Univ.-Prof. Dr. rer. nat. Rudolf Mathar

1	2	3	4	$\sum$
15	15	15	15	60

Written Examination

Fundamentals of Big Data Analytics

Monday, August 20, 2018, 11:00 a.m.

Name: ____________________ Matr.-No.: ____________

Field of study: ________________________

Please pay attention to the following:

1) The exam consists of **4 problems**. Please check the completeness of your copy. **Only** written solutions on these sheets will be considered. Removing the staples is **not** allowed.

2) The exam is passed with at least **30 points**.

3) You are free in choosing the order of working on the problems. Your solution shall clearly show the approach and intermediate arguments.

4) **Admitted materials:** The sheets handed out with the exam and a non-programmable calculator.

5) The results will be published on Monday evening, the 27.08.18, on the homepage of the institute.

The corrected exams can be inspected on Friday, 31.08.18, 10:00h. at the seminar room 333 of the Chair for Theoretical Information Technology, Kopernikusstr. 16.

Acknowledged: ________________________

(Signature)

Problem 1. (15 points)
Principal Component Analysis (PCA):
Assume that $\mathbf{A}$ is given by:

$$\mathbf{A} = \begin{pmatrix} -2 \\ 1 \\ 0 \\ 2 \end{pmatrix} \begin{pmatrix} -2 & 1 & 0 & 2 \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \\ 2 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & 2 & 2 & 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 2 \\ 2 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & 2 & 2 & 0 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & 2 & 2 & 0 \end{pmatrix}$$

a) What is the rank of $\mathbf{A}$? (1P)

b) Calculate the spectral decomposition $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}$ of $\mathbf{A}$ by determining the matrices $\mathbf{V}$ and $\mathbf{\Lambda}$. (4P)

c) Assume that $\mathbf{A}$ is a sample covariance matrix. Determine the projection matrix $\mathbf{Q}$ of the PCA to transform four-dimensional samples to one dimension. (2P)

Let $\mathbf{S}_n$ be the sample covariance matrix of n points $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^4$. Assume that it has the spectral decomposition $\mathbf{S}_n = \tilde{\mathbf{V}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{V}}^{\mathrm{T}}$ where

$$\tilde{\mathbf{V}} = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 & \mathbf{v}_4 \end{pmatrix}, \tilde{\mathbf{\Lambda}} = \begin{pmatrix} 5 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}$$

and $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4 \in \mathbb{R}^4$.

d) Given $\mathbf{v}_1 = \frac{1}{3}\begin{pmatrix} 2 & 1 & 0 & 2 \end{pmatrix}^T$ and $\mathbf{v}_2 = \frac{1}{3}\begin{pmatrix} -1 & 2 & 2 & 0 \end{pmatrix}^T$, visualize the following points in a 2D graph using PCA (4P)

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}.$$

Let $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2}{2\varepsilon})$ be the dissimilarity function used for Multidimensional Scaling (MDS).

e) Assume that $\mathbf{x}_i \neq \mathbf{x}_j$ for all $i \neq j$. If $\mathbf{M}$ denotes the transition matrix, what is the value of $\|\mathbf{M}\|_F^2$ as $\varepsilon \to 0$ and $\varepsilon \to \infty$? Justify your answer. (4P)

Problem 2. (15 points)

Classification and Clustering

A dataset is composed of six points $\mathbf{x}_1, \ldots, \mathbf{x}_6$ known to belong to one of two groups C_1 or C_2. As shown in the following table, the group assigned to $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ is known, while it is unknown for $\mathbf{x}_5$ and $\mathbf{x}_6$.

Data	Group	Data	Group
$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$	C_1	$\mathbf{x}_4 = \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}$	C_2
$\mathbf{x}_2 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}$	C_1	$\mathbf{x}_5 = \begin{pmatrix} 0 \\ -1/2 \\ -1/2 \end{pmatrix}$	?
$\mathbf{x}_3 = \begin{pmatrix} -1 \\ -1 \\ -1 \end{pmatrix}$	C_2	$\mathbf{x}_6 = \begin{pmatrix} 0 \\ 1/2 \\ 1/2 \end{pmatrix}$	?

a) Use $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ to obtain two cluster centers for k-means. (2P)

b) Use the obtained cluster centers to assign labels to $\mathbf{x}_5, \mathbf{x}_6$. (2P)

Assume that linear discriminant analysis on the dataset $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$ provides the discriminant vector

$$\mathbf{a}^* = \begin{pmatrix} -1/2 \\ 0 \\ 1 \end{pmatrix}.$$

c) Calculate the sum of squares within groups for $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$. (4P)

d) Calculate the sum of squares between groups for $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$. (4P)

e) Use the obtained $\mathbf{a}^*$ to assign a label to $\mathbf{x}_5, \mathbf{x}_6$. (3P)

Problem 3. (15 points)
Support Vector Machines:

Suppose that a training dataset is composed of vectors $\mathbf{x}_i \in \mathbb{R}^2$, $i = 1, \ldots, 6$, belonging to two classes. The class membership is indicated by the labels $y_i \in \{-1, +1\}$. Suppose that the dataset is not linearly separable. A support vector machine is used to find the maximum-margin hyperplane by solving the following dual problem:

$$\max_{\boldsymbol{\lambda}} \quad \sum_{i=1}^{6} \lambda_i - \frac{1}{2}\sum_{i=1}^{6}\sum_{j=1}^{6} y_i y_j \lambda_i \lambda_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{s.t.} \quad 0 \le \lambda_i \le 1 \qquad \text{and} \qquad \sum_{i=1}^{6} \lambda_i y_i = 0.$$

The dataset and the outputs of the optimization problem are given in the following table.

Data	Label	Solution	Data	Label	Solution
$\mathbf{x}_1 = \begin{pmatrix}1\\0\end{pmatrix}$	$y_1 = -1$	$\lambda_1^\star = 1$	$\mathbf{x}_4 = \begin{pmatrix}0\\0\end{pmatrix}$	$y_4 = 1$	$\lambda_4^\star = 1$
$\mathbf{x}_2 = \begin{pmatrix}3\\1\end{pmatrix}$	$y_2 = -1$	$\lambda_2^\star = 0$	$\mathbf{x}_5 = \begin{pmatrix}1\\-3\end{pmatrix}$	$y_5 = 1$	$\lambda_5^\star = 0.12$
$\mathbf{x}_3 = \begin{pmatrix}0\\2\end{pmatrix}$	$y_3 = -1$	$\lambda_3^\star = 0.12$	$\mathbf{x}_6 = \begin{pmatrix}-2\\-1\end{pmatrix}$	$y_6 = 1$	$\lambda_6^\star = 0$

a) Determine the support vectors. (4P)

b) Find the maximum-margin hyperplane $\mathbf{a}^{\star T}\mathbf{x} + b^\star$ by finding $\mathbf{a}^\star$ and $b^\star$. (6P)

c) Use the above support vector machine to classify $\mathbf{u} = \begin{pmatrix}1\\1\end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix}-1\\3\end{pmatrix}$. (2P)

d) Consider a polynomial kernel given by

$$K(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + 1)^3.$$

Find a feature mapping for this kernel and the dimension of the corresponding feature space. (3P)

Problem 4. (15 points)

Linear Regression for Machine Learning:

A training set with input-output pairs (x_i, y_i), $i \in \{1, 2, 3, 4\}$, is given in the following table.

i	input x_i	output y_i
$i = 1$	-5	-18
$i = 2$	-2	-9
$i = 3$	1	-1
$i = 4$	4	12

a) Use linear regression to find a linear approximation of y_i in terms of x_i. Use this model to predict the output for the input $x_5 = 0$. (8P)

Remember that for a training dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ with $x_i, y_i \in \mathbb{R}$, the matrix $\mathbf{X}$ is defined as follows:

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

b) Suppose that for a dataset the matrix $\mathbf{X}^T\mathbf{X}$ is given by

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 6 & 12 \\ 12 & 48 \end{pmatrix}.$$

Find the number of training samples, the mean value and the variance of the inputs. (4P)

c) Suppose that for the above matrix $\mathbf{X}$ and the output vector $\mathbf{y}$, we have:

$$\mathbf{X}^T\mathbf{y} = \begin{pmatrix} -3 \\ 1 \end{pmatrix}.$$

Use linear regression to find a linear approximation of the output y in terms of the input x. (3P)

Additional sheet

Problem:

Additional sheet

Problem:

Additional sheet

Problem:

Additional sheet

Problem:

Additional sheet

Problem: