Machine Learning

Program Assignment 1

Fall 2020

Instructor: Xiaodong Gu

Goals

- Implement Two Classification Algorithms
  - ✓ Naïve Bayesian
  - ✓ Logistic Regression
- Using the *sklearn* Python package for support vector machines (SVM)
- Compare the performance of several classification methods by conducting an empirical comparative study.

Dataset

Mobile Price Classification

https://www.kaggle.com/iabhishekofficial/mobile-price-classification?select=train.csv

battery_power	clock_speed	.....	wifi	Price Range
842	2.2		0	0
1021	0.5		1	1
970	1.2		0	2
1821	1.5		1	0
1988	0.8		1	3

Platform

Python notebook by Tencent's TI-ONE (Recommended)

- Consult in the WeChat group if you have any issues in using the platform.

Python notebook by Anaconda in a local machine

- For quick setup and test

Requirements and Key Points

- Data Splitting
  - Split the original 'train.csv' into 'train.csv', 'valid.csv' and 'test.csv' with the ratio of 0.8 : 0.1 : 0.1, respectively.
- Data Preprocessing
  - Convert labels into to two classes: low (0, 1) and high (2, 3)
  - For Naïve Bayes, you may need to:
    - ✓ discretize continuous attributes into intervals
    - ✓ split large number into ranges
  - You may need data normalization (i.e., scaling values of attributes to the same level, e.g., [0, 1])
- Model Implementation

  Implement Naïve Bayes and Logistic Regression using Python's primitive classes and functions. Calling third-party libraries (such as sklearn) for model building (except for SVM) will get 0 points.

Empirical Study

Compare the three methods with respect to the classification accuracy on the training set and the test set separately:

- Naive Bayesian
- SVM
- Logistic regression

Report the comparison of accuracy using figures and tables.

Report the time required by each of the methods, excluding the time needed for loading and processing data. This may be done using the *time* module. The result can be shown in a table.

Provide a description for each figure or table. Analyze the difference, pros and cons of the three methods.

Demos

https://github.com/zealptekin/Mobile-Phone-Price-Prediction/blob/master/Project_Report_github.ipynb

https://github.com/ditekunov/mobile-phones-price-prediction/blob/master/models/binary_models.ipynb

https://github.com/vikram-bhati/Mobile-price-prediction/blob/master/Machina.ipynb

https://github.com/sidmojo/Mobile-Price-Classification/blob/master/MobilePrice.ipynb

Grading Scheme

- Data separation and preprocessing [10%]
- Implement the Naïve Bayes classifier and achieve a normal accuracy [20%]
- Using SVM from the *sklearn* module [10%]
- Implement the logistic regression and achieve a normal accuracy [30%]
- Empirical study (results + analysis) [20%]
- Report quality [10%]

Submission

Submit a single file named 'StuID_NAME.ipynb' to Canvas.

Due date: Nov. 12

Tips

Your programs should be written in such a way that the TA can run them easily to verify the results reported by you.

Your report should be clear and comprehensive so that students from other departments can still understand. Reports that are too messy or brief may lose the quality score.