Data Mining Tutorial

Mark A. Austin

University of Maryland

austin@umd.edu

ENCE 688P, Fall Semester 2021

October 16, 2021

Overview

Part 03

Information Gain in Decision Trees

Mathematical Framework

Information Gain

The amount of information that is gained by knowing the value of an attribute. It equals the entropy of a distribution before a split minus the entropy of a distribution after a split.

$$IG(Y, X) = H(Y) - H(Y|X). \tag{17}$$

Here:

- Information gain $IG(X, Y)$ is the reduction of uncertainty about Y given an additional piece of information X about Y.
- $H(Y)$ is the entropy of Y (before split).
- $H(Y|X)$ is the conditional entropy of Y given the value of attribute X (after split).

Decision Trees

Design of Data Partitions for Classification Tree:

- Use information gain as measure for attribute selection.
- Pick atttribute split that maximizes information gain IG(Y,X), i.e.,

$$IG(D, A) = H(D) - \sum_{i=1}^{v} \frac{D_j}{D} H(D_j) \tag{18}$$

Here:

- D is a prescribed data partition and A is an attribute.
- Split D into v partitions (or subsets) $\{D_1, D_2, \cdots, D_j\}$, where D_j contains those tuples in D that have outcome a_j of A.

Decision Trees

Basic Algorithm (This is a greedy algorithm) ...

- Decision tree is constructed in a top-down recursive divide-and-conquer manner.
- When the construction process begins, all training examples are at the root.
- Attributes are categorical – if continuous-valued they are discretized in advance.
- Need to design a sequence of selected attributes to partition dataset recursively.
- Test attributes are selected on basic of heuristic or statistical measure (e.g., information gain).

Conditions for Stopping Partitioning

- All samples for a given node belong to the same class.

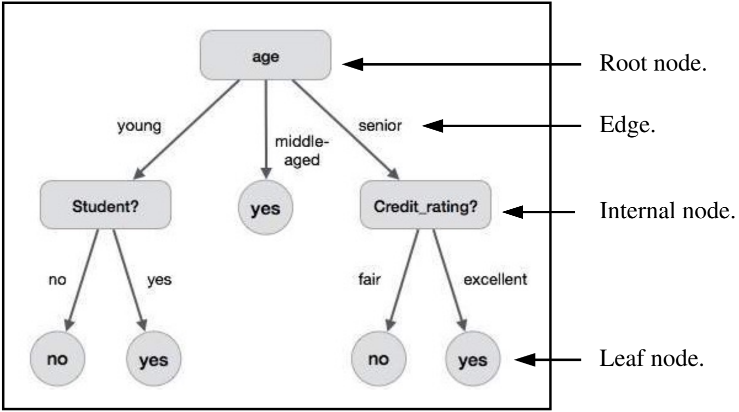
Example 1 (Buy Computer)

Initial Dataset. Will customer buy a computer?

ID	Age Group	Income	Student	Credit Rating	Buys Computer
1	young	high	no	fair	no
2	young	high	no	excellent	no
3	middle	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle	low	yes	excellent	yes
8	young	medium	no	fair	no
9	young	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	young	medium	yes	excellent	yes
12	middle	medium	no	excellent	yes
13	middle	high	yes	fair	yes
14	senior	medium	no	excellent	no

Example 1 (Buy Computer)

Sample Decision Tree: Initially split data based on age group.



Is this a good decision?

Example 1 (Buy Computer)

Entropy of Base Dataset

Purchase outcomes: $\{no = 5, yes = 9\}$.

$$H(D) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.94 \qquad (19)$$

Partitioned Dataset. Split dataset by age ...

ID	Age Group	Income	Student	Credit Rating	Buys Computer
1	young	high	no	fair	no
2	young	high	no	excellent	no
8	young	medium	no	fair	no
9	young	low	yes	fair	yes
11	young	medium	yes	excellent	yes

Example 1 (Buy Computer)

Partitioned Dataset. Split dataset by age ...

ID	Age Group	Income	Student	Credit Rating	Buys Computer
3	middle	high	no	fair	yes
7	middle	low	yes	excellent	yes
12	middle	medium	no	excellent	yes
13	middle	high	yes	fair	yes

ID	Age Group	Income	Student	Credit Rating	Buys Computer
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
10	senior	medium	yes	fair	yes
14	senior	medium	no	excellent	no

Example 1 (Buy Computer)

Entropy of Partitioned Dataset. Split by age group ...

$$
\begin{aligned}
IG(D, Age) &= H(D) - \sum_{v \in \{you, mid, sen\}} \frac{S_v}{S} H(S_v) \\
&= H(D) - \frac{5}{14} H(S_{you}) - \frac{4}{14} H(S_{mid}) - \frac{5}{14} H(S_{sen}) \\
&= 0.246.
\end{aligned} \tag{20}
$$

Remaining split options:

- IG(D,Income) $= 0.029$,
- IG(D,Student) $= 0.151$,
- IG(D,Credit Rating) $= 0.048$.

Example 1 (Buy Computer)

Conclusion

- Attribute age has the highest information gain and therefore becomes the splitting attribute at the root node.

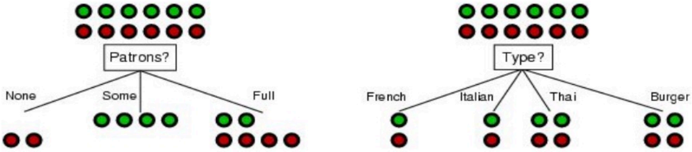
Actions

- Branches are grown for each outcome of age.
- Repeat process on lower-level nodes using split attributes of student and credit rating.

Example 2 (Customer Wait for Table at Restaurant?)

Customer Dataset (Source: Russell and Norvig, 2010)

Example	Attributes										Target
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*Wait*
X_1	T	F	F	T	Some	$$$	F	T	French	0–10	T
X_2	T	F	F	T	Full	$	F	F	Thai	30–60	F
X_3	F	T	F	F	Some	$	F	F	Burger	0–10	T
X_4	T	F	T	T	Full	$	F	F	Thai	10–30	T
X_5	T	F	T	F	Full	$$$	F	T	French	>60	F
X_6	F	T	F	T	Some	$$	T	T	Italian	0–10	T
X_7	F	T	F	F	None	$	T	F	Burger	0–10	F
X_8	F	F	F	T	Some	$$	T	T	Thai	0–10	T
X_9	F	T	T	F	Full	$	T	F	Burger	>60	F
X_{10}	T	T	T	T	Full	$$$	F	T	Italian	10–30	F
X_{11}	F	F	F	F	None	$	F	F	Thai	0–10	F
X_{12}	T	T	T	T	Full	$	F	F	Burger	30–60	T



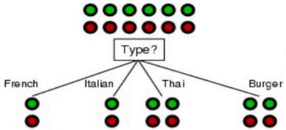
Example 2 (Customer Wait for Table at Restaurant?)

Dataset Attributes

- **Alternate:** Is there a suitable alternate restaurant nearby?
- **Bar:** Does restaurant have comfortable bar area to wait in?
- **Fri/Sat:** True on Fridays and Saturdays.
- **Hungry:** True when customer is hungry.
- **Patrons:** How many people are in the restaurant? (none, some, and full).
- **Price:** The restaurant price range ($, $$ and $$$).
- **Raining:** Is it raining outside?
- **Reservation:** Did customer make a reservation?
- **Type:** Type of restaurant (French, Italian, Thai, or Burger).
- **WaitEstimate:** Wait time estimated by host (0-10 mins, 10-30, 30-60, or > 60).

Example 2 (Customer Wait for Table at Restaurant?)

Split on Restaurant Type Attribute

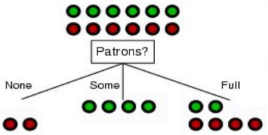


Root entropy: 0.5 * log(2) + 0.5 * log(2) = 1 bit

Leaf entropies: 2/12 * 1 + 2/12 * 1 + … = 1 bit

No reduction!

Split on Patrons Attribute



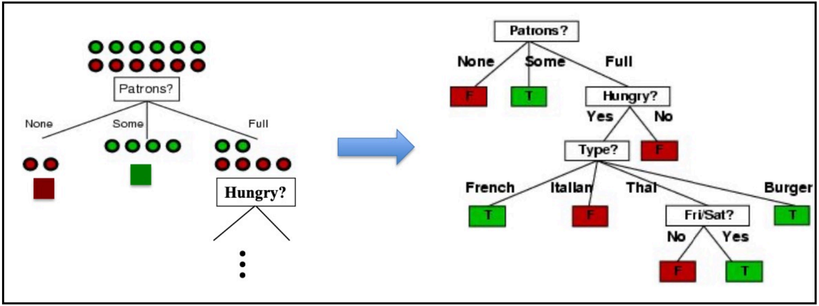
Root entropy: 0.5 * log(2) + 0.5 * log(2) = 1 bit

Leaf entropies: 2/12 * 0 + 4/12 * 0 + 6/12 * 0.9

Lower entropy after split!

Example 2 (Customer Wait for Table at Restaurant?)

Decision Tree Synthesis



Classification with Decision Trees (Summary)

Advantages

- Decision trees are simple to understand and interpret.
- Requires only a small number of observations.
- Best and expected values can be determined for different scenarios.

Disadvantages

- Difficulties in handling data with missing values.
- Information gain criterion is biased in favor of attributes with more levels.
- Calculations become complex if values are uncertain or outcomes are linked.

References

- Jaynes E.T., Information Theory and Statistical Mechanics. II, Phys. Rev. 108, 171, October 1957.
- Kapur J.N., Maximum-Entropy Models in Science and Engineering, John Wiley and Sons, 1989.
- Mitchell T.M., Machine Learning and Data Mining, Communications of the ACM, Vol. 42., No. 11, November 1999.
- Russell S., and Norvig P., Artificial Intelligence: A Modern Approach (Third Edition), Prentice-Hall, 2010.
- Shanon C.E., and Weaver W., The Mathematical Theory of Communication, University of Illinois, Urbana, Chicago, 1949.
- Witten I.H., Frank E., Hall M.A., and Pal C.J., Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2017.