3 Regularized Linear Regression Using Lasso [14 Points]

Lasso is a form of regularized linear regression, where the L1 norm of the parameter vector is penalized. It is used in an attempt to get a sparse parameter vector where features of little "importance" are assigned to zero weight. But why does lasso encourage sparse parameters? For this question, you are going to examine this.

Let $\mathbf{X}$ denote an $n \times d$ matrix where rows are training points, $\mathbf{y}$ denotes an $n \times 1$ vector of corresponding output value, $\mathbf{w}$ denotes a $d \times 1$ parameter vector and $\mathbf{w}^\star$ denotes the optimal parameter vector. To make the analysis easier we will consider the special case where the training data is whitened (i.e., $X^\top X = I$). For lasso regression, the optimal parameter vector is given by

$$\mathbf{w}^\star = \underset{\mathbf{w}}{\operatorname{argmin}}\, J_\lambda(\mathbf{w}), \tag{1}$$

where $J_\lambda(\mathbf{w})$ is the function we want to minimize, which is given by

$$J_\lambda(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - X\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1, \tag{2}$$

where $\lambda > 0$. Note that the L_1 norm for a vector $\mathbf{w} = [w_1, \ldots, w_d]^\top \in \mathbb{R}^d$ is defined as $\|\mathbf{w}\|_1 = |w_1| + \ldots + |w_d|$.

1. **[3 Points]** In 3.2 and 3.3, we will show that whitening the training data nicely decouples the features, making $\mathbf{w}_i^\star$ determined by the ith feature and the output regardless of other features. To show this, begin by writing $J_\lambda(\mathbf{w})$ in the form

$$J_\lambda(\mathbf{w}) = g(\mathbf{y}) + \sum_{i=1}^{d} f\left(X_{\cdot i}, \mathbf{y}, w_i, \lambda\right), \tag{3}$$

where $X_{\cdot i}$ is the ith column of X, g is a function of only $\mathbf{y}$ and f is a function of $X_{\cdot i}, \mathbf{y}, w_i, \lambda$

$$\begin{aligned}
J_\lambda(w) &= \tfrac{1}{2}(y - Xw)^T(y - Xw) + \lambda \sum_{i=1}^{d} |w_i| \\
&= \tfrac{1}{2}\left[y^Ty - y^T(Xw) - (Xw)^Ty + (Xw)^T(Xw)\right] + \lambda \sum_{i=1}^{d} |w_i| \\
&= \tfrac{1}{2}(y^Ty) + \tfrac{1}{2}(Xw)^T(Xw) - y^T(Xw) + \lambda \sum_{i=1}^{d} |w_i| \\
&= \underbrace{\tfrac{1}{2}(y^Ty)}_{g(y)} + \sum_{i=1}^{d} \underbrace{\left[\tfrac{1}{2}(X_{\cdot i} w_i)^T(X_{\cdot i} w_i) - y^T(X_{\cdot i} w_i) + \lambda |w_i|\right]}_{f(X_{\cdot i},\, y,\, w_i,\, \lambda)}
\end{aligned}$$

2. **[3 Points]** Assume that $w_i^* > 0$, what is the value of w_i^* in this case?

$$\frac{d}{dw_i} J_\lambda(w) = \frac{1}{2} \cdot 2 \underbrace{X_i^T X_i}_{I} w_i - y^T X_i + \lambda = 0$$

$$\Rightarrow w_i^* = y^T X_i - \lambda \quad \#$$

3. **[3 Points]** Assume that $w_i^* < 0$, what is the value of w_i^* in this case?

$$\frac{d}{dw_i} J_\lambda(w_i) = X_i^T X_i w_i - y^T X_i - \lambda = 0$$

$$\Rightarrow w_i^* = y^T X_i + \lambda \quad \#$$

4. **[3 Points]** From 2 and 3, what is the condition for w_i^* to be 0? How can you interpret that condition?

if $w_i^* = 0$

$\Rightarrow \frac{d}{dw_i} J_\lambda(w_i)$ at $w_i = 0$ can be $+1$ or -1

$\Rightarrow y^T X_i + \lambda(\text{sgn}\, w_i) = 0$, $\text{sgn}\, w_i \in [1, -1]$

$\Rightarrow |y^T X_i| = \lambda$

5. **[2 Points]** Now consider ridge regression where the regularization term is replaced by $\frac{1}{2}\lambda\|\mathbf{w}\|_2^2$. What is the condition for $w_i^* = 0$? How does it differ from the condition you obtained in 4?

$$J_\lambda(w) = \frac{1}{2}\|y - Xw\|_2^2 + \frac{1}{2}\lambda\|w\|_2^2$$

Based on Q3-1

$$J_\lambda(w) = \frac{1}{2}(y^T y) + \sum_{i=1}^{d}\Big[\frac{1}{2}(X_{\cdot i} w_i)^T (X_{\cdot i} w_i) - y^T (X_i w_i) + \frac{1}{2}\lambda w_i^2\Big]$$

$$\frac{d}{dW_i} J_\lambda(W) = W_i - y^T X_i + \lambda W_i = 0$$

$$\Rightarrow W_i^*(1+\lambda) = y^T X_i$$

$$\Rightarrow W_i^* = \frac{y^T X_i}{1+\lambda}$$

for the condition $W_i^* = 0 \Rightarrow y^T X_i = 0$

$\rightarrow$ the new $J_\lambda(w)$ is differentiable at $W_i^* = 0$

Not like in Q4 the regularization term $\lambda\|W\|_1$ was not differentiable and need to take subderivatives.

4 Logistic Regression; Improving our understanding of Convexity [25 points]

Consider a binary classification problem where the goal is to predict a class $y \in \{0, 1\}$, given an input $x \in \mathcal{R}^p$. A method that you can use for this task is *Logistic Regression*. In *Logistic Regression*, we model the log-odds as an affine function of the data and find weights to maximize the likelihood of our data under the resulting model. Let's investigate why this is a reasonable choice:

(Affine function definition: an affine function f takes the form $f(x) = w^\top x + c$ with $c \in \mathbb{R}$ and $w, x \in \mathbb{R}^n$. In other words, it is a linear function composed with a translation.)

4.1 Setup

1. **[2 points]** For a probability value $p \in (0, 1)$, what is the range of the odds, $\frac{p}{1-p}$, and the log-odds, $\log\left(\frac{p}{1-p}\right)$? Explain why this makes the log-odds a desirable transformation of our data to fit with our affine model.

$p=0 \rightarrow \frac{p}{1-p}=0$; $p=1 \rightarrow \frac{p}{1-p}=\infty$ $\Rightarrow \text{range}\left(\frac{p}{1-p}\right)=(0,\infty)$

$\log\frac{p}{1-p}=-\infty$ $\quad \log\frac{p}{1-p}=\infty \Rightarrow \text{range}\left(\log\frac{p}{1-p}\right)=(-\infty,\infty)$

The log-odds transformation makes the function covers $-\infty$ to ∞ of the affine model.

2. **[2 points]** We can proceed to model the log-odds with an affine model:

$$\log\frac{P(y_i = 1|x_i, w)}{1 - P(y_i = 1|x_i, w)} = w^\top x_i.$$

Conclude that the log-likelihood can be written as

$$\mathcal{L}(w) = \log P(y|\mathbf{X}, w) = \sum_{i=1}^{n}[y_i w^\top x_i - \log(1 + \exp(w^\top x_i))],$$

where:

- $\mathbf{X} \in \mathcal{R}^{n\times(1+p)}$ is a data matrix, with the first column composed of all ones
- $w \in \mathcal{R}^{(p+1)\times 1}$ is the weight vector, with the first index w_1 acting as the bias term
- x_i is a column vector of the i^{th} row of $\mathbf{X}$
- $y \in \mathcal{R}^{n\times 1}$ is a column vector of labels $y_i \in \{0, 1\}$

- p is the dimension of data (number of features in each observation)

$$\log \frac{P(y_i=1|x_i,w)}{1-P(y_i=1|x_i,w)} = w^T x_i$$

$$\frac{P(y_i=1|x_i,w)}{1-P(y_i=1|x_i,w)} = e^{w^T x_i} = \frac{\frac{e^{w^T x_i}}{1+e^{w^T x_i}}}{\frac{1}{1+e^{w^T x_i}}}$$

$$L(w) = \log P(y|X,w)$$

$$P(y|X,w) = \prod_{i=1}^{n} \left(P(y_i=1|x_i,w)\right)^{y_i} \left(1-P(y_i=1|x_i,w)\right)^{(1-y_i)}$$

$$\Rightarrow \log P(y|X,w) = \sum_{i=1}^{n} \Big[y_i \log\left(P(y_i=1|x_i,w)\right) + (1-y_i)\log\left(1-P(y_i=1|x_i,w)\right)\Big]$$

$$= \sum_{i=1}^{n} \left[y_i \log \frac{P(y_i=1|x_i,w)}{1-P(y_i=1|x_i,w)} + \log\left(1-P(y_i=1|x_i,w)\right)\right]$$

$$= \sum_{i=1}^{n} \left[y_i w^T x_i + \cancel{\log 1} - \log(1+e^{w^T x_i})\right]$$

$$= \sum_{i=1}^{n} \left[y_i w^T x_i - \log(1+e^{w^T x_i})\right] \#$$

4.2 Convex Optimization

Our goal is to find the weight vector w that maximizes this likelihood. Unfortunately, for this model, we cannot derive a closed-form solution with MLE. An alternative way to solve for w is to use gradient ascent, and update w step by step towards the optimal w. But we know gradient ascent will converge to the optimal solution w that maximizes the conditional log likelihood $\mathcal{L}$ when $\mathcal{L}$ is concave. In this question, you will prove that $\mathcal{L}$ is indeed a concave function.

1. **[3 points]** A real-valued function $f : S \to \mathcal{R}$ defined on a convex set S, is said to be *convex* if,

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2), \forall x_1, x_2 \in S, \forall t \in [0,1].$$

Show that a linear combination of n convex functions, $f_1, f_2, ..., f_n$, $\sum_{i=1}^{n} a_i f_i(x)$ is also a convex function $\forall a_i \in R^+$.

$$\sum_{i=1}^{n} a_i f_i(x) = \sum_{i=1}^{n} a_i f_i(tx_1 + (1-t)x_2)$$

$$= a_1 f_1(tx_1 + (1-t)x_2) + \dots + a_n f_n(tx_1 + (1-t)x_2)$$

$$\xrightarrow{a_i \in R^+} \leq a_1(tf_1(x_1) + (1-t)f_1(x_2)) + \dots + a_n(tf_n(x_1) + (1-t)f_n(x_2))$$

$$= t(a_1 f_1(x_1) + \dots + a_n f_n(x_1)) + (1-t)(a_1 f_1(x_2) + \dots + a_n f_n(x_2))$$

$$= \sum_{i=1}^{n} a_i (t f_i(x_1) + (1-t) f_i(x_2))$$

$f_1, f_2, \dots, f_n$ are convex $\Rightarrow \sum_{i=1}^{n} f_i$ is convex

a_i is nonnegative

$\sum_{i=1}^{n} a_i f_i(x)$ is also convex.

2. **[2 points]** Show that a linear combination of n concave functions, $f_1, f_2, ..., f_n$, $\sum_{i=1}^{n} a_i f_i(x)$ is also a concave function $\forall a_i \in R^+$. Recall that if a function $f(x)$ is convex, then $-f(x)$ is concave. (You can use the result from part (1))

$f(x)$ is convex when

$f(tx_1+(1-t)x_2) \le tf(x_1)+(1-t)f(x_2)$

$\xrightarrow{x(-1)} -f(tx_1+(1-t)x_2) \ge -tf(x_1)-(1-t)f(x_2)$

Let $g(x) = -f(x)$ is concave

$\Rightarrow g(tx_1+(1-t)x_2) \ge tg(x_1)+(1-t)g(x_2)$

$\Rightarrow g(x)$ is concave

$\Rightarrow$ the sum of n concave function with a nonnegative a_i $\left(\sum_{i=1}^{n} a_i f_i(x)\right)$ is also concave.

3. **[4 points]** Another property of twice differentiable convex functions is that the second derivative is non-negative. Using this property, show that $f(x) = \log(1 + \exp x)$ is a convex function. Note that this property is both sufficient and necessary. i.e. (if $f''(x)$ exists, then $f''(x) \geq 0 \iff f$ is convex)

$$f(x) = \log(1+e^x)$$

$$\xrightarrow{d} f'(x) = \frac{1}{1+e^x} \times e^x = e^x(1+e^x)^{-1}$$

$$\xrightarrow{d} f''(x) = e^x(1+e^x)^{-1} - (e^x)^2(1+e^x)^{-2}$$

$$= \frac{e^{2x}+e^x-e^{2x}}{(1+e^x)^2} = \frac{e^x}{(1+e^x)^2} \text{ (exists)}$$

$$\Rightarrow f''(x) = \frac{e^x}{(1+e^x)^2} \geq 0$$

$\Rightarrow f$ is convex #

4. **[4 points]** Let $f_i : \mathcal{S} \to \mathcal{R}$ for $i = 1, \ldots, n$ be a set of convex functions. Is $f(x) = max_i f_i(x)$ also convex? If yes, prove it. If not, provide a counterexample.

Yes

$f(tx_1 + (1-t)x_2) = f_i(tx_1 + (1-t)x_2)$ for $i = 1, \ldots, n$

$\because f_i$ is convex

$\Rightarrow f_i(tx_1 + (1-t)x_2) \leq t f_i(x_1) + (1-t) f_i(x_2)$

$\leq t \max_i f_i(x_1) + (1-t) \max_i f_i(x_2)$

$= t f(x_1) + (1-t) f(x_2)$

$\Rightarrow f(tx_1 + (1-t)x_2) \leq t f(x_1) + (1-t) f(x_2)$

$\Rightarrow f(x) = \max_i f_i(x)$ also convex

5. **[8 points]** Show that the log likelihood of *Logistic Regression* is a concave function. You may use the fact that if f and g are both convex, twice differentiable and g is non-decreasing, then $g \circ f$ is convex.

$$LL(y|X;w) = \sum_{i=1}^{n} [\underbrace{y_i w^T X_i}_{①} - \underbrace{\ln(1+e^{w^T X_i})}_{②}]$$

① $y_i w^T X_i$ is an affine function

→ is both convex and concave

② from Q4.2.3, we proved that $f(x) = \log(1+e^x)$ is convex

→ $\ln(1+e^{w^T X_i})$ is also convex. (let X be $w^T X_i$ here).

from Q4.2.2. $g = -f(x) = -\ln(1+e^{w^T X_i})$ is concave

⇒ $LL(y|X;w)$ combines the summation of two concave functions

⇒ is also concave.