Homework 1

Andrea Sofia Vallejo Budziszewski

November 2023

Problem 1

For a given $(d+1) \times 1$ weight vector $\mathbf{w}$ and a training set $S = (X, \mathbf{y})$, we want to show that the training loss for linear regression can be expressed as

$$L_S(\mathbf{w}) = \frac{1}{m}\left(\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{w}^T\mathbf{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{y}\right), \tag{1}$$

where $\mathbf{X}$ is an $m \times (d+1)$ matrix storing the input data, and $\mathbf{y}$ is an $m \times 1$ vector storing the output data. This expression for the loss is a quadratic form in $\mathbf{w}$.

The training loss for linear regression is typically expressed as the Mean Squared Error (MSE) loss, which is defined as

$$L_S(\mathbf{w}) = \frac{1}{2m}\sum_{i=1}^{m}(y_i - \mathbf{w}^T\mathbf{x}_i)^2, \tag{2}$$

where $\mathbf{w}$ is the weight vector of shape $(d+1) \times 1$, $S = (X, \mathbf{y})$ is the training set, m is the number of training examples, $\mathbf{x}_i$ is the i-th row of the matrix $\mathbf{X}$, and y_i is the i-th element of the vector $\mathbf{y}$.

Now, let's express $L_S(\mathbf{w})$ in matrix form to simplify the derivation:

$$L_S(\mathbf{w}) = \frac{1}{2m}(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}). \tag{3}$$

Expanding this expression, we get

$$\begin{aligned} L_S(\mathbf{w}) &= \frac{1}{2m}\left(\mathbf{y}^T - \mathbf{w}^T\mathbf{X}^T\right)(\mathbf{y} - \mathbf{X}\mathbf{w}) && (4)\\ &= \frac{1}{2m}\left(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{w} - \mathbf{w}^T\mathbf{X}^T\mathbf{y} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}\right). && (5) \end{aligned}$$

We can now notice that $\mathbf{y}^T\mathbf{X}\mathbf{w}$ and $\mathbf{w}^T\mathbf{X}^T\mathbf{y}$ are scalars and that they are equal to their own transposes:

$$\mathbf{y}^T\mathbf{X}\mathbf{w} = (\mathbf{y}^T\mathbf{X}\mathbf{w})^T = \mathbf{w}^T\mathbf{X}^T\mathbf{y}. \tag{6}$$

With this simplification, we have

$$L_S(\mathbf{w}) = \frac{1}{2m}\left(\mathbf{y}^T\mathbf{y} - 2\mathbf{w}^T\mathbf{X}^T\mathbf{y} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}\right). \tag{7}$$

Now, removing the constant factor $\frac{1}{2m}$ gives us the expression for the training loss as a quadratic form in $\mathbf{w}$:

$$L_S(\mathbf{w}) = \frac{1}{m}\left(\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{w}^T\mathbf{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{y}\right). \tag{8}$$

This concludes the derivation of the training loss for linear regression as a quadratic form in the weight vector $\mathbf{w}$, as expressed in equation (1).

Problem 2

Assuming that $X^T X$ is invertible, we want to show that the training loss $LS(w)$ can be written in the centered form as follows:

$$LS(w) = \frac{1}{m}\left((w-(X^TX)^{-1}X^Ty)^T X^T X(w-(X^TX)^{-1}X^Ty) + y^T(I - X(X^TX)^{-1}X^T)y\right) \tag{9}$$

Here, w_{lin} is the weight vector that minimizes the training loss $LS(w)$.
To find w_{lin}, we set the gradient of $LS(w)$ with respect to w to zero:

$$\nabla LS(w) = 0 \tag{10}$$

The training loss associated with w_{lin} is given by the value of $LS(w)$ when evaluated at w_{lin}:

$$L_{\text{lin}} = LS(w_{\text{lin}}) \tag{11}$$

Notably, for any symmetric matrix A, $A^T = A$, and for a positive definite matrix A, $v^T A v > 0$ for any non-zero vector v.

BONUS

Ridge regression minimizes the regularized objective:

$$L_{\text{ridge}}(w) = L_S(w) + \lambda\|w\|_2^2 \tag{12}$$

where $L_S(w)$ is the original training loss, λ is a positive scalar, and $\|\cdot\|_2$ is the Euclidean norm.
We previously derived the expression for $L_S(w)$ as follows:

$$L_S(w) = \frac{1}{m}\left(w^T X^T X w - 2w^T X^T y + y^T y\right) \tag{13}$$

Now, let's expand the Ridge regression objective:

$$L_{\text{ridge}}(w) = \frac{1}{m}\left(w^T X^T X w - 2w^T X^T y + y^T y\right) + \lambda\|w\|_2^2 \tag{14}$$

Expand the Euclidean norm $\|w\|_2^2$ as $w^T w$, which is the same as $w^T I w$, where I is the identity matrix. Then, add and subtract $\lambda y^T y$ for later convenience:

$$L_{\text{ridge}}(w) = \frac{1}{m}\left(w^T X^T X w - 2w^T X^T y + y^T y + \lambda w^T I w - \lambda w^T I w + \lambda y^T y - \lambda y^T y\right) \tag{15}$$

Rearrange the terms:

$$\begin{aligned} L_{\text{ridge}}(w) = \frac{1}{m}\Big(&(w-(X^TX+\lambda I)^{-1}X^Ty)^T(X^TX+\lambda I) \\ &(w-(X^TX+\lambda I)^{-1}X^Ty) + (y-\lambda(X^TX+\lambda I)^{-1}y)^T(y-\lambda(X^TX+\lambda I)^{-1}y)\Big) \end{aligned} \tag{16}$$

We have now expressed $L_{\text{ridge}}(w)$ in centered form, where w_{ridge} that minimizes this objective can be found by setting the gradient of $L_{\text{ridge}}(w)$ with respect to w to zero:

$$\nabla L_{\text{ridge}}(w) = 0 \tag{17}$$

The training loss associated with w_{ridge} is the value of $L_{\text{ridge}}(w)$ when evaluated at w_{ridge}:

$$L_{\text{ridge}} = L_{\text{ridge}}(w_{\text{ridge}}) \tag{18}$$

Similar to the previous case, we can also conclude that L_{ridge} is non-negative since $(X^T X + \lambda I)$ is positive definite for positive λ.