CSE 250B: Homework 2 Solutions

1. *Regression with one predictor variable*

   (a) We will predict the mean of the y-values: $\widehat{y} = (1+3+4+6)/4 = 3.5$. 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. The loss induced by a linear predictor $w \cdot x + b$ is

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

4. Define

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

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

5. *Discovering relevant features in regression.*

   (a) A sensible strategy is to do linear regression using the Lasso, and to choose a regularization constant λ that yields roughly 10 non-zero coefficients.

   (b) The smallest value of λ we tried that gave nonzero coefficients for 10 features is 0.4. This yielded the following features (numbering starting at 1): $2, 3, 5, 7, 11, 13, 17, 19, 23, 29$.

6. 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

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

7. *Minimizing absolute loss.* Pick any value v that is not identical to one of the data points $x^{(i)}$. Suppose that k of the data points are less than v while the remaining $n - k$ are greater than v. Then, a small change $v \leftarrow v + \epsilon$, where ϵ may be positive or negative, will change the loss

$$L(v) = \sum_{i=1}^{n} |x^{(i)} - v|$$

   by $+k\epsilon - (n-k)\epsilon = (2k - n)\epsilon$. This means that as long as $k \neq n/2$, it is always possible to change v in a way that reduces the loss. It follows that the minimum of $L(v)$ is attained at values v for which $k = n/2$, that is, when v is the *median* of the data.

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.

10. *Closed form solution for ridge regression.*

(a) The loss function can be rewritten thus:

$$\begin{aligned} L(w) &= \sum_{i=1}^{n} (y^{(i)} - w \cdot x^{(i)})^2 \ + \ \lambda \|w\|^2 \\ &= \|y - Xw\|^2 + \lambda \|w\|^2 \\ &= y^T y - 2w^T X^T y + w^T X^T X w + \lambda w^T w \end{aligned}$$

(b) Taking the derivative of the loss, we get:

$$\nabla L(w) = -2X^T y + 2X^T X w + 2\lambda w = -2X^T y + 2(X^T X + \lambda I) w.$$

Setting this to zero yields $w = (X^T X + \lambda I)^{-1} (X^T y)$.

11. *A case when Lasso finds a sparse solution.*

(a) For any w, we can write

$$\begin{aligned} LS(w) = \|y - Xw\|^2 \ &= \ \|y - Xw^* - X(w - w^*)\|^2 \\ &= \|y - Xw^*\|^2 + \|X(w - w^*)\|^2 - 2(w - w^*) X^T (y - Xw^*) \\ &= LS(w^*) + (w - w^*) X^T X (w - w^*) - 2(w - w^*)(X^T y - X^T X w^*). \end{aligned}$$

The last term is zero since $X^T X w^* = X^T X (X^T X)^{-1} X^T y = X^T y$.

(b) The simplified Lasso problem is

$$\min_{\|w\|_1 \leq 1} \|w - w^*\|^2$$

To solve this problem, imagine growing an ℓ_2 ball around the least-squares w^* until it touches the ℓ_1 unit ball. The point of first contact is the solution w.

If w^* is (say) $(2, 2)$, then this point will be $(1/2, 1/2)$, which is not sparse. If w^* is $(1, 3)$, then this point will be $(0, 1)$, which is sparse.

12. *Form of the squashing function.*

$$\begin{aligned}
\Pr(y = 1|x) = \frac{\Pr(y = 1, x)}{\Pr(x)} &= \frac{\exp(-\|x - \mu_1\|^2/2\sigma^2)}{\exp(-\|x - \mu_1\|^2/2\sigma^2) + \exp(-\|x - \mu_2\|^2/2\sigma^2)} \\
&= \frac{1}{1 + \exp((\|x - \mu_1\|^2 - \|x - \mu_2\|^2)/2\sigma^2)} \\
&= \frac{1}{1 + \exp(2x \cdot (\mu_2 - \mu_1) + \|\mu_1\|^2 - \|\mu_2\|^2)}.
\end{aligned}$$

This is of the form $s(z)$ where $s(\cdot)$ is the squashing function and z is linear in x.