

IAI²O Final Examination – Form A

Integrated Problems – Question Paper

Time: 120 minutes Total: 40 points

Instructions. Question 1 contains five multiple-select items. For each item, select *all* correct options. Questions 2–4 require clear working and reasoning. Exact forms are preferred where reasonable. Question 4 is intentionally the most challenging problem.

Question 1 – Multiple Select

10 points

For each item, select **all** correct statements. Each item is worth 2 points.

1.1 Regularization and feature scaling

2 points

Consider linear or logistic models with trainable weights. Which statements are correct?

- A. L_2 regularization can never affect a classifier's decision boundary.
- B. If a feature is replaced by $x'_j = sx_j$ with $s \neq 0$, replacing w_j by w_j/s preserves the unregularized model's predictions.
- C. Increasing the regularization parameter λ always decreases the unregularized data-fitting loss at the optimum.
- D. With redundant features, L_2 regularization can prefer one lower-norm parameterization among many parameterizations that produce the same predictions.
- E. Feature scaling can change the gradient-descent trajectory even when the set of representable prediction functions is unchanged.

1.2 Support Vector Machines

2 points

Which statements are correct for a linear SVM?

- A. Support vectors are always the training points farthest from the decision boundary.
- B. The decision boundary is $\mathbf{w}^T \mathbf{x} + b = 0$.
- C. Multiplying both $\mathbf{w}$ and b by the same positive constant preserves all predicted labels, although the parameters are no longer in the same canonical scaling.
- D. An RBF-kernel SVM must have a linear boundary in the original input space.
- E. Removing a non-support-vector training point may leave the optimal decision boundary unchanged.

1.3 Gradient boosting and XGBoost

2 points

Which statements are correct?

- A. All trees in gradient boosting can be trained independently of one another.
- B. A later tree is trained using information derived from the current ensemble's errors or loss gradients.
- C. Increasing the learning rate always improves test performance.
- D. XGBoost can regularize leaf weights and penalize additional tree complexity.
- E. A split that improves the data-fitting term can still be rejected when its complexity penalty is sufficiently large.

1.4 A* search**2 points**

Assume all edge costs are nonnegative. Which statements are correct?

- A. Every admissible heuristic is necessarily consistent.
- B. If $h(n) = 0$ for every node, A* is breadth-first search even when edge costs are unequal.
- C. A consistent heuristic satisfies $h(n) \leq c(n, n') + h(n')$ for every edge $n \rightarrow n'$.
- D. If h_1 is admissible and $0 \leq h_2(n) \leq h_1(n)$ for every node, then h_2 is also admissible.
- E. Under the standard assumptions, A* tree search with an admissible heuristic returns an optimal solution.

1.5 Neural networks and softmax**2 points**

Which statements are correct?

- A. Multiplying every softmax logit by the same positive constant leaves all softmax probabilities unchanged.
- B. Adding the same constant to every softmax logit leaves the probabilities unchanged.
- C. With cross-entropy loss and `from_logits=True`, the final layer should normally output raw logits rather than already applying softmax.
- D. A three-label multi-label problem should normally use one three-class softmax output.
- E. A stack of affine layers with only linear activations can be collapsed into a single affine transformation.

Question 2 – Classical Machine Learning**10 points****(a) Bernoulli Naive Bayes with Laplace smoothing****3 points**

A training set contains 4 examples from class $Y = 1$ and 6 from class $Y = 0$. Among the $Y = 1$ examples, $X_1 = 1$ occurs 3 times and $X_2 = 1$ occurs once. Among the $Y = 0$ examples, $X_1 = 1$ occurs once and $X_2 = 1$ occurs 4 times.

Use the empirical class priors and Laplace smoothing with $\alpha = 1$ independently for each binary feature. For a new example

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

compute $P(Y = 1 \mid X_1 = 1, X_2 = 0)$ and state the predicted class. Show the unnormalized class scores before normalizing.

(b) Soft-margin SVM and hinge loss**3 points**

A linear SVM uses

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

The four training examples are

$$A = (2, 1), y_A = +1; \quad B = (1, 0), y_B = +1; \quad C = (0, 1), y_C = -1; \quad D = (-1, 0), y_D = -1.$$

For each point compute the functional margin $m_i = y_i(\mathbf{w}^T \mathbf{x}_i + b)$ and hinge loss $\max(0, 1 - m_i)$. Then evaluate

$$\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \max(0, 1 - m_i)$$

for $C = 0.5$.

(c) XGBoost split calculation**4 points**

At a node, the gradient and Hessian sums are

$$G_P = -3, \quad H_P = 5.$$

Two candidate splits are considered:

	G_L	H_L	G_R	H_R
Split A	-5	2	2	3
Split B	-4	4	1	1

Let $\lambda = 1$. For a leaf,

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

and the split gain is

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

(i) Find the two optimal leaf weights for Split A. (ii) With $\gamma = 1$, compute the gain of both splits and choose the preferred split. (iii) Find the range of γ for which Split A has strictly positive gain.

Question 3 – Deep Learning: CNN, Softmax, and Backpropagation**10 points**

A network receives a 10×10 grayscale image. It applies a convolution with 3 filters of size 3×3 , stride 1, no padding; ReLU; 2×2 max pooling with stride 2; flattening; and a Dense layer producing 4 logits.

(a) Architecture analysis**3 points**

Find (i) the tensor shape after convolution, (ii) the shape after max pooling, (iii) the total number of trainable parameters, including biases, and (iv) the receptive-field size in the original image of one pooled activation.

(b) Forward propagation**3 points**

For one example, the flattened activation $\mathbf{a} \in \mathbb{R}^{48}$ satisfies $a_1 = 1$, $a_2 = 2$, and all other entries are zero. The final Dense layer has zero bias and only two nonzero weights:

$$W_{1,1} = \ln 4, \quad W_{2,2} = \frac{1}{2} \ln 2.$$

Compute the four logits and the exact softmax probability vector. If the true class is class 2, compute the exact cross-entropy loss.

(c) Backpropagation**4 points**

For softmax with cross-entropy, use

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

Compute $\nabla_{\mathbf{z}} L$. Then prove that for $\mathbf{z} = \mathbf{a}^T W + \mathbf{b}$,

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

Compute the first two rows of $\partial L / \partial W$ and state the remaining rows. Finally, explain from the signs of $\nabla_{\mathbf{z}} L$ how gradient descent tends to change the four logits.

Question 4 – Advanced Deep Learning: Attention Invariances

10 points

Let

$$Q = K = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad V = \begin{bmatrix} 1 & 2 \\ 3 & 0 \end{bmatrix},$$

and let $d_k = 2$. Define

$$S = \frac{QK^T}{\sqrt{d_k}}, \quad A = \text{softmax}_{\text{row}}(S), \quad O = AV.$$

(a) Forward attention calculation**3 points**Compute S , A , and O exactly. You may define

$$\alpha = \frac{e^{1/\sqrt{2}}}{e^{1/\sqrt{2}} + 1}$$

to simplify your answer.

(b) Common-key shift invariance**4 points**Let $\mathbf{c} \in \mathbb{R}^2$ be arbitrary and add the same vector to every key:

$$K' = K + \mathbf{1}\mathbf{c}^T,$$

where $\mathbf{1}$ is the all-ones column vector. Prove that the attention matrix is unchanged, $A' = A$, for every Q, K , and $\mathbf{c}$. Your proof must explicitly use the fact that row-wise softmax is invariant to adding a constant to every element of a row.

(c) Common-value shift**3 points**Now let $\mathbf{d} \in \mathbb{R}^2$ and set

$$V' = V + \mathbf{1}\mathbf{d}^T.$$

Prove that

$$O' = O + \mathbf{1}\mathbf{d}^T.$$

Explain why this property follows from the fact that every row of an attention matrix sums to 1.