CSE 151: Homework 5 Solutions

1 Regression and logistic regression

1. *Writing expressions in matrix-vector form.*
   (a) $(1/n)\mathbf{1}^T y$
   (b) XX^T
   (c) $(1/n)X^T\mathbf{1}$
   (d) $(1/n)X^TX$

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

3. (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 $c_o, \ldots, c_d$.

4. *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): $2, 3, 5, 7, 11, 13, 17, 19, 23, 27$.

5. *Inherent uncertainty.* This is somewhat subjective, but (b), (d) seem pretty clear-cut cases where perfect predictions are not possible.

6. *Logistic regression.* Since
   $$\Pr(y = 1|x) = \frac{1}{1 + e^{-(w \cdot x + b)}},$$
   we can rearrange terms to get
   $$w \cdot x + b = \ln \frac{\Pr(y = 1|x)}{1 - \Pr(y = 1|x)}$$
   (a) $w \cdot x + b = \ln 1 = 0$
   (b) $w \cdot x + b = \ln 3$
   (c) $w \cdot x + b = -\ln 3$

7. If the vocabulary is $V =$ (is, flower, rose, an, a), the representation of the sentence "a rose is a rose is a rose" is $(2, 0, 3, 0, 3)$.

8. As the margin m increases, $f(m)$, the fraction of test points with margin $\geq m$, will decrease. We would expect $e(m)$, the error rate on points with margin $\geq m$, to decrease, but this might not happen.

2 Unconstrained optimization

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

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

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

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

   (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.