Grading Rubric: Homework 4

Total: 100 points

Problem 1: PCA from Scratch (15 pts)

1.1 Eigendecomposition (7 pts)

Correct centering of data, covariance matrix computation, eigendecomposition, and reporting of eigenvalues and eigenvectors.

1.2 Projection (6 pts)

Correct projection of data onto principal components and visualization of original data versus projected data.

1.3 Verification with sklearn (2 pts)

Correct use of `sklearn.decomposition.PCA` and correct comparison of components and explained variance (up to sign).

Problem 2: Pitfalls of Principal Component Regression (15 pts)

2.1 PCA Analysis (4 pts)

Correct centering of data, PCA fit, explained variance ratios reported, and correct identification of dominant components.

2.2 PCR Using High-Variance Components (2 pts)

OLS fit using high-variance PCs, R^2 reported, and correct interpretation of poor predictive performance.

2.3 PCR Using Low-Variance Components (2 pts)

OLS fit using low-variance PCs and correct reporting of high R^2.

2.4 Conceptual Conclusion (7 pts)

Clear explanation of why selecting PCs by variance alone can be misleading.

Problem 3: Bagging from Scratch (20 pts)

3.1 Single Decision Tree (4 pts)

Correct model fit, visualization of decision boundary, and correct discussion of overfitting behavior.

3.2 Manual Bagging Implementation (10 pts)

Correct bootstrap sampling, ensemble construction, majority voting, and visualization of bagged decision boundary.

3.3 Comparison and Interpretation (6 pts)

Correct qualitative comparison between single tree and bagging ensemble with emphasis on variance reduction.

Problem 4: Gradient Boosting from Scratch (20 pts)

4.1 Base Model (OLS) (5 pts)

Correct OLS fit, MSE computation, and visualization demonstrating high bias.

4.2 Boosting Algorithm (10 pts)

Correct implementation of residual-based boosting with learning rate and weak learners.

4.3 Results and Interpretation (5 pts)

Final boosted fit shown, MSE comparison reported, and correct explanation of bias reduction.

Problem 5: Boosting vs. Random Forest (15 pts)

5.1 Model Comparison (6 pts)

Correct fitting of Random Forest and AdaBoost models and correct reporting of performance metrics.

5.2 Conceptual Explanation (9 pts)

Clear explanation of variance reduction in Random Forests and bias reduction in Boosting.

Problem 6: Clustering from Scratch (DBSCAN) (15 pts)

6.1 Algorithm Implementation (12 pts)

Correct implementation of DBSCAN-style logic, including neighborhood search, core point identification, and cluster expansion.

6.2 Visualization and Verification (3 pts)

Correct visualization of clustering results and correct comparison with `sklearn.cluster.DBSCAN`.

Grand Total: 100 points