Department of Intelligent
Cooperative Systems
Computational Intelligence Group
Prof. Dr. Rudolf Kruse, Christian Braune M.Sc.

Magdeburg, 2016-03-01

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	/12	/14	/11	/8	/60

Task 1 Descriptive and Inductive Statistics (*Points:* 8 + 5 + 2 = 15)

a) Consider the following data set. 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 data set.

x	1	1	2	3	4	5	6	7	8	9	10	10

b) Assume the data follow a normal distribution. Calculate all the necessary parameters and rescale the data so it follows a standard normal distribution.
(**Hint:** Please use just the integer part of any value you calculate for rescaling.)

c) A point can be considered an outlier, if its absolute standardized value (z-score) is larger than 2.5. Decide for the two points 26 and 14 whether they are outliers or not.

Task 2 **Classification** (***Points:*** $4 + 8 = 12$)

a) Explain the difference between a Naive Bayes Classifier and a Full Bayes Classifier.

b) Consider the following four emails you may have received:

1. Regularly paying too much for free trials?
2. Exercise as a chance for your free vehicle.
3. I have just as much fun as I need.
4. Now you have a chance to tell your girlfriend!

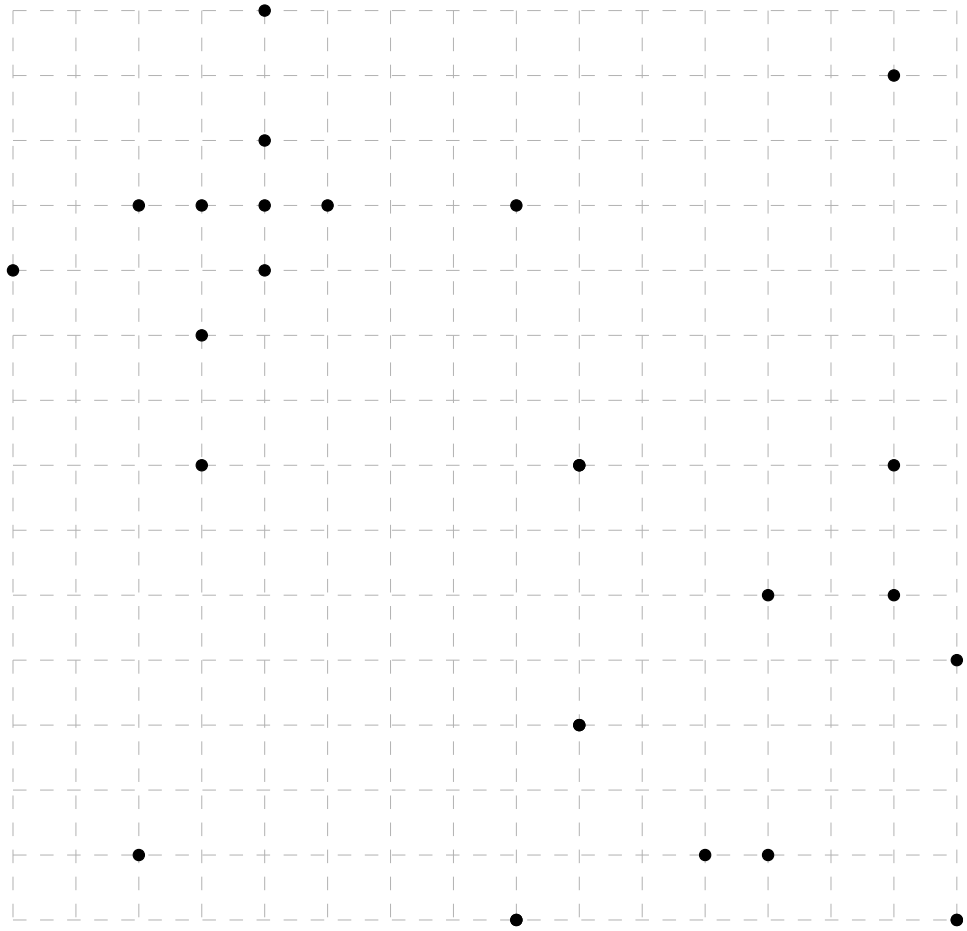
From previous emails you know, that certain words appear with different frequencies in either spam or non-spam emails. With this information use a Naive Bayes Classifier to classify each email as either spam ($C = S$) or no-spam ($C = N$). The a-priori probability for SPAM is 90%.
(**Hint:** For this case you only need to know whether $\frac{P(C=S)}{P(C=N)} \geq 1$ or $\frac{P(C=S)}{P(C=N)} \leq 1$ holds.)

Feature	Appearances in Spam	Appearances in Non-Spam
a	165	1320
advised	8	40
as	2	600
chance	55	50
clarins	2	8
exercise	6	42
for	378	2268
free	362	181
fun	48	6
girlfriend	24	6
have	392	1960
her	36	108
i	7	1435
just	204	272
much	181	362
now	183	366
paying	36	12
receive	218	109
regularly	13	78
take	132	264
tell	90	100
the	135	810
time	264	396
to	443	2215
too	59	177
trial	30	15
vehicle	35	70
viagra	34	2
you	332	664
your	440	550

Task 3 Clustering (*Points:* 6 + 8 = 14)

a) Describe the fuzzy-c-means-algorithm. Explain the role of each parameter in the clustering process and how clusters are finally found!

b) Cluster the following data set with DBSCAN and the parameters $\epsilon = 2.5$ and $minPts = 4$! Use the maximum norm as distance measure. It is sufficient to mark all core points with a circle ($\circ$) and all noise points with a cross ($\times$). Draw a line around each cluster you find!



Assume that each square is exactly 1 unit wide and tall.

Task 4 Linear Regression (*Points:* 2 + 1 + 8 = 11)

a) Describe in general how function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ can be derived via regression given a set of data points.

b) Consider the special case, where $n = 1$ and $m \geq 2$. How can you easily calculate the (linear) regression function(s)?

c) Consider the following data set. Fit an appropriate function of the form $\vec{z} = \vec{a} + \vec{b}\vec{x}$

x	0	1	2	3	4	5	6	7
z_1	0	2	2	4	4	5	6	7
z_2	-1	-1	-2	-4	-5	-7	-6	-5

Task 5 Miscellaneous (*Points:* 2 + 1½ + 2½ + 2 = 8)

For each question you will get ½ point for every correctly ticked answer and lose ½ point for every wrong answer. If you refuse to answer a question, you will neither gain nor lose points. For each subtask you will get at least 0 points.

a) **Statistics**

yes	no	
☐	☐	The variance of an estimator can be used to determine its sufficiency.
☐	☐	$\sigma^2(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is an unbiased estimator for the standard deviation of a normal distribution.
☐	☐	A Maximum-Likelihood-Estimator chooses the parameters that make the sample most likely.
☐	☐	If the null hypothesis H_0 has been rejected, the alternative H_1 has to be true.

b) **Classification & Regression**

yes	no	
☐	☐	Decision boundaries for support vector machines are always parallel to at least one data coordinate axis.
☐	☐	Decision boundaries for naive bayes classifiers are always parallel to at least one data coordinate axis.
☐	☐	Decision boundaries for decision trees are always parallel to at least one data coordinate axis.

c) **Clustering**

Assume that all data points are distinct, i.e. $\forall x_i, x_j \in \mathcal{D}, i \neq j : d(x_i, x_j) \neq 0, |\mathcal{D}| = n$.

yes	no	
☐	☐	The value of the objective function of k-means is strictly monotonically increasing when k is increased up to n (given that the optimal value for each k has been found).
☐	☐	*Average linkage* joins two clusters if their centroids are the ones that are closest to each other (given some distance measure).
☐	☐	*Complete linkage* joins two clusters if their distance is the maximal distance over the minimum of all pairwise distances between the points of all pairs of clusters.
☐	☐	For every $\omega \in [1, 2)$ used as fuzzifier in the fuzzy-c-means algorithm, fuzzy-c-means turns into k-means.
☐	☐	Hierarchical clustering can only be meaningfully applied, if euclidean distance or manhattan distance is used.

d) **Model Selection**

yes	no	
☐	☐	*Project understanding* is the first phase in the CRISP-DM Model.
☐	☐	Bootstrapping can be used to *generate* new data sets to estimate the variance of a model .
☐	☐	For ten-fold cross-validation we randomly choose ten data points as training points and use a model learned from them to classify the remaining data.
☐	☐	The principle of Occam's razor states that the simplest model is always correct.

Check boxes like this ☑ or this ☒. Correct mistakes like this: ■.