IE343: Statistical Machine Learning

Midterm Project

Professor: Seyoung Yun

(TA: Jung-Hun Kim)

- Instruction
  In this midterm project, you need to create logistic models applied to real datasets. There are two tasks for this project. Task 1 is about building a logistic regression model for a multi-class problem. In this task, the iris flower dataset is provided, and the problem is about classifying each iris data into 3 species. For the second task, you must build a binary logistic regression model for the Titanic dataset. For this dataset you need to pay attention to preprocess data for improving model accuracy. The problem is finding the survivors from the passengers. More detailed explanations of each task are in the sections below. Python is the only programming language you will be allowed to use. Also for completing code, you are not allowed to use machine learning libraries such as 'Scikit-learn', 'Keras', or 'SciPy'. However basic libraries such as 'NumPy', 'pandas', or 'math' are allowed.

- Grade policy
  50 points will be assigned to each task. For task 1, 50pts will be assigned to the report with your source code. For task 2, 20pts will be assign to the report and 30pts will be assign to the model accuracy. For measuring the model accuracy, the TA will run your code and get the model accuracy with a test dataset which is different from the distributed test dataset.

Task 1

- Data
  - The dataset has information of iris flower samples (training set: 104, test set: 23).
  - Input features: sepal length, sepal width, petal length, and petal width.
  - Target: 3 species of iris ('iris-virginica', 'iris-setosa', and 'iris-versicolor').

- Method
  The logistic regression for multi-classes is called 'multinomial logistic regression' or 'softmax regression'. The multinomial logistic function with parameter θ containing $\theta^{(j)}$'s is represented as

$$h_\theta(x) = \frac{1}{\sum_{j=1}^{K} \exp({\theta^{(j)}}^\top x)} \begin{bmatrix} \exp({\theta^{(1)}}^\top x) \\ \exp({\theta^{(2)}}^\top x) \\ \vdots \\ \exp({\theta^{(K)}}^\top x) \end{bmatrix}. \tag{0.1}$$

  where K is the number of classes, $\theta^{(j)}$ is the vector parameter for the class j, and x is the input vector.

- To-Do List
  1. Design a proper loss function for the multinomial logistic function and describe how to train your model parameter using a gradient descent method. Write down your formulas in detail.
  2. Complete the code in './App/logistic_regressor.py' file using the model described in the Method section.
  3. Describe the code you wrote in 'logistic_regressor.py' file line by line and write down the accuracy of your model.
  4. Is there any methods to apply binary logistic regression models to multi-class problems? Explain your ideas.
  5. You must submit your entire completed python files and a report which contains the answers of the question 1, 3, and 4.

Task 2

- Data
  - The dataset has information of passengers on the Titanic (train: 623, test: 134).
  - Input features: Pclass, Name, Sex, Age, ...
  - Target: Survivor(=1) or not(=0).
  - You can find more details about the data via the url: https://www.kaggle.com/c/titanic/data. There are some ambiguous features in the dataset. Discussions on the above url are helpful.
- Method
  The logistic regression for binary classes must be used for this project
- To-Do List
  1. Complete the codes in 'main.py' for data preprocessing and './App/logistic_regressor.py' for a model.
  2. Describe your methods for data preprocessing. You don't need to describe your data preprocessing code line by line but explain your ideas for data preprocessing in the report.
  3. Describe your logistic regression model with the code line by line.
  4. You must submit your entire completed python files and a report which contains the answers to question 2 and 3.