Chair for Theoretical Information Technology | RWTH AACHEN UNIVERSITY

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

1	2	3	4	$\sum$
25	25	25	25	100

Written Examination

Fundamentals of Big Data Analytics

Tuesday, March 21, 2017, 01:00 p.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 **50 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 Friday evening, the 24.03.17, on the homepage of the institute.

The corrected exams can be inspected on Monday, 27.03.17, 14:00h. at the seminar room 333 of the Chair for Theoretical Information Technology, Kopernikusstr. 16.

Acknowledged: ______________________________

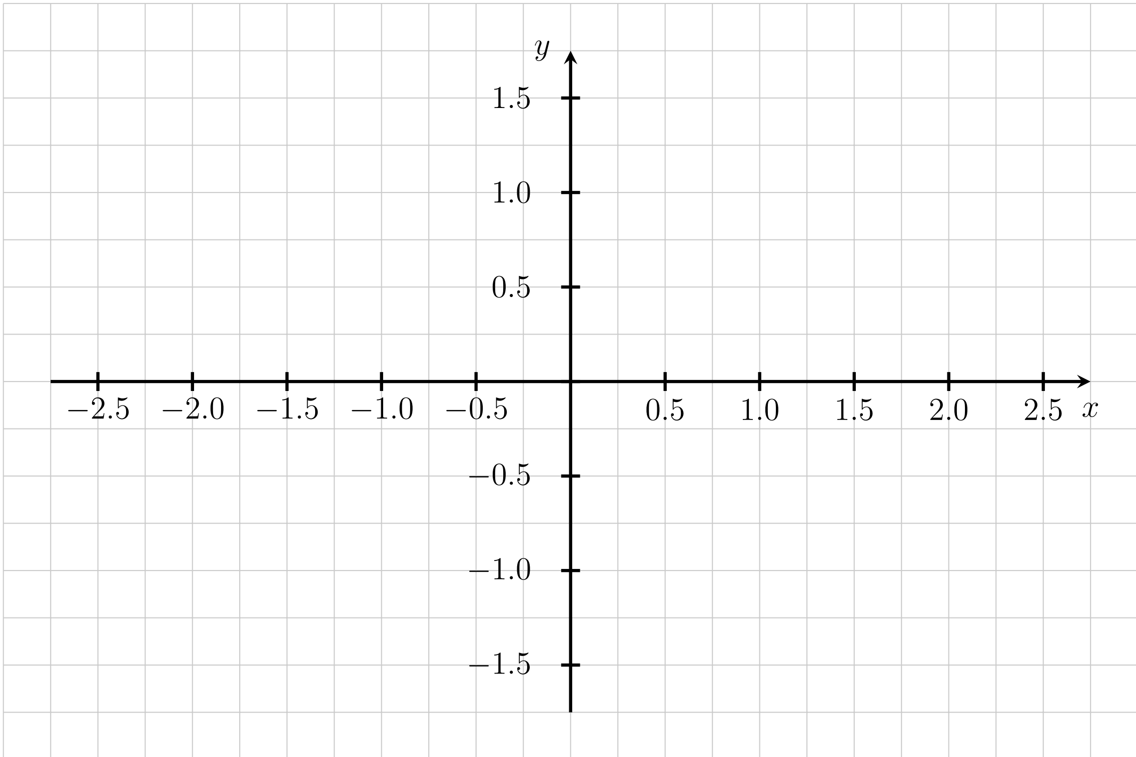
(Signature)

Problem 1. (25 points)
Analysis of Multivariate Densities and Maximum Likelihood Estimation:
Let $f_{\mathbf{Z}}(x, y)$ be the joint density of the random vector $\mathbf{Z} = (X, Y)^{\mathrm{T}}$. It is given by

$$f_{\mathbf{Z}}(x, y) = \begin{cases} c\,, & 0 \leq y \leq a \quad \text{and} \quad y \leq x \leq y + a\,, \\ 0\,, & \text{otherwise} \end{cases}$$

for a proper number c with the parameter $a > 0$.

a) Sketch the support (topview) of the density $f_{\mathbf{Z}}(x, y)$ for $a = 1$ and mark important points. (2P)



b) What is the value of c as a function of a such that $f_{\mathbf{Z}}(x, y)$ is a proper density function. (2P)

c) Calculate the missing value $\mu_X = \mathrm{E}(X)$ of the mean vector

$$\mathrm{E}(\mathbf{Z}) = \begin{pmatrix} \mu_X \\ \frac{1}{2} \end{pmatrix}$$

for $a = 1$. (3P)

d) Compute the missing values of the covariance matrix

$$\mathbf{\Sigma_Z} = \begin{pmatrix} \frac{1}{6} & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_{2,2} \end{pmatrix}$$

of the density $f_{\mathbf{Z}}$ for $a = 1$ by calculating the variance $\sigma_{2,2}$ of Y and the co-variances $\sigma_{1,2}$ and $\sigma_{2,1}$. (3+2+1P)

e) Determine the marginal densities for X and Y for $a = 1$. (2+2P)

f) Are X and Y independent random variables? Please justify the reason behind your answer. (1P)

g) Calculate the mean vector $\mathrm{E}(\mathbf{Z})$ and the covariance matrix $\mathrm{Cov}(\mathbf{Z})$ for the general case $a > 0$. (1+1P)

h) Assuming n IID random vectors $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n$ with densities $f_{\mathbf{Z}_i}(x_i, y_i) = f_{\mathbf{Z}}(x_i, y_i)$, calculate the likelihood function $L(a, \mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n)$. (1P)

i) Show that the log-likelihood function $\ell(a, \mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n) = \log\Big(L(a, \mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n)\Big)$ is strictly monotonic in a. (2P)

j) Why is the quantity $\hat{a} = \max\limits_{1 \le i \le n} \{y_i, x_i - y_i\}$ the maximum likelihood estimator for the parameter a? (2P)

Problem 2. (25 points)

Principal Component Analysis:

Part I

Assume four independent two-dimensional samples:

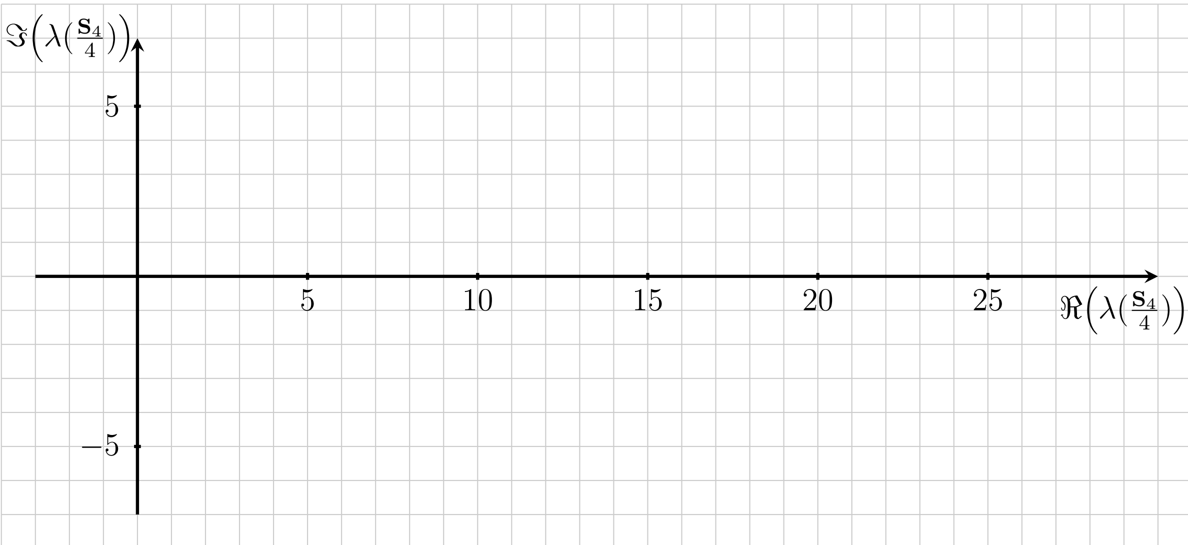
$$\mathbf{x}_1 = \sqrt{3}\begin{pmatrix}4\\4\end{pmatrix}, \qquad \mathbf{x}_2 = \sqrt{3}\begin{pmatrix}4\\-4\end{pmatrix}, \qquad \mathbf{x}_3 = \sqrt{3}\begin{pmatrix}0\\8\end{pmatrix}, \qquad \mathbf{x}_4 = \sqrt{3}\begin{pmatrix}0\\-4\end{pmatrix}.$$

We suggest to keep the symbolic value $\sqrt{3}$ *within your calculations without any conversion to a low-precision number!*

a) Calculate the sample mean $\bar{\mathbf{x}}_4$. (2P)

b) Calculate the sample covariance matrix $\mathbf{S}_4$. (4P)

c) Determine the Gerschgorin's circles of the normalized matrix $\frac{1}{4}\mathbf{S}_4$ and sketch them on the complex domain. (4+2P)



d) Is the matrix $\mathbf{S}_4$ positive definite, negative definite, or indefinite? Give a reason for your statement. (1P)

Part II

e) Now we independently take n samples in total such that the corresponding sample covariance matrix $\mathbf{S}_n$ is given by

$$\mathbf{S}_n = \begin{pmatrix}14 & -14\\-14 & 110\end{pmatrix}.$$

Calculate the spectral decomposition $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}$ of $\mathbf{S}_n$ by determining the matrices $\mathbf{V}$ and $\mathbf{\Lambda}$. (3+5P)

f) Determine the best projection matrix $\mathbf{Q}$ to transform the two-dimensional samples to a one-dimensional data and calculate the projection of $\mathbf{x}_1$. (2+1P)

g) Determine the residuum $\frac{1}{n-1}\max\limits_{\mathbf{Q}}\sum_{i=1}^{n}\|\mathbf{Q}\mathbf{x}_i - \mathbf{Q}\bar{\mathbf{x}}_n\|^2$ for the above choice of $\mathbf{Q}$. (1P)

Problem 3. (25 points)
Discriminant Analysis:

A training dataset consists of three-dimensional vectors belonging to three classes denoted by the labels $y_i \in \{1, 2, 3\}$. This dataset is given below.

Data	Label	Data	Label	Data	Label
$\mathbf{x}_1 = \begin{pmatrix}1\\1\\1\end{pmatrix}$	$y_1 = 1$	$\mathbf{x}_4 = \begin{pmatrix}1\\-1\\-1\end{pmatrix}$	$y_4 = 2$	$\mathbf{x}_7 = \begin{pmatrix}-1\\1\\1\end{pmatrix}$	$y_7 = 3$
$\mathbf{x}_2 = \begin{pmatrix}1\\2\\0\end{pmatrix}$	$y_2 = 1$	$\mathbf{x}_5 = \begin{pmatrix}0\\-2\\-1\end{pmatrix}$	$y_5 = 2$	$\mathbf{x}_8 = \begin{pmatrix}-1\\0\\2\end{pmatrix}$	$y_8 = 3$
$\mathbf{x}_3 = \begin{pmatrix}1\\1\\-1\end{pmatrix}$	$y_3 = 1$	$\mathbf{x}_6 = \begin{pmatrix}-1\\-1\\-1\end{pmatrix}$	$y_6 = 2$	$\mathbf{x}_9 = \begin{pmatrix}-1\\-1\\1\end{pmatrix}$	$y_9 = 3$

Consider only those vectors $\mathbf{x}_i$ in the dataset with labels $y_i = 1$ and $y_i = 2$.

a) Find the matrix $\mathbf{W}$, that is the matrix corresponding to the sum of squares within groups. (5P)

b) Find Fisher's linear discriminant rule for classification of vectors with labels $y_i = 1$ and $y_i = 2$. Explain each step. (3P)

Hint: For a matrix $\mathbf{A} = \begin{pmatrix}\alpha_1\mathbf{x}^{\mathrm{T}}\\ \alpha_2\mathbf{x}^{\mathrm{T}}\\ \alpha_3\mathbf{x}^{\mathrm{T}}\end{pmatrix} \in \mathbb{R}^{3\times3}$ and any $\mathbf{x} = \begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix} \in \mathbb{R}^3$, the eigenvalues and its dominant eigenvector are given by $(0, 0, \mathrm{tr}(\mathbf{A}))$ and $\begin{pmatrix}\alpha_1\\ \alpha_2\\ \alpha_3\end{pmatrix}$, respectively.

Suppose that the vectors $\mathbf{x}_i$ with label y_i are randomly drawn from a multivariate normal distribution with the covariance matrix $\mathbf{\Sigma}_i$ and the expected value $\boldsymbol{\mu}_i$ for $i = 1, 2, 3$.

c) Find the maximum likelihood estimation of the expected values. (4P)

d) If $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3 = \mathbf{\Sigma}$, find the maximum likelihood estimation of the covariance matrix $\hat{\mathbf{\Sigma}}$ by using the whole dataset (5P).

Consider only those vectors $\mathbf{x}_i$ in the dataset with labels $y_i = 2$ and $y_i = 3$.

e) Find the maximum likelihood estimation of the covariance matrix, $\hat{\mathbf{\Sigma}}$. (3P)

f) Find the Gaussian maximum likelihood discriminant rule for classification of vectors with labels $y_i = 2$ and $y_i = 3$. Explain each step. (5P)

Problem 4. (25 points)
Support Vector Machines:

a) Suppose that a training dataset is composed of vectors $\mathbf{x}_i \in \mathbb{R}^3$ belonging to two classes. The class membership is indicated by the labels $y_i \in \{-1, +1\}$. Suppose that the dataset is separable. A Support vector machine is used to find the maximum-margin hyperplane $\mathbf{a}^{\mathrm{T}}\mathbf{x} + b = 0$. The primal optimization problem gives the optimal $\mathbf{a}^\star$ as $\begin{pmatrix}1 & 3 & 0\end{pmatrix}^{\mathrm{T}}$. Two support vectors with different labels are given as :

$$\mathbf{x}_1^{\mathrm{T}} = \begin{pmatrix}1 & -1 & 1\end{pmatrix}, \qquad \mathbf{x}_2^{\mathrm{T}} = \begin{pmatrix}-1 & -1 & -1\end{pmatrix}$$

Find the optimal value $b^\star$. (3P)

Consider another training dataset that is non-separable. The following dual problem is solved for a support vector machine

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

$$\text{s.t.} \quad 0 \le \lambda_i \le 5 \qquad \text{and} \qquad \sum_{i=1}^{6} \lambda_i y_i = 0.$$

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

Data	Label	Solution	Data	Label	Solution
$\mathbf{x}_1 = \begin{pmatrix}1\\1\end{pmatrix}$	$y_1 = -1$	$\lambda_1^\star = 0$	$\mathbf{x}_4 = \begin{pmatrix}0.5\\-0.5\end{pmatrix}$	$y_4 = 1$	$\lambda_4^\star = 4.73$
$\mathbf{x}_2 = \begin{pmatrix}2\\0\end{pmatrix}$	$y_2 = -1$	$\lambda_2^\star = 0.67$	$\mathbf{x}_5 = \begin{pmatrix}-2\\1\end{pmatrix}$	$y_5 = 1$	$\lambda_5^\star = 0.94$
$\mathbf{x}_3 = \begin{pmatrix}0\\0\end{pmatrix}$	$y_3 = -1$	$\lambda_3^\star = 5$	$\mathbf{x}_6 = \begin{pmatrix}0\\-1\end{pmatrix}$	$y_6 = 1$	$\lambda_1^\star = 0$

b) Determine the support vectors. (4P)

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

d) Suppose that a kernel is given by $K(\mathbf{x}, \mathbf{y}) = (2\mathbf{x}^T\mathbf{y} + 1)^2$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. Determine the feature mapping function $\boldsymbol{\phi}(\mathbf{x})$, i.e., the function $\boldsymbol{\phi} : \mathbb{R}^p \to \mathbb{R}^d$ where $K(\mathbf{x}, \mathbf{y}) = \boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{y})$. Determine the dimension of the feature space. (6P)

Consider a training dataset that is separable and consists of vectors $\mathbf{x}_i \in \mathbb{R}^p$ with labels $y_i \in \{-1, +1\}$.

e) Write down the optimization problem for the kernel-based support vector machine using the kernel $K(\mathbf{x}, \mathbf{y}) = \exp(-\gamma\|\mathbf{x} - \mathbf{y}\|^2)$. (3P)

f) Write down the kernel-based support vector machine classifier for this dataset. (3P)

Additional sheet

Problem:

Additional sheet

Problem:

Additional sheet

Problem:

Additional sheet

Problem:

Additional sheet

Problem: