CSE514 – Fall 2023 Programming Assignment 2

This assignment is to give you hands-on experience with dimension reduction and the comparison of different classification models. It consists of a programming assignment and a report. This project can be done in groups up to three, or as individuals.

Topic

Compare, analyze, and select a classification model for identifying letters in various fonts.

Programming work

A) **Data preprocessing**

This dataset contains 26 classes to separate, but for this assignment we'll simplify to three binary classification problems.

Pair 1: H and K   Pair 2: M and Y   Pair 3: Your choice

B) **Model fitting**

For this project, you must pick 2*(size of group) from the following classification models:

1. k-nearest neighbors	2. Artificial Neural Network	3. Decision tree
4. Random Forest	5. Naïve Bayes Classifier	6. SVM

For each model, choose a hyperparameter to tune using 5-fold cross-validation. You must test at least 3 categorical values, OR at least 5 numerical values.
Hyperparameter tuning should be done separately for each classification problem; you might end up with different values for classifying H from K than for classifying M from Y.

2. **Dimension reduction**

For each model, implement dimension reduction to reduce the number of features from 16 to 4. Retrain your models using reduced datasets, including hyperparameter tuning.

IMPORTANT: You may use any packages/libraries/code-bases as you like for the project, however, you will need to have control over certain aspects of the model that may be black-boxed by default. For example, a package that trains a kNN classifier and internally optimizes the k value is not ideal if you need the cross-validation results of testing different k values.

Data to be used

We will use the Letter Recognition dataset in the UCI repository at

https://archive.ics.uci.edu/ml/datasets/letter+recognition

There are 20,000 instances in this dataset. Note that the first column is the response variable (i.e., y).

For each binary classification problem, first find all the relevant samples (ex. all the H and K samples for the first problem). Then set aside 10% of those samples for final validation of the models. This means that you cannot use these samples to train your model parameters, your model hyperparameters, or your feature selection methods.

What to submit – follow the instructions here to earn full points

- (75 pts total) The report
  - Introduction (20 pts)
    - (4 pts) Your description of the problem and the practical impacts of solving it.
    - (4 pts) What is the motivation for training and testing multiple classifiers? What factors should be considered in determining a classifier as the "best," e.g. computational complexity, validation accuracy, model interpretability, etc.
    - (4 pts) What is the motivation for dimension reduction? Which methods are "better," and what factors should be considered in determining a dimension reduction method as "good" or "bad."
    - (4 pts) Brief description of the dimension reduction method(s) you chose.
    - (4 pts) Speculate on the binary classification problems. Which pair of letters did you choose for the third problem? Which pair do you predict will be the easiest or hardest to classify?
  - Results (36 pts)
    - For each classifier:
      - (6/3/2 pts) Brief description of the classifier and its general advantages and disadvantages.
      - (6/3/2 pts) Figure: Graph the cross validation results (from fitting the classification model *without* dimension reduction) over the range of hyperparameter values you tested. There should be three sets of values, one for each binary classification problem.
      - (6/3/2 pts) Figure: Graph the cross validation results (from fitting the classification model *with* dimension reduction) over the range of hyperparameter values you tested. There should be three sets of values, one for each binary classification problem.
  - Discussion (19 pts)
    - (5 pts) Compare the performance and run time of the different classifiers on the final validation sets with either a table or a figure.
    - (5 pts) Compare the performance and run time of the different classifiers after dimension reduction on the final validation sets with either a table or a figure.
    - (9 pts) Lessons learned: What model would you choose for this problem and why? How did dimension reduction effect the accuracy and/or run times of the different classifiers? What would you do differently if you were given this same task for a new dataset? Include at least one additional topic of discussion.

- (25 pts total) Your code (in a language you choose) including:
  - (5 pts) Instructions for running your code
  - (5 pts) The code for processing the data into training/testing sets
  - (10 pts) The code for your classifiers
  - (5 pts) The code for your dimension reduction method

Due date

Wednesday, December 6 (midnight, STL time). Submission to Gradescope via course Canvas.

A one week late extension is available in exchange for a 20% penalty on your final score.

Extra credit opportunities:

Opportunities to submit sub-sections or side-goals of the project will be made available during the weeks leading up to the final submission date. In total, you can earn up to 20 bonus points on this assignment, with a cap of 110% as the maximum score.