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

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

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 $X_{\cdot i}, \mathbf{y}, w_i, \lambda$

$$J_\lambda(w) = \tfrac{1}{2}(\vec{y} - \vec{X}\vec{w})^T(\vec{y} - \vec{X}\vec{w}) + \lambda \Sigma |w_i|$$

$$= \tfrac{1}{2}\left[\vec{y}^T\vec{y} - \vec{y}^T(\vec{X}\vec{w}) - (\vec{X}\vec{w})^T\vec{y} + (\vec{X}\vec{w})^T(\vec{X}\vec{w})\right] + \lambda \Sigma |w_i|$$

$$= \tfrac{1}{2}(\vec{y}^T\vec{y}) + \tfrac{1}{2}(\vec{X}\vec{w})^T(\vec{X}\vec{w}) - \vec{y}^T(\vec{X}\vec{w}) + \lambda \Sigma |w_i|$$

$$= \underbrace{\tfrac{1}{2}(\vec{y}^T\vec{y})}_{g(\vec{y})} + \underbrace{\sum_{i=1}^{d}\left[\tfrac{1}{2}(X_{\cdot i}\vec{w})^T(X_i\vec{w}) - \vec{y}^T(X_i\vec{w}) + \lambda |w_i|\right]}_{f(X_i, \vec{y}, w_i, \lambda)}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$\Rightarrow g(x)$ 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