CSE514 – Spring 2025 Project 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, a report, and a 10min presentation.

This project can be done in groups up to three, or as individuals.

Programming work

A) Data preprocessing

Your dataset might contain response labels of two classes, multiple classes, or a numerical data type. This project requires you define two (2) classification problems. This can be done by choosing two different response variables, or by processing the response variable in two different ways.

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

I.e. if you are working alone, pick two models to train; if you're working in a pair, pick four models to train, and if you're working in a group of three, train all six models.

For each model, choose a hyperparameter to tune using k-fold cross-validation, where k > 3.

If the hyperparmeter is categorical (ex. which SVM kernel), you must test at least 3 options (ex. linear, polynomial, RBF).

If the hyperparameter is numerical (ex. pseudo-count value for Naïve Bayes) you must test at least 5 values (ex. pseudo-count = [1, 2, 3, 4, 5]).

Hyperparameter tuning should be done separately for each classification problem.

C) Dimension reduction

Implement dimension reduction to reduce the number of features in half. Retrain your models using reduced datasets, including hyperparameter tuning.

You may use the same dimension reduction method for each classification problem/model, or you may use different methods for certain cases.

IMPORTANT: None of the models for this project need to be coded from scratch; 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 an SVM classifier and internally optimizes the kernel chosen is not ideal if you need the cross-validation results of testing different k values.

Data to be used

You will be picking the dataset for this project. Some recommended places to look are:

https://archive.ics.uci.edu/ml/datasets/

https://mavenanalytics.io/data-playground

You must find a dataset that has **at least 10 features**, and enough samples that each classification problem has **at least 200 samples**. For each classification problem, set aside 10% of relevant 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.

Part A: Explore your data and define a solvable problem (13pts)

You must chose a dataset to work with. Submit to Gradescope:

- A brief definition/description of what the dataset is and what the variables are measuring
- Define a real-world motivation for fitting classification models to this dataset.
    - What would be the response variable?
    - Who would want these models and why?
- Visualize the distribution of variables.
    - Are any of them categorical?
    - Are any of them normally distributed?
- Process the dataset to create two classification problems.
    - What steps did you take to define these two problems?
    - Submit these two subsets as CSV or TSV files.

Part B: Explore your models, hyperparameters, and test cross validation (18pts)

Write code (in the programming language of your choice) to train your chosen models on your data. Submit to Gradescope:

- For each model:
    - A brief definition/description of the model
    - Two strengths and two weaknesses of the model
    - Chosen hyperparameter for the model and the values you will test
    - Function call to run the model with data and a chosen hyperparameter value
    - Pick a performance metric (or multiple metrics) to evaluate your models
- Demonstrate you can run cross-validation
    - Pick a classification problem and a model to test
    - Write code that will run k-fold cross-validation for testing hyperparameter values
    - Graph the cross validation results

Part C: Choose a dimension reduction method and fit your models (25pts)

Dimension reduction can have a positive or a negative impact on model performance, so its time to test it out. Submit to Gradescope:

- A brief description/definition of the dimension reduction method(s) you choose to use.
    - How specifically will you apply it to cut the number of features in half?
- For each model:
    - For each classification problem:
        1. Perform k-fold cross validation on the full training dataset
        2. Visualize the cross validation results
        3. Use the best hyperparameter value to train the model on the whole training dataset
        4. Use the trained model to predict on the final validation set
        5. Report the performance and the runtime of steps 3 and 4
        6. Redo all of the above with your dimensionally reduced datasets.

Final report: Explain your methods and draw conclusions (34pts)

The data mining work is mostly complete, well done! Now, you need to communicate your results. Remember that the goal of data mining is to obtain *actionable* knowledge, so your last task is to show that your work has done just that.

- Introduction: Explain the motivation behind analyzing this dataset.
- Results: Use your models to draw some conclusions.
    - Were both classification problems equally predictible?
    - Were all the tested models equally "good"?
    - Did dimension reduction have a meaningful (negative or positive) impact?
    - Given the motivation, what would you recommend?
    - Include at least one figure to support your conclusions.
- Methods: Explain/justify your methods
    - What values did you end up picking as hyperparameters? Explain how you decided
    - What dimension reduction method(s) did you end up using? Explain these methods

Presentation (10pts)

The last piece of this assignment is to deliver a 10min presentation to the class that communicates:

- Who is your "client" and why did they ask you to analyze this dataset?
- What recommended actions did you respond with, and what results do you expect?
- What exactly did you do with the data to support your answers to the previous question?

This presentation must be delivered in the same groups as the final report submission.

Due date

Tuesday, April 22 (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.