Avisek Gupta (avisek003@gmail.com)

Machine Learning

21 – Decision Trees

November 29, 2022

Measuring 'Information'

How can we quantify how **informative** an event is?

Consider a random experiment of a coin toss. The Sample Space is $\{H, T\}$.

Let $P(H) = 1$ (and $P(T) = 0$).

Is the outcome of an individual event of a coin toss **predictable**?

Measuring 'Information'

How can we quantify how **informative** an event is?

Consider a random experiment of a coin toss. The Sample Space is $\{H, T\}$.

Let $P(H) = 1$ (and $P(T) = 0$).

Is the outcome of an individual event of a coin toss **predictable**?

Possible Outcomes of n events: $H, H, H, H, H, H, ...$

Consider the same random experiment, with $P(H) = 0.5$ (and $P(T) = 0.5$).

Is the outcome of an individual event of a coin toss **less predictable** or **more predictable**?

Measuring 'Information'

How can we quantify how **informative** an event is?

Consider a random experiment of a coin toss. The Sample Space is $\{H, T\}$.

Let $P(H) = 1$ (and $P(T) = 0$).

Is the outcome of an individual event of a coin toss **predictable**?

Possible Outcomes of n events: $H, H, H, H, H, H, ...$

Consider the same random experiment, with $P(H) = 0.5$ (and $P(T) = 0.5$).

Is the outcome of an individual event of a coin toss **less predictable** or **more predictable**?

Possible Outcomes of n events: $T, H, T, T, H, T, ...$

Measuring Information

How can we quantify how **informative** an event is?

- More predictable events are less informative
  - For the coin toss event, $P(H) = 1, P(T) = 0$ is more predictable, hence less informative.
- Less predictable events are more informative
  - For the coin toss event E, $P(H) = P(T) = 0.5$ is less predictable, hence more informative.

Measuring Information: Entropy

Entropy: Let a discrete random variable X be defined to take values from $\mathcal{X}$, and X has a distribution described by $p : \mathcal{X} \to [0, 1]$, so that $p(x) = P[X = x]$.

Then the entropy of X, denoted as $H(X)$, is defined as,

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_b p(x).$$

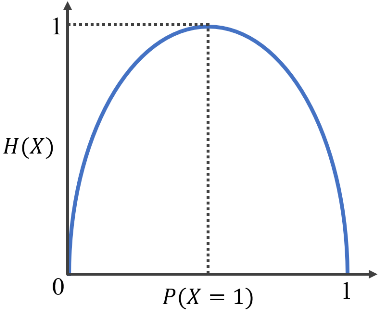
Measuring Information: Entropy

Entropy: Let a discrete random variable X be defined to take values from $\mathcal{X}$, and X has a distribution described by $p : \mathcal{X} \to [0, 1]$, so that $p(x) = P[X = x]$.

Then the entropy of X, denoted as $H(X)$, is defined as,

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_b p(x).$$

For a coin toss:



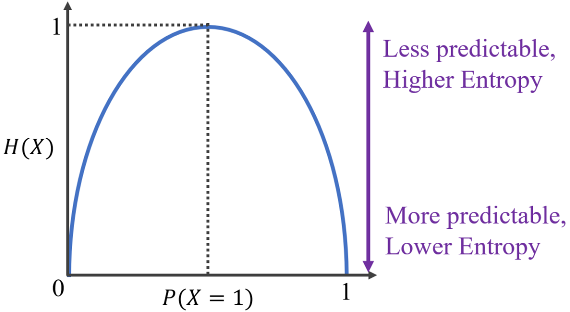
Measuring Information: Entropy

Entropy: Let a discrete random variable X be defined to take values from $\mathcal{X}$, and X has a distribution described by $p : \mathcal{X} \to [0, 1]$, so that $p(x) = P[X = x]$.

Then the entropy of X, denoted as $H(X)$, is defined as,

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_b p(x).$$

For a coin toss:



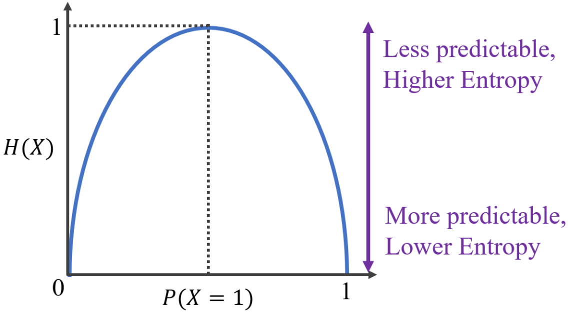
Measuring Information: Entropy

Entropy: Let a discrete random variable X be defined to take values from $\mathcal{X}$, and X has a distribution described by $p : \mathcal{X} \to [0, 1]$, so that $p(x) = P[X = x]$.

Then the entropy of X, denoted as $H(X)$, is defined as,

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_b p(x).$$

For a coin toss:



For $\mathrm{P}(X = 1) = \frac{1}{2}$,

$$H(X) = -\left[\frac{1}{2} log_2 \frac{1}{2} + \frac{1}{2} log_2 \frac{1}{2}\right] = 1$$

For $\mathrm{P}(X = 1) = 1$,

$$H(X) = -[1 log_2 1 + 0 log_2 0] = 0$$

Measuring Information: Entropy

Entropy: Let a discrete random variable X be defined to take values from $\mathcal{X}$, and X has a distribution described by $p : \mathcal{X} \to [0, 1]$, so that $p(x) = P[X = x]$.

Then the entropy of X, denoted as $H(X)$, is defined as,

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_b p(x).$$

If number of outcomes is larger, then maximum possible entropy is higher.

For a coin toss:

$$\text{Maximum possible entropy: } -\sum^{2} \left(\frac{1}{2} \log_2 \frac{1}{2} \right) = 1.$$

For a dice roll:

$$\text{Maximum possible entropy: } -\sum^{6} \left(\frac{1}{6} \log_2 \frac{1}{6} \right) = 2.585.$$

Entropy

- Proposed by Clause Shannon in 1948.
  Shannon, Claude E. (July 1948). "A Mathematical Theory of Communication". Bell System Technical Journal. 27 (3): 379-423.
- Played a vital role in the development of Information Theory and Coding Theory, to develop effective methods for compression and communication of information.

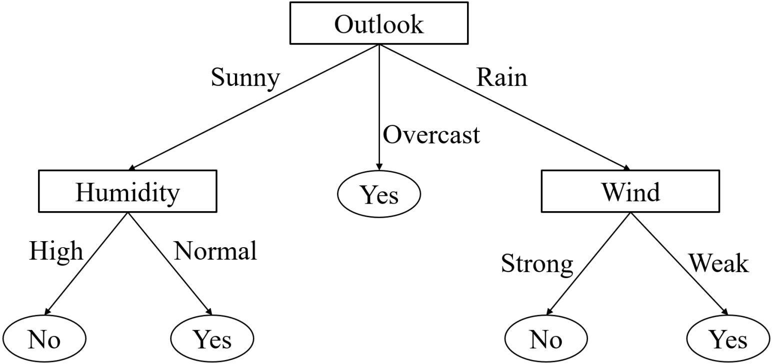
Decision Trees

Consider the following learning problem: given the variables **Outlook**, **Temperature**, **Humidity**, and **Wind**, can one learn to predict whether the weather is suitable to play the game of Tennis (target variable: **PlayTennis**)?

Day	**Outlook**	**Temperature**	**Humidity**	**Wind**	**PlayTennis**
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Decision Trees

We wish to learn a tree that predicts **PlayTennis**. The nodes of the tree will be any of the possible features, different values of the feature can lead us to either (1) a decision, or (2) other features that will lead us to a decision.



The tree **decides** the target variable, and also **shows why** it reached its decision.

Decision Trees

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Let S be the collection of positive and negative examples. Let p_+ be the proportion of positive examples, and let p_- be the proportion of negative examples.

$H(S) = -(9/14)\log_2(9/14) - (5/14)\log_2(5/14) = 0.94.$

Decision Trees

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Decision Tree learning algorithm (ID3 algorithm): At the current node, select an attribute that leads to largest reduction in entropy of the data.

Information Gain (IG): Measure of reduction in entropy on selecting attribute A:

$$IG(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v)$$

Decision Trees

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$Values(Wind) = \{Weak, Strong\}$; $S = [9+, 5-]$; $S_{weak} = [6+, 2-]$;, $S_{strong} = [3+, 3-]$

$$IG(S, Wind) = H(S) - \sum_{v \in \{Weak, Strong\}} \frac{|S_v|}{|S|} H(S_v)$$

$$= 0.94 - (8/14)H(S_{weak}) - (6/14)H(S_{strong}) = 0.048$$

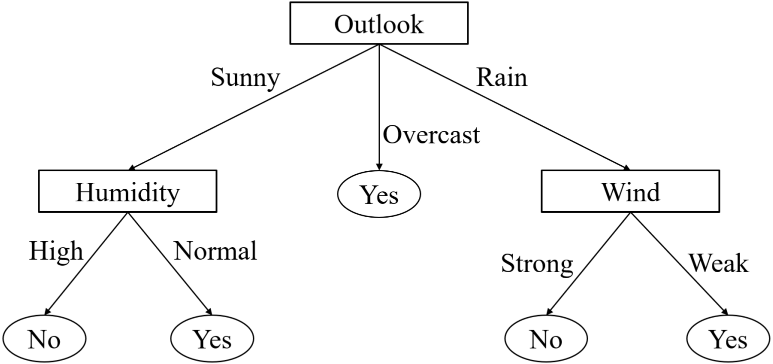
Decision Trees

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$IG(S, Outlook) = 0.246$; $IG(S, Humidity) = 0.151$; $IG(S, Wind) = 0.048$; $IG(S, Temperature) = 0.029$.

Feature that leads to largest Information Gain? - *Outlook*

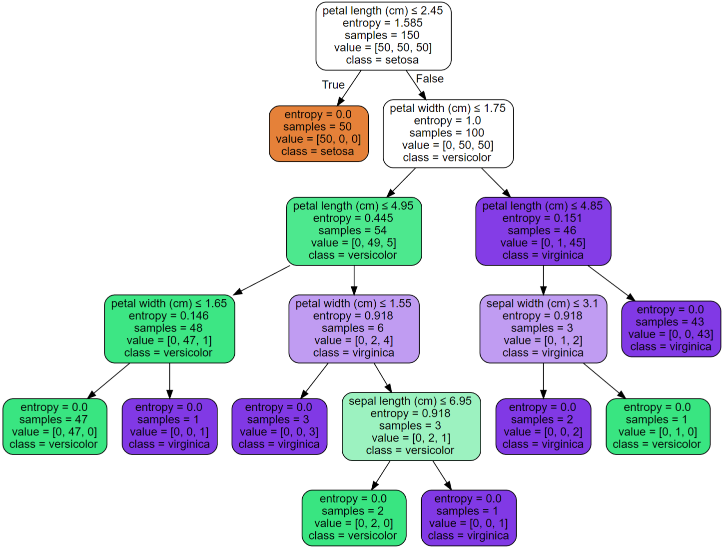
Decision Trees



Decision Tree for the data that predicts **PlayTennis**.

Decision Trees

Continuous features are handled in terms of thresholds.



Decision Trees

- Decision Trees are also prone to overfitting data. Reduced-error Pruning - Prune the tree based on a validation data.
- A number of alternate measures exist that can be used instead of Information Gain (e.g., Split Information, Gain Ratio, Ginni Index, ...).

References

[1] Tom Mitchell, Chapter 3 - Decision Tree Learning, Machine Learning. McGraw Hill, 1997.

[2] Quinlan, J.R. Induction of decision trees. Machine Learning, 1, 81-106 (1986).