DSE 220: Machine learning

Worksheet 5 — 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. *Optimality of the mean.*

   (a) $dL/ds = -2(x_1 + \cdots + x_n)/n + 2s$.

(b) Setting $dL/ds = 0$, we get $s = (x_1 + \cdots + x_n)/n$.

4. *Optimality of the median.*

   (a) 14

   (b) $152/9 = 16.8888$

   (c) $101/9 = 11.2222$.

   (d) 5.

5. We would write the loss induced by a linear predictor $w \cdot x + b$ as

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

6. *Writing expressions in matrix-vector form.*

   (a) $(1/n)\mathbf{1}^T y$

   (b) XX^T

   (c) $(1/n)X^T\mathbf{1}$

   (d) $(1/n)X^T X$

7. $b = c_o$ and $w = (c_1 - c_o, c_2 - c_o, \ldots, c_d - c_o)$.

8. (a) When $\lambda = 0$, we get the least-squares solution. As we have seen, this has loss zero, so $L(0) = 0$.

   (b) When λ increases, there is a greater penalty on $\|w\|$. Therefore $\|w_\lambda\|$ decreases.

   (c) When λ increases, and the penalty on $\|w\|$ increases, we get smaller w and larger squared loss. Therefore $L(\lambda)$ increases.

   (d) When $\lambda \to \infty$, we get $w \to 0$. For $w = 0$, the loss function of ridge regression simplifies dramatically and becomes

$$\sum_{i=0}^{d} (c_i - b)^2.$$

   This is minimized by setting b to the average of $c_o, c_1, \ldots, c_d$. The resulting loss is thus $d + 1$ times the variance of $c_o, \ldots, c_d$.

9. *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) First value of λ which gave nonzero coefficients only for 10 features is 0.4. This yielded the following features (numbering starting at 1): $2, 3, 5, 7, 11, 13, 17, 19, 23, 27$.