Introduction to Machine Learning

Fall 2019

University of Science and Technology of China

Lecturer: Jie Wang | Homework 2

Posted: Oct. 05, 2019 | Due: Oct. 12, 2019

Name: Bowen Zhang | ID: PB17000215

Notice, to get the full credits, please present your solutions step by step.

Exercise 1: Lipschitz Continuity 10pts

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, and the gradient of f is Lipschitz continuous, i.e.,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \forall\, x, y \in \mathbb{R}^n,$$

where $L > 0$ is the Lipschitz constant. Please find the relation between L and the largest eigenvalue of $\nabla^2 f(x)$.

Solution: ■

Exercise 2: Gradient Descent for Convex Optimization Problems 20pts

Consider the following problem

$$\min_x f(x), \tag{1}$$

where f is convex and its gradient is Lipschitz continuous with constant $L > 0$. Assume that f can attain its minimum.

1. Show that the optimal set $\mathcal{C} = \{y : f(y) = \min_x f(x)\}$ is convex.

2. Suppose that $d(x, \mathcal{C}) = \inf_{z \in \mathcal{C}} \|x - z\|_2$. Consider the problem (1) and the sequence generated by the gradient descent algorithm. Show that $d(x_k, \mathcal{C}) \to 0$ as $k \to \infty$.

Solution: ■

Exercise 3: Gradient Descent for Strongly Convex Optimization Problems 50pts

A function f is strongly convex with parameter μ if $f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex.

1. Show that a continuously differentiable function f is strongly convex if and only if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2, \forall\, x, y \in \mathbb{R}^n$$

2. Suppose that f is twice differentiable. Please find the relation between μ and the smallest eigenvalue of $\nabla^2 f(x)$.

Consider the following problem

$$\min_x f(x), \tag{2}$$

where f is strongly convex with convexity parameter $\mu > 0$ and its gradient is Lipschitz continuous with constant $L > 0$.

3. Show that the problem (2) admits a unique solution.

4. Show that

$$f(y) \geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_2^2, \forall\, x, y.$$

5. Consider the problem (2) and the sequence generated by the gradient descent algorithm. Suppose that x^* is the solution to the problem 2. Show that

$$f(x_k) - f(x^*) \leq (1 - \mu\alpha(2 - L\alpha))^k (f(x_0) - f(x^*)).$$

   Find the range of α such that the function values $f(x_k)$ converge linearly to $f(x^*)$.

Solution: ■

Exercise 4: Programming Exercise 20pts

We provide you with a data set, where the number of samples n is 16087 and the number of features d is 10013. Suppose that $\mathbf{X} \in \mathbb{R}^{n\times d}$ is the input feature matrix and $\mathbf{y} \in \mathbb{R}^n$ is the corresponding response vector. We use the linear model to fit the data, and thus we can formulate the optimization problem as

$$\arg\min_{\mathbf{w}} \frac{1}{n}\|\mathbf{y} - \bar{\mathbf{X}}\mathbf{w}\|_2^2, \tag{3}$$

where $\bar{\mathbf{X}} = (\mathbf{1}, \mathbf{X}) \in \mathbb{R}^{n\times(d+1)}$ and $\mathbf{w} = (w_0, w_1, \ldots, w_n)^\top \in \mathbb{R}^{d+1}$. Finish the following exercise by programming. You can use your favorite programming language.

1. Use the closed form solution to solve the problem (3), and get the solution x_1^*.

2. Use the gradient descent algorithm to solve the problem (3). Stop the iteration when $\|x_k - x_1^*\|_2 < 10^{-8}$. Plot the function values $f(x_k)$ versus the iteration step k.

Compare the time cost of the above two approaches.