Math 345 - Notes
Decision Trees II
November 7, 2018

Decison Trees - Example

Suppose we want to build a decision tree that outputs Yes (y) or No (n) for walking the dog based on the following attributes

Outlook: sunny (s), cloudy (c), rain (r)

Temperature: hot (h), mild (m), freezing (f)

Wind: yes (y), no (n)

Time: morning (m), afternoon (a), evening (e)

The labels are then $c_1 = y$ and $c_2 = n$. Recall the notation $p_1 = P(\text{label} = c_1)$ and $p_2 = P(\text{label} = c_2)$.
We are given the following 10 data points to train our decision tree:

Outlook	Temperature	Wind	Time	Label
s	m	y	m	y
c	m	y	e	n
r	f	y	e	n
r	m	y	a	y
s	h	n	a	n
r	f	n	m	n
r	h	n	m	y
c	h	n	m	y
c	m	n	e	y
c	h	n	a	y

At the root, let X be initialized as $\mathcal{D}$. Since the data set has 6 y labels and 4 n labels, $p_1 = 0.6$ and $p_2 = 0.4$. Similarly, when we condition on a certain *outlook*, we count only the data points in that specific subset.

$$
\begin{aligned}
H(X) &= -\frac{6}{10}\log_2\left(\frac{6}{10}\right) - \frac{4}{10}\log_2\left(\frac{4}{10}\right) = 0.9709 \\
\hline
H(X|\text{outlook}=s) &= -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1 \\
H(X|\text{outlook}=c) &= -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = .81 \\
H(X|\text{outlook}=r) &= -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1 \\
H(X|\text{outlook}) &= \sum_{j\in\{s,c,r\}} P(\text{outlook}=j)H(X|\text{outlook}=j) = \frac{2}{10}(1) + \frac{4}{10}(.81) + \frac{4}{10}(1) = .9245 \\
IG(X,\text{outlook}) &= H(X) - H(X|\text{outlook}) = .9709 - .9245 = .0464 \\
\hline
H(X|\text{temp}=h) &= -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = .81 \\
H(X|\text{temp}=m) &= -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = .81 \\
H(X|\text{temp}=f) &= -\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right) = 0 \\
H(X|\text{temp}) &= \sum_{j\in\{h,m,f\}} P(\text{temp}=j)H(X|\text{temp}=j) = \frac{4}{10}(.81) + \frac{4}{10}(.81) + \frac{2}{10}(0) = .648 \\
IG(X,\text{temp}) &= H(X) - H(X|\text{temp}) = .9709 - .648 = .3229 \\
\hline
H(X|\text{wind}=y) &= -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1 \\
H(X|\text{wind}=n) &= -\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) = .9182 \\
H(X|\text{wind}) &= \sum_{j\in\{y,n\}} P(\text{wind}=j)H(X|\text{wind}=j) = \frac{4}{10}(1) + \frac{6}{10}(.9182) = .9509 \\
IG(X,\text{wind}) &= H(X) - H(X|\text{wind}) = .9709 - .9509 = .0191 \\
\hline
H(X|\text{time}=m) &= -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = .81 \\
H(X|\text{time}=a) &= -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) = .9182 \\
H(X|\text{time}=e) &= -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) = .9182 \\
H(X|\text{time}) &= \sum_{j\in\{m,a,e\}} P(\text{time}=j)H(X|\text{time}=j) = \frac{4}{10}(.81) + \frac{3}{10}(.9182) + \frac{3}{10}(.9182) = .8749 \\
IG(X,\text{temp}) &= H(X) - H(X|\text{temp}) = .9709 - .8749 = .0951
\end{aligned}
$$

Since the largest information gain IG and smallest entropy H come from the attribute *temperature*, the first node will be **temp**.

We branch based on the 3 different categories of temperature:

hot (h) – 4 data points,

mild (m) – 4 data points,

freezing (f) – 2 data points.

Let X_1 be the set of outcomes restricted to temp $= f$. Then $H(X_1) = 0$ because all the labels are NO. We create a leaf coming from this branch, with the label NO.

Let X_2 be the set of outcomes restricted to temp $= h$. Then

$$
\begin{array}{rcl}
H(X_2) & = & H(X|\text{temp} = h) = .81 \\
\hline
H(X_2|\text{outlook} = s) & = & -\frac{0}{1}\log_2\left(\frac{0}{1}\right) - \frac{1}{1}\log_2\left(\frac{1}{1}\right) = 0 \\
H(X_2|\text{outlook} = c) & = & -\frac{2}{2}\log_2\left(\frac{2}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right) = 0 \\
H(X_2|\text{outlook} = r) & = & -\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right) = 0 \\
H(X_2|\text{outlook}) & = & \sum_{j\in\{m,a,e\}} P(\text{outlook} = j)H(X_2|\text{outlook} = j) = 0 \\
IG(X_2, \text{outlook}) & = & H(X_2) - H(X_2|\text{outlook}) = .81 - 0 = .81
\end{array}
$$

Note that we cannot gain more information than this, or get a smaller entropy, so we set **outlook** as the node coming from the *hot* branch of **temp**. We branch X_2 based on the 3 different categories of outlook:

sunny (s) – 1 data point (NO) $\Rightarrow$ label leaf coming from *sunny* with NO.

cloudy (c) – 2 data points (YES, YES) $\Rightarrow$ label leaf coming from *cloudy* with YES.

rain (r) – 1 data point (YES) $\Rightarrow$ label leaf coming from *rain* with YES.

Now we labeled all data that followed the *hot* branch from the root node **temp**. We return to the last branch from the root and let X_4 be the set of outcomes restricted to temp $= m$. Then

$$
\begin{array}{rcl}
H(X_4) & = & H(X|\text{temp} = m) = .81 \\
\hline
H(X_4|\text{outlook} = s) & = & -\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right) = 0 \\
H(X_4|\text{outlook} = c) & = & -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1 \\
H(X_4|\text{outlook} = r) & = & -\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right) = 0 \\
H(X_4|\text{outlook}) & = & \sum_{j\in\{m,a,e\}} P(\text{outlook} = j)H(X_2|\text{outlook} = j) = 0 + \frac{1}{2}(1) + 0 = 0.5 \\
IG(X_4, \text{outlook}) & = & H(X_4) - H(X_4|\text{outlook}) = .81 - 0.5 = .31 \\
\hline
H(X_4|\text{wind} = y) & = & -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) = .9182 \\
H(X_4|\text{wind} = n) & = & -\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right) = 0 \\
H(X_4|\text{wind}) & = & \sum_{j\in\{y,n\}} P(\text{wind} = j)H(X|\text{wind} = j) = \frac{3}{4}(.9182) + \frac{1}{4}(0) = .682 \\
IG(X_4, \text{wind}) & = & H(X_4) - H(X_4|\text{wind}) = .81 - .682 = .127 \\
\hline
H(X_4|\text{time} = m) & = & -\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right) = 0 \\
H(X_4|\text{time} = a) & = & -\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right) = 0 \\
H(X_4|\text{time} = e) & = & -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1 \\
H(X_4|\text{time}) & = & \sum_{j\in\{m,a,e\}} P(\text{time} = j)H(X|\text{time} = j) = 0 + 0 + \frac{1}{2}(1) = .5 \\
IG(X_4, \text{temp}) & = & H(X_4) - H(X_4|\text{temp}) = .81 - .5 = .31
\end{array}
$$

Since the largest information gain IG and smallest entropy H come from the attributes *time* and *outlook*, we pick one at random: *time* , so the node following the branch *mild* from the note **temp** will be **time**. We branch based on the 3 different categories of time:

morning (m) – 1 data point (YES) $\Rightarrow$ label leaf coming from *morning* with YES

afternoon (a) – 1 data point (YES) $\Rightarrow$ label leaf coming from *afternoon* with YES

evening (e) – 2 data points (YES, NO) $\Rightarrow$ call this set X_5

Let X_5 be the set of outcomes from X_4 restricted to time $= e$. Then

$$
\begin{array}{rcl}
H(X_5) & = & H(X_4|\text{time} = e) = .5 \\
\hline
H(X_5|\text{wind} = y) & = & -\frac{0}{1}\log_2\left(\frac{0}{1}\right) - \frac{1}{1}\log_2\left(\frac{1}{1}\right) = 0 \\
H(X_5|\text{wind} = n) & = & -\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right) = 0 \\
H(X_5|\text{wind}) & = & \sum_{j\in\{y,n\}} P(\text{wind} = j)H(X|\text{wind} = j) = 0 \\
IG(X_5, \text{wind}) & = & H(X_5) - H(X_5|\text{wind}) = .5
\end{array}
$$

Note that we cannot gain more information than this, or get a smaller entropy, so we set **wind** as the node coming from the *evening* branch of **time**. We branch X_5 based on the 2 different categories of windy:

yes (y) – 1 data point (NO) $\Rightarrow$ label leaf coming from *yes* with NO.

no (n) – 1 data point (YES) $\Rightarrow$ label leaf coming from *no* with YES.

We have exhausted all cases, so we get the following tree: