Assignment 2

CMSC 726: Machine Learning
September 18th, 2018 **Name:**

1. Suppose we are minimizing a convex loss function $J(\theta)$ using gradient descent with the learning rate of α. Let θ_t be the parameter values at iteration t. Further, let the largest eigenvalue of the Hessian matrix be upper bounded by λ, i.e. $\lambda(\nabla^2 J(\theta)) \leq \lambda,\ \forall\theta$. Show that: $J(\theta_t) - J(\theta^*) \leq \frac{1}{2\alpha t} \left\|\theta^* - \theta_0\right\|^2$, where θ^* is the minimizer of $J(\theta)$.

2. Prove that the cross entropy loss function used in logistic regression is convex. I.e. prove

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{m} \left[y^{(i)} \log g(\boldsymbol{\theta}^T \mathbf{x}^{(i)}) + (1 - y^{(i)}) \log(1 - g(\boldsymbol{\theta}^T \mathbf{x}^{(i)})) \right]$$

   is convex where $g(z) = \frac{1}{1+e^{-z}}$ is the sigmoid (logistic) function.

3. Let $\hat{\gamma}^*, \mathbf{w}^*$ and $\mathbf{b}^*$ be optimizers of the following optimization problem:

$$\begin{aligned} \max_{\hat{\gamma},\mathbf{w},\mathbf{b}} \quad & \frac{\hat{\gamma}}{\|\mathbf{w}\|} \\ & y^{(i)}(\mathbf{w}^T \mathbf{x}^{(i)} + \mathbf{b}) \geq \hat{\gamma}, \quad 1 \leq i \leq m. \end{aligned}$$

   Moreover, let $\mathbf{w}^{**}$ and $\mathbf{b}^{**}$ be optimizers of the following modified optimization problem:

$$\begin{aligned} \max_{\mathbf{w},\mathbf{b}} \quad & \frac{1}{\|\mathbf{w}\|} \\ & y^{(i)}(\mathbf{w}^T \mathbf{x}^{(i)} + \mathbf{b}) \geq 1, \quad 1 \leq i \leq m. \end{aligned}$$

   Show that $\hat{\gamma}^* \mathbf{w}^{**} = \mathbf{w}^*$ and $\hat{\gamma}^* \mathbf{b}^{**} = \mathbf{b}^*$. How do planes of ${\mathbf{w}^*}^T \mathbf{x} + \mathbf{b}^* = 0$ and ${\mathbf{w}^{**}}^T \mathbf{x} + \mathbf{b}^{**} = 0$ relate to one another?

4. Problem 15.4 (in Section 15.8) from the text book.

5. (Programming Assignment) Let $\mathbf{x} \in \mathbb{R}^n$. The points with labels $y = 1$ are generated from a Gaussian distributing with mean μ_1 and covariance $\mathbf{I}$, whereas points with labels $y = 0$ are generated from another Gaussian distributing with mean $\mu_2 = -\mu_1$ and covariance $\mathbf{I}$. In this assignment, we want to use stochastic gradient descent (SGD) to find a logistic regression model between $\mathbf{x}$ and y. Write a Python code to do the following:

   (a) Let $n = 5$, $\mu_1 = -\mu_2 = 3[1, 1, 1, 1, 1]^T$. Generate $4,096$ i.i.d. samples from class 1 and another $4,096$ i.i.d. samples from class 0. These points compose your training set.

   (b) Use SGD with a batch size of 32 to estimate the model parameters. Plot the cross-entropy loss vs. the number of iterations.

(c) Generate a test set of $8,192$ points using the procedure explained in (a). Use the trained model to compute the loss on the test set. In addition, compute the accuracy (the percentage of correctly classified points) on the test set.

(d) Repeat parts (a)-(c) with $\mu_1 = -\mu_2 = \frac{1}{2}[1, 1, 1, 1, 1]^T$. How have the training error, test error, and accuracy been changed? Why?