Professorship of Data Mining and Analytics
Department of Informatics
Technical University of Munich

Esolution
Place student sticker here

Note:
- During the attendance check a sticker containing a unique code will be put on this exam.
- This code contains a unique number that associates this exam with your registration number.
- This number is printed both next to the code and to the signature field in the attendance check list.

Mining Massive Datasets

Exam:	IN2323 / Endterm	**Date:**	Friday 9^{th} August, 2019
Examiner:	Prof. Dr. Stephan Günnemann	**Time:**	13:30 – 15:00

	P 1	P 2	P 3	P 4	P 5	P 6	P 7
I							

Working instructions

- This exam consists of **8 pages** with a total of **7 problems**.
  Please make sure that you received a complete copy of the exam.
- You can earn 38 points.
- **Detaching pages from the exam is prohibited!**
- Allowed resources:
  - A4 sheet of handwritten notes (two sides)
  - **no other materials (e.g. books, cell phones, calculators) are allowed!**
- Only write on the sheets given to you by supervisors. If you need more paper, ask the supervisors.
- Last two pages can be used as scratch paper.
- All sheets (including scratch paper) have to be returned at the end.
- **Only use a black or a blue pen (no pencils, red or green pens)!**
- Write your answers only in the provided solution boxes or the scratch paper.
- **For problems that say "Justify your answer" or "Show your work" you only get points if you provide a valid explanation.** Otherwise it's sufficient to only provide the correct answer.
- Exam duration - 90 minutes.

Left room from ________ to ________ / Early submission at ________

Problem 1 **AR models: correlation function (4 points)**

For each of the AR(1) and AR(2) models below find the corresponding autocorrelation function plot from Figure 1.1. Each plot was generated using one of the AR models from the list so that there is a one-to-one correspondence between the plots (1)-(4) and the AR processes (a)-(d). Everywhere $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ with a positive σ.

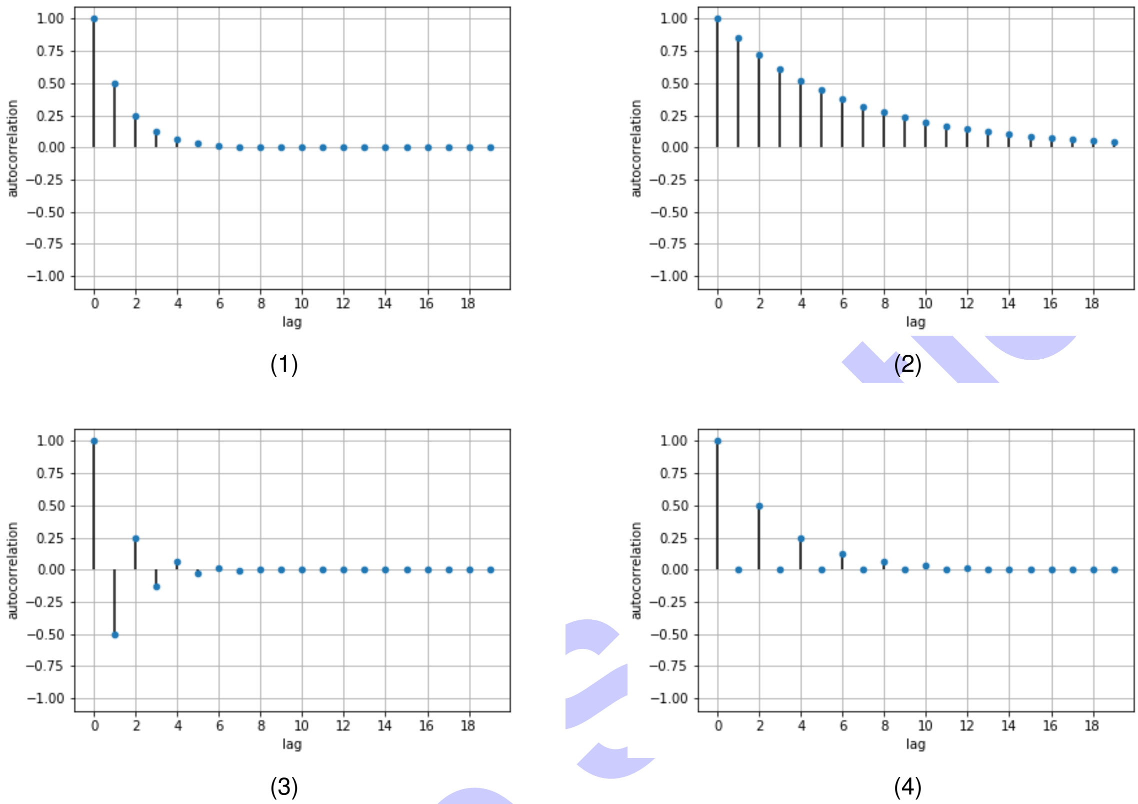


Figure 1.1: AR models: correlation function

a) $\mathcal{X}_t = -0.5\mathcal{X}_{t-1} + \epsilon_t$

☐ (1) ☐ (2) ☒ (3) ☐ (4)

b) $\mathcal{X}_t = 0.85\mathcal{X}_{t-1} + \epsilon_t$

☐ (1) ☒ (2) ☐ (3) ☐ (4)

c) $\mathcal{X}_t = 0.5\mathcal{X}_{t-2} + \epsilon_t$

☐ (1) ☐ (2) ☐ (3) ☒ (4)

d) $\mathcal{X}_t = 0.5\mathcal{X}_{t-1} + \epsilon_t$

☒ (1) ☐ (2) ☐ (3) ☐ (4)

Problem 2 Hidden Markov Models (6 points)

0 1 2 3 4 5 6

Consider the following Hidden Markov Model where Z_t are latent variables and X_t are continuous observed variables. We parametrize the prior and transition probabilities $P(Z_1 = i) = \pi_i$ and $P(Z_{t+1} = j | Z_t = i) = A_{ij}$ by:

$$\pi = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \quad A = \begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix}$$

We parametrize the emission probabilities with $X_t | Z_t = 0 \sim \mathit{Uniform}([8, 12])$ and $X_t | Z_t = 1 \sim \mathit{Uniform}([11, 13])$. We assume we observed $X = [8.5, 11, 13, 9.5]$.

Perform the Forward algorithm, compute α_t and compute the most probable state Z_t given $X_1, ..., X_t$ at every step t.

Hint 1: You can check your computations by comparing with $\alpha_4 = \begin{bmatrix} 3/3200 \\ 0 \end{bmatrix}$

Hint 2: Maintaining non-decimal fractions makes calculations easier (e.g. use $8/10$ *instead of* 0.8*).*

The probability of observing $X_1 = 8.5$ is $1/4$ given $Z_1 = 0$, 0 given $Z_1 = 1$:

$$\alpha_1 = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix} \odot \begin{bmatrix} 1/4 \\ 0 \end{bmatrix} = \begin{bmatrix} 1/8 \\ 0 \end{bmatrix}$$

The probability of observing $X_2 = 11$ is $1/4$ given $Z_1 = 0$, $1/2$ given $Z_1 = 1$:

$$\alpha_2 = \begin{bmatrix} 1/4 \\ 1/2 \end{bmatrix} \odot \left(\begin{bmatrix} 8/10 & 1/2 \\ 2/10 & 1/2 \end{bmatrix} \begin{bmatrix} 1/8 \\ 0 \end{bmatrix} \right) = \begin{bmatrix} 1/40 \\ 1/80 \end{bmatrix}$$

The probability of observing $X_3 = 11$ is 0 given $Z_1 = 0$, $1/2$ given $Z_1 = 1$:

$$\alpha_3 = \begin{bmatrix} 0 \\ 1/2 \end{bmatrix} \odot \left(\begin{bmatrix} 8/10 & 1/2 \\ 2/10 & 1/2 \end{bmatrix} \begin{bmatrix} 1/40 \\ 1/80 \end{bmatrix} \right) = \begin{bmatrix} 0 \\ 9/1600 \end{bmatrix}$$

The probability of observing $X_4 = 11$ is $1/4$ given $Z_1 = 0$, 0 given $Z_1 = 1$:

$$\alpha_4 = \begin{bmatrix} 1/4 \\ 0 \end{bmatrix} \odot \left(\begin{bmatrix} 8/10 & 1/2 \\ 2/10 & 1/2 \end{bmatrix} \begin{bmatrix} 0 \\ 9/1600 \end{bmatrix} \right) = \begin{bmatrix} 9/12800 \\ 0 \end{bmatrix}$$

The most probable states are $0, 0, 1, 0$.

Problem 3 Recurrent neural networks (8 points)

Given a sequence of integers $\{x_1, \dots, x_n\}, x_t \in \{0, 1\}$, the goal is to output 1 at step t if the sum of all the inputs up to step t is even. That is, $h_t = 1$ if $\sum_{i=0}^{t} x_i$ is even, otherwise 0. For example, the outputs for a sequence 01101 are 10110. To model this you use an RNN with the following update equations:

$$\tilde{h}_t = a \cdot h_{t-1} + b \cdot x_t$$
$$z_t = f(c \cdot h_{t-1} + d \cdot x_t)$$
$$h_t = f((1 - z_t) \cdot h_{t-1} + z_t \cdot \tilde{h}_t)$$

where $h_0 = 1$ and $a, b, c, d \in \mathbb{R}$ and $f(x) = 1$ if $x > 0.5$, otherwise 0.

0
1
2

a) Fill out the table with all the possible input values x_t and input hidden state values h_{t-1}, and calculate the corresponding output h_t according to the given requirements.

h_{t-1}	x_t	h_t
0	0	0
0	1	1
1	0	1
1	1	0

0
1
2
3
4
5
6

b) Find the values for a, b, c and d such that this model achieves 100% accuracy. Show your work.
Hint: think about what can happen with a new input and what should gate z_t do in those cases.

One solution is the following. Notice that we want to keep h_{t-1} when $x_t = 0$. That means that z_t should be 0 when $x = 0$. On the other hand, $x_t = 1$ always flips the output, therefore, we should forget about the past by setting $z_t = 1$. From here we can see that $c = 0$ and $d = 1$.
Now we know that the update equation will look like $f(h_{t-1})$ when $x_t = 0$ and $f(\tilde{h}_t)$ when $x_t = 1$. This gives equations

$$f(0a + 1b) = 1 \iff b > 0.5$$
$$f(1a + 1b) = 0 \iff a + b \leq 0.5$$

One possible solution is $a = -1$ and $b = 1$.

Problem 4 Graph laws (4 points)

You are given an Erdős-Rényi graph $G(n, p)$, where n is the number of nodes and p is the probability of an edge.

a) What is the expected number of cliques of size m?

0 1 2

$\binom{n}{m} p^{\binom{m}{2}}$

b) We add a new node and connect it to every existing node with probablity q. What is the expected number of newly created triangles?

0 1 2

$\binom{n}{2} pq^2$

Problem 5 Deep Generative Models (6 points)

The loss used in generative adversarial networks (GANs) can be written in the following form:

$$\min_{\theta} \max_{\phi} \mathcal{L}(\theta, \phi) = \min_{\theta} \max_{\phi} \mathbb{E}_{p^*(\boldsymbol{x})}[\log D_\phi(\boldsymbol{x})] + \mathbb{E}_{p(\boldsymbol{z})}[\log(1 - D_\phi(f_\theta(\boldsymbol{z})))]$$

where $p^*(\boldsymbol{x})$ is the true data distribution, $p(\boldsymbol{z})$ is the distribution of the noise, f_θ is the generator, and D_ϕ is the discriminator.

0
1
2
3

a) For a given generator (fixed parameters θ) assume there exists a discriminator $D_{\phi^*}(\boldsymbol{x})$ with parameters ϕ^* such that for all $\boldsymbol{x}$:

$$D_{\phi^*}(\boldsymbol{x}) = \frac{p^*(\boldsymbol{x})}{p^*(\boldsymbol{x}) + p_\theta(\boldsymbol{x})}$$

where $p_\theta(\boldsymbol{x})$ is the distribution learned by the generator. Show that D_{ϕ^*} is **optimal**, i.e. $\phi^* = \arg\max_\phi \mathcal{L}(\theta, \phi)$. Hint: $\max_y[a\log(y) + b\log(1-y)] = \frac{a}{a+b}$ for any $a, b \in \mathbb{R}_0^+, a + b > 0$.

$$\begin{aligned}
\arg\max_{D_\phi} \mathcal{L}(\theta, \phi) &= \arg\max_{D_\phi} \mathbb{E}_{p^*(\boldsymbol{x})}[\log D_\phi(\boldsymbol{x})] + \mathbb{E}_{p(\boldsymbol{z})}[\log(1 - D_\phi(f_{\boldsymbol{\theta}}(\boldsymbol{z})))] \\
&= \arg\max_{D_\phi} \mathbb{E}_{p^*(\boldsymbol{x})}[\log D_\phi(\boldsymbol{x})] + \mathbb{E}_{p_\theta(\boldsymbol{x})}[\log(1 - D_\phi(x))] \\
&= \arg\max_{D_\phi} \int [p^*(\boldsymbol{x}) \log D_\phi(\boldsymbol{x}) + p_\theta(\boldsymbol{x}) \log(1 - D_\phi(x))]dx \\
&= \int \arg\max_{D_\phi} [p^*(\boldsymbol{x}) \log D_\phi(\boldsymbol{x}) + p_\theta(\boldsymbol{x}) \log(1 - D_\phi(x))]dx
\end{aligned}$$

Setting $\boldsymbol{D}_\phi(\boldsymbol{x}) = y$, $p^*(\boldsymbol{x}) = a$, and $p_\theta(\boldsymbol{x}) = b$, and substituting above we get:

$$\arg\max_{D_\phi} \mathcal{L}(\theta, \phi) = D_{\phi^*}(\boldsymbol{x}) = \frac{p^*(\boldsymbol{x})}{p^*(\boldsymbol{x}) + p_\theta(\boldsymbol{x})}$$

0
1
2
3

b) Show that $\mathcal{L}(\theta, \phi^*) = -\log(4) + 2 \cdot \mathbb{JSD}(p^*(\boldsymbol{x})||p_\theta(\boldsymbol{x}))$ where

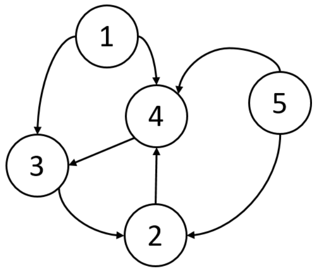
$$\mathbb{JSD}(p^*(\boldsymbol{x})||p_\theta(\boldsymbol{x})) = \frac{1}{2}\left[\mathbb{KL}(p^*(\boldsymbol{x})||m(\boldsymbol{x})) + \mathbb{KL}(p_\theta(\boldsymbol{x})||m(\boldsymbol{x}))\right]$$

is the Jensen–Shannon divergence, $\mathbb{KL}$ is the Kullback–Leibler divergence, and $m(\boldsymbol{x}) = \frac{p_\theta(\boldsymbol{x}) + p^*(\boldsymbol{x})}{2}$.

$$\begin{aligned}
\mathcal{L}(\theta, \phi^*) &= -\log(4) + 2 \cdot \mathbb{JSD}(p^*(\boldsymbol{x})||p_\theta(\boldsymbol{x})) \\
&= -\log(4) + \mathbb{KL}(p^*(\boldsymbol{x})||m(\boldsymbol{x})) + \mathbb{KL}(p_\theta(\boldsymbol{x})||m(\boldsymbol{x})) \\
&= -\log(4) + \int p^*(\boldsymbol{x}) \log \frac{2p^*(\boldsymbol{x})}{p_\theta(\boldsymbol{x}) + p^*(\boldsymbol{x})} + p_\theta \log \frac{2p_\theta}{p_\theta(\boldsymbol{x}) + p^*(\boldsymbol{x})} \\
&= \int p^*(\boldsymbol{x}) \log \frac{p^*(\boldsymbol{x})}{p_\theta(\boldsymbol{x}) + p^*(\boldsymbol{x})} + p_\theta \log \frac{p_\theta}{p_\theta(\boldsymbol{x}) + p^*(\boldsymbol{x})} \\
&= \int p^*(\boldsymbol{x}) \log \frac{p^*(\boldsymbol{x})}{p_\theta(\boldsymbol{x}) + p^*(\boldsymbol{x})} + p_\theta \log(1 - \frac{p^*(\boldsymbol{x})}{p_\theta(\boldsymbol{x}) + p^*(\boldsymbol{x})}) \\
&= \mathbb{E}_{p^*(\boldsymbol{x})}[\log D_{\phi^*}(\boldsymbol{x})] + \mathbb{E}_{p_\theta(\boldsymbol{x})}[\log(1 - D_{\phi^*}(x))] \\
&= \mathcal{L}(\theta, \phi^*)
\end{aligned}$$

Problem 6 Ranking (6 points)

Given the following graph G let $e = (s, t)$ be a new edge, i.e. an edge which is **not** yet in the graph G. We denote with G_{new} the resulting graph when adding e to G. Your task is to identify the edge e based on the following observation:
The PageRank vector of G_{new} based on a teleport set $S = \{1, 5\}$ and $\beta = 0.85$ (i.e. the probability to teleport is 0.15) is given by



0 1 2 3 4 5 6

$$\pi = [0.0750, 0.1935, 0.2668, 0.2764, 0.1884]$$

What is the edge e (specify the source node s and target node t) that has been added? Justify your answer.

Since nodes 1 and 5 have no incoming edges, their PageRank scores should equal $\frac{1-\beta}{|S|} = \frac{0.15}{2} = 0.075$. For node 1 this exactly matches, but the PageRank score for node 5 is larger, thus the target node of the new edge must be $t = 5$.

Now we only need to determine the source node.
If $s = 1$ the PageRank score of node 5 would be $0.075 + \beta\pi_1/d_1 = 0.0963 \neq \pi_5$.
If $s = 2$ the PageRank score of node 5 would be $0.075 + \beta\pi_2/d_2 = 0.1572 \neq \pi_5$.
If $s = 3$ the PageRank score of node 5 would be $0.075 + \beta\pi_3/d_3 = 0.1884 = \pi_5$.
If $s = 4$ the PageRank score of node 5 would be $0.075 + \beta\pi_4/d_4 = 0.1925 \neq \pi_5$.
If $s = 5$ the PageRank score of node 5 would be $0.075 + \beta\pi_5/d_5 = 0.1284 \neq \pi_5$.

Thus, the source node must be $s = 3$, which means the added edge is $(3, 5)$.

Problem 7 Spectral clustering (4 points)

0 1 2 3 4

You are given a graph G with $N = 1000$ nodes that is generated from a Stochastic Block Model with the following parameters

$$\pi = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \qquad \eta = \begin{bmatrix} 0.05 & 0.5 \\ 0.5 & 0.05 \end{bmatrix}$$

Assume that G is connected. Do you expect spectral clustering to recover the true communities $\boldsymbol{z}$ when applied to the graph G (yes or no)? Justify your answer.

Since the off-diagonal entries of the η matrix are much larger than the diagonal, there will be significantly more edges between the communities than within them (i.e. the cut between the true communities will be very large).

Spectral clustering on the other hand will seek a partition such that there are few edges between the communities (i.e. spectral clustering tries to minimize the cut).

Therefore, spectral clustering won't work in this scenario.

Additional space for solutions–clearly mark the (sub)problem your answers are related to and strike out invalid solutions.