OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG
INF FAKULTÄT FÜR INFORMATIK

Department of Knowledge Processing
and Language Engineering
Computational Intelligence
Prof. Dr. R. Kruse, Chr. Braune M.Sc.

Magdeburg, 2012-07-23

Written exam "Intelligent Data Analysis"

Name, first name:	Faculty:	Course:	Matriculation no.:
Type of exam: ☐ First attempt ☐ Second attempt ☐ Certificate	Signature invigilator:		#Sheets:

Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Sum
/10	/20	/10	/20	/20	/20	/100

Task 1 Probability Theory (*Points:* 10 12 min)

a) Due to a recent study 1% of women have breast cancer (and therefore 99% do not). Breast cancer can be detected by a mammography with chance of 80% if the woman actually has cancer. But the test also detects cancer in 9.6% of all scanned women, even if they are healthy. What is the actual chance of a woman having breast cancer when she got a positive screening result?

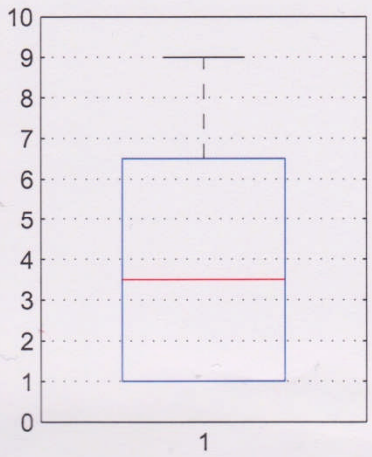
Task 2 **Descriptive Statistics** (*Points:* $6 + 4 + 10 = 20$ **24 min**)

a) Given the data set below calculate the minimal and maximal value, the arithmetic mean, the median and the mode for the attribute x. Draw a boxplot for the attribute y and a scatter plot for the data set.

x	1	1	2	2	2	3	4	4	4	5	5
y	1	2	1	2	3	3	3	4	5	4	5

b) Calculate the covariance matrix for the data set given in the previous assignment. What is the value for skewness for both attributes?

c) The boxplot on the right side describes a data set of 12 integer points. In addition to the information available from the boxplot you know that the arithmetic mean of the data set is $\overline{x} = 4.0$. The $1/3$-quantile is 1.5 and the $2/3$-quantile is 5.5. Using the biased estimator for the variance (which divides by n and not $n-1$) of the data set you know that $\sigma^2 = 25/3$. With these information given, calculate the data set, that is the source of this information.



Task 3 **Regression** (*Points:* 6 + 4 = 10 12 min)

a) Given the data set below, calculate a regression line using the method of least squares. Calculate the mean squared error as well.

x	1	1	2	2	2	3	4	4	4	5	5
y	1	2	1	2	3	3	3	4	5	4	5

b) Assume that the point $(x, y) = (20, 0)$ also belongs to the data set above. To what does the regression line change? What is the problem?

Task 4 Clustering (*Points:* 12 + 8 = 20 24 min)

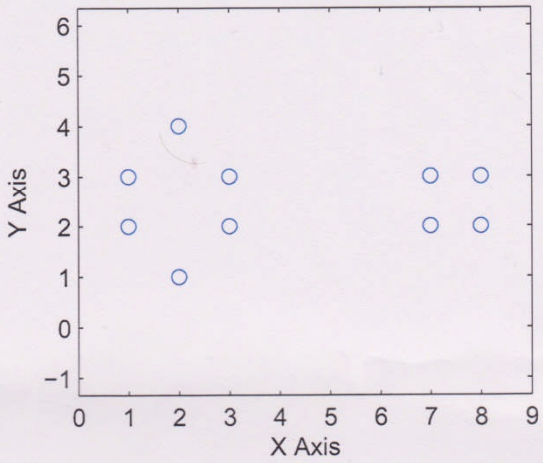
a) Find a clustering for the data set below with the k-means algorithm.

x	1	1	2	2	2	3	4	4	4	5	5
y	1	2	1	2	3	3	3	4	5	4	5

Use $k = 2$ and the Manhattan distance to calculate your result. As initial cluster prototypes use:

x	5	4
y	5	3

b) Consider the following data set for a k-means algorithm with initial prototypes $c_1 = (2, 1)$ and $c_2 = (2, 4)$ and euclidean distance. What will the resulting clustering look like, what problems of the k-means algorithm becomes obvious? How can we cope with this problem? What other problems and solutions for these can you think of?



Task 5 **Frequent Pattern Mining** (*Points:* 14 + 2 + 4 = 20 24 min)

a) Given the transaction database below, find all frequent ($s_{min} = 3$) item sets using the apriori algorithm!

b) Which of these item sets are closed?

c) Which of these item sets are maximal?

t_{ID}	items
1:	{ a,b,e }
2:	{ b,c,d }
3:	{ b,c,e }
4:	{ a,b,c,d }
5:	{ b,e }
6:	{ b,c,d }
7:	{ a,b,d,e }
8:	{ a,b,c,d }
9:	{ a,b,e }
10:	{ b,d,e }
11:	{ b,c }

Task 6 Decision Trees - Induction and Pruning (*Points:* 13 + 7 = 20 24 min)

a) Consider the data set you already know from the previous tasks, now extended by a class label. Induce a decision tree using the *rate of correctly classified example cases* as evaluation measure until no more misclassifications are made! In case of a tie (a split it both attributes is possible) choose x as attribute to use in the current node, but never use the same attribute twice in a row.

x	1	1	2	2	2	3	4	4	4	5	5
y	1	2	1	2	3	3	3	4	5	4	5
class	A	A	B	B	B	B	A	B	A	B	A

b) Prune the decision tree induced in the previous task using pessimistic pruning. Assume that you make 0.5 additional errors in each leaf. What does the tree look like?