CSE 251A: Homework 2 Solutions

1. *Regression with one predictor variable*

   (a) Suppose we predict a value v. Then the MSE is $M = \sum_{i=1}^{4}(y^{(i)} - v)^2$. Taking the derivative with respect to v:

   $$\frac{dM}{dv} = 2 \cdot \sum_{i=1}^{4}(y^{(i)} - v) \cdot (-1)$$

   This derivative is 0 when $v = (1/4)\sum_{i=1}^{4} y^{(i)}$; the double derivative is also positive at this v. Therefore, the MSE is minimized at the mean of the $y^{(i)}$'s – namely, at $v = (1/4)\sum_{i=1}^{4} y^{(i)} = (1+3+4+6)/4 = 3.5$. The MSE of this prediction is exactly the variance of the y-values, namely:

   $$\text{MSE} = \frac{(1-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (6-3.5)^2}{4} = 3.25.$$

   (b) If we simply predict x, the MSE is

   $$\frac{1}{4}\sum_{i=1}^{4}(y^{(i)} - x^{(i)})^2 = \frac{1}{4}\left((1-1)^2 + (1-3)^2 + (4-4)^2 + (4-6)^2\right) = 2.$$

   (c) We saw in class that the MSE is minimized by choosing

   $$a = \frac{\sum_i (y^{(i)} - \overline{y})(x^{(i)} - \overline{x})}{\sum_i (x^{(i)} - \overline{x})^2}$$
   $$b = \overline{y} - a\overline{x}$$

   where $\overline{x}$ and $\overline{y}$ are the mean values of x and y, respectively. This works out to $a = 1, b = 1$; and thus the prediction on x is simply $x + 1$. The MSE of this predictor is:

   $$\frac{1}{4}\left(1^2 + 1^2 + 1^2 + 1^2\right) = 1.$$

2. *Lines through the origin*

   (a) The loss function is

   $$L(a) = \sum_{i=1}^{n}(y^{(i)} - ax^{(i)})^2$$

   (b) The derivative of this function is:

   $$\frac{dL}{da} = -2\sum_{i=1}^{n}(y^{(i)} - ax^{(i)})x^{(i)}.$$

   Setting this to zero yields

   $$a = \frac{\sum_{i=1}^{n} x^{(i)} y^{(i)}}{\sum_{i=1}^{n} {x^{(i)}}^2}.$$

3. (a) Suppose the best predictor is $\sum_{i=1}^{5} a_i x_i + b$. Then the expected MSE is:

$$M = \mathbb{E}[(\sum_{i=1}^{5} a_i x_i + b - \sum_{i=1}^{10} x_i)^2]$$

Taking the partial derivative with respect to each a_i and b and setting them to zero, we get:

$$\frac{\partial M}{\partial a_i} = \mathbb{E}[((\sum_{j=1}^{5} a_j x_j + b - \sum_{j=1}^{10} x_j) \cdot x_i] = 0, i = 1, \ldots, 5 \quad (1)$$

$$\frac{\partial M}{\partial b} = \mathbb{E}[((\sum_{i=1}^{5} a_i x_i + b - \sum_{i=1}^{10} x_i) \cdot 1] = 0 \quad (2)$$

Simplifying equation 2, we get that:

$$\sum_{i=1}^{5} a_i \mathbb{E}[x_i] + b = \sum_{i=1}^{10} \mathbb{E}[x_i] \quad (3)$$

Plugging in the means $\mathbb{E}[x_i] = 1$ this gives

$$\sum_{i=1}^{5} a_i + b = 10 \quad (4)$$

Simplifying equation 1, we get that:

$$\sum_{j=1}^{5} a_i \mathbb{E}[x_i x_j] + b\mathbb{E}[x_i] = \sum_{j=1}^{10} \mathbb{E}[x_j x_i] \quad (5)$$

Since each x_i is independent of x_j for $i \neq j$, $\mathbb{E}[x_i x_j] = \mathbb{E}[x_i]\mathbb{E}[x_j] = 1$ for $i \neq j$ and $\mathbb{E}[x_i^2] = \mathbb{E}[(x_i - \mathbb{E}[x_i])^2] + \mathbb{E}[x_i]^2 = 2$. Plugging this in, we get:

$$a_i + b + \sum_{i=1}^{5} a_i = 11 \quad (6)$$

Subtracting (6) - (4), we get $a_i = 1$; plugging this into (4) gives $b = 5$. The best predictor is thus $\hat{y} = x_1 + x_2 + x_3 + x_4 + x_5 + 5$: to minimize the fluctuations due to $x_6 + \cdots + x_{10}$, we use its mean.

(b) The MSE is:

$$\mathbb{E}[(5 - x_6 - x_7 - \ldots - x_{10})^2] = \mathbb{E}[((1 - x_6) + (1 - x_7) + \ldots + (1 - x_{10}))^2]$$

Since the x_i's are independent, this is equal to

$$\sum_{i=6}^{10} \mathbb{E}[(1 - x_i)^2] = \sum_{i=6}^{10} \mathbb{E}[(x_i - \mathbb{E}[x_i])^2] = 5$$

4. The loss induced by a linear predictor $w \cdot x + b$ is

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

5. Define

$$X = \begin{bmatrix} \leftarrow x^{(1)} \rightarrow \\ \leftarrow x^{(2)} \rightarrow \\ \vdots \\ \leftarrow x^{(n)} \rightarrow \end{bmatrix}$$

$$XX^T = \begin{bmatrix} x^{(1)} \cdot x^{(1)} & x^{(1)} \cdot x^{(2)} & \dots & x^{(1)} \cdot x^{(n)} \\ x^{(2)} \cdot x^{(1)} & x^{(2)} \cdot x^{(2)} & \dots & x^{(2)} \cdot x^{(n)} \\ x^{(n)} \cdot x^{(1)} & x^{(n)} \cdot x^{(2)} & \dots & x^{(n)} \cdot x^{(n)} \end{bmatrix}$$

6. With vocabulary $V = \{is, flower, rose, a, an\}$, the bag-of-words representation of the sentence "a rose is a rose is a rose" is $(2, 0, 3, 3, 0)$.

7. 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

$$z^* = \frac{1}{n} \sum_{i=1}^{n} x^{(i)}.$$

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

   (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:

   $$w_{new} = w - \eta \nabla L(w) = (0, 0, 0, 0) - \eta(2, -4, 0, 0) = (-2\eta, 4\eta, 0, 0).$$

   (c) 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 $(-1, 1, x, x)$.

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

9. 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.