Machine Learning Final Examination

Time: 3 Hours

Total Marks: 100

Instructions:

- Answer all questions.
- Assume suitable data if missing.
- Justify all answers unless specified.

Section A: Theory & Short Answers (5 × 6 = 30 Marks)

1. **Overfitting vs. Underfitting**: Define both terms. Provide two techniques to combat overfitting in linear regression and one remedy for underfitting.
2. **Gradient Descent Variants**: Contrast Batch GD, Stochastic GD, and Mini-Batch GD. Explain how *learning rate schedules* improve Stochastic GD convergence.
3. **PCA & Eigen Decomposition**: Why is mean-centering critical in PCA? Derive the relationship between eigenvalues and explained variance.
4. **Bias-Variance Tradeoff**: Mathematically decompose the expected prediction error into bias, variance, and irreducible error.
5. **SVM Kernels**: Explain how the RBF kernel transforms non-linear data. Provide a use case where Polynomial kernels outperform RBF.
6. **Ensemble Methods**: Why does Random Forest reduce overfitting compared to a single Decision Tree? Define *OOB error*.

Section B: Derivations & Proofs (10 × 3 = 30 Marks)

1. **Ridge Regression Derivation**:
   Given the loss function $J(\mathbf{w}) = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda\|\mathbf{w}\|^2$, derive the closed-form solution $\mathbf{w}^* = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$.
2. **Naive Bayes Classifier**:
   Starting from Bayes' theorem, derive the log-probability estimation for class C_k given features $\mathbf{x}$. Explain the role of Laplace smoothing.
3. **Logistic Regression MLE**:
   Prove that maximizing the log-likelihood for binary logistic regression is equivalent to minimizing cross-entropy loss. Show the gradient update rule.

Section C: Numerical Problems (20 × 2 = 40 Marks)

1. **Gradient Descent & Regularization**:
   Dataset:

   ```text
   ```

```
X = [[1, 2], [2, 4], [3, 6], [4, 8]]
y = [3, 5, 7, 9]
```

- **(a)** Implement **one iteration** of Batch GD to find weights for linear regression (initial weights = [0, 0], learning rate = 0.01).
- **(b)** Using Ridge regression (λ = 0.1), compute the closed-form solution. Compare weights with (a).
- **(c)** Calculate **MSE** for Ridge predictions.

2. **PCA & Classification**:
   Covariance matrix:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 5.0 & 2.5 \\ 2.5 & 5.0 \end{bmatrix}$$

- **(a)** Find eigenvalues, eigenvectors, and principal components.
- **(b)** Project data point $[3.0, 3.0]$ onto the first principal component.
- **(c)** Given class labels $[0, 0, 1, 1]$ for 4 samples in PCA-transformed space (1D), compute **Gini impurity** at the root of a decision tree.

Answer Key Outline

Section A

1. **Overfitting**: High variance, fits noise. *Remedies*: Regularization, feature selection. **Underfitting**: High bias, oversimplifies. *Remedy*: Add features.
2. **Batch GD**: Full data per update; slow. **SGD**: One sample per update; noisy. **Mini-Batch**: Balance speed/accuracy. **Learning schedules**: Reduce η over time (e.g., $\eta_t = \eta_0/\sqrt{t}$).
3. **Mean-centering**: Ensures PCA axes maximize variance. Eigenvalue λ_i = variance along eigenvector $\mathbf{v}_i$. Explained variance = $\lambda_i/\sum \lambda_j$.
4. $\mathrm{E}[(y - \hat{f})^2] = \mathrm{Bias}(\hat{f})^2 + \mathrm{Var}(\hat{f}) + \sigma^2$.
5. **RBF**: Infinite-dimensional projection. **Polynomial kernel**: Preferred when feature interactions are known (e.g., physics models).
6. **Random Forest**: Aggregates decorrelated trees via bagging. **OOB error**: Validation score using unused samples during bagging.

Section B

1. **Ridge Solution**:

$$\nabla J = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + 2\lambda\mathbf{w} = 0 \implies \mathbf{w}^* = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

2. **Naive Bayes**:

$$\log P(C_k|\mathbf{x}) \propto \log P(C_k) + \sum_i \log P(x_i|C_k)$$

   Laplace smoothing: Prevents zero probabilities for unseen features.

3. **Logistic MLE**:

$$L(\mathbf{w}) = \sum_i [y_i \log \sigma(\mathbf{w}^T \mathbf{x}_i) + (1 - y_i) \log(1 - \sigma(\mathbf{w}^T \mathbf{x}_i))]$$

Gradient: $\nabla L = \mathbf{X}^T(\mathbf{y} - \sigma(\mathbf{X}\mathbf{w}))$.

Section C

1. **GD & Ridge**:
   - **(a)** Predicted $\hat{y} = [0, 0, 0, 0]$, error = $[-3, -5, -7, -9]$, gradient = $[-30, -60]$, updated weights = $[-0.3, -0.6]$.
   - **(b)** Ridge: $\mathbf{w}^* = [0.396, 0.791]$.
   - **(c)** MSE (Ridge) ≈ 0.012.
2. **PCA & Gini**:
   - **(a)** Eigenvalues: $\lambda_1 = 7.5, \lambda_2 = 2.5$; Eigenvectors: $\mathbf{v}_1 = [1, 1]^T/\sqrt{2}, \mathbf{v}_2 = [-1, 1]^T/\sqrt{2}$.
   - **(b)** Projection: $3.0 \times \frac{1}{\sqrt{2}} + 3.0 \times \frac{1}{\sqrt{2}} = 3\sqrt{2}$.
   - **(c)** Gini impurity = $1 - (0.5^2 + 0.5^2) = 0.5$.