Fall 2025 CMPE 462 Machine Learning
Assignment 1
Due: November 23 by midnight

Fruit and Vegetable Recognition

In this assignment, you will collect a fruit and vegetable classification dataset. Your dataset should include multiple modalities, such as images, categorical and numerical values, and text. Categorical and numerical data contain information like weight, color, season, and origin of the input. Text data could be a one-sentence description of the input.

Task 1: Data Collection and Feature Extraction

Each group will collect a dataset with five categories of their choice. Please collect at least 600 samples per category. You may split your dataset into a training set of size 2500 and a test set of size 500. You may create a validation set by randomly sampling 500 samples from your training set. Your dataset should satisfy the following requirements.

- Each sample should be represented by a feature vector.
- Feature dimensionality should be greater than 10 and less than 500.
- There should not be any missing values.
- Each sample should be annotated.
- The dataset should not be too easy (linearly separable) or hard (highly nonlinear) to train the classifiers in this assignment. (You may introduce some outliers and slight class imbalance to your training set to increase the complexity of your dataset.)

- You should collect at least 50 samples per class yourself. If collecting the whole dataset in the wild yourself will be time-consuming, you may generate the remaining images. However, you need to ensure that the samples in each category will be close to independent and identically distributed.

1. Please answer the following in your final report:
   (a) (5 points) Please describe your dataset. Include one example, e.g., an image, and other attributes, from each category in your report.
   (b) (5 points) Please describe your data collection procedure and your data pre-processing steps.
   (c) (30 points) Explain your feature extraction procedure.
      - If you came up with different attributes, please explain how you chose them.
      - You will need to extract features from images and text. Please investigate the best feature extraction approach for your problem. **You cannot use deep models for image feature extraction.** You may use deep learning-based word embedding models to extract word or sentence-level embeddings.
      - Propose an approach to fuse image, text, categorical, and numerical attributes.
   (d) (20 points) Please propose a strategy to report the intra-class and inter-class similarities and an approach to detect outliers. You may determine which metric to use to measure the similarity. Please comment on the difficulty of your dataset.

Task 2: Logistic Regression Classifier

1. In this task, you will train a logistic regression classifier. Use the one-vs-all approach to extend the two-class classifiers to multiclass.
   (a) (25 points) Implement the logistic regression classifier from scratch. Plot the training and validation loss. Add regularization if your model overfits. Train separate classifiers for using only images, only categorical and numerical attributes, only text, and your proposed fused features.
   (b) (10 points) Report the training and test classification metrics, such as accuracy, precision, recall, F1-score, and AUC (area under the receiver operator characteristic curve). Compare the performances of the classifiers in part (a).
   (c) (5 points) Compare your implementation's performance and runtime with the logistic regression implementation of the Scikit-learn library.

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. The PDF report will be submitted to the Turnitin assignment on Moodle. A single submission per group is sufficient. Please do not submit multiple reports per group.

4. 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.