Fall 2025 CMPE 462 Machine Learning
Assignment 2
Due: January 4th, 2026 by midnight

Fruit and Vegetable Dataset

In the second assignment, you will use the fruit and vegetable classification dataset collected in the first assignment. If you think that the dataset could be further improved, you can continue working on the data collection, pre-processing, and feature extraction in the second assignment, as well. If you have modified your dataset, please provide a clear explanation of the changes made compared to the original dataset. If you proceeded with the same dataset and features collected during the first assignment, please briefly recap the dataset statistics and the challenges your dataset may pose for a machine learning model. You may use the best feature set you obtained in the first assignment.

Task 1: Classification

You will conduct a benchmarking study as part of this task. You will compare logistic regression, logistic regression with a non-linear transformation, soft-margin SVM, soft-margin SVM with kernel trick, k-nearest neighbors classifier, naive Bayes, and random forest classifiers.

1. (10 points) Using a machine learning library, implement all the classifiers. Report their training times. Clearly explain how you determine the hyperparameters of the classifiers, non-linear transformations, and kernel functions. You may use regularization if necessary. Present all the choices you made to train the classifier in a table.

2. (15 points) Implement a soft-margin linear SVM from scratch without using a machine learning library, but using a quadratic programming solver. Find the support vectors. Find the data points that are the farthest from the hyperplane in each category.

(a) Visually inspect support vector data points. Is there anything special about them? Visually compare them with data points that are farthest in each category. Discuss your findings.

(b) Check the distances or similarities between support vectors across different classes. Rank their pairwise distances or similarities. Do the support vectors that are closest to each other belong to the categories most confused by the classifier? Comment on your findings.

3. (5 points) Report and compare the classification performance metrics of all the classifiers you trained. Based on your results, is your dataset highly non-linear or mostly behaves like linear with some outliers? Please discuss.

4. Please design an outlier detection framework for your dataset using the SVM constraints.

   (a) (20 points) Explain your proposed framework in detail. (Note: Do not propose directly training a classifier that learns a decision boundary between outliers and non-outlier points. Assume that you do not have any dataset with any annotation regarding whether a data point is an outlier or not.)

   (b) (20 points) Please also prepare a short demo video (screen-only). In the video, one group member should present their proposed idea for outlier detection. A presentation of 5-10 minutes should contain the following information:

   - Brief introduction of the idea using a figure illustrating the proposed framework.
   - Clear explanation regarding the evaluation protocol.
   - A live demo showing the detected outliers in the training set.

   NOTE: You do not need to prepare slides.

Task 2: Unsupervised Learning

You will again use your fruit and vegetable dataset in this task.

1. (10 points) Apply PCA. Explore the intrinsic dimensionality of your features by using reconstruction error and explained variance. Repeat Task 1, question 1, with the lower-dimensional features. Comment on the performance.

2. (10 points) Apply a clustering algorithm of your choice. Evaluate its outcome using external and internal metrics.

3. (10 points) Repeat question 6, part a, of Task 1 using clustering. Clearly explain your proposed outlier detection framework based on clustering and compare its output with the framework's output you designed in question 6, part a. You do not need to submit a video demo this time.

Submission

1. Please submit a PDF report comprising all the details asked in the tasks above. Please follow academic writing rules and cite your references. There is no specific format.
2. Please include a team member contribution statement at the end of your report.
3. Please include an AI use statement.
4. The PDF report will be submitted to the Turnitin assignment on Moodle. A single submission per group is sufficient. Please submit only one report per group.
5. Your datasets and implementations should be in a single zipped folder. Please include a README for how to reproduce your results. Please include a link to your zip file in your report.
6. Please include the link to the video demo in your PDF report.