Offline 3

Multivariate Logistic Regression using Gradient Descent

Problem Statement
Compute the accuracy of diabetes prediction on your test data by the logistic regression model.

Steps

1. Load Data into your preferred data structure
2. Scale features using the following formula (will help in converging faster: optional)

$$x' = \frac{x - \text{mean}(x)}{\max(x) - \min(x)}$$

3. Split data into train and test (70-30)
4. Initialize the parameter vector randomly (size = feature no. + 1)
5. Compute the value of cost function (J) using the following formula from training set

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \text{Cost}(h_\theta(x^{(i)}), y^{(i)})$$
$$= -\frac{1}{m} [\sum_{i=1}^{m} y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))]$$

   Here, m = no. of rows in training set,
   i = training set iterator
   y_i = outcome of ith row

$$h_\theta(x) = \frac{1}{1+e^{\theta^\top x}}.$$

6. While J is not close to 0:
   a. For each parameter in the parameter vector:
      i. Update the value of the parameter by gradient descent by the following formula

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

      Here, m = no. of rows in training set,
      i = training set iterator
      α = learning rate

x_j(i) = jth input data of ith training row and

$$h_\theta(x) = \frac{1}{1+e^{\theta^\top x}}.$$

  b. Compute J from the training set as shown in the previous formula

7. For each row in the test dataset:
   a. Compute your prediction using the following formula

      - if $h_\theta(x) \geq 0.5$ (i.e. $\theta^\top x \geq 0$) predict $y = 1$
      - if $h_\theta(x) < 0.5$ (i.e. $\theta^\top x < 0$) predict $y = 0$

8. Compute and print accuracy of your prediction