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 $A$ is given by:

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

a) What is the rank of $A$? (1P)

b) Calculate the spectral decomposition $VΛV^T$ of $A$ by determining the matrices $V$ and $Λ$. (4P)

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

Let $S_n$ be the sample covariance matrix of $n$ points $x_1, \ldots, x_n \in \mathbb{R}^4$. Assume that it has the spectral decomposition $S_n = ˜V ˜Λ ˜V^T$ where

$$
\begin{pmatrix}
5 & 0 & 0 & 0 \\
0 & 4 & 0 & 0 \\
0 & 0 & 3 & 0 \\
0 & 0 & 0 & 2
\end{pmatrix}
$$

and $v_1, v_2, v_3, v_4 \in \mathbb{R}^4$.

d) Given $v_1 = \frac{1}{3} \begin{pmatrix}2 & 1 & 0 & 2 \end{pmatrix}^T$ and $v_2 = \frac{1}{3} \begin{pmatrix}-1 & 2 & 2 & 0 \end{pmatrix}^T$, visualize the following points in a 2D graph using PCA (4P)

$$
x_1 = \begin{pmatrix}1 \\
0 \\
1 \\
0
\end{pmatrix},
\quad
x_2 = \begin{pmatrix}0 \\
1 \\
0 \\
1
\end{pmatrix},
\quad
x_3 = \begin{pmatrix}0 \\
0 \\
1 \\
1
\end{pmatrix}.
$$

Let $K(x_i, x_j) = \exp(-\frac{||x_i-x_j||^2}{2\epsilon})$ be the dissimilarity function used for Multidimensional Scaling (MDS).

e) Assume that $x_i \neq x_j$ for all $i \neq j$. If $M$ denotes the transition matrix, what is the value of $||M||^2_F$ as $\epsilon \to 0$ and $\epsilon \to \infty$? Justify your answer. (4P)
Problem 2. (15 points)
Classification and Clustering
A dataset is composed of six points $x_1, \ldots, x_6$ known to belong to one of two groups $C_1$ or $C_2$. As shown in the following table, the group assigned to $x_1, x_2, x_3, x_4$ is known, while it is unknown for $x_5$ and $x_6$.

<table>
<thead>
<tr>
<th>Data</th>
<th>Group</th>
<th>Data</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1 = \begin{pmatrix} 1 \ 1 \ 1 \end{pmatrix}$</td>
<td>$C_1$</td>
<td>$x_4 = \begin{pmatrix} 1 \ -1 \ -1 \end{pmatrix}$</td>
<td>$C_2$</td>
</tr>
<tr>
<td>$x_2 = \begin{pmatrix} -1 \ 1 \ 1 \end{pmatrix}$</td>
<td>$C_1$</td>
<td>$x_5 = \begin{pmatrix} 0 \ -1/2 \ -1/2 \end{pmatrix}$</td>
<td>?</td>
</tr>
<tr>
<td>$x_3 = \begin{pmatrix} -1 \ -1 \ -1 \end{pmatrix}$</td>
<td>$C_2$</td>
<td>$x_6 = \begin{pmatrix} 0 \ 1/2 \ 1/2 \end{pmatrix}$</td>
<td>?</td>
</tr>
</tbody>
</table>

a) Use $x_1, x_2, x_3, x_4$ to obtain two cluster centers for $k$-means. (2P)

b) Use the obtained cluster centers to assign labels to $x_5, x_6$. (2P)

Assume that linear discriminant analysis on the dataset \{\(x_1, x_2, x_3, x_4\)\} provides the discriminant vector
\[
a^* = \begin{pmatrix} -1/2 \\ 0 \\ 1 \end{pmatrix}.
\]

c) Calculate the sum of squares within groups for $x_1, x_2, x_3, x_4$. (4P)

d) Calculate the sum of squares between groups for $x_1, x_2, x_3, x_4$. (4P)

e) Use the obtained $a^*$ to assign a label to $x_5, x_6$. (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:

$$
\max_{\lambda} \sum_{i=1}^{6} \lambda_i - \frac{1}{2} \sum_{i=1}^{6} \sum_{j=1}^{6} y_i y_j \lambda_i \lambda_j \mathbf{x}_i^T \mathbf{x}_j \\
\text{s.t. } 0 \leq \lambda_i \leq 1 \quad \text{and} \quad \sum_{i=1}^{6} \lambda_i y_i = 0.
$$

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

<table>
<thead>
<tr>
<th>Data</th>
<th>Label</th>
<th>Solution</th>
<th>Data</th>
<th>Label</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbf{x}_1 = \begin{pmatrix} 1 \ 0 \end{pmatrix}$</td>
<td>$y_1 = -1$</td>
<td>$\lambda_1^* = 1$</td>
<td>$\mathbf{x}_4 = \begin{pmatrix} 0 \ 0 \end{pmatrix}$</td>
<td>$y_4 = 1$</td>
<td>$\lambda_4^* = 1$</td>
</tr>
<tr>
<td>$\mathbf{x}_2 = \begin{pmatrix} 3 \ 1 \end{pmatrix}$</td>
<td>$y_2 = -1$</td>
<td>$\lambda_2^* = 0$</td>
<td>$\mathbf{x}_5 = \begin{pmatrix} 1 \ -3 \end{pmatrix}$</td>
<td>$y_5 = 1$</td>
<td>$\lambda_5^* = 0.12$</td>
</tr>
<tr>
<td>$\mathbf{x}_3 = \begin{pmatrix} 0 \ 2 \end{pmatrix}$</td>
<td>$y_3 = -1$</td>
<td>$\lambda_3^* = 0.12$</td>
<td>$\mathbf{x}_6 = \begin{pmatrix} -2 \ -1 \end{pmatrix}$</td>
<td>$y_6 = 1$</td>
<td>$\lambda_6^* = 0$</td>
</tr>
</tbody>
</table>

a) Determine the support vectors. (4P)

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

<table>
<thead>
<tr>
<th>(i)</th>
<th>input (x_i)</th>
<th>output (y_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-5</td>
<td>-18</td>
</tr>
<tr>
<td>2</td>
<td>-2</td>
<td>-9</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>12</td>
</tr>
</tbody>
</table>

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 \(X\) is defined as follows:

\[
X = \begin{pmatrix}
1 & x_1 \\
1 & x_2 \\
& \vdots \\
1 & x_n
\end{pmatrix}.
\]

b) Suppose that for a dataset the matrix \(X^T X\) is given by

\[
X^T X = \begin{pmatrix}
6 & 12 \\
12 & 48
\end{pmatrix}.
\]

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

c) Suppose that for the above matrix \(X\) and the output vector \(y\), we have:

\[
X^T y = \begin{pmatrix}
-3 \\
1
\end{pmatrix}.
\]

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: