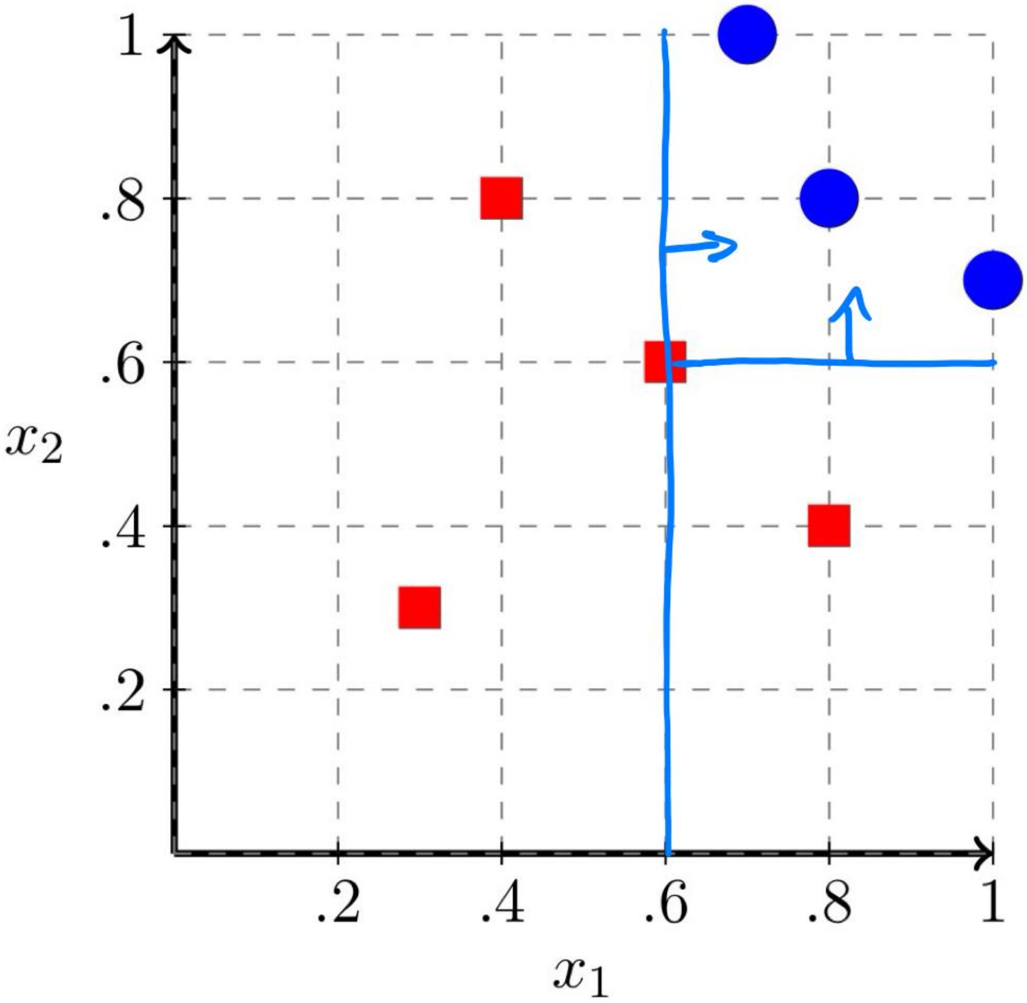
1 Machine Learning

1.1 Theory of Learning

a)



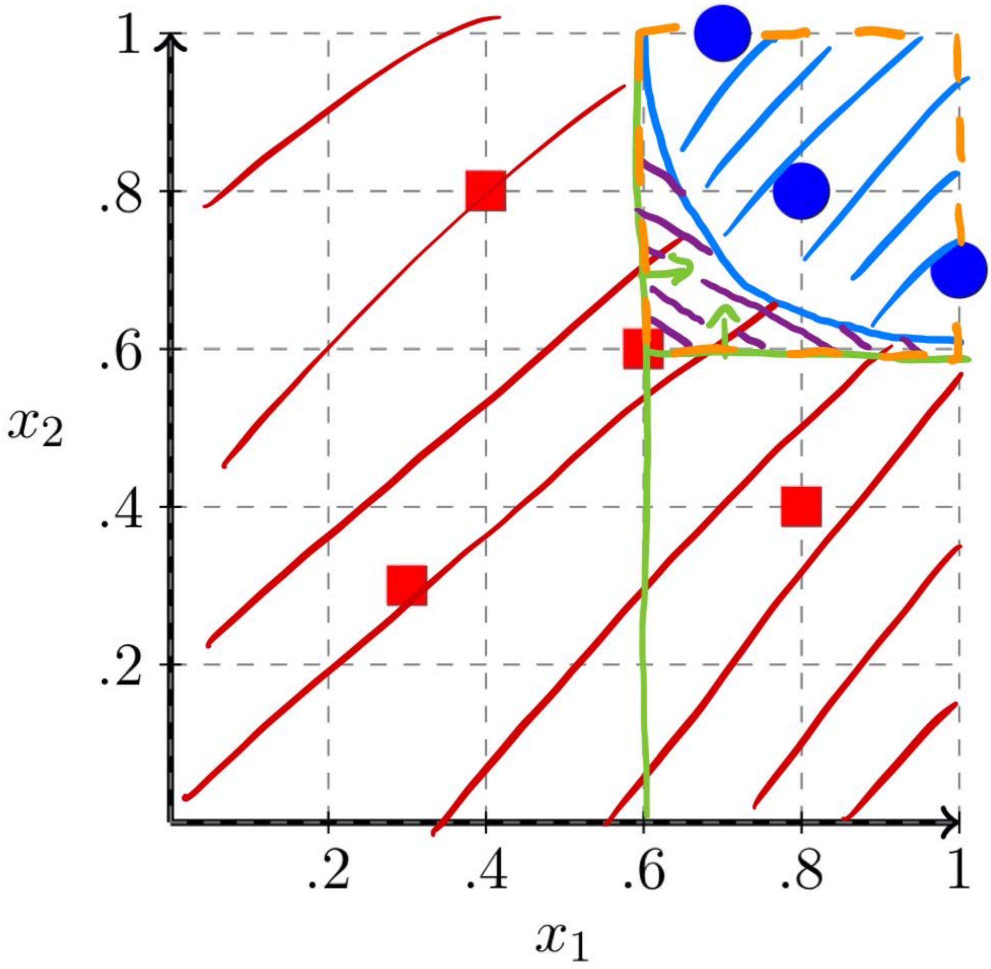
None of the data points are misclassified, the classifier is therefore good for this problem.

b)

$\hat{R}(\tilde{h}) = \frac{1}{n}\sum_{i=1}^{n} l_{01}(x_{1i}, x_{2i}, y_i, \tilde{h}) = \frac{1}{7}\sum_{i=1}^{7} l_{01}(x_{1i}, x_{2i}, y_i, \tilde{h}) = 0$

c)

As all points are uniformly distributed, the Risk $R(\tilde{h})$ will compute which portion of the overall area is misclassified.



The misclassified area is marked in purple. Its portion of the orange area can be calculated as $1 - \frac{\frac{1}{4}\pi\frac{d^2}{4}}{\frac{1}{4}d^2} = 1 - \frac{\pi}{4}$. The portion of the orange area to the whole area is $\frac{0{,}4^2}{1} = 0{,}16$. Thus, the overall portion of the misclassified area is $(1 - \frac{\pi}{4}) * 0{,}16 \approx 0{,}0341$ which is also the Risk $R(\tilde{h})$.

1.2

a)

Given are $p = 3$ positive (yes) and $n = 2$ negative (no) examples. The overall entropy before the split can be calculated as:

$$H(Output) = B\left(\frac{p}{p+n}\right) = B\left(\frac{3}{5}\right) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} \approx 0.971$$

For a split on the attribute *Major*, we can calculate its Remainder as follows:

$$Remainder(Major) = \sum_k \frac{p_k + n_k}{p+n} B\left(\frac{p_k}{p_k+n_k}\right)$$

$$= \frac{2+2}{3+2} B\left(\frac{2}{4}\right) + \frac{1+0}{3+2} B\left(\frac{1}{1}\right)$$

$$= \frac{4}{5} \cdot 1 + \frac{1}{5} \cdot 0 = 0.8$$

With the Remainder calculated, we can now compute the Information Gain for the attribute Major:

$$Gain(Major) = B\left(\frac{p}{p+n}\right) - Remainder(Major) = 0.971 - 0.8 = 0.171$$

The Information Gain of the attribute Status can be calculated similarly:

$$Remainder(Status) = \frac{2+0}{3+2}\,B\left(\frac{2}{2}\right) + \frac{1+2}{3+2}\,B\left(\frac{1}{3}\right)$$

$$= \frac{2}{5} \cdot 0 + \frac{3}{5} \cdot 0.918 \approx 0.551$$

$$Gain(Status) = 0.971 - 0.551 = 0.420$$

And finally, we calculate the Information Gain for the attribute Gender:

$$Remainder(Gender) = \frac{2+1}{5}\,B\left(\frac{2}{3}\right) + \frac{1+1}{5}\,B\left(\frac{1}{2}\right)$$

$$= \frac{3}{5} \cdot 0.918 + \frac{2}{5} \cdot 1 \approx 0.951$$

$$Gain(Gender) = 0.971 - 0.951 = 0.020$$

We can now compare the different Information Gains:

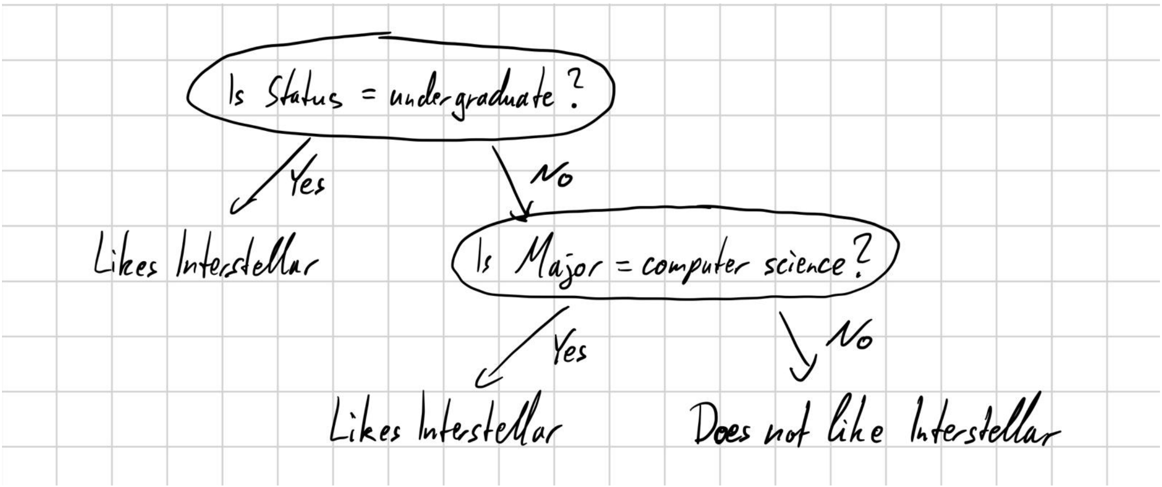
$$Gain(Status) > Gain(Major) > Gain(Gender)$$

A split on the attribute Status will yield the highest Information Gain and is therefore chosen as the first split. For the second split, we only need to consider the attributes Major and Gender. We can also compare the Remainders directly and choose the lowest one.

$$Remainder(Major) = \frac{1+0}{3}B\left(\frac{1}{1}\right) + \frac{0+2}{3}B\left(\frac{0}{2}\right) = \frac{1}{3} * 0 + \frac{2}{3} * 0 = 0$$

$$Remainder(Gender) = \frac{1+1}{3}B\left(\frac{1}{2}\right) + \frac{0+1}{3}B\left(\frac{0}{1}\right) = \frac{1}{3} * 1 + \frac{1}{3} * 0 = \frac{1}{3}$$

Thus, the second split will be on the attribute Major. Since the Remainder is 0, it is also a perfect split and the decision tree is complete.


Is Status = undergraduate ?
Yes
No
Likes Interstellar
Is Major = computer science?
Yes
No
Likes Interstellar
Does not like Interstellar

b)

$LikesInterstellar \Leftrightarrow$
$Status = undergraduate \lor (Status = graduate \land Major = computerscience)$

c)

The same attribute cant be chosen twice because it wouldn't provide any new information. If the same value or a subset was already evaluated on the path for a single attribute, evaluating it again would not change anything. If a different value for the same attribute was chosen, it would lead to an empty set of examples, which is also not useful.

Example: $A_1 = a \land A_2 = x \land A_3 = b$
If $a \subseteq b \Rightarrow A_1 = a \land A_2 = x \land A_3 = b \Leftrightarrow A_1 = a \land A_2 = x$
If $a \not\subseteq b \Rightarrow A_1 = a \land A_2 = x \land A_3 = b \Leftrightarrow false$

The algorithm never evaluates the same attribute twice because it evaluates a subtree only on the attributes of its parent minus the chosen attribute.