Korea Aerospace University

2AI4203: Machine Learning

HW02: Machine Learning Methods

Jiyoon Shin

KAU College of AI Convergence

Oct. 15 – Oct. 18, 2026

Problem 1. Decision Trees

For parts (a) and (b), a classification tree node contains 6 positive samples and 4 negative samples. Consider the following candidate splits:

Split	Left Child	Right Child
A	4 positive, 0 negative	2 positive, 4 negative
B	3 positive, 1 negative	3 positive, 3 negative

(a) [3pts] Calculate the Gini impurity of the parent node.

(b) [7pts] Calculate the weighted Gini impurity and the impurity reduction for each split. Select the better split.

Problem 1. Decision Trees

For parts (c) and (d), a regression tree splits six samples into two leaves with the following target values:

$$R_L: \{2, 4, 6\}, \qquad R_R: \{8, 10, 12\}$$

(c) [4pts] Calculate the predicted value for each leaf.

(d) [6pts] Calculate the total sum of squared errors after the split.

Problem 2. Ensembles and Gradient Boosting

(a) [4pts] Select the correct option in each statement.

- Bagging trains base models independently / sequentially.
- Boosting trains base models independently / sequentially.
- Bootstrap sampling uses sampling with / without replacement.
- Random forests randomly select candidate features at each split / only once before training the entire forest.

(b) [4pts] Three regression trees predict 4, 7, and 10 for the same sample. Calculate the bagging prediction. Also, three classification trees predict Positive, Negative, and Positive. Give the bagging prediction.

Problem 2. Ensembles and Gradient Boosting

For parts (c) and (d), consider gradient boosting with squared-error loss. The targets and current predictions for three samples are:

$$\mathbf{y} = [3, 7, 8], \qquad \hat{\mathbf{y}}^{\text{old}} = [4, 5, 6].$$

(c) [6pts] Calculate the residuals.

(d) [6pts] The new tree predicts:

$$\mathbf{h} = [-1, 2, 2].$$

Using $\eta = 0.5$, calculate the updated predictions. Also, calculate the mean squared error before and after the update.

Problem 3. Support Vector Machines and Kernels

An SVM produces the following decision scores:

Sample	True Label y_i	Decision Score $f(\mathbf{x}_i)$
A	$+1$	1.5
B	$+1$	0.4
C	-1	0.3

Problem 3. Support Vector Machines and Kernels

(a) [6pts] Calculate the hinge loss for each sample.

(b) [4pts] Assign each sample to one of the following categories:

- Correctly classified and outside the margin.
- Correctly classified but inside the margin.
- Misclassified.

(c) [4pts] Suppose $||\mathbf{w}||^2 = 4$. Using the losses from part (a), calculate the objective of the Soft-Margin SVM for $C = 0.5$ and $C = 2$. Which value of C corresponds to stronger regularization?

(d) [6pts] Two points satisfy:

$$||\mathbf{x} - \mathbf{z}||^2 = 2.$$

Calculate their RBF kernel similarity for $\gamma = 0.5$ and $\gamma = 2$.

Problem 4. Principal Component Analysis

Consider the following data matrix, where each column represents one sample:

$$X = \begin{bmatrix} 3 & 1 & 2 \\ 2 & 2 & 5 \end{bmatrix}.$$

Apply mean centering only; do not standardize the features.

Problem 4. Principal Component Analysis

(a) [4pts] Calculate the mean vector $\boldsymbol{\mu}$ and the centered data matrix X_c.

(b) [4pts] Calculate the covariance matrix.

(c) [8pts] Calculate the eigenvalues and corresponding unit eigenvectors of the covariance matrix. Order the eigenvalues from largest to smallest and identify the first principal component.

(d) [4pts] Project the centered data onto the first principal component. Also, calculate the explained variance ratio of the first principal component.

Problem 5. Clustering

For parts (a) and (b), apply K-Means to the one-dimensional data:

$$x = \{0, 2, 3, 9, 10\}.$$

Use $\mathrm{K} = 2$, Euclidean distance, and the following initial centroids:

$$\mu_1 = 0, \qquad \mu_2 = 3.$$

Each iteration consists of cluster assignment followed by centroid update. If a point is equally distant from both centroids, keep its current cluster assignment.

(a) [6pts] Perform the first iteration. List the samples in each cluster and calculate the updated centroids.

(b) [6pts] Perform the second iteration. List the samples in each cluster and calculate the updated centroids. Identify the sample whose cluster assignment changed.

Problem 5. Clustering

(c) [4pts] For a sample in a clustering result, $a(i) = 2$ and $b(i) = 5$. Calculate its silhouette score. Is the sample, on average, closer to points in its own cluster / the nearest other cluster?

(d) [4pts] Consider the two clusters of one-dimensional points:

$$A = \{0, 2\}, \qquad B = \{5, 9\}.$$

Using Euclidean distance, calculate the distance between the clusters under single linkage and complete linkage.

Korea Aerospace University

Good Luck!

jiyoonshin@kau.ac.kr