ENMG 616 Advanced Optimization & Techniques

Homework Assignment 3

Due Date: November 04, 2023

You need to show the steps followed to get the final answer (Do not just give the final result). The homework should be submitted to Moodle as ***one*** *pdf file. Please insert a copy of your Matlab code in the submitted file.*

Logistic Regression on a real Dataset

Logistic regression is a simple machine learning model that in its basic form uses a logistic function to model a binary dependent variable. More specifically, this modeling technique aims at finding significant relationship between dependent features $\mathbf{x}$ and a binary target variable y. In this problem we will use a logistic regression model to predict whether a patient diagnosed with breast cancer has a malignant or benign mass.

We are given $\mathbf{x}_i$ representing features extracted from a digitized image of a needle aspirate (FNA) of a breast mass, and y_i representing the type of breast mass, 1 for malignant and 0 for benign. To fit a logistic model of our training data, we solve the following optimization problem which minimizes the cross-entropy loss:

$$\min_{\mathbf{w},b} f(\mathbf{w},b) \triangleq \sum_{i=1}^{n} -y_i(\mathbf{w}^T\mathbf{x}_i + b) + \log\left(1+\exp(\mathbf{w}^T\mathbf{x}_i+b)\right) + \frac{\lambda}{2}\|\mathbf{w}\|^2.$$

Given a new image, we again extract the features $\mathbf{x}$. We then use our trained model $(\mathbf{w}^*, b^*)$ to determine the probability that the mass is malignant via

$$p = \frac{1}{1+e^{-((\mathbf{w}^*)^T\mathbf{x}+b^*)}}.$$

Then we set our final predictor as follows

$$\widehat{y} = \begin{cases} 1 & \text{if } p > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

We randomly choose 69 patients from the total of 569 patients and consider them as test data and choose the remaining 500 as training data. In the following steps, we will use Matlab to apply gradient descent and stochastic gradient descent to solve the corresponding optimization problem.

1. Download *Xtrain*, *Ytrain*, *Xtest*, and *Ytest* from Moodle and import the files via Matlab.

2. Normalize the features of the data. Matlab Code

3. Implement classical gradient descent algorithm with constant step-size, and find the optimal solution. Report the loss function at the optimal solution and the final accuracy.

4. Implement stochastic gradient descent algorithm with constant step-size, and find the optimal solution. Does it converge? Report the loss function at the optimal solution and the final accuracy.

5. Implement stochastic gradient descent algorithm with diminishing step-size $0.1/i$, and find the optimal solution. Does it converge? Report the loss function at the optimal solution and the final accuracy.

6. Plot the accuracy of the training and testing data for the two implemented methods. The matlab code computing the accuracy at a specific $\mathbf{w}$:

   Hint: Try different initialization to get a good final accuracy.