CSE514 – Spring 2024 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 a series of binary classification problem.

Programming work

A) Data preprocessing

Your dataset might contain response labels of two classes, multiple classes, or a numerical data type. This project requires three binary classification problems. Here are some options:

1. Dataset has two classes: Find a feature to split the dataset up into 3 populations
2. Dataset has multiple classes: Pick 3 different pairs
3. Dataset has numerical response values: Bin the values into classes. Once the variable has been processed to be categorical, refer to option 1 or option 2.

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.
If the hyperparmeter is categorical (ex. which SVM kernel), you must test at least 3 options.
If the hyperparameter is numerical (ex. # of neighbors in kNN) you must test at least 5 values.
Hyperparameter tuning should be done separately for each classification problem.

2. Dimension reduction

For each model, implement dimension reduction to reduce the number of features in half. 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

You may pick any dataset from the UCI repository at: https://archive.ics.uci.edu/ml/datasets/

You must find a dataset that has at least 10 features, and enough samples that each binary classification problem has at least 200 samples. For each binary 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 (15pts)

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 a binary classification model 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 three binary classification problems.
    - What steps did you take to define these three problems?
    - Submit these three subsets as CSV or TSV files.

Part B: Explore your models and their hyperparameters (15pts)

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
- Demonstrate you can run cross-validation
    - Pick a binary classification problem and a model to test
    - Write code that will run 5-fold cross-validation for testing hyperparameter values
    - Graph the cross validation results

Part C: Fit all models and validate results (15pts)

It's time to evaluate the models. Submit to Gradescope:

- Pick a performance metric (or multiple metrics) to evaluate your models. Explain your choice(s).
- For each model:
    - For each binary classification problem:
        1. Perform 5-fold cross validation on the 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

Part D: Test the impact of dimension reduction (15pts)

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?
- Redo Part C work with your dimensionally reduced datasets.

Final report: Explain your methods and draw conclusions (40pts)

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 all three binary 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 on those values
    - What dimension reduction method(s) did you end up using or not using? Explain how these methods work

Due date

Monday, 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.

Extra credit opportunity, can be submitted as individuals or in groups up to three (10pts):

Fit models to the dataset after defining a single multi-class classification problem.

1. Dataset has two classes, and you split the dataset up into 3 populations: Define six class labels (population1-class1, population1-class2, …, population3-class2)
2. Dataset has multiple classes: Fit multi-class classifiers to whole dataset
3. Dataset has numerical response values: Bin the values into classes. Once the variable has been processed to be categorical, refer to option 1 or option 2.

Re-write the report using these new results.