Chair for Theoretical Information Technology
RWTH AACHEN UNIVERSITY

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)
Dimensionality Reduction:

a)

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

$$= \begin{pmatrix}1\\2\\2\\0\end{pmatrix}\begin{pmatrix}1 & 2 & 2 & 0\end{pmatrix} + \begin{pmatrix}-2\\1\\0\\2\end{pmatrix}\begin{pmatrix}-2 & 1 & 0 & 2\end{pmatrix} + [\begin{pmatrix}1\\0\\0\\0\end{pmatrix} + \begin{pmatrix}0\\2\\2\\0\end{pmatrix}]\begin{pmatrix}1 & 2 & 2 & 0\end{pmatrix}$$

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

Then the rank of $\mathbf{A} = 2$.

b)

$$\mathbf{A} = (2\cdot 3\cdot 3)\cdot\frac{1}{3}\begin{pmatrix}1\\2\\2\\0\end{pmatrix}\cdot\frac{1}{3}\begin{pmatrix}1 & 2 & 2 & 0\end{pmatrix} + 9\cdot\frac{1}{3}\begin{pmatrix}-2\\1\\0\\2\end{pmatrix}\cdot\frac{1}{3}\begin{pmatrix}-2 & 1 & 0 & 2\end{pmatrix}$$

$$= \frac{1}{3}\begin{pmatrix}1 & -2\\2 & 1\\2 & 0\\0 & 2\end{pmatrix}\begin{pmatrix}18 & 0\\0 & 9\end{pmatrix}\frac{1}{3}\begin{pmatrix}1 & 2 & 2 & 0\\-2 & 1 & 0 & 2\end{pmatrix}$$

Then,

$$\mathbf{V} = \frac{1}{3}\begin{pmatrix}1 & -2\\2 & 1\\2 & 0\\0 & 2\end{pmatrix}, \mathbf{\Lambda} = \begin{pmatrix}18 & 0\\0 & 9\end{pmatrix}$$

c)

$$\mathbf{Q} = \frac{1}{9}\begin{pmatrix}1\\2\\2\\0\end{pmatrix}\begin{pmatrix}1 & 2 & 2 & 0\end{pmatrix} = \frac{1}{9}\begin{pmatrix}1 & 2 & 2 & 0\\2 & 4 & 4 & 0\\2 & 4 & 4 & 0\\0 & 0 & 0 & 0\end{pmatrix}$$

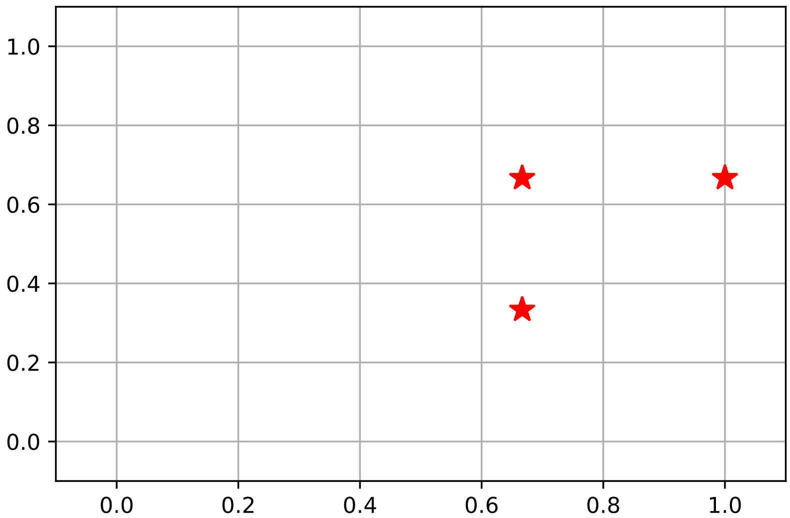
d) Components in the first dimension:

$$\mathbf{v}_1^T(\mathbf{x}_1\ \mathbf{x}_2\ \mathbf{x}_3) = \frac{1}{3}\begin{pmatrix}2 & 1 & 0 & 2\end{pmatrix}\begin{pmatrix}1 & 0 & 0\\0 & 1 & 0\\1 & 0 & 1\\0 & 1 & 1\end{pmatrix} = (\frac{2}{3}\ 1\ \frac{2}{3})$$

Components in the second dimension:

$$\mathbf{v}_2^T(\mathbf{x}_1\ \mathbf{x}_2\ \mathbf{x}_3) = \frac{1}{3}\begin{pmatrix}-1 & 2 & 2 & 0\end{pmatrix}\begin{pmatrix}1 & 0 & 0\\0 & 1 & 0\\1 & 0 & 1\\0 & 1 & 1\end{pmatrix} = (\frac{1}{3}\ \frac{2}{3}\ \frac{2}{3})$$

Then the points in 2D are $\mathbf{u}_1 = (\frac{2}{3}, \frac{1}{3}), \mathbf{u}_2 = (1, \frac{2}{3}), \mathbf{u}_3 = (\frac{2}{3}, \frac{2}{3})$.



e) As $\varepsilon \to 0$, $\mathbf{W}$ tends to $\mathbf{W} = \mathbf{I}_n$. This leads to $\deg(i) = 1$ for all i, thus $\mathbf{M} = \mathbf{W}$. Finally, we get

$$(\varepsilon \to 0) \quad \Rightarrow \quad \|\mathbf{M}\|_F^2 = \|\mathbf{W}\|_F^2 = n\,.$$

Similarlly for $\varepsilon \to \infty$ we get $\mathbf{W} = \mathbf{1}_{n\times n}$, thus $\mathbf{M} = \frac{1}{n}\mathbf{W}$. This leads to

$$(\varepsilon \to \infty) \quad \Rightarrow \quad \|\mathbf{M}\|_F^2 = \frac{1}{n^2}\|\mathbf{W}\|_F^2 = \frac{1}{n^2}(n^2) = 1\,.$$

Problem 2. (15 points)
Classification and Clustering

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)

$$\mu_1 = \frac{1}{2}(\mathbf{x}_1 + \mathbf{x}_2) = \begin{pmatrix}0\\1\\1\end{pmatrix}, \qquad \mu_2 = \frac{1}{2}(\mathbf{x}_3 + \mathbf{x}_4) = \begin{pmatrix}0\\-1\\-1\end{pmatrix}$$

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

$$\|\mathbf{x}_5 - \mu_1\|_2 = \frac{3}{2}\sqrt{2}, \qquad \|\mathbf{x}_5 - \mu_2\|_2 = \frac{1}{2}\sqrt{2} \quad \Rightarrow \quad \mathbf{x}_5 \in C_2 \tag{1}$$

$$\|\mathbf{x}_6 - \mu_1\|_2 = \frac{1}{2}\sqrt{2}, \qquad \|\mathbf{x}_6 - \mu_2\|_2 = \frac{3}{2}\sqrt{2} \quad \Rightarrow \quad \mathbf{x}_6 \in C_1 \tag{2}$$

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

$$\mathbf{a}^* = \frac{2}{\sqrt{5}} \begin{pmatrix}-1/2\\0\\1\end{pmatrix}.$$

c) Calculate the sum of squares within groups for $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$. (4P)
By definition $y_i = \mathbf{a}^T \mathbf{x}_i$, yielding

$$y_1 = \frac{2}{\sqrt{5}}\frac{1}{2} = \frac{1}{\sqrt{5}}$$

$$y_2 = \frac{2}{\sqrt{5}}\frac{3}{2} = \frac{3}{\sqrt{5}}$$

$$y_3 = -\frac{2}{\sqrt{5}}\frac{1}{2} = -\frac{1}{\sqrt{5}}$$

$$y_4 = -\frac{2}{\sqrt{5}}\frac{3}{2} = -\frac{3}{\sqrt{5}}$$

The within group averages are

$$\overline{y}_1 = \frac{1}{2}(y_1 + y_2) = \frac{2}{\sqrt{5}}, \quad \text{and} \quad \overline{y}_1 = \frac{1}{2}(y_3 + y_4) = -\frac{2}{\sqrt{5}}.$$

Lets denote the sum of of squares within groups as $\gamma_W \in \mathbb{R}$. By definition we get

$$\begin{aligned}
\gamma_W &= \sum_{l=1}^{2} \sum_{j \in C_l} (y_j - \overline{y}_l)^2 = (y_1 - \overline{y}_1)^2 + (y_2 - \overline{y}_1)^2 + (y_3 - \overline{y}_1)^2 + (y_4 - \overline{y}_2)^2 \\
&= (\frac{1}{\sqrt{5}} - \frac{2}{\sqrt{5}})^2 + (\frac{3}{\sqrt{5}} - \frac{2}{\sqrt{5}})^2 + (-\frac{1}{\sqrt{5}} + \frac{2}{\sqrt{5}})^2 + (-\frac{3}{\sqrt{5}} + \frac{2}{\sqrt{5}})^2 \\
&= \frac{4}{5}.
\end{aligned}$$

d) Calculate the sum of squares between groups for $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$. (4P)
Then the general discriminant average is $\overline{y} = \frac{1}{4}(y_1 + y_2 + y_3 + y_4) = 0$. Lets denote the sum of of squares between groups as $\gamma_B \in \mathbb{R}$. By definition we get

$$\gamma_B = \sum_{l=1}^{2} n_l (\overline{y}_l - \overline{y})^2 = \sum_{l=1}^{2} n_l\, \overline{y}_l^2 = 2\frac{2}{5} + 2\frac{2}{5} = \frac{8}{5}$$

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

$$\mathbf{a}^T(\mathbf{x} - \frac{1}{2}(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)) \gtrless 0\,.$$

Therefore we have

$$\begin{aligned}
\mathbf{a}^T(\mathbf{x}_5 - \frac{1}{2}(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)) &= & \Rightarrow \quad \mathbf{x}_5 \in C_2 \\
\mathbf{a}^T(\mathbf{x}_6 - \frac{1}{2}(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)) &= & \Rightarrow \quad \mathbf{x}_6 \in C_1
\end{aligned}$$

Problem 3. (15 points)
Support Vector Machines:

a) (4P) The support vectors are given by all vectors with $\lambda_i \neq 0$, namely:

$$\mathbf{x}_1 = \begin{pmatrix}1\\0\end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix}0\\2\end{pmatrix}, \mathbf{x}_4 = \begin{pmatrix}0\\0\end{pmatrix}, \mathbf{x}_5 = \begin{pmatrix}1\\-3\end{pmatrix}.$$

b) First let's find $\mathbf{a}^\star$:

$$\mathbf{a}^\star = \sum_{i=1}^{6} \lambda_i y_i \mathbf{x}_i = \lambda_1 y_1 \mathbf{x}_1 + \lambda_3 y_3 \mathbf{x}_3 + \lambda_4 y_4 \mathbf{x}_4 + \lambda_5 y_5 \mathbf{x}_5$$

$$\mathbf{a}^\star = -1 \times \begin{pmatrix}1\\0\end{pmatrix} - 0.12 \times \begin{pmatrix}0\\2\end{pmatrix} + 1 \times \begin{pmatrix}0\\0\end{pmatrix} + 0.12 \times \begin{pmatrix}1\\-3\end{pmatrix} = \begin{pmatrix}1\\-3\end{pmatrix} = \begin{pmatrix}-0.88\\-0.6\end{pmatrix}.$$

To find b, take two support vectors $\mathbf{x}_k$ and $\mathbf{x}_l$ with $y_k = 1$ and $y_l = -1$ with $0 < \lambda < 1$. For these support vectors, we have $y_i(\mathbf{a}^T\mathbf{x}_i + b) = 1$. Hence:

$$b^\star = \frac{-1}{2}\mathbf{a}^{\star T}(\mathbf{x}_3 + \mathbf{x}_5) = -\frac{1}{2}\begin{pmatrix}-0.88 & -0.6\end{pmatrix}(\begin{pmatrix}0\\2\end{pmatrix} + \begin{pmatrix}1\\-3\end{pmatrix}) = -\frac{1}{2}(-0.88 + 0.6) = 0.14. \tag{3}$$

c) (2P)

First see:

$$(\mathbf{a}^\star)^T\mathbf{u} + b^\star = \begin{pmatrix}-0.88 & -0.6\end{pmatrix}\begin{pmatrix}1\\1\end{pmatrix} + 0.14 = -1.34 < 0,$$

hence $y_\mathbf{u} = -1$. Finally:

$$(\mathbf{a}^\star)^T\mathbf{v} + b^\star = \begin{pmatrix}-0.88 & -0.6\end{pmatrix}\begin{pmatrix}-1\\3\end{pmatrix} + 0.14 = -0.78 < 0,$$

hence $y_\mathbf{v} = -1$.

d) (3P) The kernel function can be expanded as

$$\begin{aligned}
K(\mathbf{x}, \mathbf{y}) &= (\langle \mathbf{x}, \mathbf{y}\rangle + 1)^3 \\
&= 1 + 3\sum_{i=1}^{n} x_i y_i + 3(\sum_{i=1}^{n} x_i y_i)^2 + (\sum_{i=1}^{n} x_i y_i)^3 \\
&= 1 + 3\sum_{i=1}^{n} x_i y_i + 3(\sum_{i=1}^{n} x_i^2 y_i^2 + 2\sum_{1\le i<j\le n} x_i x_j y_i y_j) \\
&\quad + \sum_{i=1}^{n} x_i^3 y_i^3 + 3(\sum_{1\le i\neq j\le n} x_i^2 y_i^2 x_j y_j) + 6\sum_{1\le i<j<k\le n} x_i x_j x_k y_i y_j y_k
\end{aligned}$$

So a feature map can be constructed as

$$\begin{aligned}
\Phi(\mathbf{x}) =&(1, \sqrt{3}x_1, \dots, \sqrt{3}x_n, \sqrt{3}x_1^2, \dots, \sqrt{3}x_n^3, \sqrt{6}x_1x_2, \sqrt{6}x_1x_3, \dots, \sqrt{6}x_{n-1}x_n, \\
& x_1^3, \dots, x_n^3, \sqrt{3}x_1^2x_2, \sqrt{3}x_2^2x_1, \dots, \sqrt{3}x_{n-1}^2x_n, \sqrt{6}x_1x_2x_3, \dots \sqrt{6}x_{n-2}x_{n-1}x_n)
\end{aligned}$$

And the dimension is given by

$$1 + n + \binom{n}{2} + n + n + n(n-1) + \frac{n(n-1)(n-2)}{6} = \frac{(n+1)(n+2)(n+3)}{6}.$$

Problem 4. (15 points)

Linear Regression: 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) To use the linear regression algorithm, we need to find following parameters:

$$\begin{aligned}
\bar{x} &= \frac{-5-2+1+4}{4} = -0.5 \qquad (1\text{P}) \\
\bar{y} &= \frac{-18-9-1+12}{4} = -4 \qquad (1\text{P}) \\
\sigma_{xy} &= \frac{-5\times -18 - 2\times -9 + 1\times -1 + 4\times 12}{4} - (-0.5)\times -4 = 36.75 \qquad (1\text{P}) \\
\sigma_x^2 &= \frac{25+4+1+16}{4} - (-0.5)^2 = 11.25 \qquad (1\text{P})
\end{aligned}$$

Thus, the regression coefficients are

$$\begin{aligned}
\hat{\nu}_1 &= \frac{\sigma_{xy}}{\sigma_x^2} = \frac{36.75}{11.25} = 3.27, \qquad (1.5\text{P}) \\
\hat{\nu}_0 &= \bar{y} - \hat{\nu}_1 \bar{x} = -4 - 3.27 \times (-0.5) = -2.365 \qquad (1.5\text{P}).
\end{aligned}$$

The model is given by:

$$y = \hat{\nu}_1 x + \hat{\nu}_0,$$

which for $x_5 = 0$ predicts $y = -2.365$ ((1P)).

b) See that:

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2. \end{pmatrix} = \begin{pmatrix} 6 & 12 \\ 12 & 48 \end{pmatrix} = 6\begin{pmatrix} 1 & 2 \\ 2 & 8 \end{pmatrix}..$$

Hence we have $n = 6$ (1P) and

$$\bar{x} = \frac{12}{6} = 2. \quad (1\text{P})$$

$$\sigma_x^2 = \frac{48}{6} - 2^2 = 8 - 4 = 4. \quad (2\text{P})$$

c) See that first of all:

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{24}\begin{pmatrix} 8 & -2 \\ -2 & 1 \end{pmatrix}. \quad (1\text{P})$$

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

Using the above two inequalities, we can get:

$$\nu = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \frac{1}{24}\begin{pmatrix} 8 & -2 \\ -2 & 1 \end{pmatrix} \times \begin{pmatrix} -3 \\ 1 \end{pmatrix} = \frac{1}{24}\begin{pmatrix} -26 \\ 7 \end{pmatrix}.$$

Hence:

$$y = \nu_1 x + \nu_0 = \frac{7}{24}x - \frac{26}{24}. \quad \text{(2P)}$$

Additional sheet

Problem:

Additional sheet

Problem:

Additional sheet

Problem: