ENGINEERING DIVISION | NYU ABU DHABI

ENGR-UH 4560
Selected Topics in Information and Computational Systems

Machine Learning

Project 01 - Linear and logistic regression

Due: 11:59 PM AD Time, Feb. 20

Introduction - Linear regression

In statistics, linear regression is a linear approach to modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.

Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares loss function as in ridge regression (L2-norm penalty) and lasso (L1-norm penalty). Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked, they are not synonymous.

Requirements

- Implement a linear regression model with your own gradient descent module on example data (*ex1data1.mat*)
    - Calculate the parameters (a,b) in a line function y = ax+b via linear regression model.
    - Plot the output line and the input data in the same figure. Plot the cost curve.
- Implement a logistic regression model on the example data (*ex2data1.mat*)
    - Logistic regression hypothesis can be realized by sigmoid function:

      $h_\theta(x) = g(\theta^T x)$, $g(z) = \frac{1}{1+e^{-z}}$
    - Vectorize the cost function.
    - Generate the decision boundary of example data via your logistic regression model.
    - Plot the output boundary and the input data in the same figure.
- Implement a regularized logistic regression model on example data (*ex2data2.mat*)
    - Vectorize the cost function:

      $J(\theta) = \frac{1}{m}\Sigma_{i=1}^{m}[-y^i log(h_\theta(x^i)) - (1-y^i)log(1-h_\theta(x^i))] + \frac{\lambda}{2\mu}\Sigma_{j=1}^{n}\theta_j^2$
    - Generate the decision boundary of example data via your logistic regression model.
    - Plot the output boundary and the input data in the same figure.
    - Explain the difference between logistic regression and regularized logistic regression and why regulation can improve the performance.

Deliverables

A zip file containing the following:

1. a working project (source code, makefiles if needed, etc)
2. a report with the detailed description of the project
    a. explain the main aspects of your code
    b. how to run your project
    c. plots and diagrams

Before submitting your project, please make sure to test your program on the given dataset.

Notes

You may discuss the general concepts in this project with other students, but you must implement the program on your own. ***No sharing of code or report is allowed.*** *Violation of this policy can result in a grade penalty.*

Late submission is acceptable with the following penalty policy:

- ***10 points deduction for every day after the deadline***