DSE 220: Machine learning

Worksheet 7 — Solutions

1. We want to find the $z \in \mathbb{R}^d$ that minimizes
$$L(z) = \sum_{i=1}^{n} \|x^{(i)} - z\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{d}(x_j^{(i)} - z_j)^2.$$
Taking partial derivatives, we have
$$\frac{\partial L}{\partial z_j} = \sum_{i=1}^{n} -2(x_j^{(i)} - z_j) = 2nz_j - 2\sum_{i=1}^{n} x_j^{(i)}.$$
Thus
$$\nabla L(z) = 2nz - 2\sum_{i=1}^{n} x^{(i)}.$$
Setting $\nabla L(z) = 0$ and solving for z, gives us
$$z^* = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}.$$
(Aside: To confirm that z^* minimizes L, we can check to see that L is convex. Taking second partial derivatives, we have
$$\frac{\partial^2 L}{\partial z_j \partial z_k} = \begin{cases} 2n & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases}$$
Thus the Hessian of L is a diagonal matrix with every diagonal entry set to $2n$. This is positive semidefinite since $z^T H z = 2n\|z\|^2 \geq 0$ for all $z \in \mathbb{R}^d$. Therefore L is convex and z^* minimizes L.)

2. The loss function is
$$L(w) = \sum_{i=1}^{n}(w \cdot x^{(i)})^2 + \frac{c}{2}\|w\|^2.$$
   (a) $\nabla L(w) = \sum_i x^{(i)} + cw$.
   (b) Setting the derivative to zero, we get $w = -(1/c)\sum_i x^{(i)}$.

3. $L(w) = w_1^2 + 2w_2^2 + w_3^2 - 2w_3w_4 + w_4^2 + 2w_1 - 4w_2 + 4$
   (a) The derivative is
$$\nabla L(w) = (2w_1 + 2, 4w_2 - 4, 2w_3 - 2w_4, -2w_3 + 2w_4)$$
   (b) The derivative at $w = (0, 0, 0, 0)$ is $(2, -4, 0, 0)$. Thus the update at this point is:
$$w_{new} = w - \eta\nabla L(w) = (0, 0, 0, 0) - \eta(2, -4, 0, 0) = (-2\eta, 4\eta, 0, 0).$$

(a) To find the minimum value of $L(w)$, we will equate $\nabla L(w)$ to zero:

- $2w_1 + 2 = 0 \implies w_1 = -1$
- $4w_2 - 4 = 0 \implies w_2 = 1$
- $2w_3 - 2w_4 = 0 \implies w_3 = w_4$

The function is minimized at any point of the form $(-1, 1, x, x)$.

(c) No, there is not a unique solution.

4. *Local search for ridge regression.* We are interested in analyzing

$$L(w) = \sum_{i=1}^{n} (y^{(i)} - w \cdot x^{(i)})^2 + \lambda \|w\|^2.$$

(a) To compute $\nabla L(w)$, we compute partial derivatives.

$$\frac{\partial L}{\partial w_j} = \left(\sum_{i=1}^{n} -2x_j^{(i)} (y^{(i)} - w \cdot x^{(i)}) \right) + 2\lambda w_j$$

Thus

$$\nabla L(w) = -2 \sum_{i=1}^{n} (y^{(i)} - w \cdot x^{(i)}) x^{(i)} + 2\lambda w.$$

(b) The update for gradient descent with step size η looks like

$$\begin{aligned} w_{t+1} &= w_t - \eta \nabla L(w_t) \\ &= w_t(1 - 2\eta\lambda) + 2\eta \sum_{i=1}^{n} (y^{(i)} - w_t \cdot x^{(i)}) x^{(i)} \end{aligned}$$

(c) The update for stochastic gradient descent looks like the following.

$$w_{t+1} = w_t(1 - 2\eta\lambda) + 2\eta (y^{(i_t)} - w_t \cdot x^{(i_t)}) x^{(i_t)}$$

where i_t is the index chosen at time t.