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

Magdeburg, 2013-07-22

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	Sum
/15	/15	/15	/15	/15	/75

Task 1 Probability Theory (*Points:* 10 + 5 = 15 24 min)

a) According to popular belief you can hear the ocean, when you hold a seashell to your ear. Assume you went for a random walk somewhere on earth and found a seashell that you are now holding to your ear and you hear the ocean (S). You know that seashells can be found basically anywhere but are most frequently found on beaches, i.e. the chance of finding a seashell at a beach is 99%, the chance of finding a seashell anywhere else are just 5%. According to *The World Factbook* the world's coastline is $356'000km$ long, while the total land area is $148'940'000km^2$. Assume that on average the width of a beach area is exactly $\frac{7447}{890}km$. Use this information for calculating your chances of being at a beach (B).What are the chances you are actually at a beach while hearing the sea? Hint: Calculate the result using fractions to make the calculation easier.

b) Draw a Bayes network describing the dependencies of the above task.

Task 2 Descriptive Statistics (*Points:* 12 + 3 = 15 24 min)

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

x	2	2	4	4	8	8	10	10
y	10	8	10	8	2	4	2	4

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

Task 3 Regression (*Points:* 10 + 5 = 15 24 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	2	2	4	4	8	8	10	10
y	10	8	10	8	2	4	2	4

b) Assume that the point $(x, y) = (20, 200)$ also belongs to the data set above. In which way does the regression line change? What is the problem? How can we cope with such problems?

Task 4 Clustering (*Points:* 8 + 3 + 4 = 15 24 min)

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

x	2	2	4	4	8	8	10	10
y	10	8	10	8	2	4	2	4

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

x	2	4
y	10	8

b) Assume that, after you finished the above clustering, an additional point $(5, 7)$ is added to the data set. Now, instead of k-means, perform one step of the fuzzy-c-means algorithm, calculating only the membership of the new point to the previously found cluster centers.

c) Besides partioning cluster algorithms such as k-means, which other classes of clustering algorithms exist and what are some of their advantages/disadvantages.

Task 5 **Decision Trees** (***Points:* 10 + 3 + 2 = 15** **24 min)**

a) Consider the data set you already know from the previous tasks, now extended by a class label and four additional points! Induce a binary decision tree using the *rate of correctly classified example cases* as evaluation measure! In case of a tie (a split in both attributes is possible) choose x as attribute to use in the current node.

x	2	2	2	4	4	4	8	8	8	10	10	10
y	4	8	10	2	8	10	2	4	10	2	4	8
class	A	A	A	B	A	A	B	B	A	B	B	B

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 resulting tree look like?

c) Why is the resulting tree a bad predictor? What would a better predictor look like and how can it be found?