CH5019 – Mathematical Foundations of Data Science

Term Project

January-May 2018 Semester **13th April 2018**

Question 1

Write your own code to fit a logistic regression model to the data set described below in a programming language of your choice. (**IMPORTANT: DO NOT USE ANY IN-BUILT LIBRARIES**)

Description of Data Set 1:

This data set describes the operating conditions of a reactor and contains class labels about whether the reactor will operate or fail under those operating conditions. Your job is to construct a logistic regression model to predict the same.

- `q1_data_matrix.csv`: This file contains a 1000×5 data matrix. The 5 features are the operating conditions of the reactor; their corresponding ranges are described below:
  1. **Temperature:** 400-700 K
  2. **Pressure:** 1-50 bar
  3. **Feed Flow Rate:** 50-200 kmol/hr
  4. **Coolant Flow Rate:** 1000-3600 L/hr
  5. **Inlet Reactant Concentration:** 0.1-0.5 mol fraction
- `q1_labels.csv`: This file contains a 1000×1 vector of 0/1 labels for whether the reactor will operate or fail under the corresponding operating conditions.
  - 0: The reactor will operate well under the operating conditions
  - 1: The reactor fails under the operating conditions

Some General Guidelines:

1. Partition your data into a training set and a test set. Keep **70%** of your data for **training** and set aside the remaining **30%** for **testing**.
2. Fit a logistic regression model on the training set. Choose an appropriate objective function to quantify classification error. **Manually code for the gradient descent procedure** used to find optimum model parameters. (**Note:** You may need to perform multiple initializations to avoid local minima)
3. Evaluate the performance of above model on your test data. Report the **confusion matrix** and the F_1 **Score**.

Question 2

Use the same code developed in Question 1 to fit a logistic regression model to the dataset described below.

Description of Data Set 2:

This data set contains data for credit card fraud detection.

- `q2_data_matrix.csv`: This file contains a 100×5 data matrix. The 5 features and their corresponding ranges are described below:
    1. **Age:** 18-100 years
    2. **Transaction Amount:** $ 0-5000
    3. **Total Monthly Transactions:** $ 0-50000
    4. **Annual Income:** $ 30000-1000000
    5. **Gender:** 0/1 (0 - Male, 1 - Female)
- `q2_labels.csv`: This file contains a 1000×1 vector of 0/1 labels for whether the transaction is fraudulent or not.
    - 0: The transaction is legitimate
    - 1: The transaction is fraudulent

1. Report the confusion matrix and the F_1 Score for this data set.
2. Which data set gives better results better? Can you think of reasons as to why one data set gives better results than the other? (**Hint:** Think of assumptions behind the logistic regression model)
3. Can you suggest improvements to the logistic regression model to make it perform better on the unfavorable data set?
4. **Bonus Points!**: Implement your suggested improvement as a code and compare the performance of this with vanilla logistic regression.