CPEN/CPSC 436 Midterm Project

Regularized Logistic Regression with Real Dataset

Project Description

You are required to extend assignment #5 and implement regularized logistic regression using Jupyter Notebook. A real diagnostic breast cancer dataset with 30 real-valued input features is used in this project. Your goal is to achieve an accuracy rate greater than or equal to **95%**. You must at least implement the following extensions.

1) Upload and transform the dataset. Note that the first column of the dataset is an ID number and the second column represents the diagnosis with either M (malignant) or B (“benign”) (20 points).
2) Scale the dataset using z-score normalization. (10 points)
3) Implement regularization (i.e., extend the compute_cost and compute_gradient functions) (20 points)
4) Plot the learning curve (i.e., cost versus iterations) (10 points)
5) Add adequate notes (e.g., an introduction to the dataset, regularized logistic cost function and gradient descent, etc.) (10 points)
6) Adjust relevant parameters such as alpha, lambda, etc. get accuracy greater than or equal to **95%**. (20 points)
7) No compiling warnings (10 points)

Requirements

The project is open book and open notes. But you must work on it by yourself. You are not allowed to use high-level machine learning libraries like Scikit-Learn. You have one week to finish the project, but you are suggested to start early as you may encounter some unexpected errors associated with a real dataset.