First name: _____

Last name: _____

Student ID number: | | | | | | | | |

2nd Midterm Exam

course name

INTRODUCTION TO MACHINE LEARNING AND DATA MINING

Instructions:

- Write your **FIRST NAME**, **LAST NAME** and **STUDENT ID NO.** on each piece of paper with solutions;
- This midterm is composed of **5 assignments** for the total amount of **100 points**;
- Solving time is **90 minutes**;
- Only a calculator and 1 piece of paper (A4 format) – with written notes and formulas is allowed;
- All other literature, the use of Internet, laptops, mobile phones and other electronic devices is strictly forbidden!

Koper, 16th of January, 2020

Learning set:

ID	Att_1	Att_2	Att_3	Att_4	Cls
100	2000.000	n	–	x	T
101	2004.410	n	+	o	F
102	2016.623	y	+	x	F
103	2015.735	n	+	o	T
104	2002.042	n	–	x	F
105	2006.153	y	–	x	F
106	2020.000	n	–	o	T
107	2010.666	y	+	o	F
108	2013.177	y	+	x	T
109	2019.186	n	+	o	T
110	2008.797	y	–	o	F

ID: Identifier [0, ∞)

Att_1: Date in KSP format

Att_2: Nominal value {y, n}

Att_3: Nominal value {+, –}

Att_4: Nominal value {o, x}

Cls: Class, nominal value {T, F}

Test set:

ID	Att_1	Att_2	Att_3	Att_4	Cls
200	2017.115	n	+	?	T
201	2015.428	y	–	?	F
202	2011.141	?	–	x	F

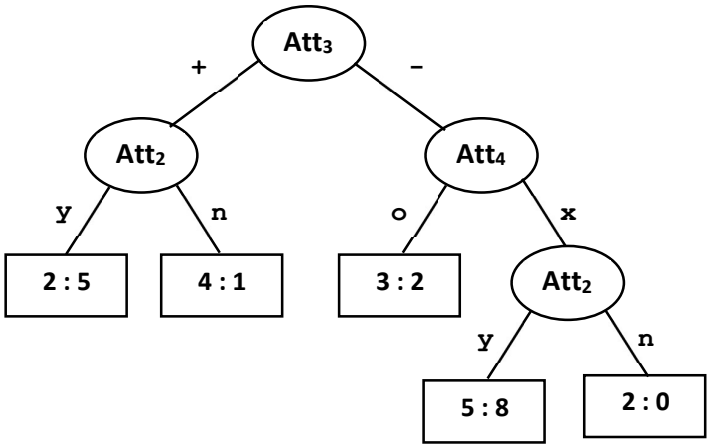
1. Decision trees

Classify the examples from the test set by using the given decision tree (depicted below)!

The distribution of examples in the leaves of this decision tree is given in the form **#(T) : #(F)**.

Write the complete probability distribution for each example in the test set!

(10 points)



2. Classification rules

The first classification rule for class value **T** »found« by the PRISM algorithm on the learning set, using just the attributes **Att_2**, **Att_3** and **Att_4**, is:

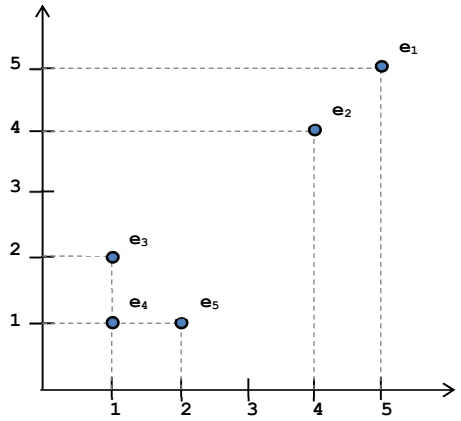
IF (Att_2 = n) AND (Att_4 = o) AND (Att_3 = -) THEN T

Find the next (second) classification rule for class value **T** on the learning set, using the same three attributes!

(20 points)

3. Clustering

Simulate the hierarchical agglomerative clustering algorithm on the set of data {e_1, e_2, e_3, e_4, e_5}, depicted in the figure below. Use the *Manhattan distance* to compute the distances between data points.



a) Use the complete linkage method when calculating distances between a single element and a group; write down each step of the agglomeration by using distance tables and draw the complete dendrogram!

(15 points)

b) How many groups does the agglomerative clustering algorithm discover? Write them down!

(5 points)

4. Nearest neighbours

Now, let **Att_2**, **Att_3**, **Att_4** and **Cls** be the attributes, and **Att_1** be the class. Use the *nearest neighbours* method (***k–NN***) with parameter ***k*** set to ***4*** (***k = 4***) to classify the (below) given example! Use the *Euclidean distance* to calculate the »vicinity« to other examples from the learning set.

(15 points)

ID	Att_1	Att_2	Att_3	Att_4	Cls
300		n	+	o	T

5. Association rules

The (below) given table represents market baskets of 5 randomly chosen customers that bought products in a supermarket. For the sake of clarity, only the initials of the product names are listed. They are sorted alphabetically and put into columns for easier counting.

TID	Products
22149	A, B, C, L, P, Q, U, X
33277	A, B, C, P, R, X
44305	A, C, T, V, X
57423	A, P, V, Z
58511	A, B, P, V, X

a) Find all *k-itemsets* with <u>support</u> at least **60%**! How many are there?
(10 points)

b) What is the total number of association rules that can be generated from the itemsets that you found in assignment 4.a)?
(5 points)

c) Which association rules generated from the itemset **{A, B, P}** have <u>confidence</u> at least **80%**?
(10 points)

d) Which of the association rules that you found in assignment 4.c) have »enough« <u>lift</u>?
(10 points)