First name	Name	Matr.-Nr.

RWTH Aachen
Lehrstuhl für Informatik 9
Prof. Dr. van der Aalst

Trial Exam
Introduction to Data Science
January 25th, 2019

Study course:

- Diplom Informatik
- Master SSE
- Master Informatik
- Master Media Informatics
- Master Data Science
- Other: ______________________

- Duration of the exam:
- Write your first name, name and Matrikelnummer on each sheet.
- Give your solutions in readable and traceable manner. **Solutions will be graded based on completeness and correctness of description/application of the algorithm/method**.
- Give your solutions on the exam sheets only. If you need extra paper, use only paper provided by the supervisors. Make sure to specify your name and Matrikelnummer on all papers.
- Please cross out those things you do not wish to be graded.
- In case of attempted deception, your exam will be graded as **failed**.
- At the end of the exam hand in your complete copy. Do not separate any sheets by removing the staples.
- **You may only use a black or blue pen, and a basic calculator; no additional material (e.g. books, cell phones, laptop, etc.) is allowed.**
- **Only answers that are given in English will be graded.**

Signature _____________________________

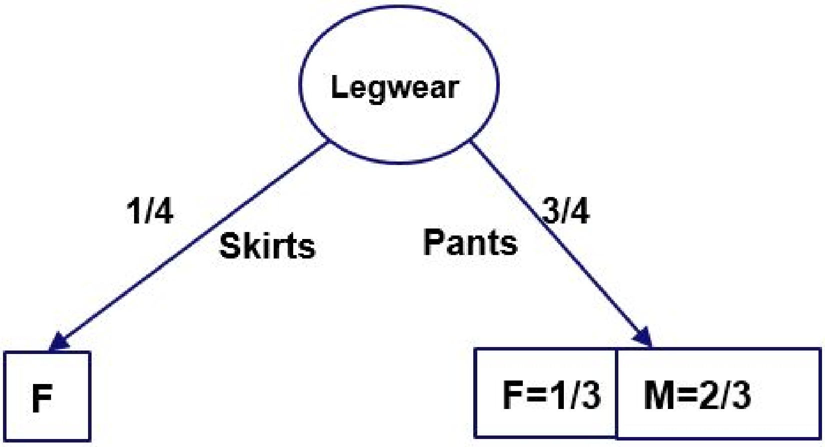
First name	Name	Matr.-Nr.

1. Decision Tree

We would like to predict the gender of a person based on two binary attributes: legwear (pants or skirts) and facial-hair (some or none). We have a data set of 1000 individuals, half male and half female. 50% of females wear skirt, and all males wear pants. 75% of males and 25% of females have facial hair. What is the information gain resulted by each attribute? And which attribute is used as the root of the decision tree when using ID3?

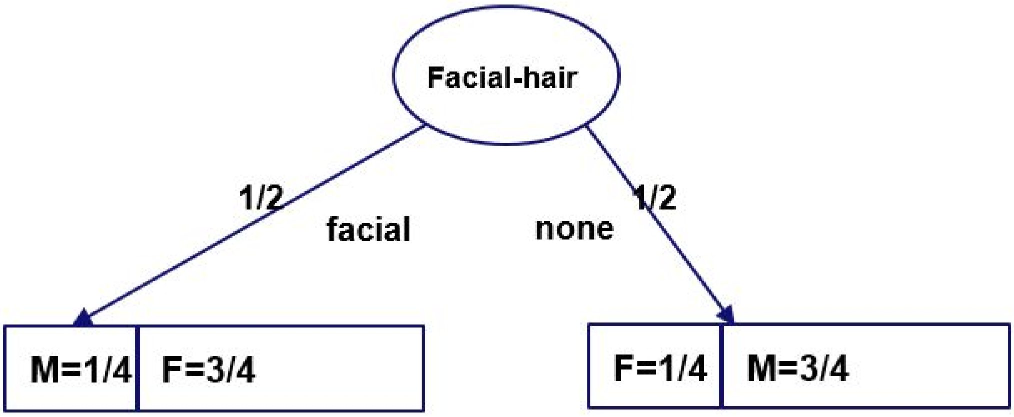
$$\log_2 \frac{1}{2} = -1 \,,\ \log_2 \frac{1}{3} = -1.58 \,,\ \log_2 \frac{2}{3} = -0.58 \,,\ \log_2 \frac{1}{4} = -2 \,,\ \log_2 \frac{3}{4} = -0.4$$

Solution:



$$Entropy = \frac{1}{4}\left(-\left(0+\frac{250}{250}log\left(\frac{250}{250}\right)\right)\right) + \frac{3}{4}\left(-\left(\frac{500}{750}log\left(\frac{500}{750}\right)+\frac{250}{750}log\left(\frac{250}{750}\right)\right)\right) = 0 + 0.69 = 0.69$$

First name	Name	Matr.-Nr.



$$
\begin{aligned}
&\mathit{Entropy} = \\
&\frac{1}{2}\left(-\left(\frac{125}{500}\mathit{log}\left(\frac{125}{500}\right)+\frac{375}{500}\mathit{lo\ g}\left(\frac{375}{500}\right)\right)\right) \\
&+\frac{1}{2}\left(-\left(\frac{375}{500}\mathit{log}\left(\frac{375}{500}\right)+\frac{125}{500}\mathit{log}\left(\frac{125}{500}\right)\right)\right) \\
&= 0.5(0.81)+0.5(0.81)=0.81
\end{aligned}
$$

First name	Name	Matr.-Nr.

2. Regression

Student	Score on trial exam	Score on final exam
1	95	85
2	85	95
3	80	70
4	70	65
5	60	70

a. You are given two linear regression functions and a dataset which shows the points five students scored for the trial and the final exam. Assume that we want to predict a students score on the final test, given the score on the trail exam. Determine which regression function fits the data better using the sum of squared errors.

Regression function 1: $y = 26 + 0.6\,x$
Regression function 2: $y = 25 + 0.7\,x$

Solution:

Student	Score on trial (*x*)	Score on final (*y*)	Predicted final score of function 1 (*y_pred*)	Error (*y-y_pred*)	Squared error
1	95	85	83	2	4
2	85	95	77	18	324
3	80	70	74	-4	16
4	70	65	68	-3	9
5	60	70	62	8	64
					Sum: 417

Student	Score on trial (*x*)	Score on final (*y*)	Predicted final score of function 1 (*y_pred*)	Error (*y-y_pred*)	Squared error

First name	Name	Matr.-Nr.

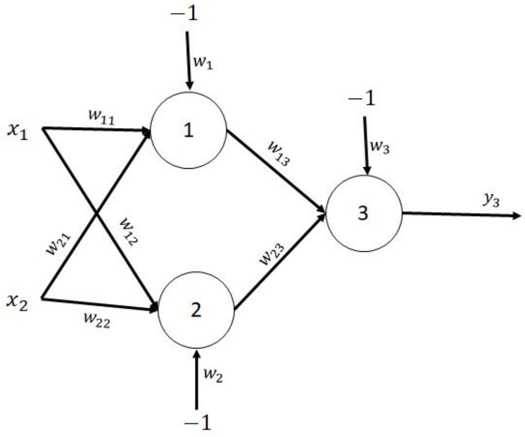
1	95	85	91.5	-6,5	42.25
2	85	95	84.5	10,5	110.25
3	80	70	81	-11	121
4	70	65	74	-9	81
5	60	70	67	3	9
					Sum: 363.5

According to the squared error, regression function number 2 better fits the data.

First name	Name	Matr.-Nr.

3. Neural Networks

Using the initial weights provided below, and the input example x_1 =0, x_2 =1, compute the output at each neuron after feed forward propagation.(Activation function is a step function with threshold zero)



Weigh t	w_1	w_{11}	w_{12}	w_{13}	w_2	w_{21}	w_{22}	w_{23}	w_3
Value	0.5	1	1	0.5	0.5	1	1	0.25	0.5

Solution:

(1) calculate $y_1 = f(-1*0.5+0*1+1*1) = 1$

(2) calculate $y_2 = f(-1*0.5+0*1+1*1) = 1$

(2) calculate $y_3 = f(-1*0.5+1*0.5+1*0.25) = 1$

4. Evaluation of Supervised Learning Algorithms

a. Compute (classwise) precision, recall and f1-score for the following confusion matrix:

		Prediction		
		Sunny	**Cloudy**	**Rainy**
Target	**Sunny**	**52**	**7**	**2**

First name	Name	Matr.-Nr.

	Cloudy	**11**	**45**	**5**
	Rainy	**3**	**6**	**48**

Solution:

Precision

Sunny: 52/(52+11+3) = 0.79

Cloudy: 45/(7+45+6) = 0.77

Rainy: 48/(2+5+48) = 0.87

Recall

Sunny: 52/(52+7+2) = 0.85

Cloudy: 45/(11+45+5) = 0.74

Rainy: 48/(3+6+48) = 0.84

F1-score

Sunny: 2*(0.79*0.85)/(0.79+0.85) = 0.82

Cloudy: 2*(0.77*0.74)/(0.77+0.74) = 0.75

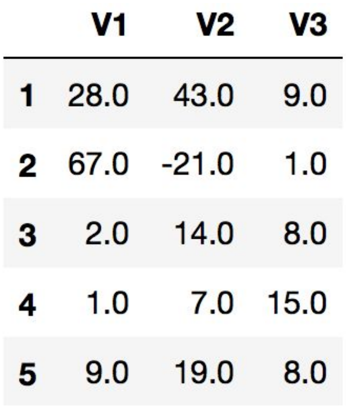
Rainy: 2*(0.87*0.84)/(0.87+0.84) = 0.85

First name	Name	Matr.-Nr.

5. Clustering

The table below shows a set of elements waiting to be clustered, where each row in this table represents an element. The first value of each row in the table represents the ID of each element. Every element in the table is characterized by three variables V1, V2 and V3.

Presume that we run k-means algorithm over this data set. Set the number of clusters to 2, i.e., there are two clusters which are cluster 1 and cluster 2. Let (10, 15, 19) be the initial centroid of cluster 1, (2, 14, 8) be the initial centroid of cluster 2. Use Euclidean Distance for the computation.



	V1	V2	V3
1	28.0	43.0	9.0
2	67.0	-21.0	1.0
3	2.0	14.0	8.0
4	1.0	7.0	15.0
5	9.0	19.0	8.0

a. Please write down the updated centroid for each cluster for each iteration. The results should be accurate up to 2 decimal places.

b. Write down which element (for each element, use its ID to represent it) has been assigned to which cluster when the k-means algorithm finishes.

First name	Name	Matr.-Nr.

Solution:

First iteration: updated centroid for cluster 1 is (47.5, 11, 5), updated centroid for cluster 2 is (4, 13.33, 10.33).
Second iteration: Nothing changes.

Cluster 1: 1, 2

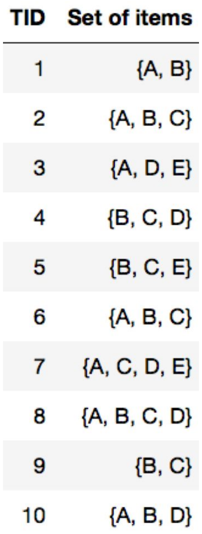
Cluster 2: 3, 4, 5

First name	Name	Matr.-Nr.

6. Frequent Items Sets and Association Rules

The table below shows a transaction set. Each set of items in the table stands for a transaction. There are totally ten transactions with transaction ID (shown by column TID) from 1 to 10. There are totally five items in the table, which are A, B, C, D, E.

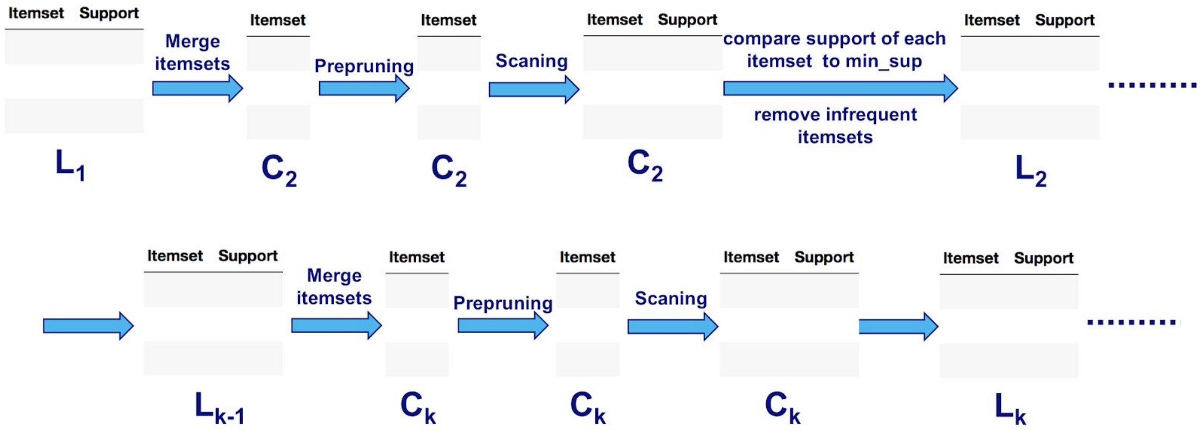
Presume that we set the absolute minimum support to 2 and run the Apriori algorithm over this transaction set for mining frequent itemsets. Let L_{k-1} ($k \geq 2$) be the frequent itemsets of length k-1 and C_k ($k \geq 2$) be the candidate itemsets of length k generated by the Apriori algorithm.



TID	Set of items
1	{A, B}
2	{A, B, C}
3	{A, D, E}
4	{B, C, D}
5	{B, C, E}
6	{A, B, C}
7	{A, C, D, E}
8	{A, B, C, D}
9	{B, C}
10	{A, B, D}

a. Please write down all the L_{k-1} ($k \geq 2$) and C_k ($k \geq 2$) that will be generated by Apriori algorithm. For the candidate itemsets. Please show all three versions of C_k generated after the itemsets merging step, the prepruning step and the scanning step as shown in the figure below.

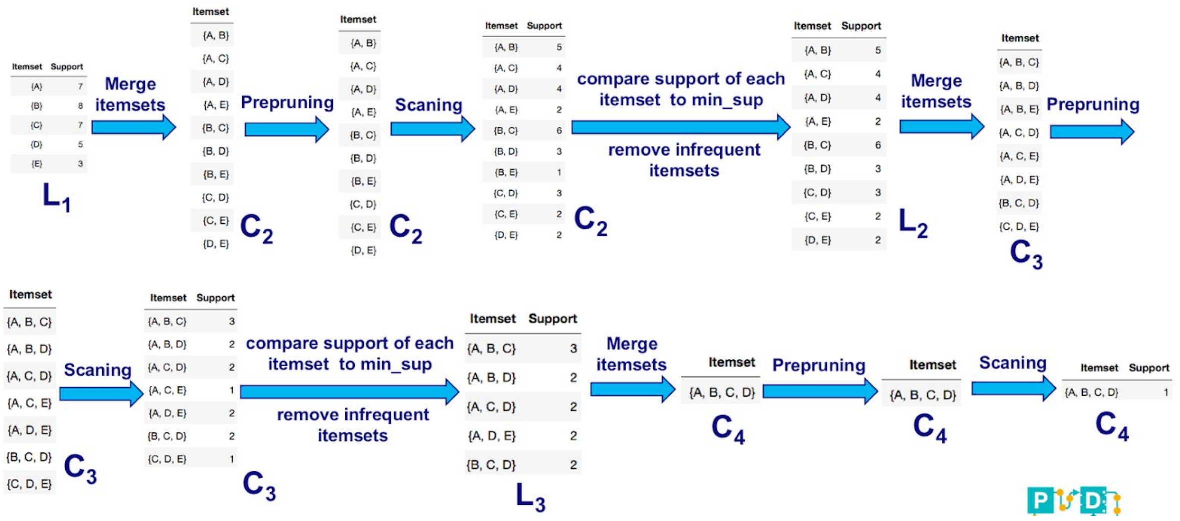
First name	Name	Matr.-Nr.



b. Presume that there are four association rules from the transaction shown in the table above, which are {A, B} => {C}, {A} => {B, D}, {B} => {C} and {B, C} => {D}. Please calculate and write down the confidence and the lift of each of the four association rules. The results should be accurate up to 3 decimal places.

First name	Name	Matr.-Nr.

Solution:



Itemset	Support
{A}	7
{B}	8
{C}	7
{D}	5
{E}	3

L_1

Merge itemsets

Itemset
{A, B}
{A, C}
{A, D}
{A, E}
{B, C}
{B, D}
{B, E}
{C, D}
{C, E}
{D, E}

C_2

Prepruning

Itemset
{A, B}
{A, C}
{A, D}
{A, E}
{B, C}
{B, D}
{B, E}
{C, D}
{C, E}
{D, E}

C_2

Scaning

Itemset	Support
{A, B}	5
{A, C}	4
{A, D}	4
{A, E}	2
{B, C}	6
{B, D}	3
{B, E}	1
{C, D}	3
{C, E}	2
{D, E}	2

C_2

compare support of each itemset to min_sup

remove infrequent itemsets

Itemset	Support
{A, B}	5
{A, C}	4
{A, D}	4
{A, E}	2
{B, C}	6
{B, D}	3
{C, D}	3
{C, E}	2
{D, E}	2

L_2

Merge itemsets

Itemset
{A, B, C}
{A, B, D}
{A, B, E}
{A, C, D}
{A, C, E}
{A, D, E}
{B, C, D}
{C, D, E}

C_3

Prepruning

Itemset
{A, B, C}
{A, B, D}
{A, C, D}
{A, C, E}
{A, D, E}
{B, C, D}
{C, D, E}

C_3

Scaning

Itemset	Support
{A, B, C}	3
{A, B, D}	2
{A, C, D}	2
{A, C, E}	1
{A, D, E}	2
{B, C, D}	2
{C, D, E}	1

C_3

compare support of each itemset to min_sup

remove infrequent itemsets

Itemset	Support
{A, B, C}	3
{A, B, D}	2
{A, C, D}	2
{A, D, E}	2
{B, C, D}	2

L_3

Merge itemsets

Itemset
{A, B, C, D}

C_4

Prepruning

Itemset
{A, B, C, D}

C_4

Scaning

Itemset	Support
{A, B, C, D}	1

C_4

Rule	Confidence	Lift
{A, B} => {C}	0.6	0.857
{A} => {B, D}	0.286	0.952
{B} => {C}	0.75	1.071
{B, C} => {D}	0.333	0.667

First name	Name	Matr.-Nr.

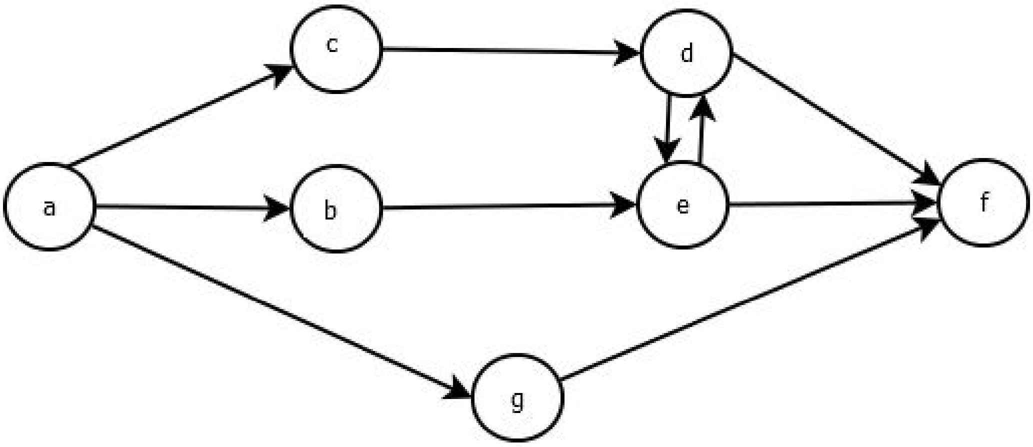
7. Process Mining

Consider the following event log: L = [<a,c,d,e,f>, <a,b,e,d,f>, <a,g,f>].

a. Draw the directly follow graph of the log, indicating the first cut of the Inductive Miner.

b. Perform Process Discovery with the Inductive Miner on the event log and obtain a process tree.

Solution:



$T = \rightarrow(a, \times(\rightarrow(\times(b, c), \wedge(d, e)), g), f)$

First name	Name	Matr.-Nr.

8. Text Mining

a. Given the following document corpus, find the tf-idf scoring of the two given queries for each document. Plural words normalize on the singular form.
Hint: recall that the tf-idf scoring for a query on a document is the sum of the tf-idf scores of each individual word of the query for that document.

D1: "Process Mining is a discipline that perform analysis on process data."
D2: "Process Mining is a subfield of Data Science."
D3: "Data Science provides methods and techniques to obtain analysis from data."
D4: "Process Mining relates to other Process Science disciplines like Business Process Management."
D5: "Process Mining and Data Mining are related disciplines."
Q1: science disciplines
Q2: process data discovery

Solution:

Scores Q1:

D1 = 0.74

D2 = 0.74

D3 = 0.74

D4 = 1.47

D5 = 0.74

Scores Q2:

D1 = 0.96

D2 = 0.64

D3 = 0.64

D4 = 0.96

D5 = 0.64

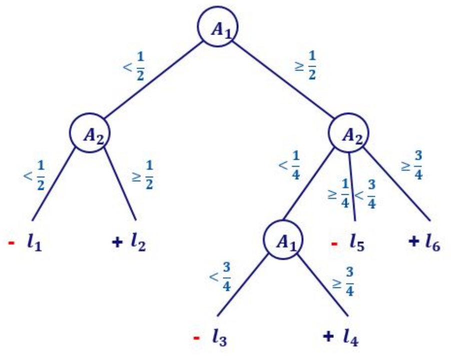
First name	Name	Matr.-Nr.

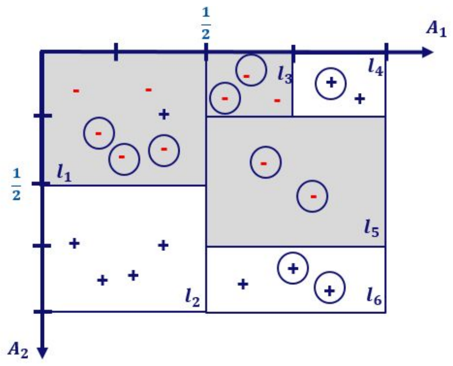
9. Responsible Data Science

In the following DT classifier, relabeling which leaf leads to the maximal reduction on the discrimination when $disc_B(C, D)$ is the discrimination measure?

Note that encircled examples are discriminatory (have B=1)

$$disc_B(C, D) := \frac{|\{x \in D \mid x.B = 0, C(x) = +\}|}{|\{x \in D \mid x.B = 0\}|} - \frac{|\{x \in D \mid x.B = 1, C(x) = +\}|}{|\{x \in D \mid x.B = 1\}|}$$





Solution:

Class	-	+	
$B = 1$	u	v	b
$B = 0$	w	x	$\bar{b}$
	n	p	a

$$If\ p < n$$

$$\Delta disc_l = -\frac{u + v}{b} + \frac{w + x}{\bar{b}}$$

$$If\ p > n$$

$$\Delta disc_l = \frac{u + v}{b} - \frac{w + x}{\bar{b}}$$

First name	Name	Matr.-Nr.

Since the effect on discrimination is just a difference between discriminatory and non-discriminatory instances, they need to find the biggest negative difference. Here the biggest negative difference causes by l_2.

$$\Delta disc_l = \frac{0}{\frac{1}{2}} - \frac{\frac{4}{20}}{\frac{1}{2}} = -0.4$$