Machine Learning Homework 1

Hunter Carroll

January 2024

Problem 1.a

For the multiple regression model:

$$f(\boldsymbol{x}) = \beta_0 + \sum_{j=1}^{n} \beta_j x_j$$

where $x_1, ..., x_n$ are inputs and $\beta_0, ..., \beta_n$ are coefficients of the model. Suppose the training set include N samples: $(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), ..., (\boldsymbol{x_N}, y_N)$ where $\boldsymbol{x}_i = (x_{i1}, ..., x_{in})^T$

(a) Show that the residual sum of squares,

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{n} \beta_j x_{ij})^2$$

can be written in matrix form as,

$$RSS(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})^T (\boldsymbol{y} - \boldsymbol{X\beta})$$

where $\boldsymbol{y} = (y_1, y_2, ..., y_N)^T$, $\boldsymbol{\beta} = (\beta_0, ..., \beta_n)^T$ and X is the matrix,

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N1} & x_{N2} & \dots & x_{Nn} \end{bmatrix}$$

My solution:

Given $f(\boldsymbol{x}) = \beta_0 + \sum_{j=1}^{n} \beta_j x_j$

Let $\boldsymbol{y}$ be a column vector of size $N \times 1$,

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

Let $\boldsymbol{\beta}$ be a column vector of size $(n+1) \times 1$,

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}$$

Finally, let X be a matrix of size $N \times (n+1)$,

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N1} & x_{N2} & \dots & x_{Nn} \end{bmatrix}$$

Then,

$$X\boldsymbol{\beta} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N1} & x_{N2} & \dots & x_{Nn} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} = \begin{bmatrix} \beta_0 + x_{11}\beta_1 + x_{12}\beta_2 + ... + x_{1n}\beta_n \\ \beta_0 + x_{21}\beta_1 + x_{22}\beta_2 + ... + x_{2n}\beta_n \\ \vdots \\ \beta_0 + x_{N1}\beta_1 + x_{N2}\beta_2 + ... + x_{Nn}\beta_n \end{bmatrix}$$

Note then the result of $X\boldsymbol{\beta}$ is a column vector of size $N \times 1$ containing our predicted values. Let $\boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}$ be the $N \times 1$ column vector containing the residuals that is, the actual values represented by column vector $\boldsymbol{y}$ and predicted values represented as the column vector $X\boldsymbol{\beta}$ such that the residuals are the difference between the two column vectors,

$$\boldsymbol{u} = \boldsymbol{y} - X\boldsymbol{\beta}$$

Therefore, we can rewrite the Residual Sum of Squares in matrix form as follows,

$$RSS(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})^T \cdot (\boldsymbol{y} - X\boldsymbol{\beta}) = u^T \cdot u$$

Thus, $u^T \cdot u$ is the sum of the squared elements of u (Residuals).

Problem 1.b

Show that the partial derivative of RSS with respect to β is,

$$\frac{\partial RSS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2X^T\boldsymbol{y} + 2X^TX\boldsymbol{\beta}$$

Use the following matrix calculus results:

$$\frac{\partial \boldsymbol{x}^T A\boldsymbol{x}}{\partial \boldsymbol{x}} = 2A\boldsymbol{x}$$

If A is symmetric

My solution:
Given $RSS(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta})$ we can expand this to get,

$$= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^TX\boldsymbol{\beta} - \boldsymbol{\beta}^TX^T\boldsymbol{y} + \boldsymbol{\beta}^TX^TX\boldsymbol{\beta}$$

also note that $\boldsymbol{y}^TX\boldsymbol{\beta} = \boldsymbol{\beta}^TX^T\boldsymbol{y}$ this is because we have that both are equivalent scalars that are $\boldsymbol{y}^TX\boldsymbol{\beta} = (1 \times N)(N \times (n+1))((n+1) \times 1)$ which results in a row vector $(1 \times (n+1))$ times column vector $((n+1) \times 1)$ and thus a scalar and taking the transpose of a scalar is the scalar itself. Implying that $\boldsymbol{\beta}^TX^T\boldsymbol{y}$ is an equivalent expression. We can now simplify the $RSS(\boldsymbol{\beta})$ to,

$$RSS(\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\beta}^TX^T\boldsymbol{y} + \boldsymbol{\beta}^TX^TX\boldsymbol{\beta}$$

When taking the partial derivative of the $RSS(\boldsymbol{\beta})$ the first term $(\boldsymbol{y}^T\boldsymbol{y})$ goes away since with respect to $\boldsymbol{\beta}$ this is considered a constant. The second term $(-2\boldsymbol{\beta}^TX^T\boldsymbol{y})$ when differentiated with respects to $\boldsymbol{\beta}$ results in $-2X^T\boldsymbol{y}$. Finally, the third term we know that X^TX will result in a symmetric $(n+1) \times (n+1)$ matrix thus we can apply the fact that,

$$\frac{\partial \boldsymbol{x}^T A\boldsymbol{x}}{\partial \boldsymbol{x}} = 2A\boldsymbol{x}$$

giving,

$$\frac{\partial \boldsymbol{\beta}^TX^TX\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = 2X^TX\boldsymbol{\beta}$$

Therefore,

$$\frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{y}^T\boldsymbol{y}) = 0 \tag{1}$$

$$\frac{\partial}{\partial \boldsymbol{\beta}}(-2\boldsymbol{\beta}^TX^T\boldsymbol{y}) = -2X^T\boldsymbol{y} \tag{2}$$

$$\frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}^TX^TX\boldsymbol{\beta}) = 2X^TX\boldsymbol{\beta} \tag{3}$$

Combining the above results we find,

$$\frac{\partial RSS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2X^T\boldsymbol{y} + 2X^TX\boldsymbol{\beta}$$

Problem 1.c

Show that the least squares solution (normal equation):

$$\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\boldsymbol{y}$$

minimizes RSS.

My solution:

First, we set the partial derivative of the Residual Sum of Squares with respect to beta equal to zero to find the value $\boldsymbol{\beta}$ that minimizes the RSS,

$$\frac{\partial RSS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$$

$$-2X^T\boldsymbol{y} + 2X^TX\boldsymbol{\beta} = 0$$

Then we can factor a 2 out,

$$2(-X^T\boldsymbol{y} + X^TX\boldsymbol{\beta}) = 0$$

Resulting in,

$$-X^T\boldsymbol{y} + X^TX\boldsymbol{\beta} = 0$$

Move $-X^T\boldsymbol{y}$ to the right hand side,

$$X^TX\boldsymbol{\beta} = X^T\boldsymbol{y}$$

Given X^TX is symmetric and the determinant is non-zero, that is the columns are not linearly dependent ensuring full rank, then we can multiply each side by the inverse of (X^TX),

$$(X^TX)^{-1}X^TX\boldsymbol{\beta} = (X^TX)^{-1}X^T\boldsymbol{y}$$

However, a matrix multiplied by its inverse is simply the identity matrix. Therefore, the $\boldsymbol{\beta}$ that minimizes the RSS is,

$$\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\boldsymbol{y}$$

as desired.

Problem 2

Consider using Ridge Regression for modeling. Use the following form of cost function:

$$J(\beta) = \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{n}\beta_j x_{ij})^2 + \alpha \sum_{i=1}^{n}\beta_i^2$$

Problem 2.a

Show that $\sum_{i=1}^{n}\beta_i^2$ can be written in matrix form as,

$$\sum_{i=1}^{n}\beta_i^2 = \beta^T A \beta$$

where A is the $(n+1) \times (n+1)$ identity matrix except with a 0 in the top left cell.

My solution:

Let $\boldsymbol{\beta}$ be a column vector of size $(n+1) \times 1$,

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}$$

Let A be a square matrix of form $(n+1) \times (n+1)$,

$$A = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix}$$

Then note the term $\alpha \sum_{i=1}^{n}\beta_i^2$ will serve as a regularization term ensuring that we keep model weights small. Further, the reason we have 0 in the top-left cell of A is to exclude β_0 in the regularization term given our sum starts at $i = 1$. We know that β is an $(n+1) \times 1$ column vector. Thus, we want to find the value of the sum of our β^2 excluding β_0. In order to do this we will transpose β resulting in a $1 \times (n+1)$ row vector resulting in the following when multiplied by A,

$$\begin{bmatrix}\beta_0 & \beta_1 & \dots & \beta_n\end{bmatrix}\begin{bmatrix}0 & 0 & \cdots & 0\\ 0 & 1 & \ddots & \vdots\\ \vdots & \ddots & \ddots & 0\\ 0 & \cdots & 0 & 1\end{bmatrix} = \begin{bmatrix}0 & \beta_1 & \beta_2 & \dots & \beta_n\end{bmatrix}$$

We now have that $\beta^T A$ is a $1 \times (n+1)$ row vector excluding β_0 which is now compatible to multiply with β the $(n+1) \times 1$ column vector. The result is a scalar that represents $\sum_{i=1}^{n} \beta_i^2$,

$$\begin{bmatrix}0 & \beta_1 & \beta_2 & \dots & \beta_n\end{bmatrix}\begin{bmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_n\end{bmatrix} = 0 \cdot \beta_0 + \beta_1^2 + ... + \beta_n^2$$

Therefore, we can represent $\sum_{i=1}^{n} \beta_i^2$ in matrix form as $\boldsymbol{\beta^T A \beta}$.

Problem 2.b

Show the closed-form solution for the Ridge Regression is given by,

$$\hat{\beta} = (X^T X + \alpha A)^{-1} X^T \boldsymbol{y}$$

My solution:

Given Residual Sum of Squares in matrix form is,

$$RSS(\beta) = \boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{y}^T X \boldsymbol{\beta} - \boldsymbol{\beta} X^T \boldsymbol{y} + \boldsymbol{\beta}^T X^T X \boldsymbol{\beta}$$

Then the partial derivative of $RSS(\boldsymbol{\beta})$ with respects to $\boldsymbol{\beta}$ is,

$$\frac{\partial RSS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2X^T \boldsymbol{y} + 2X^T X \boldsymbol{\beta}$$

Now, Ridge Regression is similar to RSS but now we add a regularization term $\alpha \sum_{i=1}^{n} \beta_i^2$ which can be written as $\alpha \boldsymbol{\beta}^T A \boldsymbol{\beta}$ where α is some constant, then we have that Ridge Regression in matrix form is,

$$J(\boldsymbol{\beta}) = \boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{y}^T X \boldsymbol{\beta} - \boldsymbol{\beta} X^T \boldsymbol{y} + \boldsymbol{\beta}^T X^T X \boldsymbol{\beta} + \alpha \boldsymbol{\beta}^T A \boldsymbol{\beta}$$

Taking the partial derivative of $J(\boldsymbol{\beta})$ with respects to $\boldsymbol{\beta}$ and using the fact that A is symmetric,

$$\frac{\partial J(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2X^T \boldsymbol{y} + 2X^T X \boldsymbol{\beta} + 2\alpha A \boldsymbol{\beta}$$

Set this equal to zero to find $\hat{\boldsymbol{\beta}}$,

$$-2X^T\boldsymbol{y} + 2X^TX\boldsymbol{\beta} + 2\alpha A\boldsymbol{\beta} = 0$$

Factor out 2 and multiply it over,

$$-X^T\boldsymbol{y} + X^TX\boldsymbol{\beta} + \alpha A\boldsymbol{\beta} = 0$$

Move $-X^T\boldsymbol{y}$ over,

$$X^TX\boldsymbol{\beta} + \alpha A\boldsymbol{\beta} = X^T\boldsymbol{y}$$

Isolate $\boldsymbol{\beta}$,

$$\boldsymbol{\beta}(X^TX + \alpha A) = X^T\boldsymbol{y}$$

Apply the inverse of $(X^TX + \alpha A)$ to both sides. Then $J(\boldsymbol{\beta})$ is minimized for,

$$\hat{\boldsymbol{\beta}} = (X^TX + \alpha A)^{-1}(X^T\boldsymbol{y})$$
