DSC 255: Machine learning

Week 5 — Solutions

Worksheet

1. $L(w) = w_1^2 + 2w_2^2 + w_3^2 - 2w_3w_4 + w_4^2 + 2w_1 - 4w_2 + 4$

   (a) We have

   $$\frac{dL}{dw_1} = 2w_1 + 2, \ \frac{dL}{dw_2} = 4w_2 - 4, \ \frac{dL}{dw_3} = 2w_3 - 2w_4, \ \frac{dL}{dw_4} = -2w_3 + 2w_4.$$

   (b) Pulling together the derivatives from (a) into a vector, we get

   $$\nabla L(w) = \begin{bmatrix} 2w_1 + 2 \\ 4w_2 - 4 \\ 2w_3 - 2w_4 \\ -2w_3 + 2w_4 \end{bmatrix}.$$

   (c) The derivative at $w = (0, 0, 0, 0)$ is $(2, -4, 0, 0)$. Thus the update at this point is:

   $$w_{new} = w - \eta \nabla L(w) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} - \eta \begin{bmatrix} 2 \\ -4 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -0.2 \\ 0.4 \\ 0 \\ 0 \end{bmatrix}.$$

   (d) 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 $w = (-1, 1, a, a)$. The resulting value is $L(w) = 1 + 2 + a^2 - 2a^2 + a^2 - 2 - 4 + 4 = 1$.

   (e) No, there is not a unique solution; the full set of solutions is given in (d).

2. The loss function is

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

   (a) $dL/dw_j = \sum_{i=1}^{n} x_j^{(i)} + cw_j$.

   (b) $\nabla L(w) = \sum_i x^{(i)} + cw$.

   (c) Setting the derivative to zero, we get $w = -(1/c) \sum_i x^{(i)}$.

3. *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.

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

   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) Here is a stochastic gradient descent algorithm:
   - Initialize $w = 0$
   - Repeatedly cycle through the data; for each point (x, y):
     - Update $w = w(1 - 2\eta\lambda) + 2\eta(y - w \cdot x)x$

4. *Convexity.*

   (a) $f''(x) = 2$: convex

   (b) $f''(x) = -2$: concave

   (c) $f''(x) = 2$: convex

   (d) $f''(x) = 0$: both convex and concave

   (e) $f''(x) = 6x$ and $x \in \mathbb{R}$: neither convex nor concave

   (f) $f''(x) = 12x^2$ and $x \in \mathbb{R}$: convex

   (g) $f''(x) = -\frac{1}{x^2}$ and $x \in \mathbb{R}$: concave

Programming lab