Decision Trees - II

Prof. Dan A. Simovici

UMB

1. Partition Entropy

2. Conditional Entropy and Entropy Gain

3. Partitions Defined on Tables

4. The rpart Package

The notion of entropy is a probabilistic concept that lies at the foundation of information theory.
Our goal is to define entropy in an algebraic setting by introducing the notion of entropy of a partition of a finite set. This approach allows us to take advantage of the partial order that is naturally defined on the set of partitions. Actually, we introduce a generalization of the notion of entropy that has the Gini index and Shannon entropy as special cases.

In classical information theory the *Shannon entropy* of a probability distribution $\mathbf{p} = (p_1, \ldots, p_m)$, where $p_i > 0$ for $1 \leqslant i \leqslant m$ and $p_1 + \cdots + p_m = 1$ is defined as

$$\mathcal{H}(p_1, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log_2 p_i. = \sum_{i=1}^{m} p_i \log_2 \frac{1}{p_i}.$$

If $\pi = \{B_1, \ldots, B_m\}$ is a partition of a set S, then a probability distribution $\mathbf{p}_\pi$ can be defined as

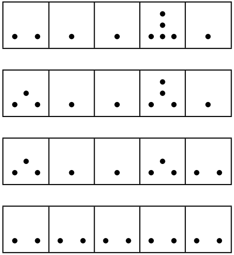
$$\mathbf{p}_\pi = \left(\frac{|B_1|}{|S|}, \cdots, \frac{|B_m|}{|S|} \right).$$

Accordingly, we can define the Shannon entropy of a partition π as:

$$\mathcal{H}(\pi) = -\sum_{i=1}^{m} \frac{|B_i|}{|S|} \log_2 \frac{|B_i|}{|S|}.$$

Example

Let S be a set containing ten elements and let $\pi_1, \pi_2, \pi_3, \pi_4$ be the four partitions shown below.



$\mathcal{H}(\pi_4) = 1.96$

$\mathcal{H}(\pi_3) = 2.04$

$\mathcal{H}(\pi_2) = 2.17$

$\mathcal{H}(\pi_1) = 2.32$

The partition π_1, which is the most uniform (each block containing two elements), has the largest entropy. At the other end of the range, partition π_4 has a strong concentration of elements in its fourth block and the lowest entropy.
The entropy can be viewed as a measure of impurity of a partition.

Definition

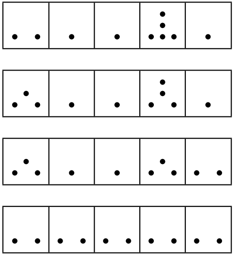
The *Gini index of* π is the number

$$\mathrm{gini}(\pi) = 1 - \sum_{i=1}^{m} \left(\frac{|B_i|}{|S|} \right)^2 .$$

Like Shannon entropy, the Gini index can be used to evaluate the uniformity of the distribution of the elements of S in the blocks of π because both $\mathcal{H}(\pi)$ and $\mathrm{gini}(\pi)$ increase with the uniformity of the distribution of the elements of S.

Example

Results concerning the Gini index are shown next:



$\text{gini}(\pi_4) = 0.68$

$\text{gini}(\pi_3) = 0.72$

$\text{gini}(\pi_2) = 0.79$

$\text{gini}(\pi_1) = 0.80$

Generalized Entropy

Definition

Let $\pi = \{B_1, \ldots, B_m\}$ be a partition of a set S and let $\beta > 1$. The β-entropy of a partition π is the number

$$\mathcal{H}_\beta(\pi) = \frac{1}{1 - 2^{1-\beta}} \cdot \left(1 - \sum_{i=1}^{m} \left(\frac{|B_i|}{|S|}\right)^\beta\right).$$

If $\beta = 2$, we obtain $\mathcal{H}_2(\pi)$, which is twice the Gini index,

$$\mathcal{H}_\beta(S, \pi) = 2 \cdot \left(1 - \sum_{i=1}^{m} \left(\frac{|B_i|}{|S|}\right)^2\right).$$

The *Gini index*, $\text{gini}(\pi) = 1 - \sum_{i=1}^{m} \left(\frac{|B_i|}{|S|}\right)^2$, is widely used in machine learning and data mining.

When we take $\lim_{\beta\to 1} \mathcal{H}_\beta(\pi)$ we obtain the Shannon entropy! Indeed, we can write:

$$\begin{aligned}
&\lim_{\beta\to 1} \mathcal{H}_\beta(\pi) \\
&= \lim_{\beta\to 1} \frac{1}{1-2^{1-\beta}} \cdot \left(1 - \sum_{i=1}^{m} \left(\frac{|B_i|}{|S|}\right)^{\beta}\right) \\
&= \lim_{\beta\to 1} \frac{-\sum_{i=1}^{m} \left(\frac{|B_i|}{|S|}\right)^{\beta} \ln \frac{|B_i|}{|S|}}{2^{1-\beta}\ln 2} \\
&\quad \text{(by l'Hôpital Rule)} \\
&= -\sum_{i=1}^{m} \frac{|B_i|}{|S|} \log_2 \frac{|B_i|}{|S|}.
\end{aligned}$$

Reminder:

Definition

Let $\pi \in \mathrm{PART}(S)$ and let $C \subseteq S$.
The trace of π on C is the partition π_C of C given by:

$$\pi_C = \{B \cap C | B \in \pi \text{ such that } B \cap C \neq \emptyset\}.$$

Clearly, $\pi_C \in \mathrm{PART}(C)$; also, if C is a block of π, then $\pi_C = \omega_C$.

Definition

Let $\pi, \sigma \in \mathrm{PART}(S)$ and let $\sigma = \{C_1, \ldots, C_n\}$. The *$\beta$-conditional entropy* of the partitions $\pi, \sigma \in \mathrm{PART}(S)$ is the function $\mathcal{H}_\beta : \mathrm{PART}(S)^2 \longrightarrow \mathbb{R}_{\geqslant 0}$ defined by

$$\mathcal{H}_\beta(\pi|\sigma) = \sum_{j=1}^{n} \left(\frac{|C_j|}{|S|} \right)^\beta \mathcal{H}_\beta(\pi_{C_j}).$$

The Shannon conditional entropy is:

$$\mathcal{H}(\pi|\sigma) = \sum_{j=1}^{n} \frac{|C_j|}{|S|} \mathcal{H}(\pi_{C_j}).$$

The Shannon conditional entropy is a limit case of the β-condional entropy, that is, $\mathcal{H}(\pi|\sigma) = \lim_{\beta \to 1} \mathcal{H}_\beta(\pi|\sigma)$.

Note that for $\pi \in \mathrm{PART}(S)$ we have:

$$\mathcal{H}_\beta(\pi|\omega_S) = \mathcal{H}_\beta(\pi)$$

and that

$$\begin{aligned}
\mathcal{H}_\beta(\omega_S|\sigma) &= \sum_{j=1}^{n} \left(\frac{|C_j|}{|S|}\right)^{\beta} \mathcal{H}_\beta(\omega_{C_j}) \\
&= 0, \\
\mathcal{H}_\beta(\pi|\alpha_S) &= \sum_{j=1}^{n} \frac{1}{|S|} \mathcal{H}(\pi_{\{x_j\}}) = 0
\end{aligned}$$

for every partition $\pi \in \mathrm{PART}(S)$.

For $\pi = \{B_1, \ldots, B_m\}$ and $\sigma = \{C_1, \ldots, C_n\}$, the conditional entropy can be written explicitly as

$$\begin{aligned}\mathcal{H}_\beta(\pi|\sigma) &= \sum_{j=1}^{n} \left(\frac{|C_j|}{|S|}\right)^{\beta} \sum_{i=1}^{m} \frac{1}{1-2^{1-\beta}} \left[1 - \left(\frac{|B_i \cap C_j|}{|C_j|}\right)^{\beta}\right] \\ &= \frac{1}{1-2^{1-\beta}} \sum_{j=1}^{n} \left(\left(\frac{|C_j|}{|S|}\right)^{\beta} - \sum_{i=1}^{m} \left(\frac{|B_i \cap C_j|}{|S|}\right)^{\beta}\right). \qquad (1)\end{aligned}$$

For the special case when $\pi = \alpha_S$, we can write

$$\mathcal{H}_\beta(\alpha_S|\sigma) = \sum_{j=1}^{n} \left(\frac{|C_j|}{|S|}\right)^{\beta} \mathcal{H}_\beta(\alpha_{C_j}) = \frac{1}{1-2^{1-\beta}} \left(\sum_{j=1}^{n} \left(\frac{|C_j|}{|S|}\right)^{\beta} - \frac{1}{|S|^{\beta-1}}\right). \qquad (2)$$

If σ, π are two partitions of S and $\pi \geqslant \sigma$, then σ is more informative than π regarding the elements of S. This intuition is captured by the following statement.

Theorem

Let S be a finite set and let $\pi, \sigma \in PART(S)$. We have $\mathcal{H}_\beta(\pi|\sigma) = \mathcal{H}_\beta(\pi)$ if and only if $\pi \geqslant \sigma$.

Proof

Suppose that $\sigma = \{C_1, \ldots, C_n\}$. If $\pi \geqslant \sigma$, each block of σ is included in a block of π and, therefore, we have $\pi_{C_j} = \omega_{C_j}$ for $1 \leqslant j \leqslant n$. Consequently, we have:

$$\mathcal{H}_\beta(\pi|\sigma) = \sum_{j=1}^{n} \left(\frac{|C_j|}{|S|}\right)^\beta \mathcal{H}_\beta(\omega_{C_j}) = 0.$$

Conversely, suppose that

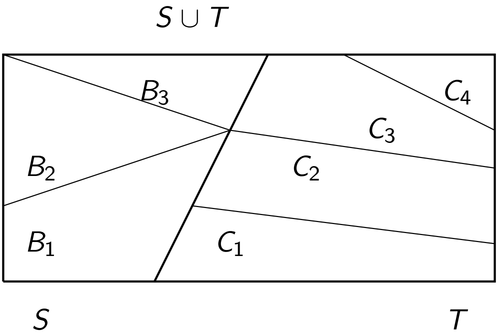
$$\mathcal{H}_\beta(\pi|\sigma) = \sum_{j=1}^{n} \left(\frac{|C_j|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{C_j}) = 0.$$

This implies $\mathcal{H}_\beta(\pi_{C_j}) = 0$ for $1 \leqslant j \leqslant n$, which means that $\pi_{C_j} = \omega_{C_j}$ for $1 \leqslant j \leqslant n$ by a previous remark. This means that every block C_j of σ is included in a block of π, so $\pi \geqslant \sigma$.

Definition

Let S, T be two disjoint sets and let $\sigma = \{B_1, \ldots, B_m\} \in \mathsf{PART}(S)$ and $\tau = \{C_1, \ldots, C_n\} \in \mathsf{PART}(T)$. The sum of the partitions σ and τ is the partition $\pi + \sigma$ of the set $S \cup T$ given by:

$$\pi + \sigma = \{B_1, \ldots, B_m, C_1, \ldots, C_n\}.$$



$$\pi = \{B_1, B_2, B_3\} \in \mathsf{PART}(S), \sigma = \{C_1, C_2, C_3, C_4\} \in \mathsf{PART}(T)$$
$$\pi + \sigma = \{B_1, B_2, B_3, C_1, C_2, C_3, C_4\} \in \mathsf{PART}(S \cup T).$$

Intersection of Two Partitions

Definition

Let now π, τ be two partitions in $\mathrm{PART}(S)$, where

$$\begin{aligned} \pi &= \{B_1, \ldots, B_m\}, \\ \tau &= \{D_1, \ldots, D_p\}, \end{aligned}$$

The partition $\pi \wedge \tau \in \mathrm{PART}(S)$ is

$$\pi \wedge \tau = \{B_i \cap D_j \mid B_i \cap D_j \neq \emptyset\}.$$

Example

Let $S = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$ and let

$$\begin{aligned} \pi &= \{\{x_1\}, \{x_2, x_3, x_4, x_5, x_6\}, \{x_7\}\} \\ \tau &= \{\{x_1, x_2, x_3\}, \{x_4, x_5, x_6, x_7\}\}. \end{aligned}$$

We have

$$\pi \wedge \tau = \{\{x_1\}, \{x_2, x_3\}, \{x_4, x_5, x_6\}, \{x_7\}\}.$$

The next statement is a generalization of a well-known property of Shannon's entropy.

Theorem

Let π and σ be two partitions of a finite set S. We have

$$\mathcal{H}_\beta(\pi \wedge \sigma) = \mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma) = \mathcal{H}_\beta(\sigma|\pi) + \mathcal{H}_\beta(\pi),$$

Proof

Let $\pi = \{B_1, \ldots, B_m\}$ and $\sigma = \{C_1, \ldots, C_n\}$ be two partitions of S. We have

$$\begin{aligned}
& \mathcal{H}_\beta(\pi \wedge \sigma) - \sum_{j=1}^{n} \left(\frac{|C_j|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{C_j}) \\
= & \frac{1}{1-2^{1-\beta}} \left(1 - \sum_i \sum_j \left(\frac{|B_i \cap C_j|}{|S|}\right)^\beta\right) \\
& - \frac{1}{1-2^{1-\beta}} \sum_j \left(\frac{|C_j|}{|S|}\right)^\beta \left(1 - \sum_i \left(\frac{|B_i \cap C_j|}{|C_j|}\right)^\beta\right) \\
= & \mathcal{H}_\beta(\sigma).
\end{aligned}$$

From the result established above

$$\mathcal{H}_\beta(\pi \wedge \sigma) = \sum_{j=1}^{n} \left(\frac{|C_j|}{|S|} \right)^\beta \mathcal{H}_\beta(\pi_{C_j}) + \mathcal{H}_\beta(\sigma),$$

we obtain

$$\mathcal{H}_\beta(\pi \wedge \sigma) = \mathcal{H}_\beta(\pi|\sigma) + \mathcal{H}_\beta(\sigma).$$

The second equality has a similar proof.

Definition

Let π, σ be two partitions of a set S, where $\sigma = \{C_1, \ldots, C_n\}$. The *$\beta$-information gain* of σ is the number

$$\mathsf{gain}_\beta(\pi, \sigma) = \mathcal{H}_\beta(\pi) - \mathcal{H}_\beta(\pi|\sigma).$$

Let $\pi \in \mathrm{PART}(S)$, and let $\sigma = \{C_1, \ldots, C_n\} \in \mathrm{PART}(S)$. If $\beta \to 1$ the information gain of the Shannon entropy is

$$\begin{aligned} \mathrm{gain}(\pi, \sigma) &= \mathcal{H}(\pi) - \mathcal{H}(\pi|\sigma) \\ &= \mathcal{H}(\pi) - \sum_{i=1}^{n} \frac{|C_j|}{|S|} \mathcal{H}(\pi_{C_j}). \end{aligned}$$

Note that $\pi \geqslant \sigma$, where $\pi, \sigma \in \mathsf{PART}(S)$, we have $\mathcal{H}_\beta(\pi|\sigma) = \mathcal{H}_\beta(\pi)$ if and only if

$$\mathsf{gain}_\beta(\pi, \sigma) = \mathcal{H}_\beta(\pi) - \mathcal{H}_\beta(\pi|\sigma) = 0,$$

by the Theorem on slide 21.

Each attribute A of a table partitions the rows of a table into blocks of rows that have equal values for that attribute. The partition that corresponds to A is denoted by π^A.

Example

The partitions of the form π^A of the Tennis table are: are

$$\begin{aligned}
\pi^{Outlook} &= \{\{1,2,8,9,11\},\{3,7,12\},\{4,5,6,10,13,14\}\},\\
\pi^{Temperature} &= \{\{1,2,3,13\},\{4,8,10,11,12,14\},\{5,6,7,9\}\},\\
\pi^{Humidity} &= \{\{1,2,3,4,8,12,14\},\{5,6,7,9,10,11,13\}\},\\
\pi^{Wind} &= \{\{1,3,4,5,8,9,10,13\},\{2,6,7,11,12,14\}\},\\
\pi^{PlayTennis} &= \{\{1,2,6,8,14\},\{3,4,5,7,9,10,11,12,13\}\}.
\end{aligned}$$

The PlayTennis attribute is the decision attribute.

To decide which attribute is the best classifier we use the information gain. Let T be a table having D as the decision attribute and let A be another attribute of T. The set of all tuples of T is denoted by $|S|$.
Let π^D be the partition determined by the attribute D. In our example,

$$\pi^{\mathrm{PlayTennis}} = \{\{1, 2, 6, 8, 14\}, \{3, 4, 5, 7, 9, 10, 11, 12, 13\}\}.$$

Suppose that $\mathrm{Dom}(A) = \{a_1, \ldots, a_k\}$. Define $S_{A=a_i}$ as the set of tuples whose A-component equals a_i:

$$S_{A=a_i} = \{t \mid t[A] = a_i\}.$$

The entropy of the decision attribute is $\mathcal{H}(\pi^D)$.

The *information gain of D relative to an attribute A*, where $\mathrm{Dom}(A) = \{a_1, \ldots, a_k\}$ is:

$$\mathrm{gain}(D, A) = \mathcal{H}(\pi^D) - \sum_{i=1}^{k} \frac{|S_{A=a_i}|}{|S|} \mathcal{H}((\pi^D)_{S_{A=a_i}}).$$

Using again the PlayTennis example

	Outlook	Temperature	Humidity	Wind	PlayTennis
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rain	mild	high	weak	yes
5	rain	cool	normal	weak	yes
6	rain	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rain	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	rain	hot	normal	weak	yes
14	rain	mild	high	strong	no

Example

For

$$\pi^{\text{PlayTennis}} = \{\{1,2,6,8,14\},\{3,4,5,7,9,10,11,12,13\}\}.$$

we have:

$$\begin{aligned}(\pi^{\text{PlayTennis}})_{S_{\text{Wind=weak}}} &= \{\{1,8\},\{3,4,5,9,10,13\}\}\\ (\pi^{\text{PlayTennis}})_{S_{\text{Wind=strong}}} &= \{\{2,6,14\},\{7,11,12\}\},\end{aligned}$$

hence

$$\begin{aligned}&\text{gain}(\text{PlayTennis},\text{Wind})\\ &= \mathcal{H}(\pi^{\text{PlayTennis}}) - \frac{8}{14}\mathcal{H}\left((\pi^{\text{PlayTennis}})_{S_{\text{Wind=weak}}}\right)\\ &\quad -\frac{6}{14}\mathcal{H}\left((\pi^{\text{PlayTennis}})_{S_{\text{Wind=strong}}}\right)\\ &= 0.940 - \frac{8}{14}\cdot 0.811 - \frac{6}{14}\cdot 1.00 = 0.048.\end{aligned}$$

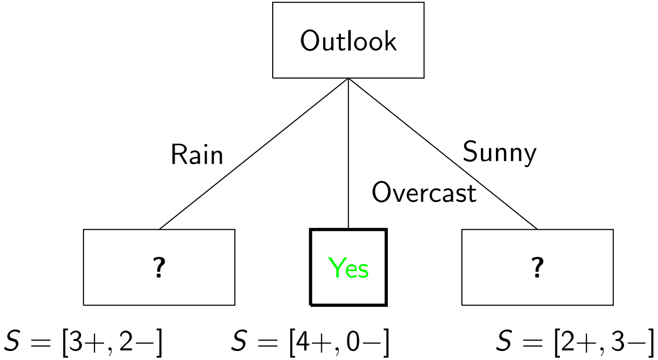
We use information gain to select the best attribute in each step in expanding the tree.
Denote by $[p+, q-]$ the composition of a set that contains p positive cases (PlayTennis is YES) and q negative cases (PlayTennis is NO).

The information gains are:

$$\begin{aligned}&\text{gain}(\text{PlayTennis}, \text{Outlook}) = 0.246;\\&\text{gain}(\text{PlayTennis}, \text{Humidity}) = 0.151;\\&\text{gain}(\text{PlayTennis}, \text{Wind}) = 0.048;\\&\text{gain}(\text{PlayTennis}, \text{Tenerature}) = 0.029;\end{aligned}$$

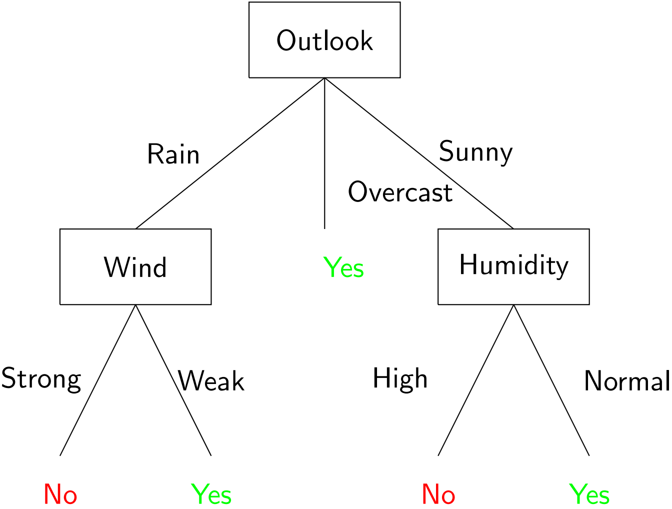
- The Outlook provides the largest information gain, and so is the best predictor for PlayTennis.
- Outlook is selected as the decision attribute and three branches are created corresponding to the values sunny, overcast, and rain.

In the partial decision tree the Overcast node has only positive examples, so it is a leaf.



gain(Sunny, Humidity) $= 0.970$;
gain(Sunny, Temperature) $= 0.570$;
gain(Sunny, Wind) $= 0.019$;

The final decision tree:



We begin by introducing data as a tibble, by loading the package `tibble`, and using the function `tribble`:

```
> d <- tribble(
~Outlook,  ~Temperature, ~Humidity, ~Wind, ~PlayTennis,
"sunny"    , "hot"  , "high"   , "weak"   , "no",
"sunny"    , "hot"  , "high"   , "strong" , "no",
"overcast" , "hot"  , "high"   , "weak"   , "yes",
"rain"     , "mild" , "high"   , "weak"   , "yes",
"rain"     , "cool" , "normal" , "weak"   , "yes",
"rain"     , "cool" , "normal" , "strong" , "no",
"overcast" , "cool" , "normal" , "strong" , "yes",
"sunny"    , "mild" , "high"   ,"weak"   , "no",
"sunny"    , "cool" , "normal" ,"weak"   , "yes",
"rain"     , "mild" , "normal" ,"weak"   , "yes",
"sunny"    , "mild" , "normal" ,"strong" , "yes",
"overcast" , "mild" , "high"   ,"strong" , "yes",
"rain"     , "hot"  , "normal" ,"weak"   , "yes",
"rain"     , "mild" , "high"   , "strong" , "no"
)
```

The tibble d is converted to a data frame using

```
> d1 <- data.frame(d)
```

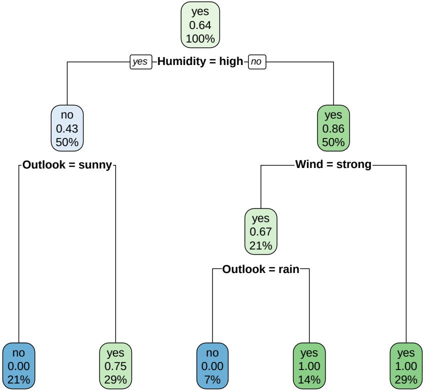
Then, we need to load the packages rpart and rpart.plot

```
> library(rpart)
> library(rpart.plot)
```

Finally, we obtain the decision tree using

```
> fit <- rpart(PlayTennis ~ Outlook + Wind + Humidity,
   + data=d1,
   + control=rpart.control(minsplit=3))
> rpart.plot(fit,type=2)
```

The decision tree can be saved in the directory of your choice as, say, a pdf file.



Note that the use of the `rpart` function may result in a decision tree different from the one we constructed by hand.

- The parameter `control` that we used in the call of the `rpart` function prescribes the minimum size of the set of tuples in a node that may be split; in our case, we will not split nodes that contain fewer than 3 records;
- The order of node splitting may be different; the function `rpart` starts with the node `Humidity` rather than `Outlook`; both provide the same gain.