CSE5ML: Machine Learning – Assignment 1

Semester 2, 2023

Overview

- This assignment contributes **20%** of your final mark in the subject. Please read this sheet carefully before doing your assignment.
- The assignment aims to consolidate your knowledge of pre-modeling processes, **Classification models,** and **Clustering models** to strengthen your Machine Learning skills.

Policies

- This is an **individual** assignment.
- Plagiarism is the submission of somebody else's work in a manner that gives the impression that the work is your own. The Department of Computer Science and Information Technology at La Trobe University treats plagiarism very seriously. When it is detected, **penalties are strictly imposed**.

Submission

- The submission of the assignment is due on **Monday 18th September 11:59 pm**.
- The assignment consists of your code file (please put them in one file) and a report of min 1000 words.
- As the assignment contributes over 15% of your final mark, you need to apply for **Special Consideration** to the University if requiring an extension. Unless special consideration is given by the University, late submission is **NOT** accepted. Please refer to the link below for more details. https://www.latrobe.edu.au/students/admin/forms/special-consideration
- A penalty of 5% per day will be imposed on all late assignments for up to 5 days. An assignment submitted more than five days after the due date **will NOT be accepted and zero marks will be assigned**.
- All submission needs to have your student name and number included.

Problem Description

The assignment requires you to first do basic pre-processing on the original dataset, and develop 2 end-to-end classification models: Logistic Regression and SVM (all have been studied during the course). Then, you are required to apply the K-Means clustering method to further understand the dataset.

The classification models should be developed to accurately predict the individual income (greater than 50k or not) based on given factors for people living in the US, and the clustering model should be used to understand the natural grouping of the data.

The dataset (*income.csv*) is collected and provided by the US Census database. We did data cleaning on the original dataset, and after cleaning, the detailed factors are as follows:

- age: the age of an individual
- workclass: a general term to represent the employment status of an individual
  - Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- education: the highest level of education achieved by an individual.
  - Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- marital-status: marital status of an individual.
  - Married, NotMarried, Separated (including divorced), Widowed,
- occupation: the general type of occupation of an individual
  - Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- relationship: represents what this individual is relative to others. For example an individual could be a Husband. Each entry only has one relationship attribute and is somewhat redundant with marital status. We might not make use of this attribute at all
  - Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- race: Descriptions of an individual's race
  - White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- sex: the biological sex of the individual
  - Male, Female
- hours-per-week: the hours an individual has reported to work per week
- income: whether or not an individual makes more than $50,000 annually.
  - 0: <=50k
  - 1: >50k

Requirements

To achieve the goal, you need to carry out the following tasks:

1. **Load the dataset, do basic data preprocessing, and split the dataset.**

   The dataset is provided for you in the .csv file format in the same folder. The first column (income) would be our target and true label (y), and the rest of the 9 columns would be used as predictors to create input X. Detailed steps will need to include (you can apply additional preprocessing if you find necessary):

   a. Describe the dataset (before and after the pre-processing), for example, variable type, and data shape. (You may also consider applying correlation analysis, optional)

   b. Deal with missing values (if there are any) and use a proper method to handle categorical variables.

   c. Remove duplicated inputs if there are any.

   d. Handle the categorical variables.
      - For the ordinal variable education, assign values 1 to 16 to the categories in this order: Preschool, 1st-4th, 5th-6th, 7th-8th, 9th, 10th, 11th, 12th, HS-grad, Some-college, Assoc-voc, Assoc-acdm, Bachelors, Masters, Prof-school, Doctorate.
      - For the binary variable sex, assign value 0 to Male and value 1 to Female
      - For the rest of the variables, apply dummy coding to deal with them

   e. Split the dataset into training and testing (with 10% of the dataset for testing).

   f. Apply normalization on X (both training and test set).

   g. In your report, describe the dataset (before and after the processing), describe how you did the above steps (you just need to describe the steps and the result you get from each step, do not attach codes in the report), and discuss why you did them.

2. **Train and evaluate the 2 classification models on the training set with the cross-validation method, optimize the models and evaluate models on the test set.**

   a. Define the two regression models, including Logistic Regression, and

SVM, with their default settings

b. Define 10-fold cross-validation to train and evaluate the two models based on the average score

c. Apply parameter finetuning steps to the two models separately to optimize the model performances and compare the cross-validated results before and after finetuning for each model.

d. Evaluate the two optimized models (with the best parameter setting from the above step for each model type) on the test set, and compare the results with what you got from 2b.

e. In your report, you need to start by explaining the basics of Logistic Regression and SVM. Then, describe the cross-validated and test results for the two models with default parameter settings, and compare and discuss the results among models. Next, describe what steps you have taken for finetuning your model (changing the parameters), describe the parameter settings that you applied in finetuning, and compare the results for each model (before and after finetuning for each model). Finally, compare the evaluation results across two optimized models on the test set, and discuss your findings. (You may consider using a table to record all the modeling results)

3. Apply K Means clustering on the normalized training input X, and understand the grouping of training data by investigating the prototype from each cluster

a. Apply clustering on the normalized training input X (you can determine the number of clusters by considering how many classes for the target y)

b. Identify how many data samples have been assigned to each cluster.

c. Extract a prototype from each cluster and investigate their similarity and difference.

d. Evaluate the clustering accuracy with the testing set and compare with the results from 2d.

e. In your report, you need to start by explaining the basics of the K means method. Then, describe how many clusters you have chosen in your data clustering and how many data samples have been assigned to each cluster with the K means model. Compare the differences and similarities between the prototype for each cluster. And finally, evaluate the accuracy of the clustering method based on the testing set, and compare the

results from the 2 models in 2d, and discuss your findings.

Your code needs to:

- ✓ Cover all the required steps in the above guidelines
- ✓ You may consider applying additional steps if you find them necessary.

Your report needs to

- ✓ Cover all the required contents in the specifications under each task
- ✓ It also requires you to have references if you have used any
- ✓ be in the pdf format

Marking Criteria

Criterion	Contribution
Data loading, data preprocessing and dataset spitting with description and explanation	30
Knowledge of the 2 classification models	10
Cross-validation, model finetuning, evaluation with description and discussion	40
Knowledge of k means clustering	5
Data clustering with description and further discussion on grouping , prototypes and comparison results	15
Total	**100**