Leren homework # 5

Date: December 6, 2014

Name (Student Number):
T.M. Meijers (10647023)

Question 1

Before adding any nodes, let us calculate the entropy:
$E(S) = -(\frac{4}{6} \cdot log(\frac{4}{6}) + \frac{2}{6} \cdot log(\frac{2}{6})) = 0.9183$

Adding node $A_1\{X_1 > 4.5 = q, X_1 < 4.5 = p\}$
looks promising, let us now calcualte the new entropy:
$E(A_1) = -\frac{3}{6}(\frac{1}{3} \cdot log(\frac{1}{3}) + \frac{2}{3} \cdot log(\frac{2}{3})) - \frac{3}{6}(1 \cdot log(1)) = 0.4591$
Our gain is: $0.9183 - 0.4591 = 0.4592$

To check if this is better than adding $B_1\{X_2 > 2.5 = p, X_2 < 2.5 = q\}$:
$E(B_1) = -\frac{5}{6}(\frac{1}{5} \cdot log(\frac{1}{5}) + \frac{4}{5} \cdot log(\frac{4}{5})) - \frac{1}{6}(1 \cdot log(1)) = 0.6016$
The gain would be lower, so let's add A_1 as our first node.

We then have one wrong prediction in our set of predicted $q's$, so let's add:
$A_2\{X_2 > 6.5 = p, X_2 < 6.5 = q\}$
$E(A_2) = 0$, since we have no wrong predictions anymore. For this problem (unrealistic dataset ofcourse) adding a second node seems like overfitting, but it is the assignment.

The decision tree (Figure 1) then gives us the result ID 7 $= q$.

Question 2

For given chances see Mitchel (ch. 6, section 6.2.1).
To find/calculate $P(cancer|2*\bigoplus)$:

$$\begin{aligned}
P(cancer|\bigoplus) &= \frac{0.0078}{0.0078+0.0298} = 0.21 \\
P(\bigoplus|cancer)P(cancer|\bigoplus) &= (0.98)\cdot 0.21 = 0.2058 \\
P(\bigoplus|\neg cancer)P(\neg cancer|\bigoplus) &= (0.03)\cdot 0.79 = 0.0.0237 \\
P(cancer|2*\bigoplus) &= \frac{P(\bigoplus|cancer)P(cancer|\bigoplus)}{P(\bigoplus|cancer)P(cancer|\bigoplus)+P(\bigoplus|\neg cancer)P(\neg cancer|\bigoplus)} \\
&= \frac{0.2058}{0.2058+0.0237} = 0.8967 \\
P(cancer|2*\bigoplus) &= 0.8967
\end{aligned}$$

Question 3

(a) $P(prijs|gegevens) = P(prijs)\prod P(gegevens|prijs)$

(b) Very vague question. The TA's did not have a good answer for what answer was required.

Figures

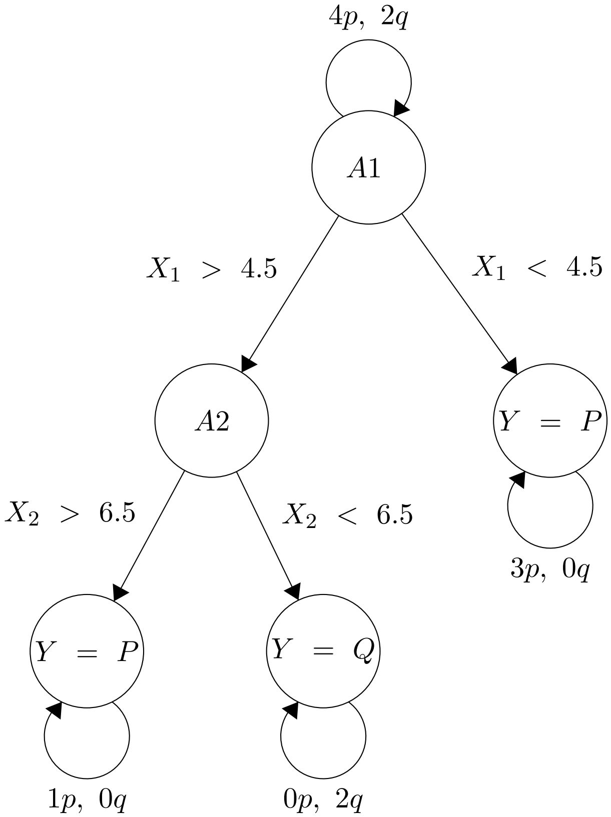

Figure 1: Q1 - decision tree