Optimization and Machine Learning, Spring 2020
Homework 1
(Due Wednesday, Mar. 18 at 11:59pm (CST))

March 19, 2020

1. Suppose that we have N training samples, in which each sample is composed of p input variables and one continuous/binary response.

   (a) Please define the input and output variables, and show a linear relationship between them. (5 points)
   Solution: Define the input variable X as a vector $x \in \mathbb{R}^p$, and the output variable Y as a continuous value $y \in \mathbb{R}$ or a binary value $y \in \{0,\ 1\}$. Their linear relationship is $Y = \beta^T X$ where $\beta \in \mathbb{R}^p$ is a linear coefficient vector.

   (b) Please define a data matrix and corresponding response vector, and find your i-th ($i = 1, ..., N$) sample with its response. (5 points)
   Solution: Define a data matrix $\mathbf{X} = \begin{bmatrix} -x_1^T- \\ \vdots \\ -x_N^T- \end{bmatrix}$, and corresponding response vector $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$.
   The i-th sample is the i-th row of $\mathbf{X}$ and its response is the i-th element of $\mathbf{y}$.

   (c) Please use the least squares to estimate the parameters of the linear model in (a) based on the dataset in (b), and explain in which case the solution is unique. (10 points)
   Solution: Using the least squares, our target is to minimize

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

   Differentiating with β and set the derivative to 0, we have

   $$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0.$$

   The unique solution $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is reachable if and only if $\mathbf{X}^T\mathbf{X}$ is invertible (nonsingular, full-ranked) [Note: other reasonable explanations are acceptable].

   (d) Is there any way to get an unique closed-form solution? If yes, please show how do you obtain the solution. (5 points)
   Solution: Since there are cases that $\mathbf{X}^T\mathbf{X}$ is singular, we cannot get a unique closed-form solution. However we can add a regularization term into the original loss function, then it becomes

   $$\mathcal{L}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta,$$

   where $\lambda > 0$. Differentiating with β,

   $$\frac{\partial \mathcal{L}(\beta)}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) + 2\lambda\beta,$$

   Set the derivative into 0, the unique closed-form solution is $\hat{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y}$. It is always reachable because adding a full-ranked matrix $\lambda\mathbf{I}_p$ makes $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p$ a full-ranked matrix, and any full-ranked matrix is invertible.

   (e) How can you select the best model in (d) based only on your training data. (5 points)
   Solution: For a set of λ, we can use K-fold cross validation method. Firstly, split the dataset into random K folds, use $K-1$ folds to train the model and the rest one for validation. And then compute the average loss over the validation set using each λ candidate, finally we pick the best λ with the least average loss.

2. Given the input variables $X \in \mathbb{R}^p$ and response variable $Y \in \mathbb{R}$, the Expected Prediction Error (EPE) is defined by

$$\text{EPE}(\hat{f}) = \mathbb{E}[L(Y, \hat{f}(X))], \tag{1}$$

where $\mathbb{E}(\cdot)$ denotes the expectation over the joint distribution $\Pr(X, Y)$, and $L(Y, \hat{f}(X))$ is a loss function measuring the difference between the estimated $\hat{f}(X)$ and observed Y.

(a) Given the squared error loss $L(Y, \hat{f}(X)) = (Y - \hat{f}(X))^2$, please derive the regression function $\hat{f}(x) = \mathbb{E}(Y|X = x)$ by minimizing $\text{EPE}(\hat{f})$ w.r.t. $\hat{f}$. (5 points)
Solution: Firstly, we rewrite

$$EPE(\hat{f}) = \mathbb{E}[L(Y, \hat{f}(X))] = \mathbb{E}_X[\mathbb{E}_{Y|X}[L(Y, \hat{f}(X))|X]],$$

it is sufficient to minimize $h = \mathbb{E}_{Y|X}[L(Y, \hat{f}(X))|X] = \mathbb{E}_{Y|X}[(Y - f(X))^2|X]$, that is $\hat{f}(x) = \underset{f}{\text{argmin}}\, \mathbb{E}_{Y|X}[(Y - f)^2|X = x]$. Then we show the detail.

$$\begin{aligned}
\frac{\partial}{\partial f}\mathbb{E}_{Y|X}[(Y - f)^2|X = x] &= \frac{\partial}{\partial f}\int [y - f]^2 \Pr(y|x)dy \\
&= \int \frac{\partial}{\partial f}[y - f]^2 \Pr(y|x)dy \\
&\Rightarrow 2\int y \Pr(y|x)dy = 2f\int \Pr(y|x)dy \\
&\Rightarrow 2\mathbb{E}[Y|X = x] = 2f \\
&\Rightarrow \hat{f}(x) = \mathbb{E}[Y|X = x].
\end{aligned}$$

(b) Please explain why the nearest neighbors is an approximation to the regression function in (a). (5 points)
Solution: The nearest neighbors method $\hat{f}(x) = \frac{1}{k}\sum_{x_i \in N_k(x)} y_i$ has two approximations. The first one is averaging over sample data to appoximate expectation, and the second one is conditioning on neighborhood to approximate conditioning on a point.

(c) Please explain how the least squares approximates the regression function in (a). (5 points)
Solution: The least square method approximates the theoretical expectation by averaging over the observed data. Using EPE in least squares, we can find the theoretical solution $\beta = \mathbb{E}(XX^T)^{-1}\mathbb{E}(XY)$, and the actual solution for least square is $\beta = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ which is an approximation for theoretical value.

(d) Please discuss the difference between the nearest neighbors and the least squares based on your results in (b) and (c). (5 points)
Solution:

- The nearest neighbors (NN) method is a locally constant function, while the least square (LS) method is a globally linear function. NN relies more on local input, while LS considers whole input.
- In terms of the number of effective parameters, LS and NN have p and N/k parameters, respectively, where N denotes the total number of training samples.
- LS usually produces high-bias and low-variance results, due to its stringent assumption on the linearity; in contrast, NN tends to make low-bias and high-variance predictions, as it has no assumption on the underlying model.

3. Given a set of observation pairs $(x_1, y_1) \cdots (x_N, y_N)$. By assuming the linear model is a reasonable approximation, we consider fitting the model via least squares approaches, in which we choose coefficients β to minimize the residual sum of squares (RSS),

$$\hat{\beta}_0,\ \hat{\beta} = \underset{\beta_0,\ \beta}{\text{argmin}}\ \sum_{i=1}^{N}(y_i - \beta_0 - \beta x_i)^2.$$

(a) Show that

$$\begin{aligned}
\hat{\beta} &= \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2}, \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}\bar{x},
\end{aligned} \tag{2}$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ are the sample means. (3 points)
Solution: Firstly, we compute β_0

$$\begin{aligned}
&\frac{\partial}{\partial \beta_0}\sum_{i=1}^{N}(y_1 - \beta_0 - \beta x_i)^2 = \sum_{i=1}^{N} -2(y_i - \beta_0 - \beta x_i) = 0 \\
&\Rightarrow \sum_{i=1}^{N}(y_i - \beta x_i) = \sum_{i=1}^{N}\beta_0 = N\beta_0 \\
&\Rightarrow \beta_0 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \beta x_i) = \frac{1}{N}\sum_{i=1}^{N} y_i - \beta\frac{1}{N}\sum_{i=1}^{N} x_i = \bar{y} - \beta\bar{x} \\
&\Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}\bar{x}.
\end{aligned}$$

Plug β_0 into $\sum_{i=1}^{N}(y_1 - \beta_0 - \beta x_i)^2$ and differentiate with β,

$$\begin{aligned}
&\frac{\partial}{\partial \beta}\sum_{i=1}^{N}(y_1 - \beta_0 - \beta x_i)^2 = \frac{\partial}{\partial \beta}\sum_{i=1}^{N}(y_1 - \bar{y} + \beta\bar{x} - \beta x_i)^2 = \sum_{i=1}^{N} 2[y_i - \bar{y} + \beta(\bar{x} - x_i)](\bar{x} - x_i) = 0 \\
&\Rightarrow \sum_{i=1}^{N}(y_i - \bar{y})(\bar{x} - x_i) = -\beta\sum_{i=1}^{N}(\bar{x} - x_i)^2 \\
&\Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}.
\end{aligned}$$

In conclusion,

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, \ \hat{\beta}_0 = \bar{y} - \hat{\beta}\bar{x}.$$

(b) Using (2), argue that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$. (2 points)
Solution: We can plug $(\bar{x}, \bar{y})$ into the equation $\hat{y} = \hat{\beta}x_i + \beta_0$, and we find $\bar{y} = \hat{\beta}\bar{x} + \bar{y} - \hat{\beta}\bar{x} = \bar{y}$ satisfies. So the least squares line always passes through the point $(\bar{x}, \bar{y})$.

4. Given a set of training data $(x_1, y_1), \cdots, (x_N, y_N)$ from which to estimate the parameters β, where each $x_i = [x_{i1}, \cdots, x_{ip}]^T$ denotes a vector of feature measurements for the ith sample. Consider a linear regression problem in which we want to "weight"different training examples differently. Specifically, suppose we aim at minimizing

$$\text{RSS}(\beta) = \frac{1}{2}\sum_{i=1}^{N} w_i(y_i - x_i^T\beta)^2. \tag{3}$$

(a) Show that $\text{RSS}(\beta) = (\mathbf{X}\beta - \mathbf{y})^T\mathbf{W}(\mathbf{X}\beta - \mathbf{y})$ for an appropriate diagonal matrix $\mathbf{W}$, and where $\mathbf{X} = [x_1, \cdots, x_N]^T$ and $\mathbf{y} = [y_1, \cdots, y_N]^T$. State clearly what $\mathbf{W}$ is. (1 points)
Solution: $\mathbf{W}$ is a diagonal matrix with its i-th diagonal element being $\frac{1}{2}w_i$. Suppose we have the predictions $\hat{\boldsymbol{y}} = \mathbf{X}\beta$, $\text{RSS}(\beta)$ is rewritten by

$$\text{RSS}(\beta) = (\mathbf{X}\beta - \mathbf{y})^T\mathbf{W}(\mathbf{X}\beta - \mathbf{y}) = (\hat{\mathbf{y}} - \mathbf{y})^T\mathbf{W}(\hat{\mathbf{y}} - \mathbf{y})$$

$$\begin{bmatrix} \hat{y}_1 - y_1 \\ \vdots \\ \hat{y}_N - y_N \end{bmatrix}^T \begin{pmatrix} \frac{1}{2}w_1 & 0 & \cdots & 0 & 0 \\ 0 & \frac{1}{2}w_2 & \cdots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2}w_{N-1} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2}w_N \end{pmatrix} \begin{bmatrix} \hat{y}_1 - y_1 \\ \vdots \\ \hat{y}_N - y_N \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{2}w_1(\hat{y}_1 - y_1) \\ \vdots \\ \frac{1}{2}w_N(\hat{y}_N - y_N) \end{bmatrix}^T \begin{bmatrix} \hat{y}_1 - y_1 \\ \vdots \\ \hat{y}_N - y_N \end{bmatrix} = \frac{1}{2}\sum_{i=1}^{N} w_i \left(y_i - x_i^T\beta\right)^2.$$

(b) By finding the derivative $\nabla_\beta \text{RSS}(\beta)$ and setting that to zero, write the normal equations to this weighted setting and give the value of β that minimizes $\text{RSS}(\beta)$ in closed form as a function of $\mathbf{X}$, $\mathbf{W}$ and $\mathbf{y}$. (2 points)

Solution:

$$\begin{aligned} \nabla_\beta \text{RSS}(\beta) &= \frac{\partial \,\text{RSS}(\beta)}{\partial \beta} \\ &= \frac{\partial}{\partial \beta}(\mathbf{X}\beta - \mathbf{y})^T\mathbf{W}(\mathbf{X}\beta - \mathbf{y}) \\ &= 2\mathbf{X}^T\mathbf{W}(\mathbf{X}\beta - \mathbf{y}) \\ &= 0 \\ \Rightarrow \hat{\beta} &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}. \end{aligned}$$

(c) Suppose the y_i's were observed with differing variances. To be specific, suppose that

$$p(y_i|x_i;\beta) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(y_i - x_i^T\beta)^2}{2\sigma_i^2}\right), \tag{4}$$

i.e., y_i has mean $x_i^T\beta$ and variance σ_i^2, where the σ_i's are fixed, known, constants). Show that finding the maximum likelihood estimate of β is equivalent to solving a weight linear regression problem. State clearly what the w_i's are in terms of the σ_i's. (4 points)

Solution: The log likelihood function is

$$\mathcal{L}(\beta) = \log\prod_{i=1}^{N} p\left(y_i|x_i;\beta\right) = \log\{\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{\left(y_i - x_i^T\beta\right)^2}{2\sigma_i^2}\right)\} = \frac{-1}{\sqrt{2\pi}\sigma_i}\sum_{i=1}^{N}\frac{\left(y_i - x_i^\top\beta\right)^2}{2\sigma_i^2}.$$

Maximizing the likelihood is equivalent to minimizing $\sum_{i=1}^{N}\frac{\left(y_i - x_i^\top\beta\right)^2}{2\sigma_i^2}$. This is equivalent to solving a weigh linear regression problem with weight $w_i = \frac{1}{2\sigma_i^2}$.

5. To perform variable selection, three classical approaches were introduced in class, including variable subset selection, forward stepwise selection and backward stepwise selection.

(a) To deepen your understanding of these approaches, please make a table to describe their key procedures as well as the pros and cons. (6 points)

(b) Suppose we perform these three approaches on a single data set. For each approach, we obtain $p + 1$ models, containing $0, 1, 2, \cdots, p$ predictors. **Explain** your answers:

- Which of the three models with k predictors has the smallest training RSS? (1 points)
- Which of the three models with k predictors has the smallest test RSS? (1 points)

(Note that: Solutions with the correct answer but without adequate explanation will not earn credit.)

Solution:

(a) We summarize the key procedures of three approaches as follows correspondingly:

Pros:

- It is a simple and conceptually appealing approach.
- In practice, it can be instructive to observe how best-subset selection could be done for small problems.

Limitations:

- In general, there are 2^p models that involve subsets of p predictors. Consequently, best subset selection becomes computationally infeasible for values of p greater than around 40.

Pros:

- It is superior to the best subset selection in terms of computation efficiency.

Algorithm 1 Best Subset Selection

1: Let $\mathcal{M}_0$ denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2: **for** $k = 1, 2 \ldots, p$ **do**
3: Fit all $\begin{pmatrix} p \\ k \end{pmatrix}$ models that contain exactly k predictors.
4: Pick the best (e.g., in terms of the smallest RSS) among these $\begin{pmatrix} p \\ k \end{pmatrix}$ models, and call it $\mathcal{M}_k$.
5: **end for**
6: Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error.

Algorithm 2 Forward Stepwise Selection

1: Let $\mathcal{M}_0$ denote the null model, which contains no predictors.
2: **for** $k = 1, 2 \ldots, p-1$ **do**
3: Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
4: Choose the best (e.g., in terms of the smallest RSS) among these $p-k$ models, and call it $\mathcal{M}_{k+1}$.
5: **end for**
6: Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error.

- It can be applied even in the high-dimensional setting where $n < p$, and so is the only viable subset method when p is very large.

Limitations:

- It is not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors.

Pros:

- Like forward stepwise selection, the backward selection approach searches through only $1+p(p+1)/2$ models, and so can be applied in settings where p is too large to apply best subset selection.

Limitations:

- It requires that the number of samples n is larger than the number of variables p.
- It is not guaranteed to yield the best model containing a subset of the p predictors.

(b)
- Best subset will have the smallest train RSS because the models will optimize on the training RSS and best subset will try every model that forward and backward selection will try.
- The best test RSS model could be any of the three. Best subset could easily over-fitting if the data has large p predictors relative to n observations. Forward and backward selection might not converge on the same model but try the same number of models and hard to say which selection process would be better.

6. Refer to [1, Ex. 3.5]. Consider the ridge regression problem

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{5}$$

where $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage. Show that problem (5) is equivalent to the problem

$$\hat{\beta}^{\text{c}} = \underset{\beta^{\text{c}}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0^{\text{c}} - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)\beta_j^{\text{c}})^2 + \lambda \sum_{j=1}^{p} (\beta_j^{\text{c}})^2 \right\}. \tag{6}$$

Give the correspondence between β^{c} and the original β in (5). Characterize the solution to this modified criterion. Moreover, show that a similar result holds for the least absolute shrinkage and selection operator (LASSO). (10 points)

Solution: Consider that the ridge expression problem (5) can be written as

$$\sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \bar{x}_j\beta_j - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2. \tag{7}$$

Algorithm 3 Backward Stepwise Selection

1: Let $\mathcal{M}_0$ denote the full model, which contains all p predictors.
2: **for** $k = p, p-1 \ldots, 1$ **do**
3: Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
4: Choose the best (e.g., in terms of the smallest RSS) among these k models, and call it $\mathcal{M}_{k-1}$.
5: **end for**
6: Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error.

By shifting that x_i's to have zero mean we have translated all points to the origin. As such only the 'intercept' of the data or β_0 is modified the 'slope's' or β_j^c for $i = 1, 2 \ldots, p$ are not modified. Define 'centered' values of β as

$$\beta_0^c = \beta_0 + \sum_{j=1}^{p} \bar{x}_j \beta_j$$
$$\beta_j^c = \beta_i, \quad i = 1, 2, \ldots, p,$$

that the above can be recast as

$$\sum_{i=1}^{N} \left(y_i - \beta_0^c - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)\beta_j^c \right)^2 + \lambda \sum_{j=1}^{p} (\beta_j^c)^2.$$

The equivalence of the minimization results from the fact that if β_i minimizes its respective functional the β_i^c's will do the same. We compute the value of β_0^c in the above expression by setting the derivative with respect to this variable equal to zero (a consequence of the expression being at a minimum). We obtain

$$\sum_{i=1}^{N} \left(y_i - \beta_0^c - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)\beta_j^c \right) = 0,$$

which implies $\beta_0^c = \frac{1}{N}\left(\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)\beta_j^c \right)$. Moreover, the same argument above can be used to show that the minimization required for the LASSO can be written in the same way (i.e., replace $(\beta_j^c)^2$ with $|\beta_j^c|$). The intercept in the centered case continues to be $\bar{\mathbf{y}}$.

7. It is well-known that ridge regression tends to give similar coefficient values to correlated variables, whereas the LASSO may give quite different coefficient values to correlated variables. We will now explore this property in a very simple setting.

   Suppose that $n = 2$, $p = 2$, $x_{11} = x_{12}$, $x_{21} = x_{22}$. Furthermore, suppose that $y_1 + y_2 = 0$ and $x_{11} + x_{21} = 0$ and $x_{12} + x_{22} = 0$, so that the estimate for the intercept in a least squares, ridge regression, or LASSO model is zero: $\hat{\beta}_0 = 0$.

   (a) Write out the ridge regression optimization problem in this setting. (2 points)

   (b) Argue that in this setting, the ridge coefficient estimates satisfy $\hat{\beta}_1 = \hat{\beta}_2$. (4 points)

   (c) Write out the LASSO optimization problem in this setting. (2 points)

   (d) Argue that in this setting, the LASSO coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are not unique—in other words, there are many possible solutions to the optimization problem in (c). Describe these solutions. (2 points)

   Solution:

   (a) In this setting, the ridge regression optimization problem reads

   $$\min_{\hat{\beta}} \; f_{\text{ridge}}(\hat{\beta}) = (y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2). \tag{8}$$

   (b) It takes the following steps to obtain the solution:

   1) Expanding the equation from (8):

   $$f_{\text{ridge}}(\hat{\beta}) = (y_1^2 + \hat{\beta}_1^2 x_{11}^2 + \hat{\beta}_2^2 x_{12}^2 - 2\hat{\beta}_1 x_{11} y_1 - 2\hat{\beta}_2 x_{12} y_1 + 2\hat{\beta}_1 \hat{\beta}_2 x_{11} x_{12}) + (y_2^2 + \hat{\beta}_1^2 x_{21}^2 + \hat{\beta}_2^2 x_{22}^2 - 2\hat{\beta}_1 x_{21} y_2 - 2\hat{\beta}_2 x_{22} y_2 + 2\hat{\beta}_1 \hat{\beta}_2 x_{21} x_{22}) + \lambda \hat{\beta}_1^2 + \lambda \hat{\beta}_2^2.$$

2) Taking the partial derivative to $\hat{\beta}_1$ and setting equation to 0 to minimize:

$$\frac{\partial f_{\text{ridge}}(\hat{\beta})}{\partial \hat{\beta}_1} = (2\hat{\beta}_1 x_{11}^2 - 2x_{11}y_1 + 2\hat{\beta}_2 x_{11}x_{12}) + (2\hat{\beta}_1 x_{21}^2 - 2x_{21}y_2 + 2\hat{\beta}_2 x_{21}x_{22}) + 2\lambda\hat{\beta}_1 = 0.$$

3) Setting $x_{11} = x_{12} = x_1$ and $x_{21} = x_{22} = x_2$ and dividing both sides of the equation by 2:

$$(\hat{\beta}_1 x_1^2 - x_1 y_1 + \hat{\beta}_2 x_1^2) + (\hat{\beta}_1 x_2^2 - x_2 y_2 + \hat{\beta}_2 x_2^2) + \lambda\hat{\beta}_1 = 0,$$

$$\Downarrow$$

$$\hat{\beta}_1(x_1^2 + x_2^2) + \hat{\beta}_2(x_1^2 + x_2^2) + \lambda\hat{\beta}_1 = x_1 y_1 + x_2 y_2.$$

4) Add $2\hat{\beta}_1 x_1 x_2$ and $2\hat{\beta}_2 x_1 x_2$ to both sides of the equation:

$$\hat{\beta}_1(x_1^2 + x_2^2 + 2x_1x_2) + \hat{\beta}_2(x_1^2 + x_2^2 + 2x_1x_2) + \lambda\hat{\beta}_1 = x_1y_1 + x_2y_2 + 2\hat{\beta}_1 x_1x_2 + 2\hat{\beta}_2 x_1x_2$$

$$\Downarrow$$

$$\hat{\beta}_1(x_1 + x_2)^2 + \hat{\beta}_2(x_1 + x_2)^2 + \lambda\hat{\beta}_1 = x_1y_1 + x_2y_2 + 2\hat{\beta}_1 x_1x_2 + 2\hat{\beta}_2 x_1x_2. \tag{9}$$

5) Because $x_1 + x_2 = 0$, we can eliminate the first two terms in (9):

$$\lambda\hat{\beta}_1 = x_1y_1 + x_2y_2 + 2\hat{\beta}_1 x_1x_2 + 2\hat{\beta}_2 x_1x_2.$$

6) Similarly by taking the partial deritive to $\hat{\beta}_2$, we can get the equation:

$$\lambda\hat{\beta}_2 = x_1y_1 + x_2y_2 + 2\hat{\beta}_1 x_1x_2 + 2\hat{\beta}_2 x_1x_2.$$

7) The left side of the equations for both $\lambda\hat{\beta}_1$ and $\lambda\hat{\beta}_2$ are the same so we have:

$$\lambda\hat{\beta}_1 = \lambda\hat{\beta}_2,$$

indicating

$$\hat{\beta}_1 = \hat{\beta}_2.$$

(c) In this setting, the LASSO regression optimization problem reads

$$\min_{\hat{\beta}} \; f_{\text{LASSO}}(\hat{\beta}) = (y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda(|\hat{\beta}_1| + |\hat{\beta}_2|).$$

(d) Following through the steps in (*b*), we get:

$$\lambda\frac{|\hat{\beta}_1|}{\hat{\beta}_1} = \lambda\frac{|\hat{\beta}_2|}{\hat{\beta}_2}.$$

So it seems that the LASSO just requires that $\hat{\beta}_1$ and $\hat{\beta}_2$ are both positive or both negative (ignoring possibility of 0...).

8. Refer to [1, Ex. 3.30]. Consider the elastic-net optimization problem:

$$\min_{\beta} \; \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda\left[\alpha\|\beta\|_2^2 + (1-\alpha)\|\beta\|_1\right]. \tag{10}$$

Show how one can turn this into a LASSO problem, using an augmented version of $\mathbf{X}$ and $\mathbf{y}$. (10 points)
Solution: For this problem note that if we augment $\mathbf{X}$ with a multiple of the $p \times p$ identity to get

$$\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X} \\ \gamma\mathbf{I} \end{bmatrix}, \tag{11}$$

then $\tilde{\mathbf{X}}\beta = \begin{bmatrix} \mathbf{X}\beta \\ \gamma\beta \end{bmatrix}$. If we next augment $\mathbf{y}$ with p zeros as

$$\tilde{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix}.$$

Then we have

$$\|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\beta\|_2^2 = \left\| \begin{bmatrix} \mathbf{y} - \mathbf{X}\beta \\ \gamma\beta \end{bmatrix} \right\|_2^2 = \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \gamma^2\|\beta\|_2^2. \tag{12}$$

Now in the this augmented space a lasso problem for β is

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(\|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\beta\|_2^2 + \tilde{\lambda}\|\beta\|_1 \right).$$

Writing this using (12) we get in the original variables the following

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \gamma^2\|\beta\|_2^2 + \tilde{\lambda}\|\beta\|_1 \right).$$

To make this match the requested expression we take $\gamma^2 = \lambda\alpha$ and $\tilde{\lambda} = \lambda(1-\alpha)$. Thus to solve the requested minimization problem given $\mathbf{y}$, $\mathbf{X}$, λ and α perform the following steps

- Augment $\mathbf{y}$ with p additional zeros to get $\tilde{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix}$.
- Augment $\mathbf{X}$ with the multiple of the $p \times p$ identity matrix $\sqrt{\lambda\alpha}\mathbf{I}$ to get $\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X} \\ \gamma\mathbf{I} \end{bmatrix}$.
- Set $\tilde{\lambda} = \lambda(1-\alpha)$.
- Solve the LASSO minimization problem with input $\tilde{\mathbf{y}}$, $\tilde{\mathbf{X}}$ and $\tilde{\lambda}$.

References

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.