Data Analytics & Machine Learning
Informatics
Technical University of Munich

Eexam
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.

Machine Learning for Graphs and Sequential Data

Exam:	IN2323 / Endterm	**Date:**	Tuesday 1st August, 2023
Examiner:	Prof. Dr. Stephan Günnemann	**Time:**	11:00 – 12:15

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

Working instructions

- This exam consists of **12 pages** with a total of **7 problems**.
  Please make sure now that you received a complete copy of the exam.
- The total amount of achievable credits in this exam is 35 credits.
- Detaching pages from the exam is prohibited.
- Allowed resources:
  - one A4 sheet of handwritten notes, two sides.
- **No other material (e.g. books, cell phones, calculators) is allowed!**
- Physically turn off all electronic devices, put them into your bag and close the bag.
- There is scratch paper at the end of the exam.
- Write your answers only in the provided solution boxes or the scratch paper.
- If you solve a task on the scratch paper, clearly reference it in the main solution box.
- All sheets (including scratch paper) have to be returned at the end.
- **Only use a black or a blue pen (no pencils, red or greens pens!)**
- **For problems that say "Justify your answer" you only get points if you provide a valid explanation.**
- **For problems that say "Derive" you only get points if you provide a valid mathematical derivation.**
- **For problems that say "Prove" you only get points if you provide a valid mathematical proof.**
- If a problem does not say "Justify your answer", "Derive" or "Prove", it is sufficient to only provide the correct answer.

Left room from __________ to __________ / Early submission at __________

Problem 1 Hidden Markov Models (4 credits)

Consider a hidden Markov model with 3 states $\{1, 2, 3\}$ and 2 possible observations $\{\mathrm{a}, \mathrm{b}\}$. The initial distribution π, transition probabilities $\boldsymbol{A}$ and emission probabilities $\boldsymbol{B}$ are

$$\pi = \begin{array}{c c} 1 & \\ 2 & \\ 3 & \end{array}\!\!\begin{pmatrix} 1/2 \\ 1/2 \\ 0 \end{pmatrix} \qquad \boldsymbol{A} = \begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 1/2 & 0 & 1/2 \\ 2 & 0 & 1/2 & 1/2 \\ 3 & 1/3 & 1/3 & 1/3 \end{array} \qquad \boldsymbol{B} = \begin{array}{c|cc} & \mathrm{a} & \mathrm{b} \\ \hline 1 & 0 & 1 \\ 2 & 1/4 & 3/4 \\ 3 & 1 & 0 \end{array},$$

where $\boldsymbol{A}_{ij}$ specifies the probability of transitioning from state i to state j.

0 1 2

a) You have observed the sequence $X_{1:3} = [\mathrm{aba}]$. Derive $\mathbb{P}(Z_3 \mid X_{1:3} = [\mathrm{aba}])$ up to a normalizing constant.

0 1 2

b) What is the most likely state sequence $\arg\max_{Z_{1:3}} \mathbb{P}(Z_{1:3} \mid X_{1:3} = [\mathrm{aba}])$? Justify your response.

Problem 2 Attention (8 credits)

Suppose we want to embed the sequence $S^{(X)} = [a, b, c, b, a, c]$ of length $N = 6$ over a vocabulary $\mathcal{V} = \{a, b, c\}$. We store the input sequence in a matrix $\boldsymbol{X} \in \{0,1\}^{N \times |\mathcal{V}|}$, where the i-th row $\boldsymbol{X}_i$ corresponds to a 1-hot representation of the i-th token. We use **masked-attention** to embed individual words which restricts the attention mechanism to a subset of other words in the sequence.
Masked attention is defined as:

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}^{(Q)} \quad \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}^{(K)} \quad \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}^{(V)}$$

$$\boldsymbol{H} = \text{masked-softmax}_{\boldsymbol{M}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}}\right)\boldsymbol{V}$$

When applying the softmax activation to obtain normalized attention scores, we ignore all entries where the mask $\boldsymbol{M} \in \{0,1\}^{N \times N}$ has zero value:

$$\text{masked-softmax}_{\boldsymbol{M}}(\boldsymbol{A})_{i,j} = \begin{cases} \frac{\exp(\boldsymbol{A}_{i,j})}{\sum_{k:\boldsymbol{M}_{i,k} \neq 0} \exp(\boldsymbol{A}_{i,k})} & \text{if } \boldsymbol{M}_{i,j} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

Also, we use linear transformations $\boldsymbol{W}^{(Q)} \in \mathbb{R}^{|\mathcal{V}| \times d}$, $\boldsymbol{W}^{(K)} \in \mathbb{R}^{|\mathcal{V}| \times d}$, $\boldsymbol{W}^{(V)} \in \mathbb{R}^{|\mathcal{V}| \times d}$ to compute queries $\boldsymbol{Q}$, keys $\boldsymbol{K}$ and values $\boldsymbol{V}$ respectively.

a) For different realizations of attention masks, we now want to analyse which words are assigned identical embeddings. You are provided with different attention masks $\boldsymbol{M}$ and your task is to find groups of input words that the corresponding masked-attention mechanism **can not** distinguish. That is, for each $\boldsymbol{M}^{(i)}$ list all groups of words of the sequence $S^{(X)}$ that will be assigned the same embedding no matter the choice of $\boldsymbol{W}^{(Q)}, \boldsymbol{W}^{(K)}, \boldsymbol{W}^{(V)}$. For example, if for $\boldsymbol{M}^{(i)}$ the first three tokens are assigned the same embedding and the last three tokens are assigned to the same embedding, your answer should be $\boldsymbol{M}^{(i)} : \{1, 2, 3\}, \{4, 5, 6\}$.

0 1 2 3

$$\boldsymbol{M}^{(1)} = \begin{pmatrix} 1&1&0&0&0&0 \\ 1&1&1&0&0&0 \\ 0&1&1&1&0&0 \\ 0&0&1&1&1&0 \\ 0&0&0&1&1&1 \\ 0&0&0&0&1&1 \end{pmatrix} \quad \boldsymbol{M}^{(2)} = \begin{pmatrix} 1&1&1&1&1&1 \\ 1&1&1&1&1&1 \\ 1&1&1&1&1&1 \\ 1&1&1&1&1&1 \\ 1&1&1&1&1&1 \\ 1&1&1&1&1&1 \end{pmatrix} \quad \boldsymbol{M}^{(3)} = \begin{pmatrix} 0&0&0&0&0&1 \\ 0&0&0&0&0&1 \\ 0&0&0&0&0&1 \\ 0&0&0&0&0&1 \\ 0&0&0&0&0&1 \\ 0&0&0&0&0&1 \end{pmatrix}$$

b) Name and briefly explain a method that is employed in practice such that the words in any sequence can be distinguished from each other by the attention mechanism regardless of the choice of the attention mask $\boldsymbol{M}$ (as long as each row $\boldsymbol{M}_{i,:}$ contains at least two non-zero elements).

0 1 2

0
1
2
3

c) In practice, masked-attention is realized by setting unnormalized scores $\boldsymbol{A}_{i,j}$ to very low values wherever the attention mask $\boldsymbol{M}_{i,j} = 0$. Assume that the attention mask $\boldsymbol{M}$ has ones everywhere except at position i, k where it is set to zero. Show that for all $j \in \{1, ..., N\}$, we effectively recover the same behaviour as masking out the corresponding value $\boldsymbol{A}_{i,k}$ in the limit. That is, for all $j \in \{1, ..., N\}$, show that $\lim_{\boldsymbol{A}_{i,k} \to -\infty} \text{softmax}(\boldsymbol{A})_{i,j} = \text{masked-softmax}_{\boldsymbol{M}}(\boldsymbol{A})_{i,j}$.

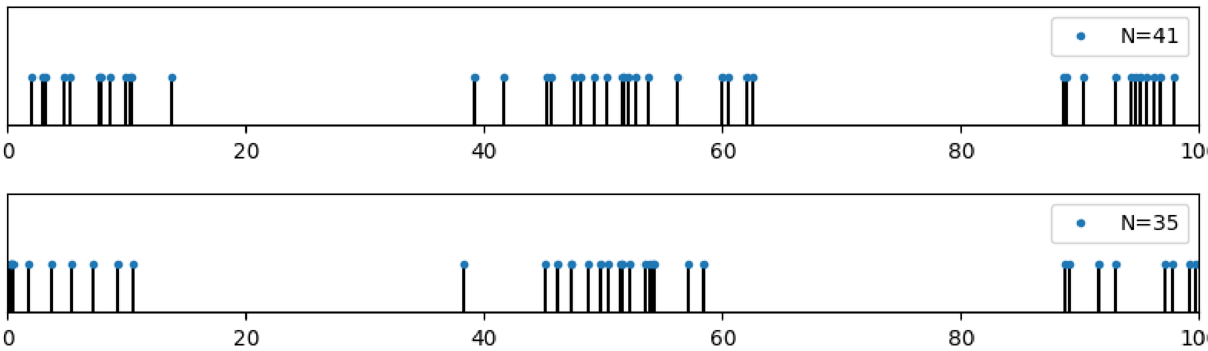
Problem 3 Temporal Point Processes (6 credits)

In the following, you are presented with five intensity functions of different temporal point processes. Your task is to subsequently match one of the intensity functions to the point process samples presented in the subtasks. Please justify your decision accordingly. Each intensity function can be only used once.

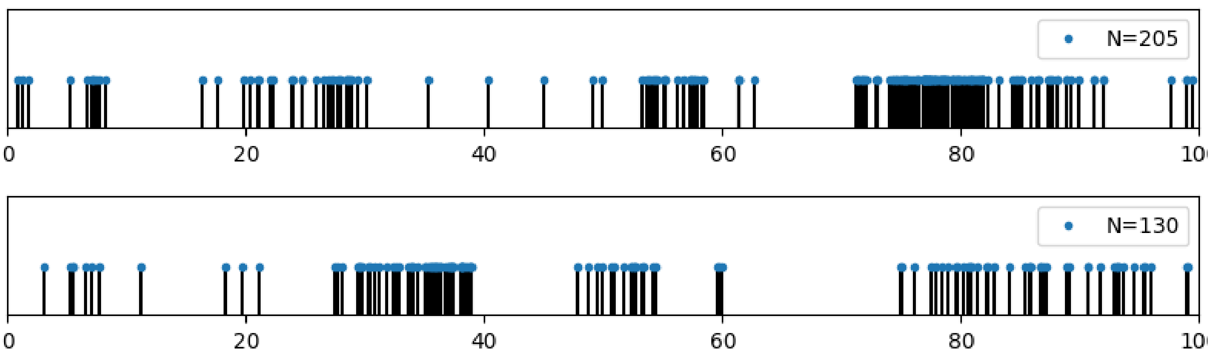
1) $\lambda^*(t) = 1.2$ **2)** $\lambda^*(t) = ReLU(0.25 + \cos(\frac{1}{25}\pi t))$ **3)** $\lambda^*(t) = 0.2 + 0.6 \sum_{t_i \in \mathcal{H}(t)} e^{-(t-t_i)}$

4) $\lambda^*(t) = 0.6$ **5)** $\lambda^*(t) = 0.2 + 0.9 \sum_{t_i \in \mathcal{H}(t)} e^{-(t-t_i)}$

a) To which of the five intensity functions do the two samples most likely belong?

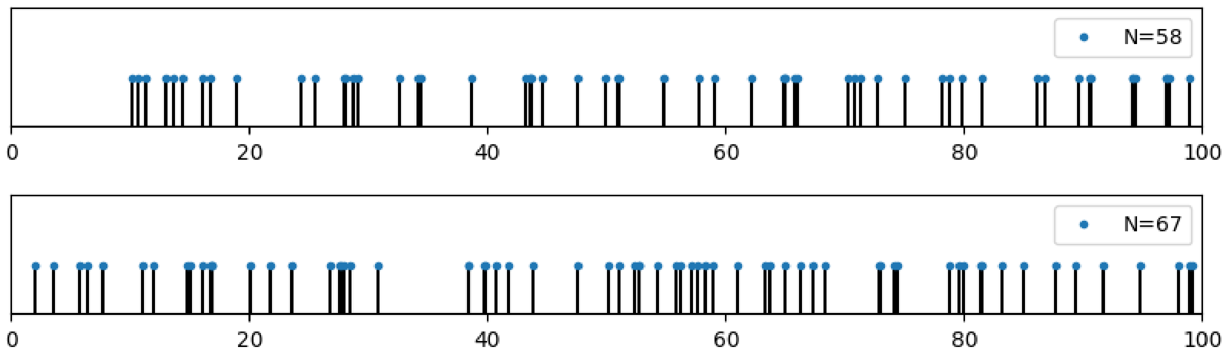


b) To which of the five intensity functions do the two samples most likely belong?



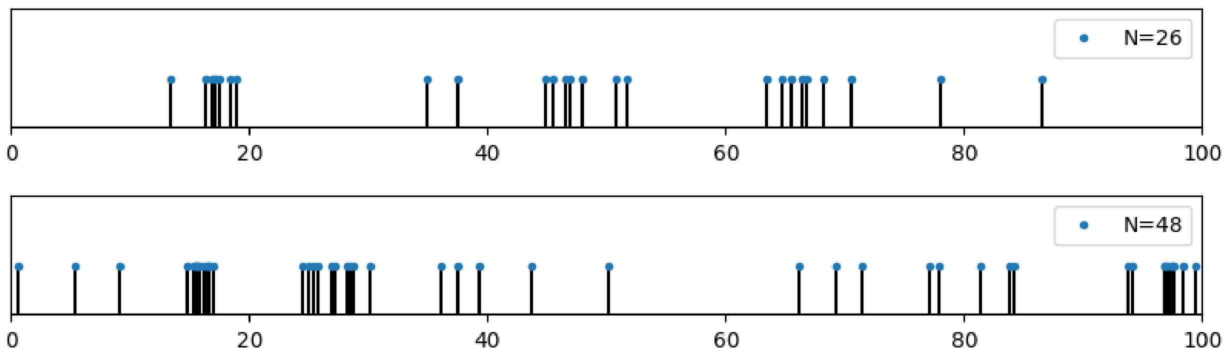
0
1

c) To which of the five intensity functions do the two samples most likely belong?



0
1

d) To which of the five intensity functions do the two samples most likely belong?



Problem 4 Spectral Clustering (5 credits)

a) Assume you are given a graph with 2 disconnected components of equal size and want to cluster the graph into 2 clusters. Provide an assignment vector $\mathbf{f}_{C_1}$ for which it holds that

$$\mathbf{f}_{C_1}^T L \mathbf{f}_{C_1} = 0 \quad \text{with } \|\mathbf{f}_{C_1}\|_2 > 0 \tag{4.1}$$

Justify.

0 1 2 3

b) Does the normalized cut minimize or maximize in-cluster associativity? Show your answer via the definition of the in-cluster associativity:

$$\text{assoc}(C_1, C_1) = \sum_{u \in C_1, v \in C_1} W_{uv} \tag{4.4}$$

0 1 2

Problem 5 **Ranking (2 credits)**

0
1
2

Assume you are given a function, which starts a random surfer without teleportation from some point v_i on a secret graph for some specified number of steps n. It returns the probability of finding the random surfer on each of the nodes after n steps. You observe that even for a sufficiently large number of steps n, you get significantly different outputs when you vary v_i. What can you say about the secret graph? How could you modify the random surfer function to solve this issue?

Problem 6 **Equivariant Machine Learning on Graphs (5 credits)**

Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ be the feature matrix of an attributed graph with n nodes. We denote the feature vectors as $\boldsymbol{x}_1, \dots, \boldsymbol{x}_d \in \mathbb{R}^n$ (the columns of $\boldsymbol{X}$). Note that $\boldsymbol{x}_i$ represents the i-th feature dimension of **all** nodes, $\boldsymbol{x}_i \in \mathbb{R}^n$.

A function f is called invariant to rescaling if

$$f(s_1\boldsymbol{x}_1, \dots, s_d\boldsymbol{x}_d) = f(\boldsymbol{x}_1, \dots, \boldsymbol{x}_d) \qquad s_i \in \mathbb{R} \setminus \{0\}$$

Consider the following function

$$g(\boldsymbol{x}_1, \dots, \boldsymbol{x}_d) = \psi\left(\bigotimes_{i=1}^{D}\left[\phi\left(\frac{\boldsymbol{x}_i}{||\boldsymbol{x}_i||}\right) + \phi\left(-\frac{\boldsymbol{x}_i}{||\boldsymbol{x}_i||}\right)\right]\right)$$

where ψ and ϕ are neural networks, $\bigotimes$ an aggregation function such as sum, and $||\cdot||$ any ℓ_p-vector norm.

Show that g is invariant to rescaling.

0
1
2
3
4
5

Problem 7 Robustness - Discrete Randomized Smoothing (5 credits)

We want to certify our message-passing Graph Neural Network f_θ against edge perturbations using discrete randomized smoothing. We define the smoothed classifier for graphs $\boldsymbol{G} = (\boldsymbol{A}, \boldsymbol{X})$ as

$$g(\boldsymbol{G})_c = \Pr_{\phi}[f(\phi(\boldsymbol{A}), \boldsymbol{X}) = c]$$

where ϕ is the sparsity-aware smoothing distribution with edge deletion probability $p_d = \frac{1}{4}$ and edge addition probability $p_a = \frac{1}{2}$. Assume we know the adjacency matrix of the clean graph $\boldsymbol{G}$ and the perturbed graph $\tilde{\boldsymbol{G}}$:

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \tilde{\boldsymbol{A}} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

0 1 2

a) Consider the two graphs $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ with the following adjacency matrices:

$$\boldsymbol{A}_1 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \qquad \boldsymbol{A}_2 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

Compute the probabilities 1. $\Pr_\phi[\boldsymbol{A}_1 \mid \boldsymbol{A}]$, 2. $\Pr_\phi[\boldsymbol{A}_1 \mid \tilde{\boldsymbol{A}}]$, 3. $\Pr_\phi[\boldsymbol{A}_2 \mid \boldsymbol{A}]$, 4. $\Pr_\phi[\boldsymbol{A}_2 \mid \tilde{\boldsymbol{A}}]$.

0 1 2

b) To compute $\Pr_\phi\left[h(\phi(\tilde{\boldsymbol{A}}), \boldsymbol{X}) = c^*\right]$ under the worst-possible classifier h^* we have to select graphs that will be classified as c^* while ensuring $\Pr_\phi[h(\phi(\boldsymbol{A}), \boldsymbol{X}) = c^*] = g_{c^*}(\boldsymbol{G})$ for clean graph $\boldsymbol{G}$. Consider a classifier h with $h(\boldsymbol{G}_1) = c_{\text{other}}$ and $h(\boldsymbol{G}_2) = c^*$. Can h be a worst-case classifier? Why or why not? Justify your answer.
Hint: If you did not solve the previous exercise you can use $\Pr_\phi[\boldsymbol{A}_1 \mid \boldsymbol{A}] = 0.2$, $\Pr_\phi[\boldsymbol{A}_1 \mid \tilde{\boldsymbol{A}}] = 0.1$, $\Pr_\phi[\boldsymbol{A}_2 \mid \boldsymbol{A}] = 0.1$, $\Pr_\phi[\boldsymbol{A}_2 \mid \tilde{\boldsymbol{A}}] = 0.2$.

0 1

c) Are the flipping probabilities $p_d = \frac{1}{4}$ and $p_a = \frac{1}{2}$ a good choice in practice? Why or why not?

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