IAI^{2}O Final Examination - Form A

Machine Learning & Deep Learning

Question Paper

Name: _____

Contestant ID: _____

Total: 100 points

Instructions. Section I has 25 single-choice questions (1 point each). Section II has 10 multiple-select questions (3 points each); full credit requires selecting all and only the correct choices. Section III has 5 code-reading/code-completion questions (5 points each). Section IV has 2 proof/derivation questions (10 points each). Show sufficient work for Sections III and IV.

Useful values when needed: $e \approx 2.718$, $\sigma(2) \approx 0.881$, and $H(5/6) \approx 0.65$ for binary entropy in bits.

Section I - Single Choice **25 points**

1. Consider $f_w(x) = wx$ with training examples $(1,2), (2,4), (3,6)$ and

$$J(w) = \frac{1}{2m} \sum_{i=1}^{m} (wx^{(i)} - y^{(i)})^2.$$

Gradient descent uses $w_{t+1} = w_t - \alpha J'(w_t)$. For every initial $w_0 \neq 2$, which range of α guarantees $w_t \to 2$? **[1 point]**

A. $0 < \alpha < \frac{3}{14}$

B. $0 < \alpha < \frac{3}{10}$

C. $0 < \alpha < \frac{3}{7}$

D. $0 < \alpha \leq \frac{3}{7}$

E. $0 < \alpha < \frac{6}{7}$

2. A regularized logistic-regression model uses $p = \sigma(wx)$ and

$$J(w) = J_{\text{data}}(w) + \frac{\lambda}{2m} w^2.$$

A new feature $z = x/5$ is introduced and $v = 5w$ is used so that $vz = wx$. If the reparameterized objective is $J'(v) = J'_{\text{data}}(v) + \frac{\lambda'}{2m} v^2$, which λ' makes the entire objective identical for corresponding parameters? **[1 point]**

A. $\lambda/25$

B. $\lambda/5$

C. λ

D. 5λ

E. 25λ

3. A classifier has human-level error 0.5%, training error 0.8%, and dev error 8.5%. Which diagnosis and response is most appropriate? **[1 point]**

A. Primarily high bias; reduce the training set size.

B. Primarily high bias; increase regularization strongly.

C. Primarily data mismatch; change only the test metric.

D. Primarily high variance; more training data and/or stronger regularization may help.

E. The model is already near Bayes optimal on the dev set.

4. Training data come from web images, while dev/test data come from a mobile app. Errors are: training 1.0%, train-dev 1.3% (same distribution as training), dev 7.5%, test 7.7%. What is the dominant issue? **[1 point]**

A. Avoidable bias

B. Data mismatch between the training distribution and the dev/test distribution

C. Severe overfitting to the dev set

D. Test-set leakage

E. Underfitting caused by too much dropout

5. For exponentially weighted averages, let $v_0 = 0$, $\beta = 0.9$, and observations be $\theta_1 = 10$, $\theta_2 = 0$. Using $v_t = \beta v_{t-1} + (1-\beta)\theta_t$ and bias correction $\hat{v}_t = v_t/(1-\beta^t)$, what is $\hat{v}_2$? **[1 point]**

A. 0.90

B. 1.00

C. 4.50

D. 5.00

E. $\frac{90}{19} \approx 4.74$

6. Ignore the numerical ϵ term in Batch Normalization. For a mini-batch of pre-activations z, define $\hat{z} = (z-\mu)/\sigma$. A preceding layer is changed so that every pre-activation becomes $z' = 3z + 7$. What happens to the Batch-Normalized values before the learned scale/shift parameters? **[1 point]**

A. They are multiplied by 3.

B. They are shifted by 7.

C. They are unchanged.

D. Their sign is reversed.

E. They become all zeros.

7. A three-class softmax model has logits $(1000, 999, 998)$. Without explicitly evaluating the exponentials, what is the ratio p_1/p_3? **[1 point]**

A. e

B. e^2

C. $2e$

D. $1/e^2$

E. It is undefined because e^{1000} overflows.

8. A fully connected network has architecture $20 \to 8 \to 4 \to 3$, with a bias for every neuron in the three non-input layers. How many trainable scalar parameters does it have? **[1 point]**

A. 192

B. 203

C. 211

D. 219

E. 227

9. A network maps $x \in \mathbb{R}^4$ through a 2-unit hidden layer and then to a 3-dimensional output. Both layers use linear activations. Let W_{eff} be the 4×3 effective weight matrix after collapsing the layers. Which statement must be true? **[1 point]**

A. W_{eff} must be diagonal.

B. W_{eff} must be symmetric.

C. $\text{rank}(W_{\text{eff}}) = 3$.

D. Every 4×3 matrix can be represented.

E. $\text{rank}(W_{\text{eff}}) \leq 2$.

10. In inverted dropout, a hidden activation is $a = 10$ and the keep probability is 0.8. Which statement correctly describes the activation if the unit is retained during training, its expectation over dropout masks, and test-time behavior? **[1 point]**

A. Training retained value 12.5; training expectation 10; test value 10 with no dropout scaling.

B. Training retained value 8; training expectation 6.4; test value 12.5.

C. Training retained value 10; training expectation 8; test value 10.

D. Training retained value 12.5; training expectation 12.5; test value 8.

E. Training retained value 8; training expectation 10; test value 8.

11. A convolutional layer receives a $32 \times 32 \times 3$ input and uses 5 filters of size 5×5, stride 2, padding 2, with one bias per filter. Which pair gives the output shape and number of trainable parameters? **[1 point]**

A. $15 \times 15 \times 5$ and 375

B. $16 \times 16 \times 3$ and 380

C. $16 \times 16 \times 5$ and 380

D. $17 \times 17 \times 5$ and 375

E. $16 \times 16 \times 5$ and 1,900

12. A 1×1 convolution maps a $14 \times 14 \times 64$ activation volume to 32 output channels, using one bias per output channel. What are the output shape and parameter count? **[1 point]**

A. $14 \times 14 \times 64$, 2,048

B. $14 \times 14 \times 32$, 2,080

C. $14 \times 14 \times 32$, 32

D. $1 \times 1 \times 32$, 2,080

E. $14 \times 14 \times 32$, 28,672

13. Ignore bias and boundary effects. A standard 3×3 convolution maps 32 input channels to 64 output channels. A depthwise-separable alternative uses a 3×3 depthwise convolution followed by a 1×1 pointwise

convolution. Approximately what fraction of the multiplications of the standard convolution does the separable version use? **[1 point]**

A. 1.6%

B. 5.6%

C. 9.0%

D. 12.7%

E. 50%

14. A residual block has compatible input/output dimensions and computes $a^{[l+2]} = \mathrm{ReLU}(F(a^{[l]}) + a^{[l]})$. If the residual-branch weights are set so that $F(a^{[l]}) = 0$, which statement is most accurate? **[1 point]**

A. The block outputs zero for every input.

B. The block is always exactly the identity map, even for negative components.

C. The skip connection is removed.

D. The block becomes a softmax layer.

E. The block computes $\mathrm{ReLU}(a^{[l]})$, which is the identity on nonnegative components.

15. Two same-class predicted boxes are $A = [0, 0, 4, 4]$ and $B = [2, 1, 5, 5]$, where coordinates are $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$. Box A has the higher confidence. If non-max suppression uses IoU threshold 0.25, what happens? **[1 point]**

A. IoU $= 3/11 \approx 0.273$, so B is suppressed.

B. IoU $= 3/11 \approx 0.273$, so both are kept.

C. IoU $= 1/2$, so B is suppressed.

D. IoU $= 1/4$, so both are kept.

E. IoU $= 6/16$, so A is suppressed.

16. Triplet loss is $L = \max\{0, d(A, P) - d(A, N) + \alpha\}$. If the squared embedding distances are $d(A, P) = 0.60$, $d(A, N) = 0.70$, and $\alpha = 0.20$, what is the loss? **[1 point]**

A. 0

B. 0.05

C. 0.10

D. 0.30

E. 0.50

17. In a scalar RNN, suppose the recurrent weight is $w = 2$ and, along a particular sequence, the activation derivative at each of six recurrent steps is exactly 0.25. What multiplicative factor relates a perturbation in the initial hidden state to the hidden state after six steps? **[1 point]**

A. 1/16

B. 1/64

C. 1/32

D. 2

E. 64

18. For an LSTM cell, which gate pattern most directly preserves the previous cell-state information while writing almost no new candidate information at the current step? **[1 point]**

A. Forget gate ≈ 0, input/update gate ≈ 1

B. Forget gate ≈ 0, input/update gate ≈ 0

C. Forget gate ≈ 1, input/update gate ≈ 1

D. Forget gate ≈ 1, input/update gate ≈ 0

E. Output gate ≈ 0 is sufficient regardless of the other gates

19. Suppose embeddings are $e_{\text{man}} = (1, 0)$, $e_{\text{woman}} = (0, 1)$, and $e_{\text{king}} = (2, 0)$. The analogy vector $e_{\text{king}} - e_{\text{man}} + e_{\text{woman}}$ is compared by cosine similarity with the following candidate embeddings. Which is the exact match? **[1 point]**

A. Prince $(3, 0)$

B. Girl $(0, 2)$

C. Apple $(-1, 0)$

D. Orange $(0, -1)$

E. Queen $(1, 1)$

20. In machine translation, the human reference y^* has $\log P(y^* \mid x) = -4.0$ under the trained model, while beam search returns $\hat{y}$ with $\log P(\hat{y} \mid x) = -4.6$. The reference is judged clearly better. What does this evidence primarily indicate? **[1 point]**

A. A search (beam-search) error: the model preferred y^* but the search failed to find it.

B. A modeling error: the model assigns too low a probability to y^*.

C. A BLEU implementation error must have occurred.

D. Length normalization can never help this case.

E. The beam width must already be infinite.

21. Scaled dot-product attention uses $\text{softmax}(qK^T/\sqrt{d_k})V$. Let $d_k = 2$, $q = (\sqrt{2}, 0)$, $k_1 = (1, 0)$, $k_2 = (0, 1)$, $v_1 = (1, 0)$, and $v_2 = (0, 1)$. What is the attention output? **[1 point]**

A. $(1/2, 1/2)$

B. $(1, 0)$

C. $\left(\frac{e}{e+1}, \frac{1}{e+1}\right)$

D. $\left(\frac{e^2}{e^2+1}, \frac{1}{e^2+1}\right)$

E. $(0, 1)$

22. A decision-tree node contains 10 examples, 5 positive and 5 negative, so its entropy is 1 bit. A candidate split sends 6 examples to the left child (5 positive, 1 negative) and 4 examples to a pure-negative right child. Using $H(5/6) \approx 0.65$, what is the information gain? **[1 point]**

A. 0.39

B. 0.61

C. 0.65

D. 0.75

E. 1.00

23. For one-dimensional K-means with points $\{0, 2, 8, 10\}$ and $K = 2$, initial centroids are 0 and 10. After one assignment step and one centroid-update step, what is the K-means objective $J = \frac{1}{m}\sum_i \|x^{(i)} - \mu_{c^{(i)}}\|^2$? **[1 point]**

A. 0

B. 0.5

C. 0.75

D. 1

E. 2

24. A content-based recommender scores an item by the dot product $v_u^T v_i$. For a user embedding $v_u = (1, 2)$, which candidate item receives the largest score? **[1 point]**

A. $v_A = (2, 0)$

B. $v_B = (0, 2)$

C. $v_C = (1, 1)$

D. $v_D = (3, -1)$

E. $v_E = (-1, 3)$

25. For one action in an MDP, the immediate reward is 1 and $\gamma = 0.5$. The action leads with probability 0.75 to a next state whose optimal next-action value is 4, and with probability 0.25 to a next state whose optimal next-action value is 8. What is the Bellman target for $Q(s, a)$? **[1 point]**

A. 2.5

B. 3.0

C. 3.5

D. 4.0

E. 5.0

Section II - Multiple Select **30 points**

For each question, select all correct choices. Full credit requires selecting all and only the correct choices.

26. Which statements about softmax are correct? **[3 points]**

A. Multiplying every logit by the same positive constant leaves all softmax probabilities unchanged.

B. Adding the same constant to every logit leaves all softmax probabilities unchanged.

C. The class with the largest logit also has the largest softmax probability.

D. Softmax probabilities sum to 1.

E. Adding arbitrary different constants to different logits cannot change the predicted class.

27. Which statements about bias/variance diagnostics are correct? **[3 points]**

A. A large train-dev gap with very high training error is the signature of pure high bias only.

B. Increasing L_2 regularization generally decreases bias and increases variance.

C. Low training error but much higher dev error is evidence of high variance.

D. More training data often helps a high-variance problem.

E. Increasing model capacity can help when the dominant problem is high avoidable bias.

28. Which statements about convolutional networks are correct? **[3 points]**

A. A pooling layer necessarily contains trainable weights.

B. In a standard convolution, each filter spans all channels of the previous layer.

C. The number of parameters in a convolutional layer grows linearly with the spatial height and width of its input.

D. A 1×1 convolution can change the number of channels.

E. Depthwise-separable convolution can substantially reduce computation compared with standard convolution.

29. Which statements about object detection and semantic segmentation are correct? **[3 points]**

A. Non-max suppression should always compare boxes across all classes together.

B. Intersection-over-Union measures overlap area divided by union area.

C. Anchor boxes can help represent multiple objects associated with the same grid cell.

D. Semantic segmentation predicts a class label for each pixel.

E. Transposed convolution can be used for upsampling in a U-Net-style segmentation network.

30. Which statements about recurrent networks are correct? **[3 points]**

A. A bidirectional RNN uses only past context and never future context.

B. A bidirectional RNN can use context from both earlier and later positions in a sequence.

C. GRU/LSTM gates are designed to help preserve useful information over long ranges.

D. In the simplified GRU update $c_t = \Gamma_u \tilde{c}_t + (1 - \Gamma_u) c_{t-1}$, $\Gamma_u \approx 0$ preserves the previous state.

E. Exploding gradients cannot occur in recurrent networks.

31. Which statements about word embeddings and attention are correct? **[3 points]**

A. One-hot word vectors directly encode semantic similarity through Euclidean distance.

B. Negative sampling converts the word-context training task into a set of binary classification problems.

C. Cosine similarity compares the direction of two embedding vectors and is insensitive to a common positive scaling of either vector.

D. Self-attention forms each output as a weighted combination of value vectors.

E. Positional encoding supplies order/location information that pure self-attention otherwise lacks.

32. Which statements about decision trees and tree ensembles are correct? **[3 points]**

A. Binary entropy is largest when all examples belong to the same class.

B. Information gain compares parent entropy with the weighted entropy of the child nodes.

C. Random forests combine bootstrap sampling with random subsets of candidate features at splits.

D. Boosted trees place more emphasis on examples that earlier trees handled poorly.

E. A single fully grown decision tree is guaranteed to have lower variance than a random forest.

33. Which statements about anomaly detection are correct? **[3 points]**

A. Supervised classification is always preferable when there are only a handful of known anomaly examples.

B. Anomaly detection is well suited when positive anomalies are rare and future anomaly types may differ from the few seen during training.

C. Transforming a highly skewed feature (for example with a log transform) may make a Gaussian density model more appropriate.

D. Under the independent-feature Gaussian model, $p(x) = \prod_j p(x_j)$.

E. If an example is declared anomalous when $p(x) < \epsilon$, increasing ϵ generally causes more examples to be flagged.

34. Which statements about recommender systems are correct? **[3 points]**

A. Collaborative filtering requires every item to have hand-designed interpretable features.

B. Mean normalization can improve recommendations for a new user who has not rated any items.

C. Content-based filtering can use explicit user and item features.

D. Large-scale recommenders often separate candidate retrieval from ranking.

E. Optimizing only for engagement can create undesirable or unethical incentives.

35. Which statements about reinforcement learning are correct? **[3 points]**

A. A discount factor greater than 1 is the standard choice for making distant rewards dominate.

B. $Q(s, a)$ represents expected return when starting in s, taking action a, and then behaving optimally.

C. An ϵ-greedy policy deliberately explores with nonzero probability.

D. In deep Q-learning, a separate target network and soft updates can improve training stability.

E. Reinforcement learning is generally easier to deploy safely on a physical robot than in simulation.

Section III - Code Reading and Code Completion **25 points**

36. Read the following code without running it. Write the exact value of `A1`, then the printed rounded value of `A2`, and finally `pred`. **[5 points]**

```
import numpy as np

def sigmoid(z):
    return 1 / (1 + np.exp(-z))

X = np.array([[1., -1.],
              [2.,  0.]])

W1 = np.array([[1., -1.],
               [1.,  1.]])
b1 = np.array([0., 0.])

W2 = np.array([[ 1.],
               [-2.]])
b2 = np.array([0.])

A1 = np.maximum(0, X @ W1 + b1)
A2 = sigmoid(A1 @ W2 + b2)
pred = (A2 >= 0.5).astype(int)

print(A1)
print(np.round(A2, 3))
print(pred)
```

37. Complete the blanks to build, train, and use a numerically stable 10-class classifier. Use ReLU hidden layers, raw logits at the output, Adam with learning rate 10^{-3}, and the recommended loss configuration. **[5 points]**

```
import tensorflow as tf
from tensorflow.keras import Sequential
from tensorflow.keras.layers import Dense
from tensorflow.keras.losses import SparseCategoricalCrossentropy

model = Sequential([
    Dense(25, activation=_________________),
    Dense(15, activation=_________________),
    Dense(10, activation=_________________)
])

model.compile(
    optimizer=tf.keras.optimizers.Adam(learning_rate=_________________),
    loss=SparseCategoricalCrossentropy(from_logits=_________________)
)

model.fit(X_train, y_train, epochs=_________________)

logits = ______________________________
probabilities = ______________________________
predictions = ______________________________
```

38. Complete this NumPy implementation of a valid 2D cross-correlation (the operation commonly called convolution in CNN libraries) with arbitrary positive integer stride and no padding. **[5 points]**

```
import numpy as np

def conv2d_valid(X, K, stride=1):
    H, W = X.shape
    kH, kW = K.shape

    H_out = ____________________________________________
    W_out = ____________________________________________

    Z = np.zeros((H_out, W_out))

    for i in range(H_out):
        for j in range(W_out):
            patch = X[____________________________________________]
            Z[i, j] = ____________________________________________

    return Z
```

39. Complete the vectorized forward pass of a basic many-to-many RNN. Here `X[t]` is the input at time t, `W_ax` maps input to hidden state, `W_aa` is recurrent, and `W_ya` maps hidden state to class logits. Use a numerically stable softmax. **[5 points]**

```
import numpy as np

def softmax(z):
    z = z - ____________________________________________
    e = np.exp(z)
    return ____________________________________________

def rnn_forward(X, W_ax, W_aa, b_a, W_ya, b_y):
    T = X.shape[0]
    n_h = W_aa.shape[0]
    a = np.zeros(n_h)
    outputs = []

    for t in range(T):
        a = np.tanh(_____________________________________________)
        logits = ____________________________________________
        yhat = _____________________________________________
        outputs.append(yhat)

    return np.stack(outputs), a
```

40. Complete one K-means iteration. `X` has shape (m, n) and `centroids` has shape (K, n). Preserve an old centroid if a cluster receives no points. Finally compute the mean squared distance to the newly updated assigned centroids. **[5 points]**

```
import numpy as np

def kmeans_step(X, centroids):
    K = centroids.shape[0]

    dist2 = np.sum(
        (X[:, None, :] - centroids[None, :, :])**2,
        axis=_________________
    )

    assignment = np.argmin(dist2, axis=_________________)
    new_centroids = np.zeros_like(centroids)

    for k in range(K):
        mask = ______________________________
        if np.any(mask):
            new_centroids[k] = ______________________________
        else:
            new_centroids[k] = ______________________________

    J = np.mean(np.sum(
        (X - new_centroids[assignment])**2,
        axis=_________________
    ))

    return assignment, new_centroids, J
```

Section IV - Proof and Derivation **20 points**

41. For K classes, define softmax probabilities

$$p_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}},$$

and suppose the true class is r, so the cross-entropy loss is $L = -\log p_r$.

(a) Prove that adding the same scalar c to every logit leaves all p_j unchanged. [2]

(b) Derive, from first principles, $\dfrac{\partial L}{\partial z_j} = p_j - \mathbf{1}(j = r)$. [5]

(c) Prove that $\sum_{j=1}^{K} \frac{\partial L}{\partial z_j} = 0$ and explain how this is connected to the invariance in part (a). [3]

[10 points]

42. Consider the scalar recurrent network

$$h_t = \tanh(w h_{t-1} + u x_t + b), \qquad t = 1, \dots, T,$$

with loss $L = \frac{1}{2}(h_T - y)^2$.

(a) Derive an explicit product formula for $\dfrac{\partial L}{\partial h_0}$ in terms of w and the hidden states $h_1, \dots, h_T$. [5]

(b) Show that if $|w| \le \rho < 1$, then $\left|\dfrac{\partial L}{\partial h_0}\right| \le |h_T - y|\rho^T$, and state the implication as T grows. [2]

(c) Explain why $|w| > 1$ does not by itself guarantee exploding gradients with a tanh RNN. Give a sufficient condition, expressed in terms of the per-step factors $|w|(1 - h_t^2)$, under which the magnitude of the gradient grows exponentially with T. [3]

[10 points]