Pregatire IAIO

Probleme Teoretice

الأولمبياد الدولي للذكاء الاصطناعي

International AI Olympiad

1 – Introduction to AI

2 – Societal impact of AI (ethics, fairness)

What is the primary legal concern for the New York Times in this lawsuit?

A) Unauthorized use of copyrighted content.

B) The potential decrease in their readership.

C) The risk of AI-generated content competing with human journalists.

D) The ethical implications of AI creating content.

(Correct: A)

What ethical concern does the use of data without permission raise?

A) AI replacing human creativity.

B) The exploitation of proprietary content without compensation.

C) The environmental impact of training large AI models.

D) The quality and accuracy of AI-generated content.

(Correct: B)

How could OpenAI address the copyright concerns raised by the New York Times?

A) By purchasing a license to use the content.

B) By claiming fair use for training purposes.

C) By only using public domain data in future models.

D) By providing a share of profits from AI-generated content.

(Correct: A, D)

What broader implications does this lawsuit have for the AI industry?

- A) Increased scrutiny and regulation of data usage practices.
- B) A potential decline in AI research due to legal risks.
- C) Development of new technologies to protect intellectual property.
- D) A shift towards more transparent and ethical AI development practices.

(Correct: A, D)

If the court rules in favor of the New York Times, what might OpenAI need to change in its data collection practices?

A) Exclude all copyrighted material from training datasets.

B) Implement more rigorous data licensing agreements.

C) Develop proprietary data for model training.

D) Rely on user-generated content for training purposes.

(Correct: A, B, C)

3 – Kernel Methods

Sample question

This question concerns implementing a learning strategy in a kernel defined feature space. Suppose we are given a kernel function:

$$\kappa(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$$

and a labelled training set

$$S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\},$$

where $y_i \in \{-1, +1\}$, $i = 1, \ldots, m$. Consider some vector

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i \phi(\mathbf{x}_i)$$

in the kernel defined feature space.

Sample question

- Write down the expression to evaluate the inner product $\langle \mathbf{w}, \phi(\mathbf{x}) \rangle$, for an input $\mathbf{x}$. [1 mark]

$$\langle \mathbf{w}, \phi(\mathbf{x}) \rangle = \sum_{i=1}^{m} \alpha_i \kappa(\mathbf{x}_i, \mathbf{x})$$

- Write down the expression for the average $\mu_{S'}$ of the inner products with $\mathbf{w}$ over a subset $S' \subseteq S$. [1 mark]

$$\mu_{S'} = \frac{1}{|S'|} \sum_{\mathbf{x} \in S'} \sum_{i=1}^{m} \alpha_i \kappa(\mathbf{x}_i, \mathbf{x})$$

Sample question

- Hence or otherwise, write down the optimisation criterion for choosing $\mathbf{w}$ with norm 1 to maximise $\mu_{S^+} - \mu_{S^-}$, where $S^+ = \{(\mathbf{x}, y) \in S : y = +1\}$ and similarly for S^-. [2 marks]

$$\begin{aligned}
\mathbf{w} &= \operatorname{argmax}_{\mathbf{w}:\|\mathbf{w}\|=1} \mu_{S^+} - \mu_{S^-} \\
&= \operatorname{argmax}_{\alpha} \frac{1}{|S^+|} \sum_{\mathbf{x} \in S^+} \sum_{i=1}^{m} \alpha_i \kappa(\mathbf{x}_i, \mathbf{x}) - \frac{1}{|S^-|} \sum_{\mathbf{x} \in S^-} \sum_{i=1}^{m} \alpha_i \kappa(\mathbf{x}_i \\
&= \operatorname{argmax}_{\alpha:\alpha' K \alpha = 1} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i b_j \kappa(\mathbf{x}_i, \mathbf{x}_j),
\end{aligned}$$

where $b_j = \begin{cases} 1/|S^+|; \text{ if } y_j = 1 \\ 1/|S^-|; \text{ otherwise.} \end{cases}$

Sample question

- Write down the solution to the optimisation problem. [3 marks]

$$\begin{aligned}\alpha &= \operatorname{argmax}_{\alpha} \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i b_j \kappa(\mathbf{x}_i, \mathbf{x}_j), \\ &= \operatorname{argmax}_{\alpha:\alpha' K \alpha = 1} \left\langle \sum_{i=1}^{m} \alpha_i \phi(\mathbf{x}_i), \sum_{j=1}^{m} b_j \phi(\mathbf{x}_j) \right\rangle\end{aligned}$$

$$\text{where } b_j = \begin{cases} 1/|S^+|; \text{ if } y_j = 1 \\ 1/|S^-|; \text{ otherwise.} \end{cases}$$

Inner product is maximised by choosing unit vector parallel to $\sum_{j=1}^{m} b_j \phi(\mathbf{x}_j)$, implying

$$\alpha_i = b_i / \mathbf{b}' K \mathbf{b}$$

4 – Working with Data

Question

- Connie is using cross-validation to select the best value for a hyperparameter λ in her regularization term. She has provided the validation errors for each fold and the corresponding fitted parameters for the training data points. The errors and parameters for each fold across three choices of λ are shown in the tables below. Assume that the training and validation sets for each fold are consistent across all three choices of λ.

Fold Num	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
1	1	0.5	9
2	3	1.5	1

Fold Num	Training Data	θ_1	θ_2
1	Rows 1 and 2	-1	1
2	Rows 3 and 4	0	2

- Given this information, which value of λ should Connie choose to achieve the lowest validation error?

Answer

Let's calculate the average error for each λ:

Fold Num	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
1	1	0.5	9
2	3	1.5	1

For $\lambda = 1$:

$\text{Average error} = \frac{1+3}{2} = \frac{4}{2} = 2$

For $\lambda = 2$:

$\text{Average error} = \frac{0.5+1.5}{2} = \frac{2}{2} = 1$

For $\lambda = 3$:

$\text{Average error} = \frac{9+1}{2} = \frac{10}{2} = 5$

Based on these calculations, the average errors for each λ are:

- $\lambda = 1$: 2
- $\lambda = 2$: 1
- $\lambda = 3$: 5

Connie should choose $\lambda = 2$ as it has the lowest average error, indicating the best performance among the given options.

Question

- A dataset contains 5000 samples and a categorical feature with five possible values: {Apple, Banana, Cherry, Date, Elderberry}. Answer the following:

- Question: If you apply one-hot encoding to this feature, how many binary columns will be created?
- Answer: Binary columns created: 5 (one for each possible value).

- Question: For an observation with the value 'Cherry', provide its one-hot encoded representation.
- Answer: One-hot encoded representation for 'Cherry': [0, 0, 1, 0, 0].

5 – Deployed deep generative models

Exercise 1

- A language model uses a Transformer architecture with a vocabulary size of V=10,000 discrete tokens. Each token is represented by an embedding vector of dimensionality d=256.
- a. Calculate the total number of parameters in the embedding matrix used for encoding these tokens.
  Answer: The embedding matrix has dimensions equal to the vocabulary size and the embedding dimension:

$$\text{Total Parameters} = V \times d = 10{,}000 \times 256 = 2{,}560{,}000$$

Exercise 1

B. The feed-forward network within each Transformer layer consists of two linear transformations: one with input dimension d and output dimension f=1024, and another with input dimension f and output dimension d. Compute the total number of parameters for the feed-forward network in a single Transformer layer.

Answer: The feed-forward network consists of two linear transformations:

1. From d to f
2. From f back to d

- Parameters for the first linear transformation:
  $d \times f = 256 \times 1024 = 262{,}144$
- Parameters for the second linear transformation:
  $f \times d = 1024 \times 256 = 262{,}144$
- Total Parameters for Feed-Forward Network:
  $262{,}144 + 262{,}144 = 524{,}288$

Exercise 2

Consider a BERT model with 12 Transformer layers, each having a hidden size of 768. The model uses a vocabulary of 30,000 tokens, and each token is represented by an embedding vector of dimensionality 768.

a. Calculate the total number of parameters in the embedding matrix used for encoding the tokens.

- Answer: The embedding matrix has dimensions equal to the vocabulary size and the hidden size:
- Total Parameters=Vocabulary Size× Hidden Size=30,000×768=23,040,000

Exercise 2

B. Each Transformer layer in BERT consists of a multi-head self-attention mechanism with 12 heads, where each head has a dimensionality of dk=64. Calculate the total number of parameters for the projection matrices used to compute the queries, keys, and values in a single Transformer layer.

Answer: Each head has its own projection matrices (3 matrices) for queries, keys, and values, and there are 12 heads in total.

- Parameters for each projection matrix=Hidden Size×dk=768×64=49,152

Since there are 3 projection matrices (queries, keys, and values) per head and 12 heads:

- Total Parameters=3×49,152×12=1,769,472

6 – Supervised learning

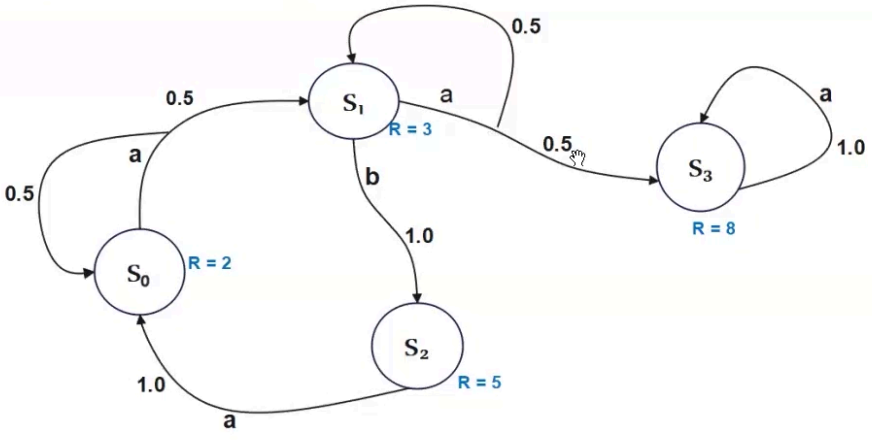
7 – Ai-Search

8 – Learning Evaluation

9 – Reinforcement learning

Question

1. Consider the MDP given in the figure below. Assume $\gamma = 0.9$. The R-values are rewards that the agent gets when in that state. The numbers on arrows are probabilities. Except state s_1 with actions (a&b), other states have only one action (a) for each state.



Question

- Calculate the optimal value of state S_1. (Hint: first find the optimal policy for S_3 using the closed form solution. Assume the optimal policy from state S_1 is to take action a) .
  $V(S_3) = 8 + 0.9 * V(S_3) = 80$
  $V(S_1) = 3 + max(0.9(0.5 * V(S_1) + 0.5 * V(S_3)), 0.9(V(S_2)))$
  Since a_* from state s_1 is a therefore 1^{st} item is greater than 2^{nd} item
  $V(S_1) = 3 + 0.9(0.5 * V(S_1) + 40) = 3 + 0.45 * V(s_1) + 36$
  $0.55 * V(S_1) = 39$
  $V(s_1) = 70.90$

Question

2. Suppose $\gamma = 0.6$ and the following sequence of rewards is received $R_1 = 1, R_2 = 2, R_3 = 6$, and $R_4 = 2$, with $T = 4$. What are $G_0, G_1, ..., G_4$? Hint: Work backwards.

$$G_4 = R_5 + R_6 + \ldots = 0$$

$$G_3 = R_4 + \gamma * G_4 = 2$$

$$G_2 = R_3 + \gamma * G_3 = 7.2$$

$$G_1 = R_2 + \gamma * G_2 = 6.32$$

$$G_0 = R_1 + \gamma * G_1 = 4.792$$

Question

3. How do we know if the history is important or not? Is there any intuition or heuristic?
   The importance of history is typically determined by the structure of the environment and the task at hand. If the environment has the Markov property, where the future state depends only on the current state and action, then history is not crucial. However, if the environment lacks the Markov property or has partial observability, history becomes important.

10 – Introduction to solving a real world machine learning problem on the Zindi platform