Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

Intelligent Systems

Exercise 9 - Classification and Anomalies

Simon Reichhuber

January 06, 2019

University of Kiel, Winter Term 2019

Table of Content

DBSCAN and Outlier Detection

A. **Calculate the *Local Outlier Factor* (*LOF*) of the points A_1 and N in the figure**

B. Draw the distribution of ascending minimum *kdists* of every point with $k = 1, 2, 3$.

C. How can you estimate the parameter ϵ by given a percentage of noise?

D. Find parameters $\epsilon > 0$, min_pts $\in \mathbb{N}$ s.t.
   - $A_i, i = 1, 2, 3$ is clustered as a cluster
   - $B_j, j = 1, 2$ is clustered as a cluster
   - N is marked as noise.

E. Find parameters $\epsilon > 0$, min_pts $\in \mathbb{N}$, and points C_k s.t.
   - $A_i, i = 1, 2, 3, B_j, j = 1, 2$ is clustered as a cluster.
   - N is marked as noise.

Calculate the *Local Outlier Factor* (*LOF*) of the points A_1 and N in the figure

What exactly is the *Local Outlier Factor?*

Novelty detection algorithms (11)

Christian-Albrechts-Universität zu Kiel

LOF – Equations

- k = number of neighbours considered
- $kdist(x)$ = distance to k-th neighbour
- $N_k(x) =$ ordered set of k-nearest neighbours of x
- $reachability\_dist_k(x, y) = \max\big(kdist(y), dist(x, y)\big)$
- $lrd_k(x) = \frac{k}{\sum_{y \in N_k(x)} reachability\_dist_k(x,y)}$ = local reachability density
- $LOF_k(x) = \frac{\sum_{y \in N_k(x)} \frac{lrd_k(y)}{lrd_k(x)}}{k}$

Metric undefined $\rightarrow$ Choose Euclidean metric

$dist(x, y) = ||x - y||_2$

For $k = 1$ it follows:

$$1 - dist(A_1) = 1$$

$$N_1(A_1) = \{first : A_2\}$$

$$lrd_1(A_1) = \frac{1}{\sum_{y \in N_1(A_1)} reachability\_dist_1(A_1, y)}$$

$$= \frac{1}{reachability\_dist_1(A_1, A_2)} = \frac{1}{max(1 - dist(A_2), dist(A_1, A_2))}$$

$$= \frac{1}{max(1, 1)} = 1$$

Remark:

For $k = 1, k - dist$ is noted as $1 - dist$.

$$LOF_1(A_1) = \frac{\sum_{y \in N_1(A_1)} \frac{lrd_1(y)}{lrd_1(A_1)}}{1} = \frac{1}{\sum_{y \in N_1(A_2)} reachability\_dist_1(A_2, y)}$$

$$= \frac{1}{\sum_{y \in \{first:A_1\}} reachability\_dist_1(A_2, y)}$$

$$= \frac{1}{reachability\_dist_1(A_2, A_1)}$$

$$= \frac{1}{max(1 - dist(A_1), dist(A_2, A_1))} = \frac{1}{max(1, 1)} = 1$$

Remark:

The function $reachability\_dist_k(x, y)$ is non-symmetric, because $kdist(x, y)$ is non-symmetric.

$$LOF_1(N) = \frac{\sum_{y \in N_1(N)} \frac{lrd_1(y)}{lrd_1(N)}}{1}$$

$$\sum_{y \in \{first:A_2\}} \frac{lrd_1(y)}{lrd_1(N)}$$

$$\frac{lrd_1(A_2)}{lrd_1(N)} = \frac{\sum_{y \in N_1(N)} reachability\_dist_1(N, y)}{\sum_{y \in N_1(A_2)} reachability\_dist_1(A_2, y)}$$

$$= \frac{reachability\_dist_1(N, A_2)}{reachability\_dist_1(A_2, A_1)}$$

$$= \frac{max(1 - dist(A_2), dist(N, A_2))}{max(1 - dist(A_1), dist(A_2, A_1))}$$

$$= \frac{max(1, 3)}{max(1, 1)} = 3$$

→ Because $LOF_1(A_1) = 1$ and $LOF_1(N) = 3$ it is more likely that N is an outlier.

Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

For $k = 2$ it follows:

$$2-dist(A_1) = \sqrt{2}$$

$$N_2(A_1) = \{first : A_2, second : A_3\}$$

$$lrd_2(A_1) = \frac{2}{\sum_{y \in N_2(A_1)} reachability\_dist_2(A_1, y)}$$

$$= \frac{2}{\sum_{y \in \{first:A_2, second:A_3\}} reachability\_dist_2(A_1, y)}$$

$$= \frac{2}{reachability\_dist_2(A_1, A_2) + reachability\_dist_2(A_1, A_3)}$$

$$= \frac{2}{max(2-dist(A_2), dist(A_1, A_2)) + max(2-dist(A_3), dist(A_1, A_3))}$$

$$= \frac{2}{max(1, 1) + max(\sqrt{2}, \sqrt{2})} = \frac{2}{1+\sqrt{2}} \approx 0.828$$

1. A LOF CALCUATIONS (6)

C|A|U
Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

$$
\begin{aligned}
lrd_2(A_2) &= \frac{2}{\sum_{y \in N_2(A_2)} reachability\_dist_2(A_2, y)} \\
&= \frac{2}{\sum_{y \in \{first: A_1, second: A_3\}} reachability\_dist_2(A_2, y)} \\
&= \frac{2}{max(2 - dist(A_1), dist(A_2, A_1)) + max(2 - dist(A_3), dist(A_2, A_3))} \\
&= \frac{2}{max(\sqrt{2}, 1) + max(\sqrt{2}, 1)} = \frac{1}{\sqrt{2}} \approx 0.707
\end{aligned}
$$

1. A LOF calculations (7)

$$lrd_2(A_3) = \frac{2}{\sum_{y \in N_2(A_3)} reachability\_dist_2(A_3, y)}$$

$$lrd_2(A_3) = \frac{2}{\sum_{y \in \{first: A_2, second: A_1\}} reachability\_dist_2(A_3, y)}$$

$$lrd_2(A_3) = \frac{2}{max(2 - dist(A_2), dist(A_3, A_2)) + max(2 - dist(A_1), dist(A}$$

$$= \frac{2}{max(1, 1) + max(\sqrt{2}, \sqrt{2})} = \frac{2}{1 + \sqrt{2}} \approx 0.828$$

1. A LOF calculations (8)

CAU
Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

$$
\begin{aligned}
LOF_2(A_1) &= \frac{\sum_{y \in N_2(A_1)} \frac{lrd_2(y)}{lrd_2(A_1)}}{2} \\
&= \frac{\sum_{y \in \{first:A_2, second:A_3\}} \frac{lrd_2(y)}{lrd_2(A_1)}}{2} \\
&= \frac{\frac{lrd_2(A_2)}{lrd_2(A_1)} + \frac{lrd_2(A_3)}{lrd_2(A_1)}}{2} \\
&= \frac{\frac{0.707}{0.828} + \frac{0.828}{0.828}}{2} \approx 0.439
\end{aligned}
$$

$$2-dist(N)=3$$

$$N_2(N)=\{first:A_2, second:B_1\}$$

$$lrd_2(N)=\frac{2}{\sum_{y\in N_2(N)} reachability\_dist(N,y)}$$

$$=\frac{2}{\sum_{y\in\{first:A_2,second:B_1\}} reachability\_dist_2(N,y)}$$

$$=\frac{2}{max(2-dist(A_2),dist(N,A_2))+max(2-dist(B_1),dist(N,B_1))}$$

$$=\frac{2}{max(1,3)+max(3,3)}=\frac{1}{3}$$

1. A LOF calculations (10)

$$lrd_2(A_2) \approx 0.707$$

$$lrd_2(B_1) = \frac{2}{\sum_{y \in N_2(B_1) reachability\_dist_2(B_1, y)}}$$

$$= \frac{2}{\sum_{y \in \{first:B_2, second:N\}} reachability\_dist(B_1, y)}$$

$$= \frac{2}{max(2 - dist(B_2), dist(B_1, B_2)) + max(2 - dist(N), dist(B_1, N))}$$

$$= \frac{2}{max(4, 1) + max(3, 3)} = \frac{2}{7}$$

1. A LOF CALCULATIONS (11)

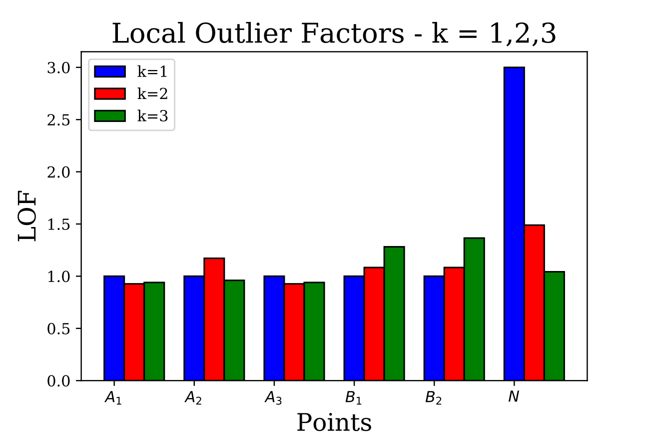
$$LOF_2(N) = \frac{\sum_{y \in N_2(N)} \frac{lrd_2(y)}{lrd_2(N)}}{2}$$

$$= \frac{\sum_{y \in \{first:A_2, second:B_1\}} \frac{lrd_2(y)}{lrd_2(N)}}{2}$$

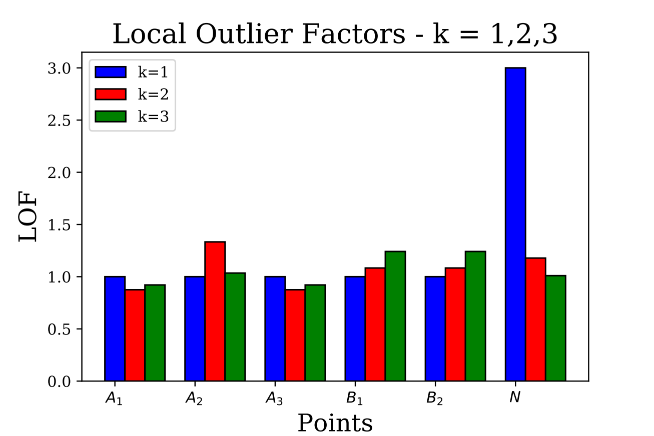
$$= \frac{\frac{0.707}{\frac{1}{3} + \frac{\frac{2}{7}}{\frac{1}{3}}}}{2} \approx 1.489$$

$\rightarrow LOF_2(A_1) < LOF_2(N)$, i.e. $0.439 < 1.489$, hence N is more likely an outlier.

Choosing Euclidean distance metric:



Choosing Euclidean distance metric:



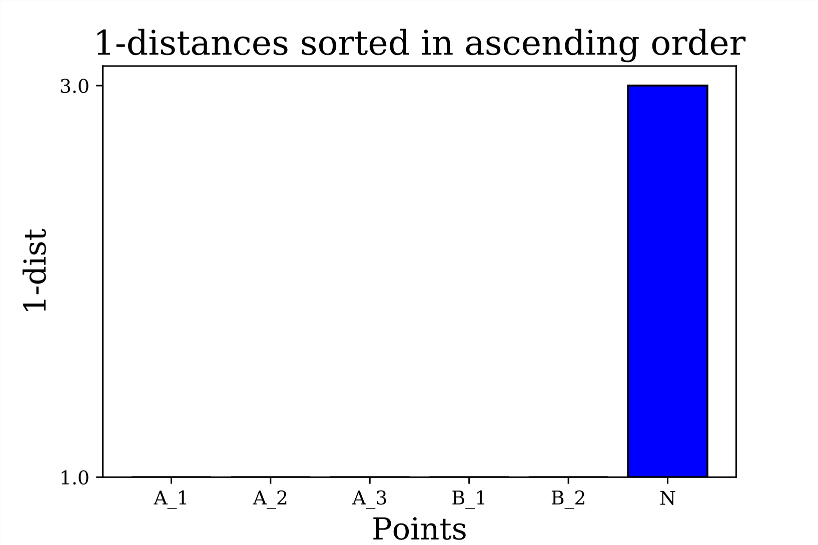
A. Calculate the *Local Outlier Factor* (*LOF*) of the points A_1 and N in the figure
B. Draw the distribution of ascending minimum *kdists* of every point with $k = 1, 2, 3$.
C. How can you estimate the parameter ϵ by given a percentage of noise?
D. Find parameters $\epsilon > 0$, min_pts $\in \mathbb{N}$ s.t.
    - $A_i, i = 1, 2, 3$ is clustered as a cluster
    - $B_j, j = 1, 2$ is clustered as a cluster
    - N is marked as noise.
E. Find parameters $\epsilon > 0$, min_pts $\in \mathbb{N}$, and points C_k s.t.
    - $A_i, i = 1, 2, 3, B_j, j = 1, 2$ is clustered as a cluster.
    - N is marked as noise.

Draw the distribution of ascending minimum *kdists*

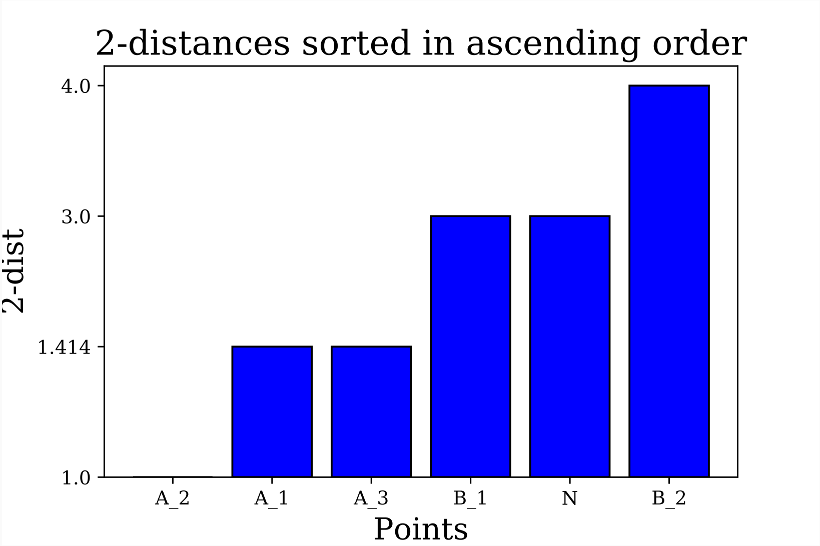
$$min_y(kdist(x, y)),$$

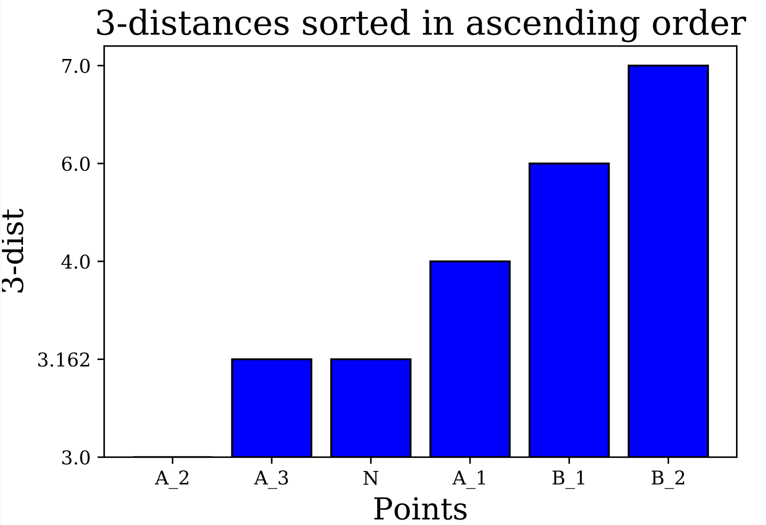
of every point with $k = 1, 2, 3$.

1. B 1-DISTANCES



Find the proportion of noise by using the ellbow method.


2-distances sorted in ascending order
2-dist
4.0
3.0
1.414
1.0
A_2
A_1
A_3
B_1
N
B_2
Points


3-distances sorted in ascending order
3-dist
7.0
6.0
4.0
3.162
3.0
A_2
A_3
N
A_1
B_1
B_2
Points

A. Calculate the *Local Outlier Factor* (*LOF*) of the points A_1 and N in the figure

B. Draw the distribution of ascending minimum *kdists* of every point with $k = 1, 2, 3$.

C. How can you estimate the parameter ϵ by given a percentage of noise?

D. Find parameters $\epsilon > 0$, min_pts $\in \mathbb{N}$ s.t.
- $A_i, i = 1, 2, 3$ is clustered as a cluster
- $B_j, j = 1, 2$ is clustered as a cluster
- N is marked as noise.

E. Find parameters $\epsilon > 0$, min_pts $\in \mathbb{N}$, and points C_k s.t.
- $A_i, i = 1, 2, 3, B_j, j = 1, 2$ is clustered as a cluster.
- N is marked as noise.

How can you estimate the parameter ϵ by given a percentage of noise?

The parameter ϵ can be determined by using the the sorted $k - distances$. Points with $k - distance > \epsilon$ will be treated as noise. In the following the noise proportion is set to $1/6$

k=1:

$\epsilon = 2$

k=2:

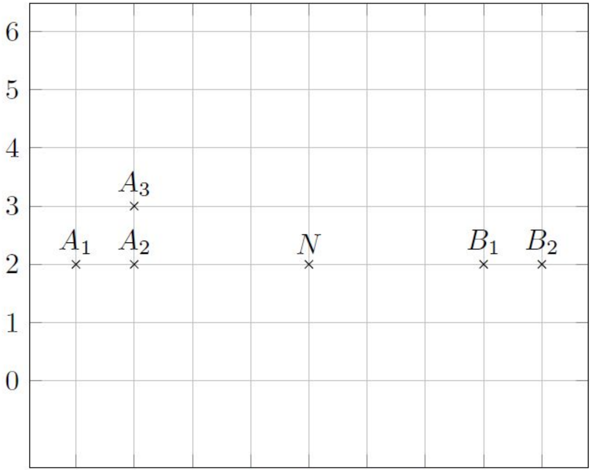
$\epsilon = 3.5$

k=3:

$\epsilon = 6.5$

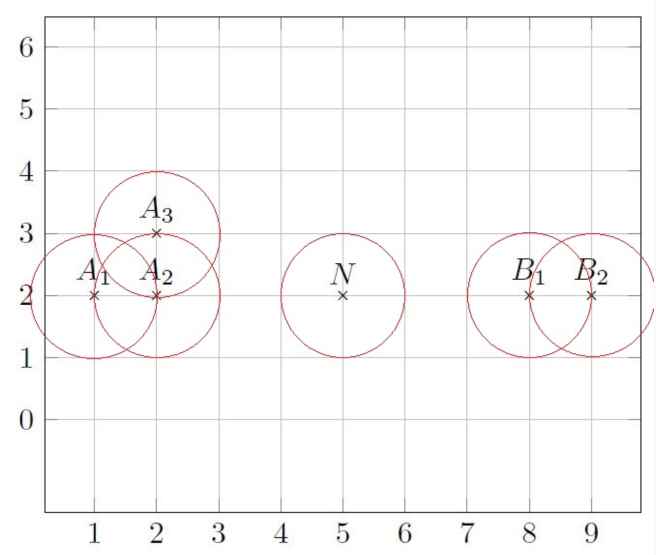
A. Calculate the *Local Outlier Factor* (*LOF*) of the points A_1 and N in the figure
B. Draw the distribution of ascending minimum *kdists* of every point with $k = 1, 2, 3$.
C. How can you estimate the parameter ϵ by given a percentage of noise?
D. Find parameters $\epsilon > 0$, min_pts $\in \mathbb{N}$ s.t.
   - $A_i, i = 1, 2, 3$ **is clustered as a cluster**
   - $B_j, j = 1, 2$ **is clustered as a cluster**
   - N **is marked as noise.**
E. Find parameters $\epsilon > 0$, min_pts $\in \mathbb{N}$, and points C_k s.t.
   - $A_i, i = 1, 2, 3, B_j, j = 1, 2$ is clustered as a cluster.
   - N is marked as noise.

Find parameters $\epsilon > 0$, min_pts $\in \mathbb{N}$ s.t.
Cluster_A $= \{A_1, A_2, A_3\}$, Cluster_B $= \{B_1, B_2\}$, and N is marked as noise.



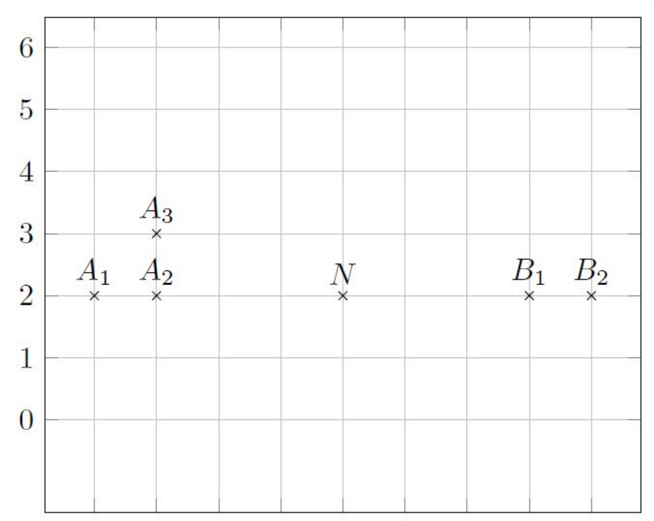
Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

Choose Euclidean metric, *min_pts* $= 1, \epsilon = 1$

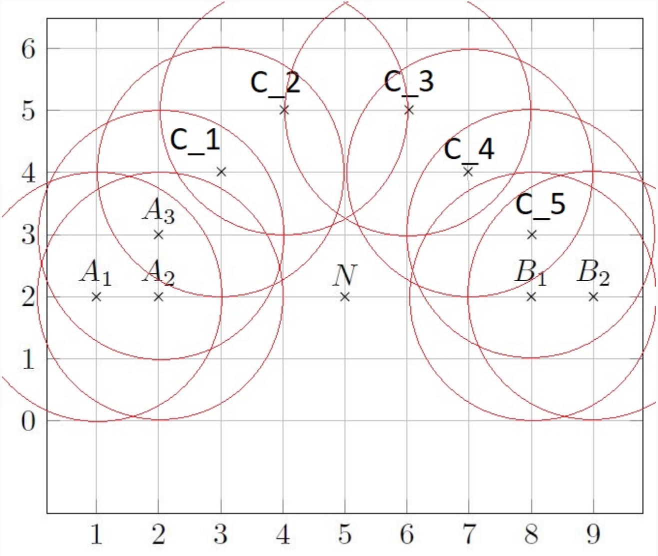


A. Calculate the *Local Outlier Factor* (*LOF*) of the points A_1 and N in the figure
B. Draw the distribution of ascending minimum *kdists* of every point with $k = 1, 2, 3$.
C. How can you estimate the parameter ϵ by given a percentage of noise?
D. Find parameters $\epsilon > 0$, min_pts $\in \mathbb{N}$ s.t.
   - $A_i, i = 1, 2, 3$ is clustered as a cluster
   - $B_j, j = 1, 2$ is clustered as a cluster
   - N is marked as noise.
E. **Find parameters $\epsilon > 0$, min_pts $\in \mathbb{N}$, and points C_k s.t.**
   - $A_i, i = 1, 2, 3, B_j, j = 1, 2$ **is clustered as a cluster.**
   - N **is marked as noise.**

Find parameters $\epsilon > 0$, min_pts $\in \mathbb{N}$, and points $C_k \in \mathbb{R}^2$ s.t. *Cluster* $= \{A_1, A_2, A_3, B_1, B_2\}$, and N is marked as noise.



Choose Euclidean metric, *min_pts* $= 2, \epsilon = 2$



Classification algorithms

A. **Observe the data set in the table. First, create a 1-R Classifier that is able to predict whether a person is going to visit the party this evening by using the information of his/her amount of money, whether he/she writes an exam tomorrow, or if his/her heartthrob will come to the party.**
B. Extend your 1-R Classifier to a Decision Tree. Which features should be placed on higher levels of the tree?
C. Apply the Naïve Bayes Classifier on the same data set. Calculate also the probabilities $P(Yes|E1)$ and $P(No|E6)$.

2. A

Observe the data set in the table. First, create a 1-R Classifier that is able to predict whether a person is going to visit the party this evening by using the information of his/her amount of money, whether he/she writes an exam tomorrow, or if his/her heartthrob will come to the party.

Sample	**Money**	**Exam**	**Heartthrob**	**Party**
E1	10	Yes	Yes	Yes
E2	13	No	Yes	Yes
E3	11	Yes	No	No
E4	12	No	No	Yes
E5	7	Yes	Yes	Yes
E6	5	Yes	No	No
E7	6	No	Yes	Yes
E8	8	No	No	No

Sample	Manifestation	Yes	No
Money	>9	3	1
Money	<=9	2	2
Exam	Ja	2	2
Exam	Nein	3	1
Heartthrob	Ja	4	0
Heartthrob	Nein	1	3

Figure 6: 1R Feature selection

Money	>9	3	1
Money	<=9	2	2
Exam	Ja	2	2
Exam	Nein	3	1
Heartthrob	Ja	4	0
Heartthrob	Nein	1	3

Figure 7: 1R Feature selection

The feature Heartthrob produces a minimal number of prediction errors (1).

Rule: If Heartthrob == ”Yes” then Party, elsif Heartthrob == ”No” then no Party.

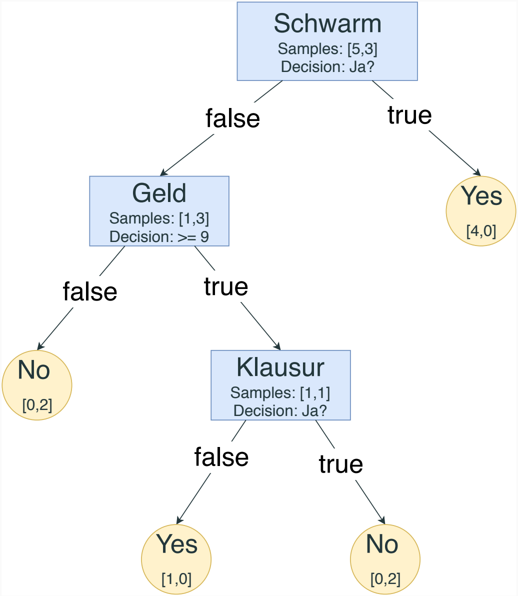
2. B

A. Observe the data set in the table. First, create a 1-R Classifier that is able to predict whether a person is going to visit the party this evening by using the information of his/her amount of money, whether he/she writes an exam tomorrow, or if his/her heartthrob will come to the party.

B. Extend your 1-R Classifier to a Decision Tree. Which features should be placed on higher levels of the tree?

C. Apply the Naïve Bayes Classifier on the same data set. Calculate also the probabilities $P(Yes|E1)$ and $P(No|E6)$.

Extend your 1-R Classifier to a Decision Tree. Which features should be placed on higher levels of the tree?


Schwarm
Samples: [5,3]
Decision: Ja?
false
true
Geld
Samples: [1,3]
Decision: >= 9
Yes
[4,0]
false
true
No
[0,2]
Klausur
Samples: [1,1]
Decision: Ja?
false
true
Yes
[1,0]
No
[0,2]

A. Observe the data set in the table. First, create a 1-R Classifier that is able to predict whether a person is going to visit the party this evening by using the information of his/her amount of money, whether he/she writes an exam tomorrow, or if his/her heartthrob will come to the party.

B. Extend your 1-R Classifier to a Decision Tree. Which features should be placed on higher levels of the tree?

C. Apply the Naïve Bayes Classifier on the same data set. Calculate also the probabilities $P(Yes|E1)$ and $P(No|E6)$.

Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

Apply the Naïve Bayes Classifier on the same data set.
Calculate also the probabilities $P(Yes|E1)$ **and** $P(No|E6)$**.**

Apply Bayes Rule on **P(Ja|E1)**=

- $\frac{P(E1|Yes)*P(Yes)}{P(E1)} = \frac{P(Money=Viel|Yes)*P(Exam=Yes|Yes)*P(Heartthrob=Yes|Yes)*P(Yes)}{P(E1|Yes)*P(Yes)+P(E1|No)*P(No)}$
- $\frac{\frac{3}{5}*\frac{2}{5}*\frac{4}{5}*\frac{5}{8}}{P(E1|Yes)+P(E1|No)*P(No)}$
- $\frac{\frac{3}{25}}{\frac{3}{25}+P(E1|No)*P(No)}$

And for **P(No|E1)** =

- $\frac{\frac{1}{3}*\frac{2}{3}*\frac{0}{3}*\frac{5}{8}}{\frac{3}{25}+P(E1|No)*P(No)}$

Zero Frequency Problem: If any manifestation is missing in the data, we virtually add an artificial value to it and assume that it occurs at least one time. Usually, the value is one.

Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

Again, use Bayes rule: **P(Yes|E1)** =

- $\frac{\frac{3+1}{5+1}*\frac{2+1}{5+1}*\frac{4+1}{5+1}*\frac{5}{8}}{P(E1|Yes)*P(Yes)+P(E1|No)*P(No)}$
- $\frac{\frac{25}{144}}{\frac{25}{144}+\frac{9}{256}} = 0.832$

And for **P(No|E1)** =

- $\frac{\frac{1+1}{3+1}*\frac{2+1}{3+1}*\frac{0+1}{3+1}*\frac{3}{8}}{\frac{25}{144}+P(E1|No)*P(No)}$
- $\frac{\frac{9}{256}}{\frac{25}{144}+\frac{9}{256}} = 0.168$

Use Bayes rule: **P(No|E6)** =

- $\frac{\frac{2+1}{3+1} * \frac{1+1}{3+1} * \frac{3+1}{3+1} * \frac{3}{8}}{P(E6|No) * P(No) + P(E6|Yes) * P(Yes)}$
- $\frac{\frac{9}{64}}{\frac{9}{64} + \frac{5}{96}} = 0.73$

And for **P(Yes|E6)** =

- $\frac{\frac{2+1}{5+1} * \frac{2+1}{5+1} * \frac{1+1}{5+1} * \frac{5}{8}}{\frac{9}{64} + P(E1|Yes) * P(Yes)}$
- $\frac{\frac{5}{96}}{\frac{9}{64} + \frac{5}{96}} = 0.27$