Re-exam Data Mining, IBI008

April 9, 2020

- Write your name and student number on every sheet.
- Use **paper and pen** (not pencil), as if you're taking an actual non-digital exam: we will not accept typed answers.
- The number between brackets before every (sub)question specifies the maximum number of points to be earned (150 in total).
- Make sure to **explain your answers**, even when not explicitly asked for. You can do this with catchwords ("steekwoorden"), but do make sure that your answer is readable!
- The exam is **open book**: you're allowed to use any printed and written material you like. You're not allowed to make use of digital devices, apart from a non-graphical calculator.
- By taking the exam, **you declare that no plagiarism is or will be committed**. If we have the suspicion that fraud has been committed, we will contact you and, if needed, redirect your case to the Examination Board.
- During the exam, you **stay connected** to the Zoom meeting, with your camera and microphone on (you may want to turn your volume off, not be disturbed by sound from others).
- You can **communicate through private chat** with the teacher. If Zoom fails, you immediately send an email to `tomh@cs.ru.nl`.
- You have **3 hours** to make the exam, unless of course you are entitled to extra time.
- When you're done, you make a **picture or scan** of all your sheets and upload these to Brightspace.
- After you've managed to do this, notify the teacher through chat. He can then check that your sheets are fine and sign you off. Until you're signed off, you **stay connected** to the Zoom meeting.
- Make sure to **keep your sheets**, including scrap papers, so that we can collect them later if needed.
- **Good luck!**

No.	Attribute description	Abbrev.
x_1	Room temperature	Temperature
x_2	Room humidity	Humidity
x_3	Light intensity in room	Light
x_4	CO2 concentration in room	CO2
x_5	Feature-transformed variable	HumidityRatio
y	Is the room occupied?	Occupancy

Table 1: Attributes of the Occupancy dataset. The dataset contains 8143 measurements of rooms made over time, summarized in 5 input attributes $x_1, \ldots, x_5$, and one output variable y, indicating whether the room is occupied or not.

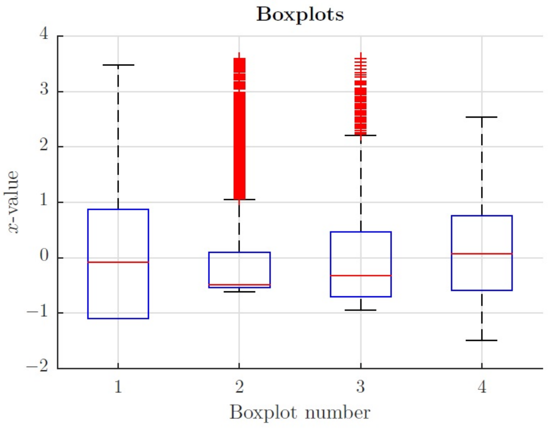


Figure 1: Boxplots corresponding to the attributes x_1, x_2, x_3, x_4 of the Occupancy dataset from Table 1, but not necessarily in that order. All attributes have been standardized.

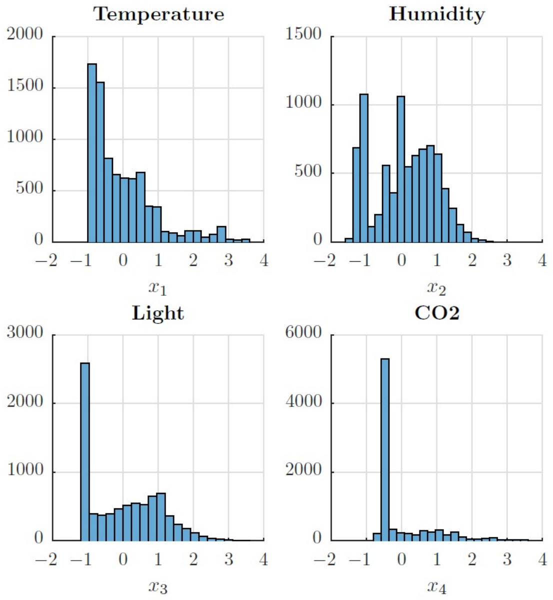


Figure 2: Histograms of four attributes of the Occupancy dataset from Table 1.

1. [**12**] Figures 1 and 2 give boxplots and histograms, respectively, of four attributes of the Occupancy dataset from Table 1. The attributes have been standardized for this question. Each histogram corresponds to one of the boxplots.

   Specify for each boxplot (1 through 4) to which histogram (x_1 through x_4) it corresponds. Explain your answer.

2. A principal component analysis is carried out on the Occupancy dataset based on the attributes $x_1, \ldots, x_5$ found in Table 1. We standardize the data by subtracting the mean from each attribute and dividing each attribute by its standard deviation to form the standardized data matrix $\tilde{\mathbf{X}}$ of size 8143 rooms × 5 attributes and apply a singular value decomposition $\tilde{\mathbf{X}} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ where

$$\mathbf{S} = \begin{pmatrix} 149 & 0 & 0 & 0 & 0 \\ 0 & 118 & 0 & 0 & 0 \\ 0 & 0 & 53 & 0 & 0 \\ 0 & 0 & 0 & 42 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix} \quad \text{and} \quad \mathbf{V} = \begin{pmatrix} -0.3 & -0.5 & 0.7 & 0.2 & 0.2 \\ -0.4 & 0.6 & -0.0 & 0.2 & 0.7 \\ -0.4 & -0.4 & 0.7 & 0.4 & -0.0 \\ -0.6 & -0.1 & -0.1 & -0.8 & 0.1 \\ -0.5 & 0.4 & 0.2 & 0.2 & -0.7 \end{pmatrix}.$$

$\mathbf{S}$ and $\mathbf{V}$ have been rounded to zero and one decimal, respectively. The columns of the matrix $\mathbf{V}$ are the principal components and the diagonal elements of the matrix $\mathbf{S}$ give the standard deviation of the data in the direction of these principal components. The variance explained by each principal component is proportional to the square of the corresponding diagonal element in $\mathbf{S}$.

a) [**5**] What percentage of the variance is explained by the first two principal components? Provide your calculation.

b) [**5**] The first principal component clearly discriminates between warm, humid, light rooms with a high CO2 concentration and high humidity ratio versus cold, dry, dark rooms with a low CO2 concentration and low humidity ratio. Along the same lines: which, if any, of the principal components primarily discriminates between warm, dark rooms versus cool, light rooms? Explain your answer.

x	2	4	8	11	15	19	20	27	30	31

Table 2: A one-dimensional dataset with 10 observations.

3. [**12**] We wish to apply K-means clustering to the dataset with $N = 10$ observations shown in Table 2. The K = 3 one-dimensional cluster centers are initialized to the first three data points, i.e., $m_1 = x_1 = 2$, $m_2 = x_2 = 4$ and $m_3 = x_3 = 8$. K-means clustering iteratively assigns all data points to their nearest cluster center and then updates each cluster center to the mean of the data points assigned to it.

What are the final (approximate) cluster centers m_1, m_2, and m_3 after termination of the K-means clustering algorithm? Clearly indicate each of the intermediate results (clusterings and cluster centers) obtained by the K-means algorithm.

4. Binarizing observations from the Occupancy dataset, we arrive at two binary vectors $\mathbf{o}_1 = (0,1,1,0,1)$ and $\mathbf{o}_2 = (0,0,1,0,0)$. Recall the definitions for the Simple Matching Coefficient,

$$\text{SMC}(\mathbf{x},\mathbf{y}) = \frac{\text{Number of matching attribute values}}{\text{Number of attributes}} ,$$

the Jaccard similarity,

$$J(\mathbf{x},\mathbf{y}) = \frac{\text{Number of matching presences}}{\text{Number of attributes not involved in 00 matches}}$$

and the cosine similarity

$$\cos(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} .$$

Compute

a) **[2]** $SMC(o_1, o_2)$,

b) **[2]** $J(o_1, o_2)$,

c) **[2]** $\cos(o_1, o_2)$.

Indicate for each of the similarity measures

d) **[2]** $SMC(\cdot,\cdot)$,

e) **[2]** $J(\cdot,\cdot)$,

f) **[2]** $\cos(\cdot,\cdot)$,

whether it is suited for computing the similarity between two long, sparse binary vectors. Explain your answers.

5. A database contains the following information about companies in the Netherlands; the number of employees in the company (denoted *Employees*, e.g., 80), the revenue of the company (denoted *Revenue*, e.g., 120.000 euro), the tax registration number of the company (denoted *VAT*, e.g., NL000099998B57), and the year the company was founded (denoted *Year*, e.g., 1987).

Indicate and explain for each of the variables

a) **[3]** *Employees*,

b) **[3]** *Revenue*,

c) **[3]** *VAT*,

d) **[3]** *Year*

whether it is nominal, ordinal, interval, or ratio.

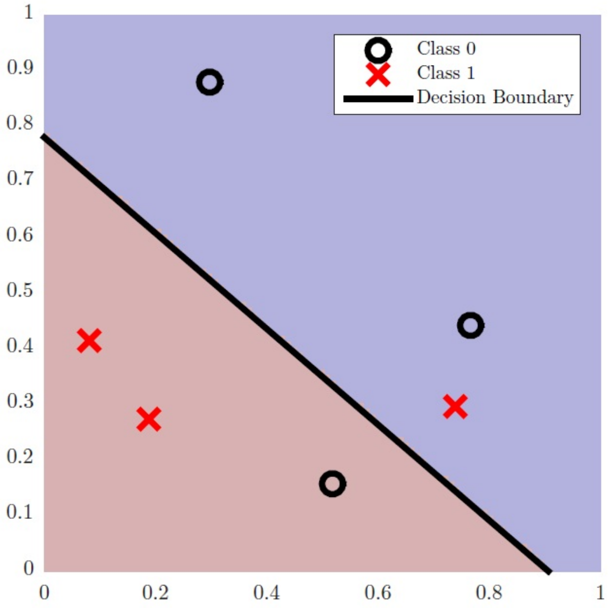


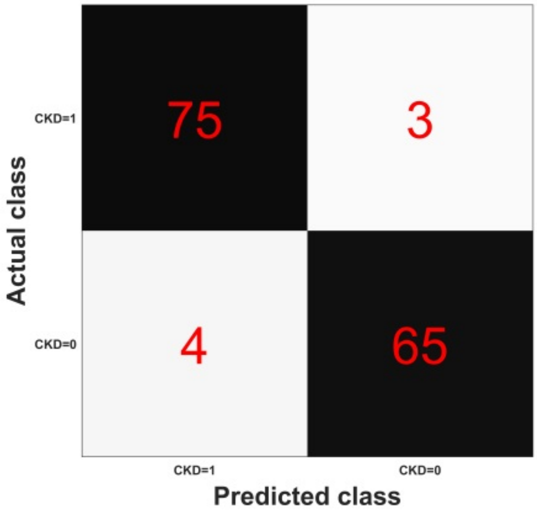
Figure 3: A binary classification problem and the decision boundary obtained by a logistic regression classifier. Observations left of the boundary are classified as belonging to the positive class 1 (x) and observations right of the boundary to the negative class 0 (○).

6. We consider the logistic regression classifier in Figure 3. The black line indicates the decision boundary obtained by thresholding at 0.5 when trained on a small 2-class dataset composed of a negative class (black circles) and a positive class (red crosses). The observations to the left of the boundary are classified in the positive class and to the right of the boundary in the negative class. By lowering the threshold, the decision boundary moves up and to the right, whereas by increasing the threshold, the decision boundary moves down and to the left. The slope of the decision boundary always stays the same.

   Recall that the true positive rate (TPR) gives the number of correctly classified positive examples divided by the total number of positive examples. The false positive rate (FPR) gives the number of negative examples incorrectly classified to be positive divided by the total number of negative examples.

   a) [**2**] Compute the TPR and the FPR given the decision boundary in Figure 3.

   b) [**8**] Draw the ROC (receiver operating characteristic) curve by plotting the TPR as a function of the FPR when the decision boundary moves from the lower left corner to the upper right corner.



Actual class
Predicted class

	CKD=1	CKD=0
CKD=1	75	3
CKD=0	4	65

Figure 4: Confusion matrix of a logistic regression classifier tested on part of the Chronic Kidney Disease dataset.

7. The confusion matrix of a logistic regression classifier evaluated on test data from the Chronic Kidney Disease dataset is given in Figure 4. To generate the confusion matrix, data has been split in two: half of the data is used for training the classifier and the other half for testing. We will consider CKD $= 1$ as the positive class and CKD $= 0$ as the negative class of the classifier. Precision measures how precise the classifier is in predicting the positive class. Recall considers how many of the positive subjects (i.e., patients) are predicted to be positive.

   Compute

   a) [**3**] the accuracy,

   b) [**3**] the error rate,

   c) [**3**] precision,

   d) [**3**] recall

   of this classifier.

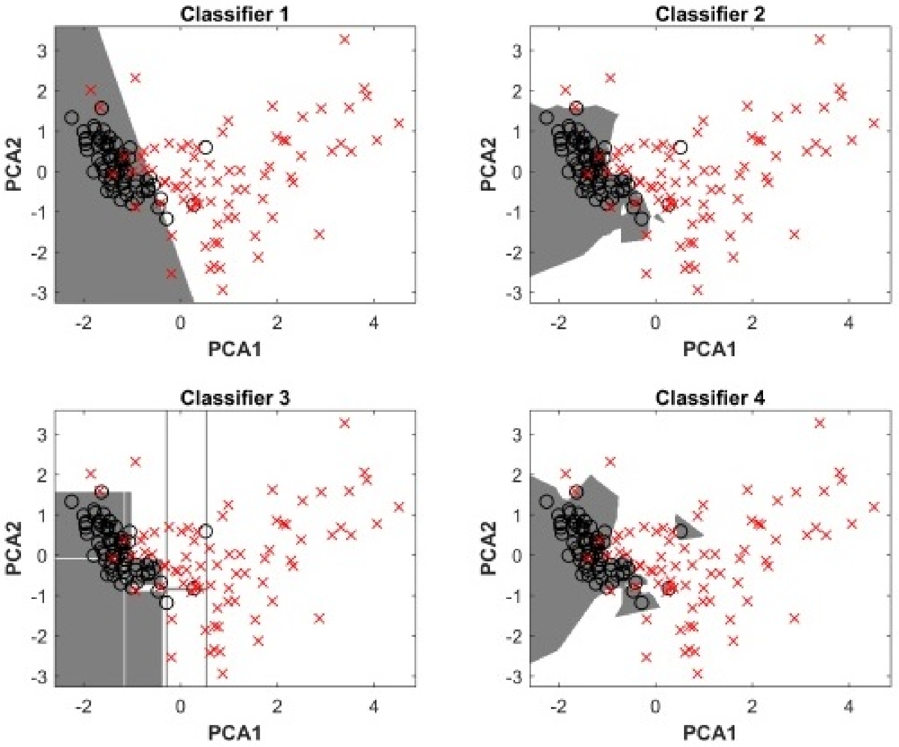


Figure 5: Four classifiers trained on part of the Chronic Kidney Disease dataset using the first two principal components as inputs for each classifier.

8. [**16**] Figure 5 shows the decision boundaries obtained by training four different classifiers on part of the so-called Chronic Kidney Disease dataset, where projections onto the first two principal components are taken as the input of the four classifiers. Circles (healthy controls: CKD = 0) and crosses (patients: CKD = 1) show the training objects. The gray area corresponds to the part of the input space that is classified as zero.

   Each of these subplots is one of the following four classifiers, in alphabetic order.

   DecTree: a decision tree,

   LogReg: a logistic regression model (equivalent to an artificial neural network without any hidden units),

   OneNN: a one-nearest-neighbor classifier,

   ThreeNN: a three-nearest-neighbor classifier.

   Specify for each of Classifiers 1 through 4 in the different subplots which of the above listed classifiers it corresponds to. Explain your answer.

	RBC	PC	PCC	HTN	DM	CAD	PE
O_1	0	0	0	0	1	0	0
O_2	0	1	1	1	0	0	1
O_3	0	0	0	0	0	0	0
O_4	0	1	0	0	1	0	1
O_5	0	1	1	1	1	0	0
O_6	1	1	1	1	1	0	0
O_7	1	1	1	1	1	0	1
O_8	0	1	1	1	1	1	1
O_9	0	1	0	1	0	0	0
O_{10}	1	1	0	0	0	1	0
O_{11}	0	0	0	0	1	0	0
O_{12}	0	0	0	0	0	0	0
O_{13}	0	0	0	0	0	0	0
O_{14}	0	0	0	0	0	0	0
O_{15}	0	0	0	0	0	0	0

Table 3: Dataset on chronic kidney disease with $N = 15$ subjects and $M = 7$ binary features. O_1 through O_9 do have chronic kidney disease (CKD = 1), O_{10} through O_{15} do not.

9. We consider the fifteen subjects given in Table 3. We will treat this dataset as a market basket problem in which the subjects have various combinations of the seven items denoted RBC, PC, PCC, HTN, DM, CAD, PE. We apply the Apriori algorithm to the dataset in Table 3 with the so-called minsup threshold set to 30%, so that only itemsets with support above of 30% or higher are considered frequent. The Apriori algorithm uses a "generate-and-count" strategy. Candidate itemsets of size $k + 1$ are generated by combining two frequent itemsets of size k. A candidate itemset is pruned when one or more of its (other) subsets happens to be infrequent. If a candidate itemset cannot be pruned, the algorithm goes once again through the dataset to compute the support of the candidate itemset and to check whether it is (still) frequent.

   a) [**9**] List all frequent itemsets with support higher than 30% found by the Apriori algorithm.

   b) [**3**] Which **candidate** itemsets of size 3 are generated by the Apriori algorithm?

   c) [**6**] Indicate for each of these candidate itemsets whether it is pruned, found to be infrequent (but not pruned), or found to be frequent. Explain your answer.

10. [**10**] Nine of the fifteen subjects in Table 3 have chronic kidney disease (O_1 through O_9 given in blue) whereas six of the observations do not have chronic kidney disease (O_{10} through O_{15} given in black). We would like to predict whether a subject has chronic kidney disease or not using the data in Table 3 and the attributes RBC, PC, DM, and CAD in column 1, 2, 5, and 6. We will apply a naïve Bayes classifier that uses simple frequency estimates (instead of the more involved Laplace estimates) for estimating the required probabilities and conditional probabilities.

    Recall that the naïve Bayes classifier applies Bayes' rule,

    $$P(y|x_1,\ldots,x_n) = \frac{P(x_1,\ldots,x_n|y)P(y)}{P(x_1,\ldots,x_n|y=0)P(y=0)+P(x_1,\ldots,x_n|y=1)P(y=1)},$$

    for arbitrary values of the binary class y and attributes x_1 through x_n. It further makes the naïve assumption that the attributes are conditionally independent given the class, i.e., that

    $$P(x_1,\ldots,x_n|y) = P(x_1|y) \times \ldots \times P(x_n|y) \,,$$

    and estimates the conditional probabilities $P(x_1|y)$ through $P(x_n|y)$ as well as the probability $P(y)$ from the available data.

    Given a subject with RBC = 1, PC = 1, DM = 1, and CAD = 1, what is the (approximate) probability that this subject has chronic kidney disease according to the naïve Bayes classifier, i.e., what is $P(\mathrm{CKD}=1|\mathrm{RBC}=1,\mathrm{PC}=1,\mathrm{DM}=1,\mathrm{CAD}=1)$?

	O1	O2	O3	O4	O5	O6	O7
O1	0	69	55	117	50	326	36
O2	69	0	36	128	104	303	85
O3	55	36	0	129	94	314	78
O4	117	128	129	0	85	220	91
O5	50	104	94	85	0	303	23
O6	326	303	314	220	303	0	307
O7	36	85	78	91	23	307	0

Table 4: Pairwise Euclidean distance between 7 subjects in the Chronic Kidney Disease dataset.

11. [**14**] Table 4 gives the Euclidean distances between 7 subjects in a study on chronic kidney disease. Hierarchical clustering starts with all subjects having their own group and then consecutively merges pairs of groups into a new group. After each merge, the distances of the already existing groups to this new group are calculated. How these are computed depends on the type of "linkage". With **complete** (also called maximum) linkage, the distance between two groups of subjects is defined to be the maximum distance between the subjects in the two groups.

    Apply hierarchical clustering to cluster these 7 subjects with **complete** linkage. Draw a dendrogram that clearly shows which (groups of) subjects are merged. Give for each merge the distance at which they are merged, either by clearly annotating this in your dendrogram or by specifying this separately.

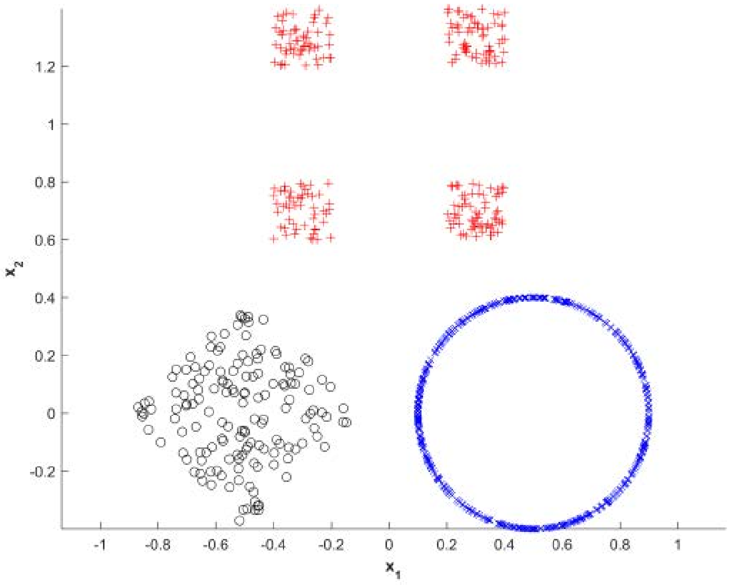


Figure 6: A classification problem consisting of three classes: red plusses (+), black circles (∘), and blue crosses (x).

12. We consider the data given in Figure 6 containing three classes given by red plusses (+), black circles (∘) and blue crosses (x).

    Using standard Euclidean distance as proximity measure, indicate for each of the following clustering approaches whether it will flawlessly separate the data into the three classes +, ∘, and x. Explain your answers.

    a) **[4]** Hierarchical clustering using single linkage.

    b) **[4]** K-means with $K = 3$ clusters (when properly initialized).

    c) **[4]** DBSCAN (with a clever setting of its parameters).

DONE!

Make a picture or scan of your answer sheets, check that these are sufficiently readable, preferably combine them in a single pdf, and then upload this to Brightspace. Stay connected to Zoom until you received confirmation that your sheets are fine.

You are not allowed to distribute the exam, so delete it from your computer. We will make it available again for exam inspection.