SAARLAND UNIVERSITY
Prof. Jana Koehler

ARTIFICIAL INTELLIGENCE

Theoretical Exercise Sheet 8

Solutions due Tuesday, June 30, 18:00.
Total points of the sheet: 42

Exercise 1: Harry's Pizza-Party — 20 points

Harry wants to invite his friends to a Pizza-Party. He does not want to ask them which pizza preferences they have in order to keep the party secret. He asks other people to get an impression what pizza preferences most people have.

1. Use the Decision-Tree-Learning algorithm from the lecture to build up a decision tree for his observations depicted in the table below. Make sure to show your calculations. For each step, begin with the calculation of the entropy, and then after that, calculate the gain of each ingredient, calculating them in the same order as they are presented in the table ($Pineapple \rightarrow Mushrooms \rightarrow Ham \rightarrow Sweetcorn$). Explain which variable you take in each step and why. Draw the resulting Decision-Tree

2. Harry gets bored of all the calculations and decides to just pick variables at random. Why is this a bad idea. Specify what the resulting problem could be, and what the Decision-Tree-Learning algorithm tries to do.

Pineapple	Mushrooms	Ham	Sweetcorn	Rating
1	1	1	1	true
1	0	1	0	false
1	1	0	1	false
1	1	1	0	false
0	1	0	1	true
0	0	1	1	true
0	1	1	1	true
0	0	0	1	false
0	1	1	0	true
0	1	0	1	true

Solution:

1. *We use the following abbreviations for the ingredients: P is Pineapple, M is Mushrooms, H is Ham, S is Sweetcorn.*
   Using Decision-Tree-Learning algorithm from the lecture notes.

Step 1

$$p = 6, n = 4 \Rightarrow B\left(\frac{6}{6+4}\right) = 0.97 = B(pos)$$

Pineapple

$$\begin{aligned}
E_1: \quad & p = 1, n = 3 \\
& \Rightarrow B(\frac{1}{1+3}) = 0.81 \\
E_0: \quad & p = 5, n = 1 \\
& \Rightarrow B\left(\frac{5}{6}\right) = 0.65 \\
\Rightarrow R(P) =& \frac{4}{10} * 0.81 + \frac{6}{10} * 0.65 = 0.71 \\
\Rightarrow G(P) =& B(Pos) - R(P) = 0.26
\end{aligned}$$

Mushrooms

$$\begin{aligned}
E_1: \quad & p = 5, n = 2 \\
& \Rightarrow B\left(\frac{5}{7}\right) = 0.86 \\
E_0: \quad & p = 1, n = 2 \\
& \Rightarrow B\left(\frac{1}{3}\right) = 0.92 \\
\Rightarrow R(M) =& \frac{7}{10} * 0.86 + \frac{3}{10} * 0.92 = 0.88 \\
\Rightarrow G(M) =& B(Pos) - R(M) = 0.09
\end{aligned}$$

Ham

$$
\begin{aligned}
E_1: \quad & p = 4, n = 2 \\
& \Rightarrow B\left(\frac{4}{6}\right) = 0.92 \\
E_0: \quad & p = 2, n = 2 \\
& \Rightarrow B\left(\frac{2}{4}\right) = 1 \\
\Rightarrow R(H) =& \frac{6}{10} * 0.92 + \frac{4}{10} * 1 = 0.95 \\
\Rightarrow G(H) =& B(Pos) - R(H) = 0.02
\end{aligned}
$$

Sweetcorn

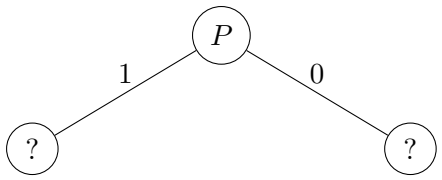
$$
\begin{aligned}
E_1: \quad & p = 5, n = 2 \\
& \Rightarrow B\left(\frac{5}{7}\right) = 0.86 \\
E_0: \quad & p = 1, n = 2 \\
& \Rightarrow B\left(\frac{1}{3}\right) = 0.92 \\
\Rightarrow R(S) =& \frac{7}{10} * 0.86 + \frac{3}{10} * 0.92 = 0.88 \\
\Rightarrow G(S) =& B(Pos) - R(S) = 0.09
\end{aligned}
$$

Result of step 1

$$G(P) > G(M) = G(S) > G(H)$$

$$0.26 > 0.09 = 0.09 > 0.02$$

Choose the ingredient with the highest information. As a result, we have Pineapple as our first node in the decision tree:



Step 2 - Investigation of the right side

$$p = 5, n = 1 \Rightarrow B\left(\frac{5}{5+1}\right) = 0.65 = B(pos)$$

Mushrooms

$$\begin{aligned}
E_1: \quad & p = 4, n = 0 \\
& \Rightarrow B\left(\frac{4}{4}\right) = 0 \\
E_0: \quad & p = 1, n = 1 \\
& \Rightarrow B\left(\frac{1}{2}\right) = 1 \\
\Rightarrow R(M) = & \frac{4}{6} * 0 + \frac{2}{6} * 1 = 0.33 \\
\Rightarrow G(M) = & B(Pos) - R(M) = 0.32
\end{aligned}$$

Ham

$$\begin{aligned}
E_1: \quad & p = 3, n = 0 \\
& \Rightarrow B\left(\frac{3}{3}\right) = 0 \\
E_0: \quad & p = 2, n = 1 \\
& \Rightarrow B\left(\frac{2}{3}\right) = 0.92 \\
\Rightarrow R(H) = & \frac{3}{6} * 0 + \frac{3}{6} * 0.92 = 0.46 \\
\Rightarrow G(H) = & B(Pos) - R(H) = 0.19
\end{aligned}$$

Sweetcorn

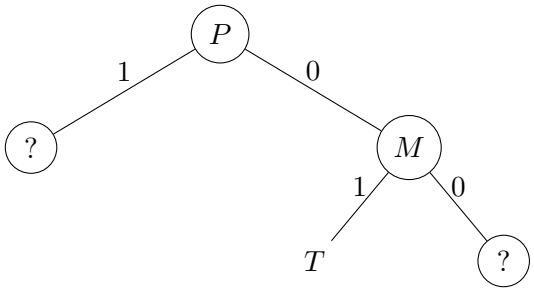
$$
\begin{aligned}
E_1: \quad & p = 4, n = 1 \\
& \Rightarrow B\left(\frac{4}{5}\right) = 0.72 \\
E_0: \quad & p = 1, n = 0 \\
& \Rightarrow B\left(\frac{1}{1}\right) = 0 \\
\Rightarrow R(S) =& \frac{5}{6} * 0.72 + \frac{1}{6} * 0 = 0.6 \\
\Rightarrow G(S) =& B(Pos) - R(S) = 0.05
\end{aligned}
$$

Result of step 2

$$G(M) > G(H) > G(S)$$

$$0.32 > 0.19 > 0.05$$

Choose the ingredient with the highest information. As a result, we have Mushrooms as next Node:



Step 3

$$p = 1, n = 1 \Rightarrow B\left(\frac{1}{1+1}\right) = 1 = B(pos)$$

Only Ham and Sweetcorn remain as attributes.

Ham

$$
\begin{aligned}
E_1: \quad & p = 1, n = 0 \\
& \Rightarrow B\left(\frac{1}{1}\right) = 0 \\
E_0: \quad & p = 0, n = 1 \\
& \Rightarrow B\left(\frac{0}{1}\right) = 0 \\
\Rightarrow R(H) =& \frac{1}{2} * 0 + \frac{1}{2} * 0 = 0 \\
\Rightarrow G(H) =& B(Pos) - R(H) = 1
\end{aligned}
$$

Sweetcorn

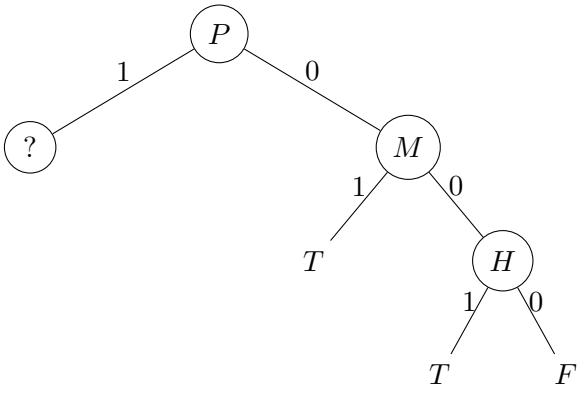
$$
\begin{aligned}
E_1: \quad & p = 1, n = 1 \\
& \Rightarrow B\left(\frac{1}{2}\right) = 1 \\
E_0: \quad & p = 0, n = 0 \\
& \Rightarrow B\left(\frac{0}{0}\right) = 0 \\
\Rightarrow R(S) =& \frac{2}{2} * 1 + \frac{0}{2} * 0 = 1 \\
\Rightarrow G(S) =& B(Pos) - R(S) = 0
\end{aligned}
$$

Result of step 3

$$G(H) > G(S)$$

$$1 > 0$$

Choose the ingredient with the highest information. As a result, we have Ham as next Node:



Step 4 - Investigation of the left side

Only Mushroom, Ham and Sweetcorn remain as attributes.

$$p = 1, n = 3 \Rightarrow B\left(\frac{1}{1+3}\right) = 0.81 = B(pos)$$

Mushrooms

$$\begin{aligned}
E_1: \quad & p = 1, n = 2 \\
& \Rightarrow B\left(\frac{1}{3}\right) = 0.91 \\
E_0: \quad & p = 0, n = 1 \\
& \Rightarrow B\left(\frac{0}{1}\right) = 0 \\
\Rightarrow R(H) = & \frac{3}{4} * 0.91 + \frac{1}{4} * 0 = 0.68 \\
\Rightarrow G(H) = & B(Pos) - R(H) = 0.13
\end{aligned}$$

Ham

$$
\begin{aligned}
E_1: \quad & p = 1, n = 2 \\
& \Rightarrow B\left(\frac{1}{3}\right) = 0.91 \\
E_0: \quad & p = 0, n = 1 \\
& \Rightarrow B\left(\frac{0}{1}\right) = 0 \\
\Rightarrow R(H) =& \frac{3}{4} * 0.91 + \frac{1}{4} * 0 = 0.68 \\
\Rightarrow G(H) =& B(Pos) - R(H) = 0.13
\end{aligned}
$$

Sweetcorn

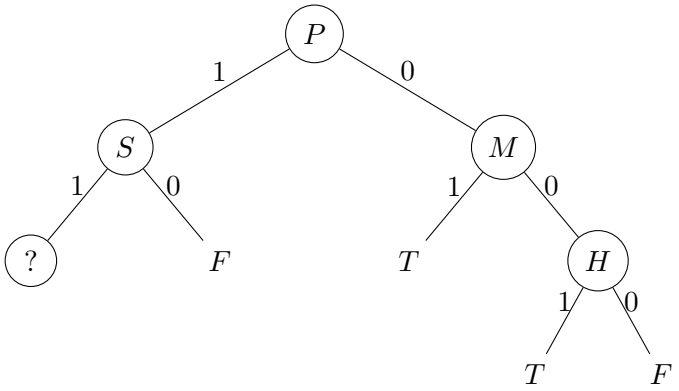
$$
\begin{aligned}
E_1: \quad & p = 1, n = 1 \\
& \Rightarrow B\left(\frac{1}{2}\right) = 1 \\
E_0: \quad & p = 0, n = 2 \\
& \Rightarrow B\left(\frac{0}{2}\right) = 0 \\
\Rightarrow R(S) =& \frac{2}{4} * 1 + \frac{2}{4} * 0 = 0.5 \\
\Rightarrow G(S) =& B(Pos) - R(S) = 0.31
\end{aligned}
$$

Result of step 4

$$G(S) > G(M) = G(H)$$

$$0.31 > 0.13 = 0.13$$

Choose the ingredient with the highest information. As a result, we have Sweetcorn as next Node:



Step 5

Only Ham and Mushrooms remain as attributes.

$$p = 1, n = 1 \Rightarrow B\left(\frac{1}{1+1}\right) = 1 = B(pos)$$

Mushrooms

$$\begin{aligned}
E_1: \quad & p = 1, n = 1 \\
& \Rightarrow B\left(\frac{1}{2}\right) = 1 \\
E_0: \quad & p = 0, n = 0 \\
& \Rightarrow B\left(\frac{0}{0}\right) = 0 \\
\Rightarrow R(H) =& \frac{2}{2} * 1 + \frac{0}{2} * 0 = 1 \\
\Rightarrow G(H) =& B(Pos) - R(H) = 0
\end{aligned}$$

Ham

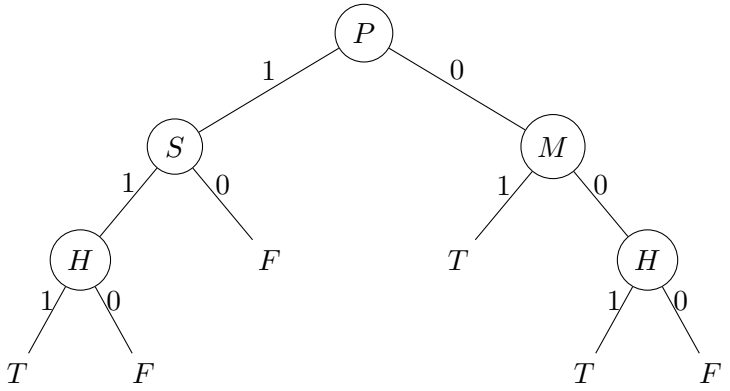
$$
\begin{aligned}
E_1: \quad & p = 1, n = 0 \\
& \Rightarrow B\left(\frac{1}{1}\right) = 0 \\
E_0: \quad & p = 0, n = 1 \\
& \Rightarrow B\left(\frac{0}{1}\right) = 0 \\
\Rightarrow R(H) =& \frac{1}{2} * 0 + \frac{1}{2} * 0 = 0 \\
\Rightarrow G(H) =& B(Pos) - R(H) = 1
\end{aligned}
$$

Result of step 5

$$G(H) > G(M)$$

$$1 > 0$$

Choose the ingredient with the highest information. As a result, we have Ham as next Node:



2. Whereas the resulting tree will correctly classify all given examples, it will not say much about other cases. It just memorizes the observations and does not generalize. This is called "Overfitting".

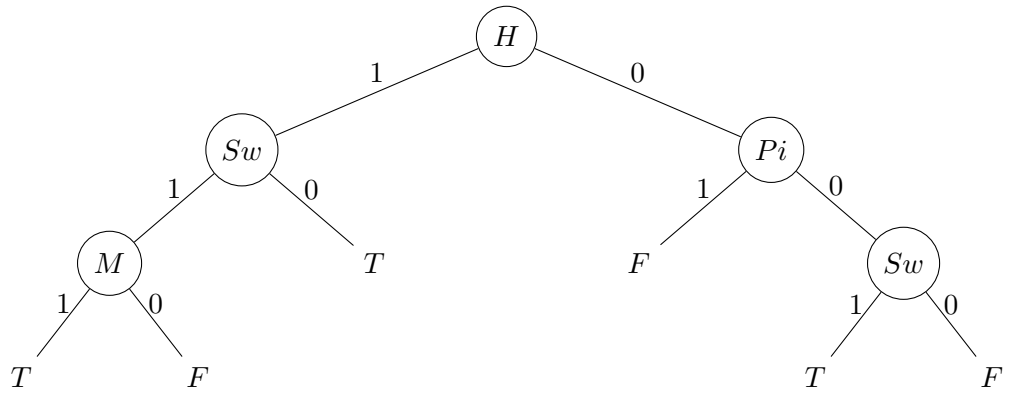
Total points: 20

Exercise 2: Harry's Pizza-Party Part II 5 points

In the table below the favorite pizzas of Harry's friends are listed. For each of the pizzas, use the resulting tree from question 1 to say if Harry orders this pizza or not. Do not forget to declare your steps. If you did not manage to build a tree in exercise 1, use the provided backup tree (Note that the backup tree is not the solution to the Question 1, but if you use the backup tree in this question instead of the solution to Question 1 you will still receive full marks). We use the following abbreviations for the ingredients: Pi is Pineapple, M is Mushrooms, H is Ham, Sw is Sweetcorn.

Hermione	Mushroom, Ham, Sweetcorn and Onion
Neville	Pineapple, Sweetcorn and Mushrooms
Ron	Salami and Sweetcorn
Ginny	Ham
Luna	Salami, Bolognese sauce and Pepper

The backup tree:



Solution:
For the tree from exercise 1:

	Ordered?
Hermione	*Yes*
Neville	*No*
Ron	*No*
Ginny	*Yes*
Luna	*No*

For the backup tree:

	Ordered?
Hermione	*Yes*
Neville	*No*
Ron	*Yes*
Ginny	*Yes*
Luna	*No*

Total points: 5

Exercise 3: Fitting Functions 6 points

Look at this example, introduced in the lecture. Assuming the points are measurements from some experiment and the functions are fitting functions to predict further measurements. The left one is a linear function and the right one is a polynomial function. According to Ockham's razor, which one of these functions is the better pick. Justify your answer briefly (in 2-3 sentences), highlighting why Ockhams razor prefers one of the functions and what the problem of overfitting would be in this case.

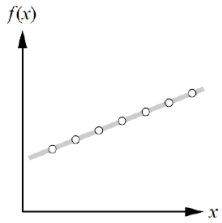


Figure 1: The linear function

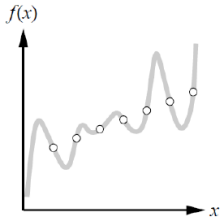


Figure 2: The polynomial function

Solution: Ockham's razor would pick the Linear Function. It prefers the simpler solution, and in this case, the polynomial function is more complex than the linear function. The polynomial function overfitts the past measurements, being correct for the current data points, but not following the general curve, so it will likely delivering worse predictions for future measurements.

Total points: 6

Exercise 4: Spam Detection 6 points

Suppose you are working on a spam detection system. You formulated the problem as a classification task where "Spam" is the positive class and "not Spam" is the negative class. Your training set contains m = 1000 emails.

	Predicted Spam	Predicted not Spam
Actual Spam	8	2
Actual not Spam	16	974

Calculate the Accuracy, Recall, Specificity and Precision in this example.

How do your answers change if we now consider "Spam" the negative class and "not Spam" the positive class?

Solution:

- Accuracy $\frac{8+974}{1000} = 0.982$
- Recall $\frac{8}{8+2} = 0.80$
- Specificity $\frac{974}{974+16} = 0.9838$
- Precision $\frac{8}{8+16} = 0.3333$

If we consider "Spam" the negative class and "not Spam" the positive class then the original Specificity becomes Recall and vice versa, so new Recall is $\frac{974}{974+16} = 0.9838$ and new Specificity is $\frac{8}{8+2} = 0.80$. Accuracy remains the same, Accuracy $\frac{974+8}{1000} = 0.982$. Precision becomes $\frac{974}{976} = 0.9980$ (In tutorial note the difference in Precision values).

Total points: 6

Exercise 5: Entropy 5 points

Consider a fair six-sided die and another loaded die such that the probability of obtaining an outcome of 6 is 50%, the probabiliy of obtaining 5 is 25% and the remaining outcomes are equally likely. Calculate the entropy for both dice. How do the results differ? Explain!

Solution:

Fair die:

$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$

$$H_{fair} = \sum_{1}^{6} \frac{1}{6} \cdot log_2 \left(\frac{1}{\frac{1}{6}} \right) \approx 2.58$$

Loaded die:
$P(6) = \frac{1}{2},\ \ P(5) = \frac{1}{4},\ \ P(1) = P(2) = P(3) = P(4) = \frac{1}{16}$

$$H_{loaded} = \frac{1}{2} \cdot log_2\left(\frac{1}{\frac{1}{2}}\right) + \frac{1}{4} \cdot log_2\left(\frac{1}{\frac{1}{4}}\right) + 4 \cdot \left(\frac{1}{16} \cdot log_2\left(\frac{1}{\frac{1}{16}}\right)\right) = 2$$

Entropy is a measure of disorder. Since the outcome 6 or 5 is more certain for the loaded die the entropy is lower than for the fair die.

Total points: 5

Submission Instructions

Solutions need to be packaged into a `.zip` file and uploaded on the AI CMS page. The file should be named using the matriculation number (MNr) of all students who submit together, i.e.,

`AI2020_TE8_<MNr S1>_<MNr S2>_<MNr S3>.zip`

Note, that all students which submit together have to be in the SAME tutorial.
The `.zip` file needs to contain the following files

- a PDF file with the solutions to the theoretical questions of this sheet

Note, that only one student of each group needs to do the submission!