Assignment 7

Problem 1

Define

$$J_1(\boldsymbol{w}) = |\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}|^2 + \lambda_2|\boldsymbol{w}|^2 + \lambda_1|\boldsymbol{w}|_1$$

and

$$J_2(\boldsymbol{w}) = |\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{w}|^2 + c\lambda_1|\boldsymbol{w}|_1$$

where $c = (1 + \lambda_2)^{-\frac{1}{2}}$ and

$$\tilde{\boldsymbol{X}} = c\begin{pmatrix} \boldsymbol{X} \\ \sqrt{\lambda_2}\boldsymbol{I}_d \end{pmatrix}, \tilde{\boldsymbol{y}} = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0}_{d\times 1} \end{pmatrix}$$

Show

$$\arg\min J_1(\boldsymbol{w}) = c(\arg\min J_2(\boldsymbol{w}))$$

i.e. $J_1(c\boldsymbol{w}) = J_2(\boldsymbol{w})$ and hence that one can solve an elastic net problem using a lasso solver on modified data.

Problem 2

Let $RSS(w) = ||Xw - y||_2^2$ be the residual sum of squares.

a. Show that

$$\frac{\partial}{\partial w_k} RSS(w) = a_k w_k - c_k$$

$$a_k = 2\sum_{i=1}^{n} x_{ik}^2 = 2||x_{:,k}||^2$$

$$c_k = 2\sum_{i=1}^{n} x_{ik}(y_i - w_{-k}^T x_{i,-k}) = 2x_{:,k}^T r_k$$

where $w_{-k} = w$ without component k, $x_{i,-k}$ is x_i without component k, and $r_k = y - w_{-k}^T x_{:,-k}$ is the residual due to using all the features except feature k. Hint: Partition the weights into those involving k and those not involving k.

b. Show that if $\frac{\partial}{\partial w_k} RSS(w) = 0$, then

$$\hat{w}_k = \frac{x_{:,k}^T r_k}{||x_{:,k}||^2}$$

Hence when we sequentially add features, the optimal weight for feature k is computed by computing orthogonally projecting $x_{:,k}$ onto the current residual.