Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur
CSL 7620: Machine Learning
Major Examination

Total Marks: 60 **Set C - Answer Key** **Time: 3 hours**

---

1. An AdaBoost classifier makes errors on 3 out of 10 weighted samples. The total error is 0.3. If the initial weight of each sample is 0.1, what is the updated weight of the misclassified samples after this iteration? **[2 marks]**
   i. 2.723
   ii. 0.072
   iii. 0.923
   iv. None of the others
2. Which statements are TRUE for a fully-connected neural network? **[1 mark]**
   i. It always produces non-linear output
   ii. It sometimes uses a loss function for training
   iii. None of the others
   iv. Applies gradient descent for parameter learning
3. Which of the following is/are true? **[1 mark]**
   i. Batch gradient descent is always guaranteed to converge to the global optimum of any loss function.
   ii. Stochastic gradient descent is always guaranteed to converge to the global optimum of any loss function.
   iii. For convex loss functions (i.e. with a bowl shape), batch gradient descent and stochastic gradient descent are guaranteed to eventually converge to the global optimum.
   iv. None of the others
4. Which of the following correctly compares Linear Discriminant Analysis (LDA) and Independent Component Analysis (ICA)? **[1 mark]**
   i. LDA is supervised, while ICA is unsupervised.
   ii. LDA maximizes class separability, while ICA maximizes the likelihood of observations
   iii. LDA assumes Gaussian-distributed classes, while ICA assumes non-Gaussian independent sources.
   iv. None of the others
5. Find out the correct statements. **[1 mark]**
   i. Boosting reweights data samples during training
   ii. Boosting requires all weak learners to be identical from every perspective
   iii. None of the others
   iv. Bagging and Boosting both rely on fully random feature selection
6. Match the following ANN Components (left column) with their correct descriptions (right column) **[1 mark]**

A. MLP	1. Not differentiable everywhere
B. ReLU	2. Fully connected feedforward network
C. Backpropagation	3. Enables gradient-based weight update

D. Max-pooling	4. Used in CNNs and not in fully connected networks

i. A-1, B-2, C-3, D-4
ii. A-2, B-1, C-3, D-4
iii. A-2, B-1, C-4, D-3
iv. None of the others

7. Which of the following discriminative models make use of distance in their decision process? **[1 mark]**

i. k-NN
ii. SVM
iii. Decision Tree
iv. None of the others

8. From an application perspective, autoencoders are primarily used for: **[1 mark]**

i. Feature extraction
ii. Dimensionality reduction
iii. Reconstructing input data
iv. None of the others

9. Find the best match between the left and the right columns: **[1 mark]**

A. k-means B. GMM C. DBSCAN	1. Can deal with clusters of different densities, but prone to outlier. 2. Uses soft assignments derived from posterior probabilities 3. The inclusion of the points at the border of clusters may significantly depend on initialization

i. A-1, B-2, C-3
ii. A-1, B-3, C-2
iii. A-2, B-3, C-1
iv. None of the others

10. Two observed signals x1, x2 are linear mixtures of two independent non-Gaussian sources s1, s2. Consider

x=As and A = [1 2] [3 1].

If A^{-1} is used to unmix the signals, what is the resulting unmixing matrix $W=A^{-1}$? **[2 marks]**

i. [−0.2 0.4] [0.6 −0.2]
ii. None of the others
iii. [0.25 −0.5] [−0.75 0.25]
iv. [0.2 −0.4][−0.6 0.2]

11. For an input x, consider the activation function $\sigma(x) = sigmoid(x)$. If $\frac{d}{dx}\sigma(x) = f(x)$, the value of $f(x)$ is **[1 mark]**

i. $\sigma(x)^2$
ii. $\sigma(x) log\ \sigma(x)$

iii. $\sigma(x)/(1 + \sigma(x))$

iv. None of the others

12. In an autoencoder architecture, the input layer has 784 neurons (flattened 28 × 28 grayscale image). The encoder compresses the input through layers of sizes 128 and 64 down to a bottleneck layer of 32 neurons. The decoder then symmetrically reconstructs the input through layers of sizes 64 and 128, ending with an output layer of 784 neurons. The layer dimensions are:

    784 → 128 → 64 → 32 → 64 → 128 → 784

    All hidden layers use ReLU activation, and the output layer uses tanh activation. **No bias parameters are used in any layer**, and trainable parameters consist **only of weights**. Calculate the total number of trainable weight parameters in this autoencoder. **[2 marks]**

    i. 222,384
    ii. 221,184
    iii. 222,184
    iv. None of the others

13. Consider a convolutional neural network designed for downsampling, consisting of two convolution layers. The network takes an RGB image with a spatial size of 127×127 as input. The first convolutional layer uses a stride of 2 and a padding of 1, producing an output feature map of size 62×62 with 16 channels. The second convolution layer also uses a stride of 2 but with zero padding and produces a feature map of size 29×29 with 32 channels. Determine the kernel size k1 used in the first convolution layer and the kernel size k2 used in the second convolution layer.

    [2 marks]

    i. k1=7, k2=6
    ii. k1=6, k2=7
    iii. k1=7, k2=7
    iv. None of the others

14. Match the architecture with its characteristics **[1 mark]**

A. SVM B. Neural Network	1. Margin maximization 2. Requires activation functions 3. Eventually, training always depends only only a subset of training data 4. Cross-entropy loss

    i. A-1, B-2, B-3, B-4
    ii. A-1, B-2, A-3, B-4
    iii. A-1, B-2, A-3, A-4
    iv. None of the others

15. Consider a hard-margin linear Support Vector Machine (SVM) trained on the following five labeled points in $\mathbb{R}^2$(labels are given by y)
    (1,0),y=+1; (2,1), y=+1; (0,1), y=-1; (-1,2), y=-1; (0,2),y=-1.
    It is known that the data are linearly separable and that the optimal SVM has obtained the following parameters in canonical form:

$w = [1 \ -1]^T$, $b = 0$

Using the given w and b, compute the total width of the margin (length between boundaries of two classes) for this classifier. **[2 marks]**

i. $w = 2$

ii. $w = \sqrt{2}$

iii. $w = -\sqrt{2}$

iv. None of the others

16. Find out the correct statements **[1 mark]**

i. The slack variable in SVM is used only when the decision boundary is known to be non-linear

ii. A kernel is used for transforming data from a lower-dimensional feature space to a finite and higher-dimensional feature space

iii. SVM is used only for binary classification

iv. None of the others

17. According to Bayes decision theory, the objective of a classifier is to minimize the expected (1) _____, which represents the average cost associated with classification errors. The optimal decision rule that achieves this is known as the (2) _____ rule, which assigns an observation to the class with the highest posterior probability. When the class-conditional probabilities follow (3) _____ distributions, the resulting discriminant functions can be linear or quadratic depending on the covariance matrices. In parameter estimation, the (4) _____ approach determines parameters by maximizing the likelihood of the observed data, while the (5) _____ approach combines prior information with the likelihood to produce posterior estimates. **[1 mark]**

i. (1) Risk (2) Bayes Decision (3) Normal (Gaussian) (4) Maximum Likelihood (5) Bayesian

ii. (1) Error Rate (2) Minimum Error Rate (3) Exponential (4) Bayesian (5) Maximum Likelihood

iii. (1) Risk (2) Bayesian (3) Uniform (4) Maximum Likelihood (5) Bayes Decision

iv. None of the others

18. Select all correct statements: **[1 mark]**

i. In a neural network, the choice of regularization affects not only weight values but also the learned feature representations in hidden layers

ii. L1 regularization tends to shrink weights toward zero but rarely makes them exactly zero.

iii. L2 regularization is better suited than L1 to prevent overfitting in high-dimensional input spaces.

iv. None of the others

19. You are training a multiple linear regression model of the form $y = w_1 x_1 + w_2 x_2 + b$ on the dataset: $(x_1, x_2, y) = \{(1, 1, 4), (2, 0, 5)\}$. The model parameters are initialized as $w_1(0) = 0$, $w_2(0) = 0$, $b(0) = 0$, and gradient descent is performed to minimize the Mean Squared Error (MSE) using a learning rate of $\eta = 0.1$. Perform two full batch gradient descent steps. Which of the following statements are correct after two steps? **[1 mark]**

i. **After the first update, the parameter values are approximately w 1 (1) = 0.7 , w 2 (1) = 0.2 , b (1) = 0.45.**
ii. After the second update, the parameters become approximately w 1 (2) = 1.79 , w 2 (2) = 0.53 , b (2) = 1.16.
iii. **The loss decreases after both updates.**
iv. None of the others

20. Match the technique to its respective characteristics: **[1 mark]**

A. PCA	1. Fails if source signals are Gaussian and not independent
B. LDA	2. Can extract at most C−1 discriminant components for C classes
C. ICA	3. May retain high-variance directions irrelevant to class separation

i. **A-3, B-2, C-1**
ii. A-1, B-2, C-3
iii. A-2, B-3, C-1
iv. None of the others

21. In a class of 140 students, you record the following age distribution: 40 students are 21 years old, 45 students are 22 years old, 30 students are 23 years old, and 25 students are 24 years old.

Two models are proposed for the age distribution:

Model M_1: categorical over ages {21, 22, 23, 24} with probabilities [0.30, 0.35, 0.20, 0.15].

Model M_2: uniform over ages 21–27, so each observed age (21–24) has probability 1/7.

Based on the observed data and these models, which of the following statements is **correct**?

[2 marks]

i. **Model M_1 has a higher likelihood than M_2.**
ii. Model M_2 has a higher likelihood than M_1.
iii. The likelihoods of both models are exactly equal.
iv. It is not possible to decide without additional information.

22. PCA reduces dimensionality by: **[1 mark]**

i. **Projecting data onto new axes with maximum variance**
ii. Removing features with zero variance
iii. None of the others
iv. Clustering features by similarity

23. Consider a Gaussian Mixture Model with 2 components (K=2) in a 1-dimensional feature space. The initial parameters are: Mixing coefficients: $\pi_1 = 0.6$, $\pi_2 = 0.4$; Means: $\mu_1 = 2$, $\mu_2 = 5$; Variances: $\sigma_1^2 = 1$, $\sigma_2^2 = 4$. Given a single observation $x = 3$, find $p(x|\mu_2, \sigma_2^2)$ (in three decimal places) **[1 mark]**

i. 0.241
ii. 0.121
iii. 0.083

iv. None of the others

24. Which of the following statements about the kernel trick is TRUE? **[1 mark]**
    - **i. It allows linear algorithms to operate in implicitly transformed feature spaces.**
    - ii. None of the others
    - iii. It requires explicit computation of the high-dimensional features.
    - **iv. It can be used with SVMs and PCA.**
25. You are given a grayscale image of size 5 × 5 and a convolution kernel of size 3 × 3. The convolution is applied with: No zero padding (valid convolution), Stride = 2, No bias, and No activation function

    The 5 x 5 input image ***I*** is:

$$I = \begin{bmatrix} 1 & 2 & 3 & 0 & 1 \\ 4 & 1 & 0 & 2 & 3 \\ 2 & 3 & 4 & 1 & 0 \\ 1 & 0 & 2 & 3 & 4 \\ 2 & 1 & 3 & 0 & 2 \end{bmatrix}$$

    The 3 x 3 kernel ***K*** is:

$$K = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$$

    Compute the output feature map **[2 marks]**

    i. $\begin{bmatrix} 4 & 0 \\ -5 & 1 \end{bmatrix}$ ii. $\begin{bmatrix} 3 & -1 \\ 0 & 2 \end{bmatrix}$

    iii. $\begin{bmatrix} 1 & 0 \\ -4 & -2 \end{bmatrix}$ iv. None of the others

26. Soft-margin SVM can be used with **[1 mark]**
    - **i. Perfectly linearly separable training data**
    - **ii. Linearly non-separable training data**
    - **iii. High-dimensional data**
    - iv. None of the others
27. A reliability engineer tests **six** identical electronic sensors from the same manufacturing batch. All six sensors fail during operation at times **3, 4, 6, 5, 9, and 8 hours**. The engineer assumes that the sensor lifetimes follow an **Exponential distribution** with an unknown rate parameter $\lambda>0$, and decides to use **maximum likelihood estimation (MLE)** to determine λ

    Hint: **Exponential density** for a lifetime $t \geq 0$: $f(t;\lambda) = \lambda e^{-\lambda t}$. Find λ **[2 marks]**

i. 0.20

ii. 0.25

iii. 0.28

iv. None of the others

28. Suppose I have a dataset which I want to divide in K clusters. Consider the following python code snippet

ind = np.random.choice(data.shape[0], K, replace=False)
a = data[initial_indices]

Assume that the above code snippet shows the initialization of a clustering algorithm. The variable '*data'* has a shape of 100×2 and it contains the two-dimensional dataset to be clustered. Ideally, this code snippet should be an initialization step for **[1 mark]**

i. K-medoids clustering

ii. K-means clustering

iii. GMM Clustering

iv. DBSCAN

29. Which of the following statements about the bias-variance tradeoff are TRUE? **[1 mark]**

i. As model complexity increases, bias decreases, and variance increases

ii. Increasing the number of training data always reduces the variance

iii. Underfitting results in high bias and low variance

iv. None of the others

30. Let S_W and S_B denote the within-class and between-class scatter matrices in Linear Discriminant Analysis. **[1 mark]**

i. **LDA seeks directions that maximize $|W^T S_B W| / |W^T S_W W|$.**

ii. **If S_W is singular, LDA cannot be directly applied.**

iii. When all class means are equal, LDA reduces to PCA.

iv. **LDA can yield at most rank(S_B) discriminant components.**

31. Find out the correct statements **[1 mark]**

i. All neural networks are non-linear

ii. All neural network training requires the optimization of a loss (objective) function

iii. All supervised learning problems with neural networks require one-hot encoding of labels

iv. None of the others

32. _(a)__ typically consist of dense layers connecting every neuron to every input. In contrast, __(b)__ exploit local spatial patterns in structured data such as images. They achieve parameter efficiency through weight sharing and typically include pooling layers to downsample features. _(c)__ can extract hierarchical spatial features, reducing the risk of overfitting compared to __(d)__, which requires far more parameters for the same input size. Thus, while __(e)__ treat images as flattened vectors, __(f)__ preserve spatial locality and translation invariance. **[1 mark]**

i. a) Fully-connected ANN, b) CNN, c) Fully-connected ANN, d) CNN, e) CNN, f) Fully-connected ANN

ii. a) Fully-connected ANN, b) CNN, c) Fully-connected ANN, d) CNN, e) Fully-connected ANN, f) CNN

iii. a) Fully-connected ANN, b) CNN, c) CNN, d) Fully-connected ANN, e) Fully-connected ANN, f) CNN

iv. None of the others

33. For a 2-class classification problem with equal loss for all misclassifications, what does the Bayes decision rule do? **[1 mark]**

i. Assigns data to the class with higher prior probability regardless of features

ii. Randomly assigns data to either class

iii. Assigns data to the class with the highest posterior probability p(y|x)

iv. None of the others

34. Find the best match of the following SVM Concepts (left column) with their correct descriptions (right column) **[1 mark]**

A. Kernel Trick	1. Slack variables
B. Soft-Margin SVM	2. Implicit high-dimensional mapping
C. Dual Form	3. Uses Lagrange multipliers
D. Primal Form	4. Works directly in input space

i. A-1, B-2, C-3, D-4

ii. A-2, B-1, C-3, D-4

iii. A-2, B-1, C-4, D-3

iv. None of the others

35. A node in a decision tree contains 15 samples: 5 belong to class X, 7 belong to class Y, and 3 belong to class Z. What is the entropy of this node (in two decimal places; consider base 2 in logarithm)? **[1 mark]**

i. 1.98

ii. 0.98

iii. -1.98

iv. None of the others

36. Find out the correct statements **[1 mark]**

i. The latent space of a normal autoencoder is continuous

ii. An autoencoder can be used for data compression

iii. The latent layer of a fully connected autoencoder always has a smaller dimension compared to its input layer

iv. None of the others

37. A standardized dataset has a covariance matrix with eigenvalues:

$\lambda1=7.5$, $\lambda2=2.5$, $\lambda3=0.5$, $\lambda4=0.5$. If you keep only the top **two principal components**, what percentage of total variance is preserved? **[1 mark]**

i. None of the others

ii. 80.5%

iii. 83.3%

iv. **90.9%**

38. Find out the correct statements. **[1 mark]**

i. CNNs generally use convolution and pooling layers

ii. None of the others
iii. CNNs exploit spatial locality
iv. CNNs enable parameter sharing

39. In AdaBoost, we need to calculate **[1 mark]**
    i. Weights of features
    ii. Importance of stumps
    iii. Importance of nodes
    iv. None of the others
40. Both Autoencoders and PCA can be used for dimensionality reduction. __(a)__ performs linear projections, whereas __(b)__ can learn nonlinear mappings through neural networks. While __(c)__ directly computes principal directions via eigenvalue decomposition, __(d)__ relies on gradient-based optimization and backpropagation. Unlike __(e)__, __(f)__ can include additional constraints such as sparsity or denoising objectives, making them more flexible but computationally more expensive. **[1 mark]**
    i. a) Autoencoder, b) PCA, c) PCA, d) Autoencoder, e) PCA, f) Autoencoder
    ii. a) PCA, b) Autoencoder, c) PCA, d) Autoencoder, e) PCA, f) Autoencoder
    iii. a) PCA, b) Autoencoder, c) PCA, d) Autoencoder, e) Autoencoder, f) PCA
    iv. None of the others
41. Consider a feedforward neural network with two input neurons x_1 and x_2, a single hidden neuron, and one output neuron. The hidden neuron computes a pre-activation $a_h = w_1 x_1 + w_2 x_2 + b_1$ and uses a ReLU activation $z_1 = max(0, a_h)$. The output neuron computes a pre-activation $a_0 = w_3 z_1 + b_2$ and applies a sigmoid activation $y = \sigma(a_0)$. Given the training sample $x_1 = 2$, $x_2 = 3$ with target $t = 1$, initial weights $w_1 = 0.5$, $w_2 = -0.3, w_3 = 0.8$, biases $b_1 = b_2 = 0.2$, learning rate $\eta = 0.1$, ReLU at hidden, sigmoid at output, and loss $L = \frac{1}{2}(y - t)^2$, compute the updated weights after one gradient-descent update. For all numerical values, round to 3 decimal places. **[2 marks]**
    i. $w_1 = 0.5,\ w_2 = -0.3, w_3 = 0.8$
    ii. $w_1 = 0.409,\ w_2 = -0.285, w_3 = 0.802$
    iii. $w_1 = 0.489,\ w_2 = 0.285, w_3 = 0.782$
    iv. None of the others