

Tutorial 9:

Current Best Learning and Decision Trees

Name	Student no.

For this tutorial, you can work in groups of 1 or 2. Submit the answers to each of the 2 Questions

Question 1

This problem asks you to search a *hypothesis space*. Assume that we have four examples (see below) where attributes take values as indicated in the header. The target attribute/concept is *EnjoySport*.

	Sky (Sunny, Cloudy, Rain)	AirTemp (Warm, Cold)	Humidity (Normal, High)	Wind (Weak, Strong)	Water (Warm, Cool)	Forecast (Same, Change)	EnjoySport (Yes, No)
1	Sunny	Warm	Normal	Strong	Warm	Same	No
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rain	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

A hypothesis regarding what it means to *EnjoySport* is represented as a pattern $\langle v_1, v_2, v_3, v_4, v_5, v_6 \rangle \leftrightarrow \mathbf{Yes}$, where v_1 can take the values *Sunny*, *Cloudy* or *Rain* (according to the first attribute *Sky*) or a wildcard value * (matching all possible values), v_2 can take the values of *AirTemp* plus *, and so forth for all the attributes. All examples matching the pattern are classified as *Yes* (i.e. part of the concept *EnjoySport*), those not matching the pattern are classified as *No*. In addition there is one special pattern $\langle \mathbf{nil} \rangle \leftrightarrow \mathbf{Yes}$ which classifies all examples as *No*.

- a) What size is the hypothesis space? (ie: how many possible hypotheses are there, following the rules above).
- b) Rank the following hypotheses according to *EnjoySport=Yes* specificity (1-most specific, 5-least specific):

A= $\langle *, *, \mathbf{High}, *, *, * \rangle \leftrightarrow \mathbf{Yes}$.

B= $\langle *, *, \mathbf{Normal}, \mathbf{Weak}, *, \mathbf{Change} \rangle \leftrightarrow \mathbf{Yes}$.

C= $\langle \mathbf{Rain}, \mathbf{Cold}, \mathbf{High}, \mathbf{Strong}, \mathbf{Warm}, \mathbf{Change} \rangle \leftrightarrow \mathbf{Yes}$.

D= $\langle \mathbf{nil} \rangle \leftrightarrow \mathbf{Yes}$.

E= $\langle *, \mathbf{Cold}, \mathbf{Normal}, \mathbf{Strong}, \mathbf{Cool}, \mathbf{Same} \rangle \leftrightarrow \mathbf{Yes}$.

- c) Using Current-best-hypothesis learning, track the currently best (most specific) hypothesis, going through examples 2-4, starting with

H1 = < *, **Cold**, **High**, *, *, * > ↔ Yes (consistent with example 1)

Question 2

Ernie's entertainment park has a merry-go-round that makes some people sick. Ernie has recently collected data to help resolve which attributes cause this condition.

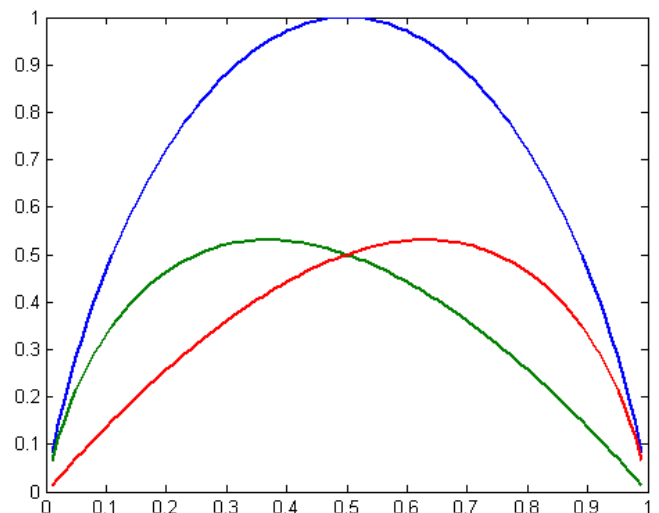
	Height	Weight	Age	Gender	Sick?
1	Tall	Fat	Young	Female	Yes
2	Tall	Thin	Middleage	Male	Yes
3	Short	Medium	Old	Male	No
4	Medium	Medium	Old	Female	No
5	Medium	Fat	Young	Male	No
6	Tall	Thin	Young	Female	Yes
7	Short	Medium	Middleage	Male	No
8	Medium	Fat	Young	Female	No
9	Tall	Thin	Old	Female	Yes
10	Tall	Thin	Young	Female	Yes
11	Short	Medium	Middleage	Female	No
12	Tall	Medium	Young	Male	Yes
13	Tall	Fat	Young	Female	Yes
14	Short	Thin	Old	Male	No
15	Medium	Thin	Old	Female	Yes
16	Tall	Fat	Young	Female	Yes
17	Tall	Thin	Middleage	Male	Yes
18	Short	Thin	Young	Male	No
19	Medium	Fat	Old	Female	No
20	Tall	Thin	Young	Male	Yes

$$Gain(A) = I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) - Remainder(A)$$

$$Remainder(A) = \sum_{i=1}^v \frac{p_i + n_i}{p+n} I\left(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i}\right)$$

$$I(P(v_1), \dots, P(v_m)) = \sum_{i=1}^m -P(v_i) \log_2 P(v_i)$$

$$I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n}$$



a) Which attribute has the greatest information gain with respect to the classification of motion sickness?

b) What is the greatest information gain?

c) Create a full ID3 decision tree.

d) Do you think that this decision tree would be sufficient for Ernie to warn nearly everyone at risk? Why / why not?