Homework #2: Solution

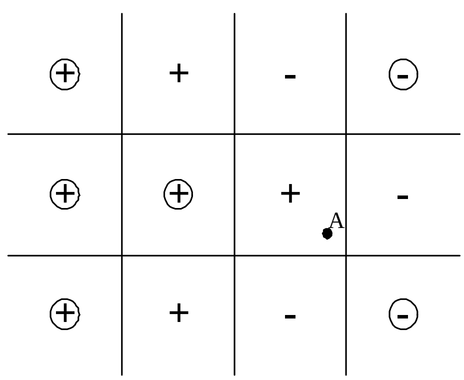
1. A and C are independent given B, so

$$
\begin{aligned}
P(A|B,\neg C) &= P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|\neg A)P(\neg A)} \\
&= \frac{0.8*0.5}{0.8*0.5+0.2*0.5} = 0.8
\end{aligned}
$$

Another way to solve the problem is,

$$
\begin{aligned}
P(A|B,\neg C) &= \frac{P(A,B,\neg C)}{P(B,\neg C)} \\
&= \frac{P(A)P(B|A)P(\neg C|B)}{P(A,B,\neg C)+P(\neg A,B,\neg C)} \\
&= \frac{P(A)P(B|A)P(\neg C|B)}{P(A)P(B|A)P(\neg C|B)+P(\neg A)P(B|\neg A)P(\neg C|B)} \\
&= \frac{P(A)P(B|A)}{P(A)P(B|A)+P(\neg A)P(B|\neg A)} \\
&= \frac{P(A)P(B|A)}{P(B)} \\
&= P(A|B)
\end{aligned}
$$

2.(a)(c)



(b) The nearest-neighbor algorithm will predict for A a positive example, the 3-nearest-neighbor algorithm will predict it a negative example.

3. The information gain of any attribute can be computed by:

$$Gain(S,A) = H(S) - \sum_{v} P(A=v) \cdot H(S|A=v)$$

In perfect case, the attribute A can successfully classify the training set into class Red, Green, and Blue, where $H(S|A=v)$ in the above equation is equal to 0 for any v value. Thus, the maximum possible information gain of attribute A is,

$$\begin{aligned} max(Gain(S,A)) &= H(S) \\ &= \sum_{v \in \{Red, Green, Blue\}} -P(H=v) \lg P(H=v) \\ &= 1.5 \end{aligned}$$

4. Step1: calculating the decision tree accuracy over the validation set of subtrees rooted at W, X, and Y respectively:

$$\begin{aligned} Au(T_W) &= Au(A) \cdot P(W=w_1) + Au(D) \cdot P(W=w_2) = 0.4 \\ Au(T_X) &= Au(T_W) \cdot P(X=x_1) + Au(C) \cdot P(X=x_2) = 0.35 \\ Au(T_Y) &= Au(B) \cdot P(Y=y_1) + Au(T_X) \cdot P(Y=y_2) = 0.6875 \end{aligned}$$

Step2: calculating the increase of the accuracy over the validation set if the subtrees are removed respectively:

$$\begin{aligned} Inc(T_W) &= Au(W) - Au(T_W) = -0.1 \\ Inc(T_X) &= Au(X) - Au(T_X) = 0.05 \\ Inc(T_Y) &= Au(Y) - Au(T_Y) = -0.1875 \end{aligned}$$

So remove the node X and the subtree rooted at it, and loop again from step 1. This time we only need to calculate the increase of the accuracy if node Y is removed:

$$\begin{aligned} Au(T_Y) &= Au(B) \cdot P(Y=y_1) + Au(X) \cdot P(Y=y_2) = 0.7 \\ Inc(T_Y) &= Au(Y) - Au(T_Y) = -0.2 < 0 \end{aligned}$$

The pruning will be harmful, so the process stops. Subtree rooted at X is removed.