Problem 1

For the multiple regression model:

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

where $x_1, \ldots, x_n$ are inputs and $\beta_0, \ldots, \beta_n$ are coefficients of the model. Suppose the training set include N samples: $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)$, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{in})^T$.

(a) Show that the residual sum of squares

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

can be written in matrix form as

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

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

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

(b) Show that the partial derivative of RSS with respect to $\boldsymbol{\beta}$ is:

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

Use the following matrix calculus results:

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

if A is symmetric.

(c) Show that the least squares solution (normal equation):

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

minimizes RSS.

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

$$X = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,n} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{N,1} & X_{N,2} & \cdots & X_{N,n} \end{bmatrix}$$

$$RSS(\beta) = \sum_{i=1}^{N} \left(y_i \underbrace{- \beta_0 - \sum_{j=1}^{n} \beta_j x_{i,j}}_{-f(x)} \right)^2$$

$$RSS(\beta) = \sum_{i=1}^{N} \left(y_i - (\beta_0 + \sum_{j=1}^{n} \beta_j X_{i,j}) \right)^2$$

$$X\beta = \beta_0 + \sum_{j=1}^{n} \beta_j x_j$$

$\therefore$

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - X\beta)^2$$

$$r = (y - X\beta) \rightarrow r^2 = r^T r$$

$$RSS(\beta) = (y - X\beta)^2$$
$$= (y - X\beta)^T (y - X\beta)$$

(b) $\frac{\partial x^T A x}{\partial x} = 2Ax$

$$(y - X\beta)^T (y - X\beta) = y^T y - \beta^T X^T y - y^T X\beta + \beta^T X^T X \beta$$
$$= y^T y - 2\beta^T X^T y + \beta^T X^T X \beta$$

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

(c)

$$-2X^T y + 2X^T X\beta = 0$$
$$\Rightarrow -X^T y + X^T X \beta = 0$$

$$\therefore \; X^T X \beta = X^T y$$

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

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

Problem 2

Consider using Ridge Regression for modeling. Use the following form of cost function (equivalent to Eqn. 5):

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

where all the notations from Problem 1 are still valid here.

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

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

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

(b) Show that the closed-form solution for the Ridge Regression is given by

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

(a)

$$\underbrace{\begin{bmatrix} \beta_0 & \beta_1 & \cdots & \beta_n \end{bmatrix}}_{\beta^T} \underbrace{\begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1 & & 1 \\ \vdots & & \ddots & \vdots \\ 0 & & & 1 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}}_{\beta} = \underbrace{(\beta_1^2 + \dots + \beta_n^2)}_{\sum^{n} \beta_i^2}$$

$'\beta^T$ A $'\beta$ $\sum_{i=1} \beta_i^2$

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

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

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

$$= (y - x\beta^T)(y - x\beta) + \alpha \beta^T A \beta$$

$$= y^T y - \beta^T x^T y - y^T x \beta + \beta^T x^T x \beta + \alpha \beta^T A \beta$$

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

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

set to 0

$$0 = -2x^T y + 2X^T x \beta + 2 x A \beta$$
$$= -X^T y + X^T X \beta$$

$$\therefore X^T X \beta + \alpha A \beta = X^T y$$

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

Problem 3 (MATH 5388 Only)

Consider the Ridge Regression problem where the cost function is given in Problem 2. Show that the problem is equivalent to the problem

$$\hat{\boldsymbol{\beta}}^c = \arg\min_{\boldsymbol{\beta}^c} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0^c - \sum_{j=1}^{n} (x_{ij} - \bar{x}_j)\beta_j^c \right)^2 + \alpha \sum_{j=1}^{n} (\beta_j^c)^2 \right\}$$

where $\bar{x}_j$ denotes the mean of the jth feature of all samples:

$$\bar{x}_j = \frac{1}{N} \sum_{i=1}^{N} x_{ij}$$

Give the correspondence between $\boldsymbol{\beta}^c$ and the original $\boldsymbol{\beta}$ in the cost function $J(\boldsymbol{\beta})$ in Problem 2.

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

to center data:

Let $x_{ij} = x_{ij} - \bar{x}_j$

Then $\beta_0^c = \bar{y}$

Then $\beta_0^c = \bar{y}$

because $\beta_0^c = \beta_0 + \sum_{j=1}^{n} \bar{x}_j \beta_j = \frac{1}{n} \sum_{j=1}^{n} y_j$

$\}$ $\beta_j^c = \beta_j$

$\therefore \sum_{i=1}^{N} (y_i - (\beta_0 + \sum_{j=1}^{n} \bar{x}_j \beta_j) - \sum_{j=1}^{n} (x_{ij} - \bar{x}_j)\beta_j)^2 + \alpha \sum_{j=1}^{n} \beta_j^2$

Thus $\hat{\beta}^c = \operatorname{argmin} \left\{ \sum_{i=1}^{n} (y_i - \beta_0^c - \sum_{j=1}^{n} (x_{ij} - \bar{x}_j)\beta_j^c)^2 + \alpha \sum_{j=1}^{n} (\beta_j^c)^2 \right\}$

$$\therefore = (y - x^c\beta^c)^T (y - x^c\beta^c) + \alpha \beta^{cT} A \beta^c$$

$$= y^T y - 2\beta^{cT} x^{cT} y + \beta^{cT} x^{cT} x^c \beta^c + \alpha \beta^{cT} A \beta^c$$

$$\frac{\partial J(\beta^c)}{\partial \beta^c} = 2x^{cT} y + 2x^{cT} x^c \beta^c + 2\alpha A \beta^c = 0$$

$$(x^{cT} x^c + \alpha A)\beta^c = x^{cT} y$$

$$\hat{\beta}^c = (x^{cT} x^c + \alpha A)^{-1} x^{cT} y$$

$\hat{\beta}_c$ has same regression coefficient.

different intercept.