University of Southern Denmark
IMADA

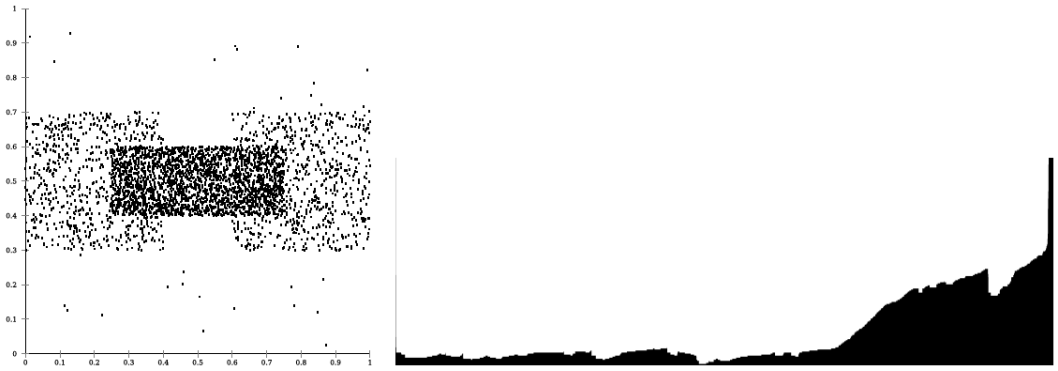
DM566: Data Mining and Machine Learning

Spring term 2022

Exercise 9

Exercise 9-1 OPTICS Plot

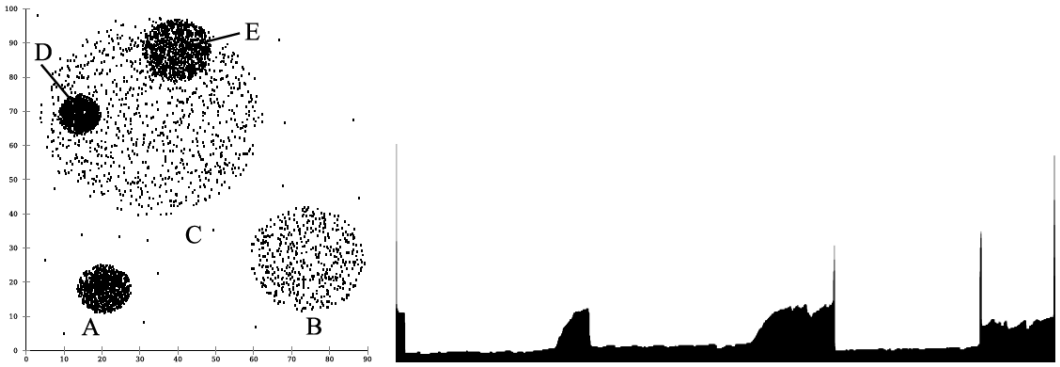
1. For the data below we got computed the reachability diagram to the right.



With a näıve understanding of hierachical clustering, wouldn't we have expected three valleys in the plot?
Explain, why this is not the case and why the plot, instead, looks as it does and accurately describes the density structure of the data.

Suggested solution:
The clusters to the left and to the right are density-connected by the more dense, central cluster. OPTICS aims at finding clusters starting from the more dense parts. Once the seed list contains objects from the densest part, it can extend the structure towards both sides simultaneously.
This is an extreme case of the "density-linkage" effects (for single-link: the single-link-effect, in OPTICS, choosing a larger minPts is a way to reduce this effect but only excluding densities below the threshold).
Understanding the density-based cluster-model, we indeed have here only one cluster of lower density that contains a cluster of higher density.

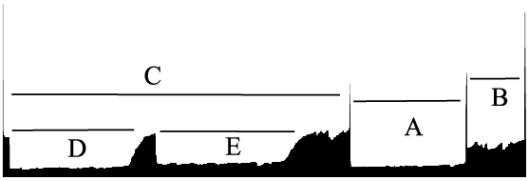
2. For this dataset (left) we have the reachability plot (right).



Mark in the reachability plot which areas relate to the clusters A, B, C, D, and E.

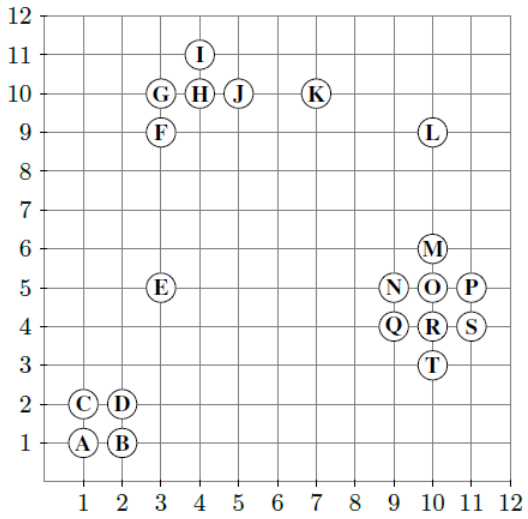
Suggested solution:

Solution:



Exercise 9-2 Outlier Scores

Given the following 2 dimensional data set:



As distance function, use Manhattan distance $L_1(a, b) = |a_1 - b_1| + |a_2 - b_2|$.
Compute the following (without including the query point when determining the kNN):

- LOF using $k = 2$ for the points E, K and O.
- LOF using $k = 4$ for the points E, K and O.
- kNN distance using $k = 2$ for all points.
- kNN distance using $k = 4$ for all points.
- aggregated kNN distances for $k = 2$ and $k = 4$ for all points
  (aggregated kNN distance = averaged sum of the distances to all the kNN!)

Suggested solution:
We find kNN and kNN distance for all points using $k = 2$ and $k = 4$ respectively. We then calculate the local reachability density (lrd).

	2NN	2d.	4NN	4d.	lrd_2	lrd_4
A	B C	1	B C D E	6	-	-
B	A D	1	A C D E	5	-	$1/(\frac{6+5+4+5}{4}) = 0.2$
C	A D	1	A B D E	5	-	$1/(\frac{6+5+4+5}{4}) = 0.2$
D	B C	1	A B C E	4	$1/(\frac{1+1}{2}) = 1$	$1/(\frac{6+5+5+5}{4}) \approx 0.1905$
E	D F	4	B C D F G	5	$1/(\frac{4+4}{2}) = 0.25$	$1/(\frac{5+5+4+4+5}{5}) \approx 0.217$
F	G H	2	G H I J	3	$1/(\frac{1+2}{2}) \approx 0.667$	$1/(\frac{2+2+3+3}{4}) = 0.4$
G	F H	1	F H I J	2	-	$1/(\frac{3+2+3+2}{4}) = 0.4$
H	G I J	1	F G I J	2	$1/(\frac{1+2+2}{3}) = 0.6$	$1/(\frac{3+2+3+2}{4}) = 0.4$
I	G H J	2	F G H J	3	-	$1/(\frac{3+2+2+2}{4}) \approx 0.444$
J	G H I K	2	G H I K	2	$1/(\frac{2+1+2+3}{4}) = 0.5$	$1/(\frac{2+2+3+4}{4}) \approx 0.364$
K	H J	3	G H I J L	4	$1/(\frac{2+3}{2}) = 0.4$	$1/(\frac{4+3+4+2+5}{5}) \approx 0.2778$
L	K M O	4	K M N O P R	5	-	$1/(\frac{4+3+5+4+5+5}{6}) \approx 0.231$
M	N O P R	2	N O P R	2	$1/(\frac{2+2+1+2}{4}) \approx 0.5714$	$1/(\frac{2+1+2+2}{4}) \approx 0.5714$
N	O Q	1	M O P Q R	2	$1/(\frac{1+1}{2}) = 1$	$1/(\frac{2+1+2+2+2}{5}) \approx 0.556$
O	M N P R	1	M N P R	1	$1/(\frac{2+1+1+1}{4}) = 0.8$	$1/(\frac{2+2+2+1}{4}) \approx 0.5714$
P	O S	1	M N O R S	2	$1/(\frac{1+1}{2}) = 1$	$1/(\frac{2+2+1+2+2}{5}) \approx 0.556$
Q	N R	1	N O R S T	2	-	-
R	O Q S T	1	O Q S T	1	$1/(\frac{1+1+1+2}{4}) = 0.8$	$1/(\frac{1+2+2+2}{4}) \approx 0.5714$
S	P R	1	O P Q R T	2	-	-
T	O Q R S	2	O Q R S	2	-	-

Recall the formula for the local outlier factor (LOF):

$$LOF_k(p) = \frac{\sum_{o \in kNN(p)} \frac{lrd_k(o)}{lrd_k(p)}}{Cardinality(kNN(p))}$$

LOF scores for $k = 2$:

$$\begin{aligned} LOF_2(E) &= \frac{\frac{lrd_2(D)}{lrd_2(E)} + \frac{lrd_2(F)}{lrd_2(E)}}{2} \\ &= \frac{\frac{1}{0.25} + \frac{0.667}{0.25}}{2} \\ &\approx 3.333 \\ LOF_2(K) &= \frac{\frac{lrd_2(H)}{lrd_2(K)} + \frac{lrd_2(J)}{lrd_2(K)}}{2} \\ &= \frac{\frac{0.6}{0.4} + \frac{0.5}{0.4}}{2} \\ &= 1.375 \end{aligned}$$

$$
\begin{aligned}
LOF_2(O) &= \frac{\frac{lrd_2(M)}{lrd_2(O)} + \frac{lrd_2(N)}{lrd_2(O)} + \frac{lrd_2(P)}{lrd_2(O)} + \frac{lrd_2(R)}{lrd_2(O)}}{4} \\
&= \frac{\frac{0.5714}{0.8} + \frac{1}{0.8} + \frac{1}{0.8} + \frac{0.8}{0.8}}{4} \\
&\approx 1.054
\end{aligned}
$$

LOF scores for $k = 4$:

$$
\begin{aligned}
LOF_4(E) &= \frac{\frac{lrd_4(B)}{lrd_4(E)} + \frac{lrd_4(C)}{lrd_4(E)} + \frac{lrd_4(D)}{lrd_4(E)} + \frac{lrd_4(F)}{lrd_4(E)} + \frac{lrd_4(G)}{lrd_4(E)}}{5} \\
&= \frac{\frac{0.2}{0.217} + \frac{0.2}{0.217} + \frac{0.1905}{0.217} + \frac{0.4}{0.217} + \frac{0.4}{0.217}}{5} \\
&\approx 1.2816 \\
LOF_4(K) &= \frac{\frac{lrd_4(G)}{lrd_4(K)} + \frac{lrd_4(H)}{lrd_4(K)} + \frac{lrd_4(I)}{lrd_4(K)} + \frac{lrd_4(J)}{lrd_4(K)} + \frac{lrd_4(L)}{lrd_4(K)}}{5} \\
&= \frac{\frac{0.4}{0.2778} + \frac{0.4}{0.2778} + \frac{0.444}{0.2778} + \frac{0.364}{0.2778} + \frac{0.231}{0.2778}}{5} \\
&\approx 1.324 \\
LOF_4(O) &= \frac{\frac{lrd_4(M)}{lrd_4(O)} + \frac{lrd_4(N)}{lrd_4(O)} + \frac{lrd_4(P)}{lrd_4(O)} + \frac{lrd_4(R)}{lrd_4(O)}}{4} \\
&= \frac{\frac{0.5714}{0.5714} + \frac{0.556}{0.5714} + \frac{0.556}{0.5714} + \frac{0.5714}{0.5714}}{4} \\
&\approx 0.9865
\end{aligned}
$$

Aggregated kNN distances for $k = 2$ and $k = 4$:

Recall that aggregated kNN distance = averaged sum of the distances to all the kNN.

	2NN	2d.	agg. 2d.	4NN	4d.	agg. 4d
A	B C	1	$\frac{1+1}{2} = 1$	B C D E	6	$\frac{1+1+2+6}{4} = 2.5$
B	A D	1	$\frac{1+1}{2} = 1$	A C D E	5	$\frac{1+2+1+5}{4} = 2.25$
C	A D	1	$\frac{1+1}{2} = 1$	A B D E	5	$\frac{1+2+1+5}{4} = 2.25$
D	B C	1	$\frac{1+1}{2} = 1$	A B C E	4	$\frac{2+1+1+4}{4} = 2$
E	D F	4	$\frac{4+4}{2} = 4$	B C D F G	5	$\frac{5+5+4+4+5}{5} = 4.6$
F	G H	2	$\frac{1+2}{2} = 1.5$	G H I J	3	$\frac{1+2+3+3}{4} = 2.25$
G	F H	1	$\frac{1+1}{2} = 1$	F H I J	2	$\frac{1+1+3+3}{4} = 2$
H	G I J	1	$\frac{1+1+1}{3} = 1$	F G I J	2	$\frac{2+1+1+1}{4} = 1.25$
I	G H J	2	$\frac{2+1+2}{3} \approx 1.667$	F G H J	3	$\frac{3+2+1+2}{4} = 2$
J	G H I K	2	$\frac{2+1+2+2}{4} = 1.75$	G H I K	2	$\frac{2+1+2+2}{4} = 1.75$
K	H J	3	$\frac{3+2}{2} = 2.5$	G H I J L	4	$\frac{4+3+4+2+4}{5} = 3.4$
L	K M O	4	$\frac{4+3+4}{3} \approx 3.667$	K M N O P R	5	$\frac{4+3+5+4+5+5}{6} \approx 4.333$
M	N O P R	2	$\frac{2+1+2+2}{4} = 1.75$	N O P R	2	$\frac{2+1+2+2}{4} = 1.75$
N	O Q	1	$\frac{1+1}{2} = 1$	M O P Q R	2	$\frac{2+1+2+1+2}{5} = 1.6$
O	M N P R	1	$\frac{1+1+1+1}{4} = 1$	M N P R	1	$\frac{1+1+1+1}{4} = 1$
P	O S	1	$\frac{1+1}{2} = 1$	M N O R S	2	$\frac{2+1+2+1+2}{5} = 1.6$
Q	N R	1	$\frac{1+1}{2} = 1$	N O R S T	2	$\frac{1+2+1+2+2}{5} = 1.6$
R	O Q S T	1	$\frac{1+1+1+1}{4} = 1$	O Q S T	1	$\frac{1+1+1+1}{4} = 1$
S	P R	1	$\frac{1+1}{2} = 2$	O P Q R T	2	$\frac{2+1+2+1+2}{5} = 1.6$
T	O Q R S	2	$\frac{2+2+1+2}{4} = 1.75$	O Q R S	2	$\frac{2+2+1+2}{4} = 1.75$

Exercise 9-3 Evaluation of Outlier Scores

A data set with known outliers + was evaluated using two outlier detection methods S_1 and S_2. The results of the methods are given in the table below:

Object	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
Label	−	−	−	−	+	−	−	−	+	−
S_1	1.0	1.1	1.1	1.3	3.0	2.0	1.5	0.9	1.4	1.2
S_2	.80	.80	.10	.81	.89	.50	.50	.91	.90	.20

Evaluate both outlier detection methods S_1 and S_2 using the following metrics:

- Precision, Recall and F-Measure,
  assuming that the top $k = 2$ ranked outliers were classified as outliers.
- Average Precision for $k = 1...4$,
  assuming that the top k ranked outliers were classified as outliers.
- Draw the ROC curve, and compute the area under curve (AUC) measure.

Suggested solution:

Sorted w.r.t. S_1:

Label	+	−	−	+	−	−	−	−	−	−
S_1	3.0	2.0	1.5	1.4	1.3	1.2	1.1	1.1	1.0	0.9

Sorted w.r.t. S_2:

Label	−	+	+	−	−	−	−	−	−	−
S_2	.91	.90	.89	.81	.80	.80	.50	.50	.20	.10

Precision, Recall and F-Measure:

$$\text{Precision}(h,i) = \frac{|\{o \in h_i | h(o) = f(o)\}|}{|h_i|}$$

$$\text{Precision}(S_1) = \frac{1}{2}$$

$$\text{Precision}(S_2) = \frac{1}{2}$$

$$\text{Recall}(h,i) = \frac{|\{o \in f_i | h(o) = f(o)\}|}{|f_i|}$$

$$\text{Recall}(S_1) = \frac{1}{2}$$

$$\text{Recall}(S_2) = \frac{1}{2}$$

$$F_1(h,i) = \frac{2 \cdot \text{Recall}(h,i) \cdot \text{Precision}(h,i)}{\text{Recall}(h,i) + \text{Precision}(h,i)}$$

$$F_1(S_1) = \frac{2 \cdot \frac{1}{2} \cdot \frac{1}{2}}{1} = \frac{0.5}{1} = \frac{1}{2}$$

$$F_1(S_2) = \frac{2 \cdot \frac{1}{2} \cdot \frac{1}{2}}{1}$$
$$= \frac{0.5}{1} = \frac{1}{2}$$

Average Precision for $k = 1...4$:

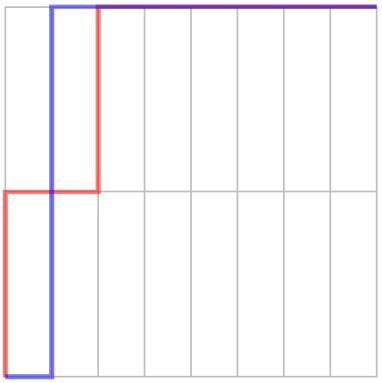
$S_1 : \frac{1+\frac{1}{2}+\frac{1}{3}+\frac{2}{4}}{4} = \frac{7}{12}$

$S_2 : \frac{0+\frac{1}{2}+\frac{2}{3}+\frac{2}{4}}{4} = \frac{5}{12}$

ROC curves:

Recall for ROC curves:

- each TP in the ranking: one step up
- each FP in the ranking: one step to the right
- comparison of two rankings: area under the curve (ROC AUC)



Area in both cases: $\frac{14}{16} = 0.875$

Exercise 9-4 Decision Trees

Predict the risk class of a car driver based on the following attributes:

- Time since getting the driving license ($1-2$ years, $2-7$ years, >7 years)
- Gender (male, female)
- Residential area (urban, rural)

For your analysis you have the following manually classified training examples:

Person	Time since license	Gender	Area	Risk class
1	$1-2$	m	urban	low
2	$2-7$	m	rural	high
3	>7	f	rural	low
4	$1-2$	f	rural	high
5	>7	m	rural	high
6	$1-2$	m	rural	high
7	$2-7$	f	urban	low
8	$2-7$	m	urban	low

1. Construct a decision tree based on this training dataset. Use information gain for selecting the split attributes. Build a separate branch for each attribute. The decision tree shall stop when all instances in the branch have the same class, you do not need to apply a pruning algorithm.

   Suggested solution:

   Remember: split of T by selection of attribute A in partitions $T_1 \dots T_m$:

   $$entropy(T) = -\sum_{i=1}^{k} p_i \cdot \log_2 p_i$$

   where p_i is the probability of randomly selecting an example in class i, and k is the number of classes.

   $$\text{information-gain}(T, A) = entropy(T) - \sum_{i=1}^{m} \frac{|T_i|}{|T|} entropy(T_i)$$

   We can see that $\frac{4}{8}$ are `low` and $\frac{4}{8}$ are `high`. Calculating the entropy, we get

   $$entropy(T) = -\left(\left(\frac{4}{8} \cdot \log_2 \frac{4}{8}\right) + \left(\frac{4}{8} \cdot \log_2 \frac{4}{8}\right)\right) = 1$$

   Now we want to calculate the information gain for each attribute.

   - IG *'Time since license'*
     - 1-2 years: $T_1 =$ persons 1, 4, 6

     $$p(R = low) = \frac{1}{3}$$
     $$p(R = high) = \frac{2}{3}$$
     $$entropy(T_1) = -\left(\left(\frac{1}{3} \cdot \log_2 \frac{1}{3}\right) + \left(\frac{2}{3} \cdot \log_2 \frac{2}{3}\right)\right)$$
     $$\approx 0.918$$

– 2-7 years: T_2 = persons 2, 7, 8

$$\begin{aligned} p(R = low) &= \frac{2}{3} \\ p(R = high) &= \frac{1}{3} \\ entropy(T_2) &= entropy(T_1) \\ &\approx 0.918 \end{aligned}$$

– > 7 years: T_3 = persons 3, 5

$$\begin{aligned} p(R = low) &= \frac{1}{2} \\ p(R = high) &= \frac{1}{2} \\ entropy(T_3) &= -\left(\left(\frac{1}{2} \cdot \log_2 \frac{1}{2}\right) + \left(\frac{1}{2} \cdot \log_2 \frac{1}{2}\right)\right) \\ &= 1 \end{aligned}$$

Thus, we can calculate the information-gain for *'Time since license'*.

$$\begin{aligned} \text{information-gain}(T, time) &= entropy(T) - \sum_{i=1,2,3} \frac{|T_i|}{|T|} entropy(T_i) \\ &= 1 - \left(\frac{3}{8} \cdot 0.918 + \frac{3}{8} \cdot 0.918 + \frac{2}{8} \cdot 1\right) \\ &\approx 0.06 \end{aligned}$$

· IG *'Gender'*

– m: T_1 = persons 1, 2, 5, 6, 8

$$\begin{aligned} p(R = low) &= \frac{2}{5} \\ p(R = high) &= \frac{3}{5} \\ entropy(T_1) &\approx 0.971 \end{aligned}$$

– f: T_2 = persons 3, 4, 7

$$\begin{aligned} p(R = low) &= \frac{2}{3} \\ p(R = high) &= \frac{1}{3} \\ entropy(T_2) &= 0.918 \end{aligned}$$

Thus, we can calculate the information-gain for *'Gender'*.

$$\begin{aligned} \text{information-gain}(T, gender) &= entropy(T) - \sum_{i=1,2} \frac{|T_i|}{|T|} entropy(T_i) \\ &= 1 - \left(\frac{5}{8} \cdot 0.971 + \frac{3}{8} \cdot 0.918\right) \\ &\approx 0.05 \end{aligned}$$

· IG *'Area'*

– urban: T_1 = persons 1, 7, 8

$$\begin{aligned} p(R = low) &= 1 \\ p(R = high) &= 0 \\ entropy(T_1) &= 0 \end{aligned}$$

– rural: T_2 = persons 2, 3, 4, 5, 6

$$\begin{aligned} p(R = low) &= \frac{1}{5} \\ p(R = high) &= \frac{4}{5} \\ entropy(T_2) &\approx 0.722 \end{aligned}$$

Thus, we can calculate the information-gain for *'Area'*.

$$\begin{aligned} \text{information-gain}(T, area) &= entropy(T) - \sum_{i=1,2} \frac{|T_i|}{|T|} entropy(T_i) \\ &= 1 - \left(\frac{3}{8} \cdot 0 + \frac{5}{8} \cdot 0.722 \right) \\ &\approx 0.55 \end{aligned}$$

Area has the largest information-gain.

Split 2, right branch: $T = \{2, 3, 4, 5, 6\}$
$entropy(T) = -\left(\left(\frac{1}{5} \cdot \log_2 \frac{1}{5} \right) + \left(\frac{4}{5} \cdot \log_2 \frac{4}{5} \right) \right) \approx 0.722$

· IG *'Time since license'*

– 1-2 years: T_1 = persons 4, 6

$$\begin{aligned} p(R = high) &= 1 \\ entropy(T_1) &= 0 \end{aligned}$$

– 2-7 years: T_2 = person 2

$$\begin{aligned} p(R = high) &= 1 \\ entropy(T_2) &= 0 \end{aligned}$$

– $>$ 7 years: T_3 = persons 3, 5

$$\begin{aligned} p(R = low) &= \frac{1}{2} \\ p(R = high) &= \frac{1}{2} \\ entropy(T_3) &= 1 \end{aligned}$$

$$\begin{aligned}\text{information-gain}(T, time) &= entropy(T) - \sum_{i=1,2,3} \frac{|T_i|}{|T|} entropy(T_i) \\ &= 0.722 - \left(0 + 0 + \frac{2}{5} \cdot 1\right) \\ &= 0.322\end{aligned}$$

- IG *'Gender'*
  - m: T_1 = persons 2, 5, 6

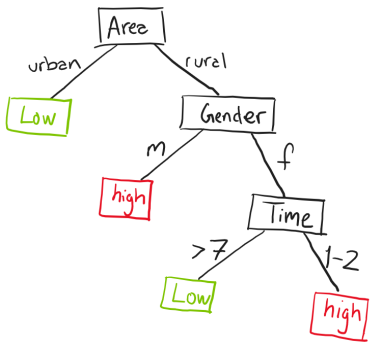
$$\begin{aligned}p(R = high) &= 1 \\ entropy(T_1) &= 0\end{aligned}$$

  - f: T_2 = persons 3, 4

$$\begin{aligned}p(R = low) &= \frac{1}{2} \\ p(R = high) &= \frac{1}{2} \\ entropy(T_2) &= 1\end{aligned}$$

$$\begin{aligned}\text{information-gain}(T, gender) &= entropy(T) - \sum_{i=1,2} \frac{|T_i|}{|T|} entropy(T_i) \\ &= 0.722 - \left(0 + \frac{2}{5} \cdot 1\right) \\ &= 0.322 \\ &= \text{information-gain}(T, time)\end{aligned}$$

You can choose one of the two attributes arbitrarily, and sketch the resulting tree.



2. Apply the decision tree to the following drivers:
   Person A: 1 – 2, f, rural
   Person B: 2 – 7, m , urban
   Person C: 1 – 2, f, urban

   Suggested solution:
   Person A: 1 – 2, f, rural → high
   Person B: 2 – 7, m , urban → low
   Person C: 1 – 2, f, urban → low