Introduction to Machine Learning

Spring 2020

University of Science and Technology of China

Lecturer: Jie Wang | Homework 3

Posted: Mar. 26, 2020 | Due: Apr. 2, 2020

Name: San Zhang | ID: PBXXXXXXXX

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

Exercise 1: Local minima of convex function 10pts

Let $f : \mathbb{R}^n \to (-\infty, +\infty]$ be a convex function. Consider the following problem

$$\min_{\mathbf{x} \in D} f(\mathbf{x}), \tag{1}$$

where $D \subseteq \mathbf{dom}\ f \subseteq \mathbb{R}^n$ is a convex set. We assume that the problem (1) is solvable. Then, if $\mathbf{x}$ is a local optimum, it is also a global optimum.

1. Show the above result by NOT using contradiction.
2. Show the above result by contradiction.

Solution: ■

Exercise 2: Gradient Descent for Convex Optimization Problems 30pts

Consider the following problem

$$\min_{\mathbf{x}} f(\mathbf{x}), \tag{2}$$

where $f : \mathbb{R}^n \to (-\infty, +\infty)$ is convex, first order continuously differentiable, and its gradient is Lipschitz continuous with constant $L > 0$. Suppose that f can attain its minimum.

1. Show that the optimal set $\mathcal{C} = \{\mathbf{y} : f(\mathbf{y}) = \min_{\mathbf{x}} f(\mathbf{x})\}$ is closed and convex.
2. We define the distance from $\mathbf{x}$ to the optimal set $\mathcal{C}$ by $d(\mathbf{x}, \mathcal{C}) = \inf_{\mathbf{z}\in\mathcal{C}} \|\mathbf{x}-\mathbf{z}\|_2$. Prove that there exists $\mathbf{z} \in \mathcal{C}$ such that $d(\mathbf{x}, \mathcal{C}) = \|\mathbf{x} - \mathbf{z}\|_2$ for any fixed $\mathbf{x}$.
3. Consider the problem (2) and the sequence generated by the gradient descent algorithm with the step size $\alpha \in (0, \frac{2}{L})$. Show that $d(\mathbf{x}_k, \mathcal{C}) \to 0$ as $k \to \infty$.
4. Consider the problem (2) and the sequence generated by the gradient descent algorithm. Given $\alpha > 0$, suppose that $\{\mathbf{x}_k\}$ is convergent. Show that $d(\mathbf{x}_k, \mathcal{C}) \to 0$ as $k \to \infty$.

Solution: ■

Exercise 3: Projection operator 30pts

Consider the following problem

$$\min_{x \in D} f(x), \tag{3}$$

where $f : \mathbb{R}^n \to (-\infty, +\infty]$ is continuously differentiable and strongly convex with convexity parameter $\mu > 0$. We assume that $D \subseteq \mathbf{dom}\ f$ is closed and nonempty.

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

For a nonempty, closed, and convex set $C \subseteq \mathbb{R}^n$, the projection of an arbitrary point $x \in \mathbb{R}^n$ onto C is defined by

$$\mathbf{P}_C(\mathbf{x}) = \mathop{\mathbf{argmin}}_{\mathbf{z} \in C} \|\mathbf{x} - \mathbf{z}\|_2.$$

We call $\mathbf{P}_C(\mathbf{x})$ the projection of the point $\mathbf{x}$ onto the convex set C.

2. Show that $\mathbf{P}_C(\mathbf{x})$ always exists and is unique.
3. Show that $\mathbf{y} = \mathbf{P}_C(\mathbf{x})$ if and only if $\mathbf{y} \in C$ and

$$\langle \mathbf{z} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle \leq 0, \forall\, \mathbf{z} \in C.$$

4. Show that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\|\mathbf{P}_C(\mathbf{x}) - \mathbf{P}_C(\mathbf{y})\|_2 \leq \|\mathbf{x} - \mathbf{y}\|_2.$$

Solution: ■

Exercise 4: Projected Gradient Descent 30pts

Consider the problem (3). We assume that the feasible set $D \subseteq \mathbf{dom}\ f$ is closed and nonempty and the gradient of f is Lipschitz with constant $L > 0$. A commonly used approach to solve the constrained optimization problem (3) is the so-called *projected gradient descent*, in which each iteration improves the current estimation $\mathbf{x}_k$ of the optimum by

$$\mathbf{x}_{k+1} = \mathbf{P}_D(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)),$$

where $\alpha > 0$ is the step size.

1. Show that

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

2. Consider the problem (3) and the sequence generated by the *projected gradient descent* algorithm. Suppose that $\mathbf{x}^*$ is the solution to the problem (3).
   (a) Find the range of α such that the function values $f(\mathbf{x}_k)$ converge linearly to $f(\mathbf{x}^*)$.
   (b) When does the (projected) gradient descent always achieve the optimal solution in one iteration wherever the intial point $\mathbf{x}_0$ is?

Solution: ■

Exercise 5: 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

$$\arg\min_{\mathbf{w}} f(\mathbf{w}) = \frac{1}{n}\|\mathbf{y} - \bar{\mathbf{X}}\mathbf{w}\|_2^2, \tag{4}$$

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

1. Normalize the columns $\mathbf{x}_i$ of $\mathbf{X}$ $(1 \le i \le d)$ as follows:

$$z_{ij} \leftarrow \frac{x_{ij} - \min(\mathbf{x}_i)}{\max(\mathbf{x}_i) - \min(\mathbf{x}_i)},$$

   where x_{ij} denotes the jth entry of $\mathbf{x}_i$. Similarly, z_{ij} denotes the jth entry of $\mathbf{z}_i \in \mathbb{R}^n$ and $\bar{\mathbf{Z}} = (\mathbf{1}, \mathbf{z}_1, \ldots, \mathbf{z}_d) \in \mathbb{R}^{n\times(d+1)}$. The problem (4) becomes

$$\arg\min_{\mathbf{u}} g(\mathbf{u}) = \frac{1}{n}\|\mathbf{y} - \bar{\mathbf{Z}}\mathbf{u}\|_2^2, \tag{5}$$

   where $\mathbf{u} = (u_0, u_1, \ldots, u_d)^\top \in \mathbb{R}^{d+1}$.

2. Please find the Lipschitz constants of $f(\mathbf{w})$ and $g(\mathbf{u})$ respectively.

3. Use the closed form solution to solve the problem (5), and get the solution $\mathbf{u}^*$ and the corresponding optimal value $g^* = g(\mathbf{u}^*)$.

4. Use the gradient descent algorithm to solve the problem (5). Stop the iteration until $|g(\mathbf{u}_k) - g^*| < 0.01$. Please use $\mathbf{u}_k$ to recover $\mathbf{w}_k$ and plot $g(\mathbf{u}_k)$ and $f(\mathbf{w}_k)$ versus the iteration step k.

5. Compare the time cost of the two approaches in 3. and 4.

Solution: ■