Amir Hossein Eyvazkhani
1747696

Ex8

Task 1)

@ In Gradient Descent, the algorithm simultaneously updates all the parameter in each step, and moves toward the direction of the negative gradient. However, in coordinate descent, the algorithm updates one parameter per time. By that I mean in each iteration, a parameter is selected and the algorithm updates only that Parameter based on the negative gradient of the cost function with respect to that parameter.

Task 1) (b)

Coordinate descent (X, y, k, β_0) : $X \in R^{n \times m}$, $y \in R^n$

for $k = 1 \dots K$:

for $i = 0 \dots m-1$:

$$\beta_i = X_i^T (y_i - X_i \beta_{-i})$$

if (coverage)
return β

$\beta_0 = (1,1,1)$, $X = \begin{pmatrix} 1 & 1.5 & 2 \\ 1 & 3 & 2.5 \\ 1 & 4.5 & 3 \end{pmatrix}$, $y = (10 \quad 15.5 \quad 21)$

$k = 1$:

$$\beta_0 = (1,1,1)\left[\begin{pmatrix} 10 \\ 15.5 \\ 21 \end{pmatrix} - \begin{pmatrix} 1.5 & 2 \\ 3 & 2.5 \\ 4.5 & 3 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right] = (1,1,1)\begin{pmatrix} 6.5 \\ 10 \\ 13.5 \end{pmatrix} = 30$$

$$\beta_1 = (1.5, 3, 4.5)\left[\begin{pmatrix} 10 \\ 15.5 \\ 21 \end{pmatrix} - \begin{pmatrix} 1 & 2 \\ 1 & 2.5 \\ 1 & 3 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right] = (1.5,3,4.5)\begin{pmatrix} 7 \\ 12 \\ 17 \end{pmatrix} = 123$$

$$\beta_2 = (2 \quad 2.5 \quad 3)\left[\begin{pmatrix} 10 \\ 15.5 \\ 21 \end{pmatrix} - \begin{pmatrix} 1 & 1.5 \\ 1 & 3 \\ 1 & 4.5 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right] = (2 \quad 2.5 \quad 3)\begin{pmatrix} 7.5 \\ 11.5 \\ 15.5 \end{pmatrix} = 90.25$$

$$\text{Loss} = \sum_{i=1}^{3} (x_i \beta^T - y_i)^2 = (385)^2 + (609.1)^2 + (833.25)^2 = 1,213,569.82$$

$k = 2$:

$\beta_0 = -1737.375$

$\beta_1 = -2280$

$\beta_2 = -3055.75$

Loss $= 839,403,977.53$

Task 2)

(a) $g(x) = |x_1 x_2| + 0.1(x_1 + x_2)$

if we want to update x_1 at a step, we consider x_2 like a constant,

So we would have: $g(x_1) = \underbrace{|x_2|}_{\text{constants}} x_1 + 0.1(x_2)\, x_1$

because of the absolute function, g is not differentiable at 0. The same goes to x_2. It means that the optimum choice in the step above is not to move! Therefore the algorithm stops.

(b) $L(x) = f(x) + \|x\|_1 = f(x) + |x|$

$$\frac{\partial L}{\partial x} = \frac{df}{dx} + \frac{d(|x|)}{dx} = 0$$

Since $|x|$ is not differentiable at 0, we use subdifferentials of it from the last question of last sheet:

$$\frac{\partial L}{\partial x} = \frac{df}{dx} + g$$

$$g = \begin{cases} 1 & x_i > 0 \\ -1 & x_i < 0 \\ [-1,1] & x_i = 0 \end{cases}$$

this shows that the algorithm can reach the optimal solution x^*.