid \text{TempBin}, \text{Date} = \text{O}) = {} & \frac{4}{6}\left[-\left(\frac{2}{4}\log_2\frac{2}{4}\right) - \left(\frac{2}{4}\log_2\frac{2}{4}\right)\right] + \dots \\ & \frac{2}{6}\left[-\left(\frac{1}{2}\log_2\frac{1}{2}\right) - \left(\frac{1}{2}\log_2\frac{1}{2}\right)\right] + \dots \\ & \frac{0}{6}[\,] = 1 \end{aligned} \tag{30}$$

$$\mathbf{I(Win; TempBin \mid Date = O) = 1 - 1 = 0} \tag{31}$$

$$\tag{32}$$

So at Date=E we'll choose to ask [Home?]. The resulting extended tree is shown in Figure 2.

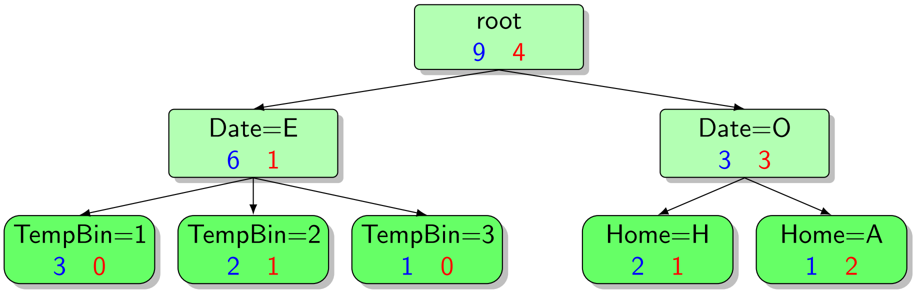

Figure 2: Tree after asking first question. Blue is number of Wins, Red the number of Losses at that node.

2.5 Choosing the final set of questions

At this point there are 6 leaves. Two of them, Date=E→TempBin=1 and Date=E→TempBin=3 are pure, so they will remain leaves. The other three nodes still have examples in both categories, and there still remains one question to ask in each case.

For instance, at Date=E→TempBin=2, the attribute Home still remains. Since this is the only attribute left, we don't need to calculate any IG at this point. The same thing applies to the other two nodes as well. In a few cases we'll need to break a tie, and will use majority vote on the full training set. After asking the remaining questions and applying labels we get the tree shown in Figure 3. There is one node which still is not pure, but we've run out of questions to ask.

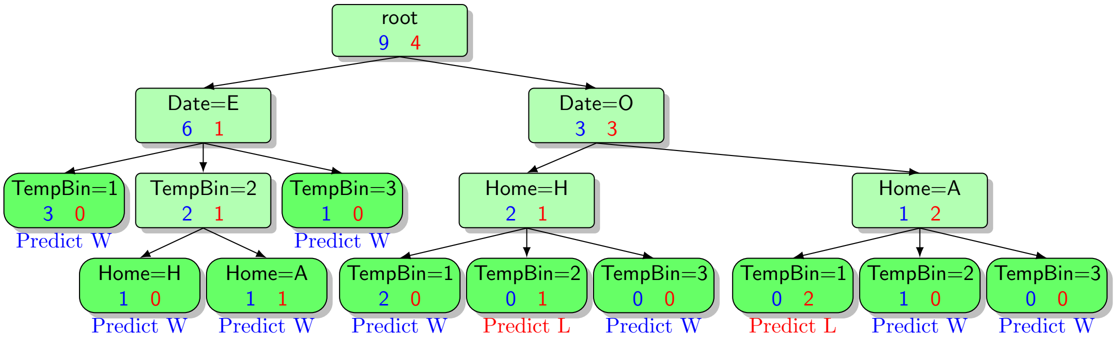

Figure 3: Tree after asking final questions. Blue is number of Wins, Red the number of Losses at that node.

3 Pruning

There are 6 internal nodes which are candidates for pruning. We show our pruning algorithm in action below. The accuracy shown is on the tuning set, repeated here.

Instance	Date	Home	TempBin	Win
1	O	H	2	W
2	E	A	2	L
3	E	H	1	W

Table 9: Tuning Set (m=3)

node pruned	tune acc.
none	1/3
Root	2/3
Date=E	1/3
Date=E→TempBin=2	1/3
Date=O	2/3
Date=O→Home=H	2/3
Date=O→Home=A	1/3

Table 10: First round of pruning.

Since our Tuning Set is so small, we get many prunes with the same Tuning Set Accuracy. If we break ties by selecting the prune which removes the most nodes, we end up with a tree that consists only of our Root node, which means none of our attributes really helped us with classification and all of this work has really been for naught, except hopefully to demonstrate how to build and prune a DT.

4 Testing

We could now ask "How well do our full and pruned trees generalize?". We can use the 2011 season as a test set (Table 11). Both our full tree and our pruned tree have a Test Set Accuracy of 15/16. In this case our tree seems to generalize well even without pruning. But I wouldn't place any bets on it. (If we look at the 2013 season, 3 games in when this was written, our full tree is 0/3 and our pruned tree is only 1/3.)

Game	DayOfMonth	**Date**	**Home**	Temp	**TempBin**	Win
01	8	E	H	68	2	W
02	18	E	A	64	2	W
03	25	O	A	61	2	W
04	2	E	H	64	2	W
05	9	O	A	67	2	W
06	16	E	H	70	2	W
07	23	O	A	57	2	W
08	6	E	A	70	2	W
09	14	E	H	70	2	W
10	20	E	H	32	1	W
11	24	E	A	70	2	W
12	4	E	A	52	2	W
13	11	O	H	43	1	W
14	18	E	A	50	2	L
15	25	O	H	37	1	W
16	1	O	H	31	1	W

Table 11: Packers 2011 Regular Season data