Introduction to Machine Learning
Lab Assignment 3: Classification with Dimensionality Reduction and Ensemble Learning

1. **Image Classification with Dimensionality Reduction [50 points]:** You will evaluate how using the unsupervised learning tool, PCA, to reduce the feature set size impacts the performance of two classification algorithms on two different datasets.

   (a) Prepare two image classification datasets for this study: (Code)

      i. Load two datasets: MNIST (number recognition), LFW (face recognition).

      ii. Split each dataset into a 75/25 train/test split.

   (b) For each dataset, train and evaluate two classification algorithms (e.g., SVM, Naive Bayes, Decision Tree, KNN) with at least 25 different feature dimension sizes. To achieve the different feature dimension sizes, you will need to vary the number of principal components you keep using the unsupervised dimensionality reduction algorithm PCA. (Code)

   (c) Report the predictive performance of each classifier on each dataset as a function of the feature dimension size (i.e., number of principal components). As a result, your write-up should include four plots to show results for both classification algorithms on both datasets. (Write-up)

   (d) Write a discussion analyzing the influence of applying PCA on the classification performance. For example, what feature dimension sizes were better/worse and why do you think so? What can you infer by observing the classification performance across the different datasets and different classification algorithms? Your discussion should consist of two to four paragraphs. (Write-up)

2. **Ensemble Learning [50 points]**: You will analyze the effects of using different types of ensembles for the classification task.

   (a) Load two text-based classification datasets (e.g., from NLTK or Kaggle) and preprocess as necessary to arrive at a numerical representation of the text. (Code)

   (b) For each dataset, evaluate these classifiers using 10-fold cross validation: (Code)

      i. Three different classifiers (e.g., SVM, Naive Bayes, Decision Tree, KNN).

      ii. Majority vote classifier that uses the three classifiers from previous step.

      iii. Bagging method.

      iv. Boosting method.

   (c) Report in a table the mean accuracy of each classifier you evaluated in the previous step for each dataset. Consequently, your report should include two tables that each show six accuracy scores for the six methods. (Write-up)

   (d) Write a discussion about the performance of the different ensemble methods. For example, what classification approaches did better/worse and why do you think so? How did ensemble methods compare to non-ensemble methods? What can you infer by observing the classification performance across the different datasets? Your discussion should consist of two to four paragraphs. (Write-up)

How to Submit Lab Assignment 3: Please submit a pdf that provides the code (or hyperlinks to the code) and answers to the questions, as deemed appropriate for the task. The pdf file should be named using your first and last name; i.e., firstname_lastname.pdf. The material you submit must be your own.