

2026 IAI²O AI Innovators Challenge (AIIC) Final

Part 1 - AI Theoretical Foundations

Time Allowed: 120 minutes

Total Marks: 40

Name: _____

Email: _____

Instructions

- Answer **all four questions**. Show sufficient working for Questions 2–4.
- For Question 1, **select ALL correct choices**. Partial credit is available for partially correct selections; incorrect selections may reduce the mark. No subquestion can receive a negative score.
- **Calculators are not permitted**. Unless otherwise stated, natural logarithms are used.
- You may use the reference information below without derivation.

Reference Information - No Calculator**Useful values**

$$\begin{aligned}
 e &\approx 2.718, & \ln 2 &\approx 0.693, \\
 \ln 3 &\approx 1.099, & \ln 4 &\approx 1.386, \\
 \sqrt{2} &\approx 1.414, & \sqrt{3} &\approx 1.732, \\
 \sqrt{5} &\approx 2.236.
 \end{aligned}$$

Classical ML

$$P(Y | X) = \frac{P(X | Y)P(Y)}{P(X)}$$

Naive Bayes score:

$$\text{score}(c) = P(Y = c) \prod_j P(X_j | Y = c).$$

Binary Laplace smoothing:

$$P(X_j = 1 | Y = c) = \frac{n_{jc} + \alpha}{n_c + 2\alpha}.$$

SVM geometric margin:

$$\frac{y_i(\mathbf{w}^T \mathbf{x}_i + b)}{\|\mathbf{w}\|}.$$

Neural Networks

$$n_{\text{out}} = \left\lceil \frac{n + 2p - k}{s} \right\rceil + 1$$

Conv parameters:

$$(k_h k_w C_{\text{in}} + 1) C_{\text{out}}$$

$$p_j = \frac{e^{z_j}}{\sum_k e^{z_k}}, \quad L = -\ln p_y$$

$$\nabla_{\mathbf{z}} L = \mathbf{p} - \mathbf{y}$$

Attention:

$$\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V.$$

Do not open or begin until instructed by the proctor.

Question 1 Multiple Select

10 marks

Each subquestion is worth **2 marks**. Select **ALL** correct answers.

1.1 Recurrent Neural Networks and LSTMs

Consider a standard recurrent neural network with

$$\mathbf{h}_t = \tanh(W_h \mathbf{h}_{t-1} + W_x \mathbf{x}_t + \mathbf{b}).$$

Which statements are correct?

- A.** During backpropagation through time, gradients with respect to earlier hidden states involve products of Jacobian matrices across multiple time steps.
- B.** If the relevant Jacobian factors repeatedly have spectral norm substantially below 1, gradients propagated over many time steps may become very small.
- C.** Replacing $\tanh$ with ReLU guarantees that exploding and vanishing gradients can no longer occur.
- D.** In an LSTM, the additive cell-state update provides a path through time whose derivative is controlled directly by the forget gates, which can help preserve gradients over long sequences.
- E.** If an LSTM forget gate outputs values close to 0 for many consecutive time steps, information in the cell state is more likely to be preserved over long distances.

1.2 Reinforcement Learning and Q-Learning

Consider tabular Q-learning with update rule

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[r + \gamma \max_{a'} Q(s', a') - Q(s, a) \right].$$

For one non-terminal transition, suppose

$$Q(s, a) = 2, \quad r = 1, \quad \gamma = 0.5, \quad \alpha = 0.25,$$

and

$$\max_{a'} Q(s', a') = 6.$$

Which statements are correct?

- A.** The temporal-difference target for this transition is 4.
- B.** After this update, $Q(s, a) = 2.5$.
- C.** If s' were instead a terminal state with the same reward $r = 1$, the temporal-difference target would still be 4.
- D.** Q-learning is off-policy because its update target uses the greedy action value at the next state even if the behavior policy selected actions using an exploratory strategy such as ϵ -greedy.
- E.** Decreasing ϵ in an ϵ -greedy behavior policy directly changes the numerical TD target for this particular transition, even when all Q -values, r , and γ remain unchanged.

1.3 A* Search and Heuristic Design

Let h_1 and h_2 be admissible heuristics for the same search problem, and let $0 \leq \lambda \leq 1$. Which statements are correct?

- A.** The heuristic $h_{\max}(n) = \max\{h_1(n), h_2(n)\}$ is admissible.

- B.** The heuristic $h_{\Sigma}(n) = h_1(n) + h_2(n)$ must be admissible.
- C.** If both h_1 and h_2 are consistent, then $h_{\max}(n) = \max\{h_1(n), h_2(n)\}$ is also consistent.
- D.** If both h_1 and h_2 are admissible, then $h_{\min}(n) = \min\{h_1(n), h_2(n)\}$ must also be consistent.
- E.** The convex combination $h_{\lambda}(n) = \lambda h_1(n) + (1 - \lambda)h_2(n)$ is admissible.

1.4 Softmax, Temperature, and Cross-Entropy

For logits $z_1, \dots, z_K$ and temperature $T > 0$, define

$$p_i(T) = \frac{e^{z_i/T}}{\sum_{j=1}^K e^{z_j/T}}.$$

Which statements are correct?

- A.** Adding the same constant to every logit leaves all probabilities $p_i(T)$ unchanged.
- B.** Multiplying every logit by the same positive constant leaves all probabilities $p_i(T)$ unchanged.
- C.** For any fixed $T > 0$, a class with the largest logit also has the largest softmax probability.
- D.** As $T \rightarrow \infty$, the softmax distribution approaches the uniform distribution over the K classes.
- E.** For one-hot cross-entropy, the gradient with respect to the original logits z_i is always exactly $p_i(T) - y_i$, independent of T .

1.5 Recommender Systems

A recommender system predicts user preferences using matrix factorization. For user u and item i ,

$$\hat{r}_{ui} = \mathbf{p}_u^T \mathbf{q}_i,$$

where $\mathbf{p}_u$ and $\mathbf{q}_i$ are learned latent-factor vectors. Which statements are correct?

- A.** In an explicit-feedback dataset, every missing entry in the user–item rating matrix should normally be treated as a rating of 0 during training.
- B.** Matrix factorization can learn similar latent representations for users who exhibit similar preference patterns, even if they have not rated exactly the same set of items.
- C.** If the model uses only user IDs and item IDs, a completely new item with no observed interactions creates a cold-start problem because its latent representation cannot be reliably learned from interaction data.
- D.** If there is no regularization, replacing every user vector by $\mathbf{p}'_u = c\mathbf{p}_u$ and every item vector by $\mathbf{q}'_i = \frac{1}{c}\mathbf{q}_i$, where $c \neq 0$, leaves all dot-product predictions unchanged.
- E.** Increasing the dimension of the latent-factor vectors must improve prediction accuracy on unseen users and items.

Question 2 Classical Machine Learning**10 marks**

A research team is comparing several classical machine-learning approaches for a binary classification task.

(a) Naive Bayes**[3 marks]**

A training dataset contains 10 examples:

$$P(Y = 1) = 0.4, \quad P(Y = 0) = 0.6.$$

Among the four examples with $Y = 1$:

- $X_1 = 1$ occurs 3 times;
- $X_2 = 1$ occurs 1 time.

Among the six examples with $Y = 0$:

- $X_1 = 1$ occurs 1 time;
- $X_2 = 1$ occurs 3 times.

Assume X_1 and X_2 are binary and conditionally independent given Y . Use **Laplace smoothing with $\alpha = 1$** .

For a new observation

$$X_1 = 1, \quad X_2 = 0,$$

answer the following:

1. Compute $P(Y = 1 \mid X_1 = 1, X_2 = 0)$.
2. State the predicted class.
3. Suppose X_1 and X_2 are later found to remain strongly correlated even after conditioning on Y . Which Naive Bayes assumption is violated? Does this necessarily imply that the classifier's prediction is wrong?

(b) Support Vector Machine**[3 marks]**

A linear SVM has

$$\mathbf{w} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad b = -4.$$

Consider the four training points

$$A = (2, 1), \quad y_A = +1, \quad B = (3, 0), \quad y_B = +1,$$

$$C = (1, 1), \quad y_C = -1, \quad D = (0, 3), \quad y_D = -1.$$

(i) Compute the functional margin

$$y_i(\mathbf{w}^T \mathbf{x}_i + b)$$

for each point. Identify the points lying on the canonical support hyperplanes.

(ii) Now define

$$\mathbf{w}' = 3\mathbf{w}, \quad b' = 3b.$$

Without retraining the model, explain what happens to: (1) the decision boundary; (2) the functional margins; and (3) the geometric margins.

(c) XGBoost**[4 marks]**

At one node in an XGBoost tree,

$$G_P = -2, \quad H_P = 4.$$

Two candidate splits are considered.

	Left child	Right child
Split A	$G_L = -5, H_L = 2$	$G_R = 3, H_R = 2$
Split B	$G_L = -3, H_L = 1$	$G_R = 1, H_R = 3$

Let $\lambda = 1$. The optimal weight of a leaf is

$$w^* = -\frac{G}{H + \lambda}.$$

The gain of a split is

$$\text{Gain} = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G_P^2}{H_P + \lambda} \right] - \gamma.$$

1. Compute the optimal left and right leaf weights for **Split A**.
2. Compute the gain of Split A and Split B when $\gamma = 1$. Which split should be preferred?
3. Determine the range of γ for which Split A has **strictly positive gain**.

Question 3 CNN, Softmax, and Backpropagation**10 marks**

A convolutional neural network receives an 8×8 grayscale image. The architecture is

$$8 \times 8 \times 1 \longrightarrow \text{Conv}(2 \text{ filters}, 3 \times 3, \text{ stride } 1, \text{ no padding}) \longrightarrow \text{ReLU} \\ \longrightarrow 2 \times 2 \text{ Max Pooling, stride } 2 \longrightarrow \text{Flatten} \longrightarrow \text{Dense}(3 \text{ logits}).$$

(a) Architecture Analysis**[3 marks]**

Determine:

1. the output shape immediately after the convolution;
2. the output shape after max pooling;
3. the total number of trainable scalar parameters in the entire network, including biases;
4. the receptive-field size, in the original input image, of one activation after the max-pooling layer.

Show your reasoning.

(b) Forward Propagation**[3 marks]**

For one particular input image, suppose the flattened activation vector is

$$\mathbf{a} \in \mathbb{R}^{18}, \quad a_1 = 1, \quad a_2 = 2, \quad a_3 = \dots = a_{18} = 0.$$

The final Dense layer computes

$$\mathbf{z} = \mathbf{a}^T W + \mathbf{b}.$$

Assume $\mathbf{b} = \mathbf{0}$, and the **only nonzero entry** of W is

$$W_{1,1} = \ln 2.$$

1. Compute the three logits.
2. Compute the exact softmax probabilities.
3. If the true class is **class 2**, compute the exact cross-entropy loss.

Classes are indexed as 1, 2, 3.

(c) Backpropagation**[4 marks]**

For softmax followed by cross-entropy loss,

$$\frac{\partial L}{\partial z_j} = p_j - y_j,$$

where $\mathbf{y}$ is the one-hot encoded target. Using your answer from part (b):

1. Compute $\nabla_{\mathbf{z}} L$.
2. Prove that

$$\boxed{\frac{\partial L}{\partial W} = \mathbf{a} (\nabla_{\mathbf{z}} L)^T}$$

for this Dense layer.

3. Compute the **first two rows** of $\partial L / \partial W$. What are rows 3 through 18?
4. Based only on the signs of the components of $\nabla_{\mathbf{z}} L$, explain how gradient descent tends to change the three logits.

Question 4 Advanced Deep Learning: Self-Attention**10 marks****This is the most difficult question on the examination.**

Consider two input-token representations

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

A single-head self-attention layer uses

$$W_Q = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad W_K = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad W_V = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Define

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V.$$

Since $d_k = 1$, the scaled dot-product attention score reduces to

$$S = QK^T.$$

Row-wise softmax gives

$$A = \text{softmax}(S),$$

and the attention output is

$$O = AV.$$

(a) Forward Pass**[3 marks]**

Compute explicitly

$$Q, \quad K, \quad V, \quad S, \quad A, \quad O.$$

You may leave your answer in terms of e .**(b) Permutation Equivariance****[4 marks]**Let P be any permutation matrix that reorders the input tokens, and define

$$X' = PX.$$

The same parameter matrices W_Q, W_K, W_V are used.

Prove that self-attention without positional information satisfies

$$\boxed{O' = PO}.$$

Your proof must establish

$$Q' = PQ, \quad K' = PK, \quad V' = PV,$$

and then

$$S' = PSP^T.$$

You must also justify why row-wise softmax satisfies

$$\boxed{\text{softmax}(PSP^T) = P \text{softmax}(S) P^T}.$$

Use these results to prove $O' = PO$.

(c) Positional Information and Causal Masking**[3 marks]**

Answer both parts.

(i) Explain why the result

$$O' = PO$$

shows that self-attention **without positional information cannot distinguish sequences purely from token order**. Explain how positional encodings address this limitation.

(ii) Now suppose a fixed causal mask M is added before softmax:

$$A = \text{softmax}(S + M).$$

The input tokens are permuted using P , but the **same fixed mask** M is used afterward.

Is

$$O' = PO$$

still guaranteed for an arbitrary permutation P ? Justify your answer mathematically.

End of Part 1

Check that you have attempted all four questions.