

Exercise11:

The goal is to decide if someone buys a computer or not. Derive the best decision tree by calculating a little by hand (Shannon). At least the first split.

Calculating the Entropy and Gain for Decision Tree

Buy Computer

Buy Computer		
yes	no	sum
12	8	20

Age

	Buy computer	Buy computer	
	yes	no	sum
<30	2	6	8
31...40	6	0	6
>40	4	2	6

Income

	Buy computer	Buy computer	
	yes	no	sum
high	3	2	5
medium	5	3	8
low	4	3	7

Student

	Buy computer	Buy computer	
	yes	no	sum
yes	8	1	9
no	4	7	11

Credit rating

	Buy computer	Buy computer	
	yes	no	sum
Fair	7	3	10
Excellent	5	5	10

Calculated with ID3 method from http://www.saedsayad.com/decision_tree.htm

Entropy Buy Computer

$$\begin{aligned}E(\text{BuyComputer}) &= E(12,8) \\ &= 0.971\end{aligned}$$

Entropy(BuyComputer, Age)

$$\begin{aligned}E(\text{BuyComputer, Age}) &= P(<30)*E(2,6) + P(31..40)*E(6,0) + P(>40)*E(4,2) \\ &= (8/20)*0.811 + (6/20)*0 + (6/20)*0.918 \\ &= 0.6\end{aligned}$$

Entropy(BuyComputer, Income)

$$\begin{aligned}E(\text{BuyComputer, Income}) &= P(\text{high})*E(3,2) + P(\text{medium})*E(5,3) + P(\text{low})*E(4,3) \\ &= (5/20)*E(3,2) + (8/20)*E(5,3) + (7/20)*E(4,3) \\ &= 0.25*0.971 + 0.4*0.954 + 0.35*0.9855 \\ &= 0.96\end{aligned}$$

Entropy(BuyComputer, Student)

$$\begin{aligned}E(\text{BuyComputer, Student}) &= P(\text{IsStudent})*E(8,1) + P(\text{noStudent})*E(4,7) \\ &= (9/20)*E(8,1) + (11/20)*E(4,7) \\ &= 0.45*0.5044 + 0.55*0.9457 \\ &= 0.747\end{aligned}$$

Entropy(BuyComputer, CreditRating)

$$E(\text{BuyComputer, CreditRating}) = P(\text{Fair})*E(7,3) + P(\text{Excellent})*E(5,5)$$

$$= (10/20)*E(7,3) + 10/20*E(5,5)$$

$$= 0.5*0.881 + 0.5*1$$

$$= 0.941$$

Calculating the GAINS

$$G(\text{BuyComputer}, \text{Age}) = E(\text{BuyComputer}) - E(\text{BuyComputer}, \text{Age})$$

$$= 0.971 - 0.6$$

$$= 0,371$$

$$G(\text{BuyComputer}, \text{Income}) = E(\text{BuyComputer}) - E(\text{BuyComputer}, \text{Income})$$

$$= 0.971 - 0.96$$

$$= 0,011$$

$$G(\text{BuyComputer}, \text{Student}) = E(\text{BuyComputer}) - E(\text{BuyComputer}, \text{Student})$$

$$= 0.971 - 0.747$$

$$= 0,224$$

$$G(\text{BuyComputer}, \text{Creditrating}) = E(\text{BuyComputer}) - E(\text{BuyComputer}, \text{Creditrating})$$

$$= 0.971 - 0.941$$

$$= 0,03$$

Resume

So the most important impact is the person's age, followed by is a student or not.

The most unimportant property in this example is the person's income, **a surprising result for me.**

Entropy (Age 31..40) is 0, therefore group Age 31..40 is a leaf node, means every person from this group buys a computer

So our decision tree should start with age, followed by property 'is student', then person's creditrating and at least person's income.

→→→

You'll find a Jupyter notebook at GitHub:

<https://github.com/mahlsvede/Exercises/blob/master/Exercise11DET.ipynb>