Exercise 1 :

Consider that we have the following data-points:

x_1	x_2	y
1	1	2
0	0	0
1	0	3
2	1	2

What is the Mean Squared Error (MSE) error for the following sets of parameters:

i) $w_0 = 0, w_1 = 1, w_2 = -1$

ii) $w_0 = 1, w_1 = 1, w_2 = 0$

Which model is preferable according to the MSE error function?

Solution:

i) First, we compute the value of $\hat{y}$

x_1	x_2	y	$\hat{y}$
1	1	2	0
0	0	0	0
1	0	3	1
2	1	2	1

The error is then: $2^2 + 0^2 + 2^2 + 1^2 = 9$

ii)

x_1	x_2	y	$\hat{y}$
1	1	2	2
0	0	0	1
1	0	3	2
2	1	2	3

The error is then: $0^2 + 1^2 + 1^2 + 1^2 = 3$

Therefore, it is preferable the second model, as it has less error

Exercise 2 :

Consider a database of cars represented by the five training examples below. The target attribute *Acceptable*, which can have values `yes` and `no`, is to be predicted based on the other attributes of the car in question. These attributes indicate a) the age of the car (*Age* having values < 5 `years` and ≥ 5 `years`), b) the make of the car (*Make* having states `Toyota` and `Mazda`), c) the number of previous owners (*#Owners* having values `1`, `2` and `3`), d) the number of kilometers (*#Kilometers* having values $> 150k$ and $\leq 150k$) and e) the number of doors (*#Doors* having values 3 and 5).

	Attributes					Target
	Age	*Make*	*#Owners*	*#Kilometers*	*#Doors*	*Acceptable*
1	< 5	Mazda	1	$> 150k$	3	yes
2	≥ 5	Mazda	3	$> 150k$	3	no
3	≥ 5	Toyota	1	$\leq 150k$	3	no
4	≥ 5	Mazda	3	$> 150k$	5	yes
5	≥ 5	Toyota	2	$\leq 150k$	5	yes

a) Calculate the entropy for the attribute *#Owners*.[1]

b) Show the decision/classification tree that would be learned by the learning algorithm assuming that it is given the training examples in the database.

c) Show the value of the information gain for each candidate attribute at each step in the construction of the tree.

Solution: For question (a):

$$ENT(\#Owners) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) = 1.521 \tag{1}$$

For question (b) we need to calculate the information gain for each of the features. That is, for the generic feature X we should calculate

$$\mathit{Gain}(X) = \mathit{Ent}(\text{Accept}) - \mathit{ExpectedEntropy}(\text{Accept}|X),$$

where *ExpectedEntropy*(Accept$|X$) is the expected entropy of *Accept* wrt. X.

For *Ent*(Accept) we get

$$\mathit{Ent}(\text{Accept}) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) = 0.971$$

Considering now the features in sequence from left to right, we have that the expected entropy of *Age* is

$$\begin{aligned}
&\mathit{ExpectedEntropy}(\text{Accept}|\text{Age}) \\
&= P(\text{Age} < 5)\mathit{Ent}(\text{Accept}|\text{Age} < 5) + P(\text{Age} \geq 5)\mathit{Ent}(\text{Accept}|\text{Age} \geq 5) \\
&= \frac{1}{5}\mathit{Ent}(\text{Accept}|\text{Age} < 5) + \frac{4}{5}\mathit{Ent}(\text{Accept}|\text{Age} \geq 5).
\end{aligned}$$

Here *Ent*(Accept$|$Age < 5) and *Ent*(Accept$|$Age ≥ 5) denote the entropy of *Accept* when only considering the instances restricted to Age < 5 and Age ≥ 5, respectively. Thus we get

$$\begin{aligned}
\mathit{Ent}(\text{Accept}|\text{Age} < 5) &= -\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right) = 0 \\
\mathit{Ent}(\text{Accept}|\text{Age} \geq 5) &= -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1
\end{aligned}$$

[1] Note that $\log_2(x) = \frac{\log_{10}(x)}{\log_{10}(2)}$.

Plugging these values into the expression for the expected entropy above, we get

$$\textit{ExpectedEntropy}(\text{Accept}|\text{Age}) = \frac{1}{5} \cdot 0 + \frac{4}{5} \cdot 1 = \frac{4}{5}$$

and the information gain therefore becomes

$$\textit{Gain}(\text{Age}) = 0.971 - \frac{4}{5} = 0.171.$$

Doing the same calculations for the remaining attributes we end up with

$$\begin{aligned} Gain(\text{Make}) &= 0.0202 \\ Gain(\text{\#Owners}) &= 0.171 \\ Gain(\text{\#Kilo}) &= 0.0202 \\ Gain(\text{\#Doors}) &= 0.4202 \end{aligned}$$

Since #Doors has the highest information gain we put that feature at the top of the tree.

For the branch of the tree with #Doors = 5 there is nothing more to do, since both of the training examples with this feature value has the same value for *Accept*, namely *yes*. For #Doors = 3, we have two instances with *Accept* being *yes* and one instance with *Accept* being *no*. Thus, we need to make another feature test for instances where #Doors = 3. To do that we proceed along the same lines as above, expect that we now only consider the instances consistent with #Doors = 3. For this restricted data set we find the entropy for *Accept*:

$$Ent(\text{Accept}|\text{\#Doors} = 3) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) = 0.918$$

Based on this value we can calculate the information gain for each of the remaining features (this involves calculating the expected entropy for these features using the same calculation scheme as above):

$$\begin{aligned} Gain(\text{Age}) &= 0.918 - 0 \\ Gain(\text{Make}) &= 0.918 - 2/3 \\ Gain(\text{\#Owners}) &= 0.918 - 2/3 \\ Gain(\text{\#Kilo}) &= 0.918 - 2/3 \end{aligned}$$

Hence, for #Doors$=$ 3 we pick *Age* as the next node. The classification tree can now directly be constructed based on the results above:

Exercise 3 :

Solve Exercise 7.3 (except sub-question f) in PM.

Solution:

1. (a) The optimal decision tree with one node predicts Likes = false. It has 5 errors.
2. (b) The optimal prediction is to predict likes with probability 5/12. It has sum-of- squares error $5 \cdot (7/12)^2 + 7 \cdot (5/12)^2 = 2.92$.
3. (c) The optimal (with respect to sum of absolute errors) decision tree of depth 2 is: *if lawyers then Likes=true else Likes=false*. It has 3 errors. At the root are all of the examples $(e_1, \ldots, e_{13})$. Filtered to the *lawyers* $=$ *true* node are $e_2, e_3, e_4, e_8, e_9, e_{10}$. Filtered to the *lawyers* $=$ *false* node are $e_1, e_5, e_6, e_7, e_{11}, e_{12}$.
4. (d) The optimal (with respect to sum-of-squares error) decision tree of depth 2 is: *if lawyers then likes with probability 2/3 else likes with probability 1/6* The error is $4 \cdot (1/3)^2 + 2 \cdot (2/3)^2 + 5 \cdot (1/6)^2 + 1 \cdot (5/6)^2 = 2.83$.
5. (e) The smallest tree that correctly classifies all training examples is: *if guns then lawyers else comedy* The information gain split gives a more complicated tree that represents the same function. To construct the tree follow the same procedure as for the preceding exercise.
6. (g) It is not linearly separable. The examples e_3 and e_{10} must be on the same side of a hyperplane (as they are both true on Likes). Therefore any linear interpolation also must be on the same side, but e_4 is between these in the input categories, but has a different classification.

Exercise 4 :

Given below is a trainings data set about esoteric programming languages. We want to predict whether people are willing to learn the languages depending on Form, Usefulness, and Torment.

ID	Usefulness	Form	Torment	Appealing
1	useful	Other	pleasant	No
2	useless	Text	torture	No
3	useless	Text	torture	No
4	useless	Images	pleasant	Yes
5	useful	Images	pleasant	Yes
6	useful	Other	torture	No
7	useless	Images	torture	Yes
8	useless	Text	pleasant	Yes
9	useless	Text	pleasant	Yes
10	useful	Other	torture	No

(a) Compute the information gain associated with choosing the attribute Form as root.

(b) Draw a decision tree with Form as root.

(c) Use your decision tree to predict the following records:

Usefulness	Form	Torment
useful	Other	pleasant
useless	Text	torture

Solution: Solution 0:

(a)

$$\begin{aligned} \text{Gain}(A) &= B\left(\frac{p}{p+n}\right) - \text{Remainder}(A) \\ \text{Remainder}(A) &= \sum_{k=1}^{d} \frac{p_k+n_k}{p+n} B\left(\frac{p_k}{p_k+n_k}\right) \\ B(q) &= -q\log_2 q - (1-q)\log_2(1-q) \end{aligned}$$

We know for Form:

Form	#Yes	#No	Σ
Other	0	3	3
Text	2	2	4
Images	3	0	3

$$\begin{aligned} \text{Remainder(Form)} &= \frac{3}{10}B\left(\frac{0}{3}\right) + \frac{4}{10}B\left(\frac{2}{4}\right) + \frac{3}{10}B\left(\frac{3}{3}\right) \\ &= \frac{2}{5} \\ \text{Gain(Form)} &= B\left(\frac{5}{10}\right) - \text{Remainder(Form)} \\ &= \frac{3}{5} \end{aligned}$$

(b)

(c)

Usefulness	Form	Torment	Appealing
useful	Other	pleasant	No
useless	Text	torture	No

Exercise 5 :

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: ***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

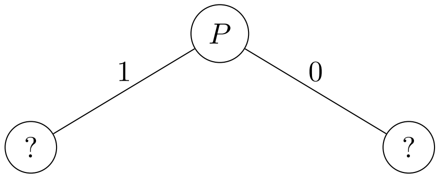
$$\begin{aligned}
E_1: \quad & p=5, n=2 \\
& \Rightarrow B\left(\frac{5}{7}\right)=0.86 \\
E_0: \quad & p=1, n=2 \\
& \Rightarrow B\left(\frac{1}{3}\right)=0.92 \\
\Rightarrow R(S)= & \frac{7}{10}*0.86+\frac{3}{10}*0.92=0.88 \\
\Rightarrow G(S)= & B(Pos)-R(S)=0.09
\end{aligned}$$

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

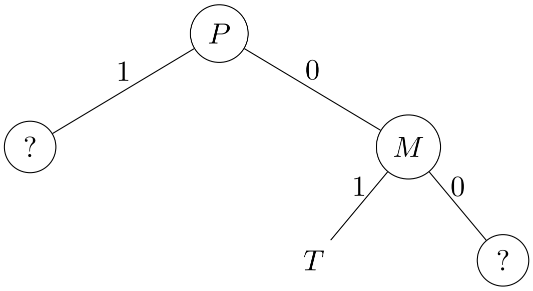
$$\begin{aligned}
E_1: \quad & p = 4, n = 1 \\
& \Rightarrow B\left(\frac{4}{5}\right) = 0.72 \\
E_0: \quad & p = 1, n = 0 \\
& \Rightarrow B\left(\frac{1}{1}\right) = 0 \\
\Rightarrow R(S) =& \frac{5}{6} * 0.72 + \frac{1}{6} * 0 = 0.6 \\
\Rightarrow G(S) =& B(Pos) - R(S) = 0.05
\end{aligned}$$

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

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

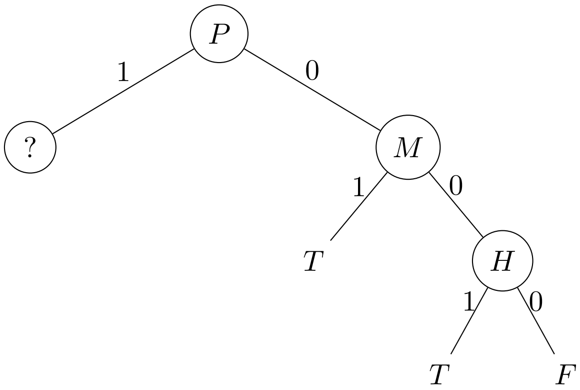
Sweetcorn

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

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

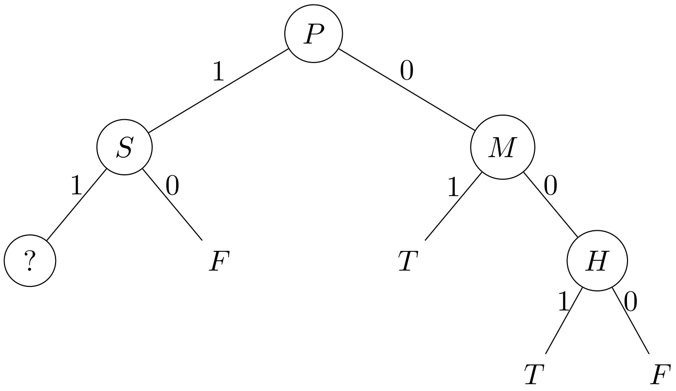
$$\begin{aligned}
E_1 : \quad & p = 1, n = 1 \\
& \Rightarrow B\left(\frac{1}{2}\right) = 1 \\
E_0 : \quad & p = 0, n = 2 \\
& \Rightarrow B\left(\frac{0}{2}\right) = 0 \\
\Rightarrow R(S) =& \frac{2}{4} * 1 + \frac{2}{4} * 0 = 0.5 \\
\Rightarrow G(S) =& B(Pos) - R(S) = 0.31
\end{aligned}$$

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

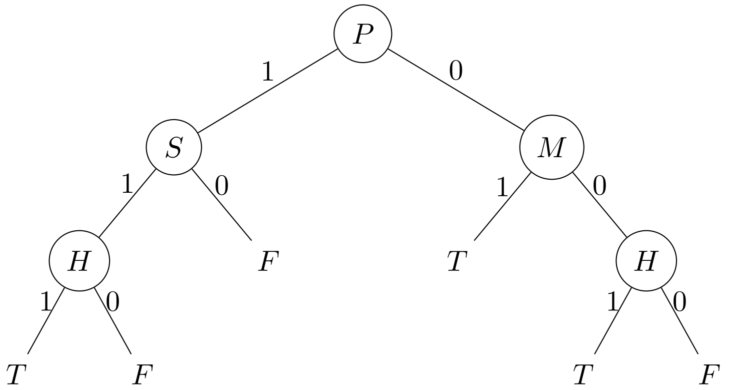
$$\begin{aligned}
E_1: \quad & p = 1, n = 0\\
& \Rightarrow B\left(\frac{1}{1}\right) = 0\\
E_0: \quad & p = 0, n = 1\\
& \Rightarrow B\left(\frac{0}{1}\right) = 0\\
\Rightarrow R(H) =& \frac{1}{2} * 0 + \frac{1}{2} * 0 = 0\\
\Rightarrow G(H) =& B(Pos) - R(H) = 1
\end{aligned}$$

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:

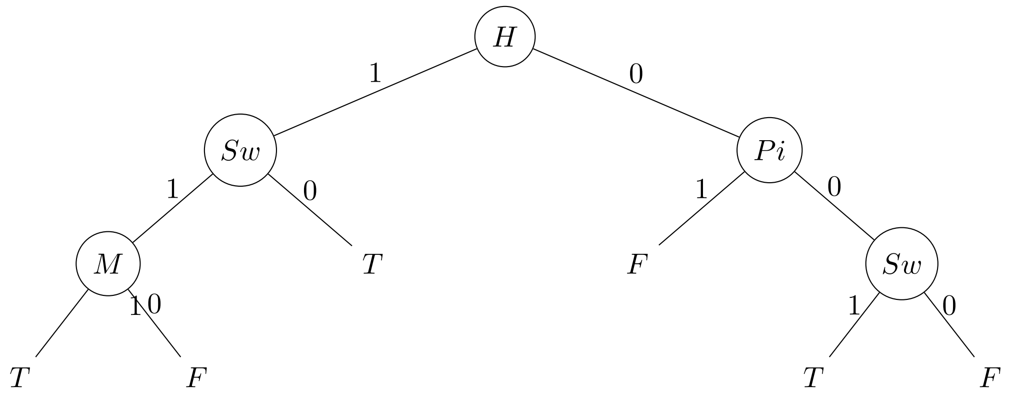


2. The Decision-Tree-algorithm tries to generate ”smallish” trees. If Harry just picks random variables, the resulting tree could be much bigger, and therefore harder to use.

Exercise 6 :

In the table below the favorite pizzas of Harry’s friends are listed. For each of the pizzas, use the following tree to say if Harry orders this pizza or not. (Note that this is not necessarily the solution to Exercise 5). Do not forget to declare your steps. 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



Solution: ***Solution:***

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

Exercise 7 :

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.

Exercise 8 :

Download and install the WEKA data-mining toolbox:

http://www.cs.waikato.ac.nz/ml/weka/

WEKA provides several user-interfaces. Select the ’Explorer’ interface from the ’Applications’ menu, and try the following:

- Load the ’Iris’ dataset. This dataset contains measurements from 150 indi- vidual plants of the genus Iris, belonging to 3 different species ’Iris setosa’, ’Iris versicolor’, and ’Iris virginica’. The machine learning task associated with this dataset is: predict the species from the four measurement values.
- Use the ’Visualize’ tab to get an overview of the attribute values and their relation to the class label. Sketch by hand a small decision tree for predicting the class label.
- Use WEKA’s decision tree construction methods to build a decison tree (under the ’Classify’ tab select e.g. J48 or the SimpleCart classifier). Compare with your own proposed decision tree.

Exercise 9 (Optional):

- Download the Pregnancy dataset. Note that the format of this file does not follow the standard file-format used by Weka. When trying to load the file you will therefore have to use the ’converter’ suggested by Weka.
- Construct a decision tree for classification. Try to reason about the structure of the tree. Hint: have a look at the underlying Bayesian network model (which can be found here) that we have previously looked at in the course.
