CS 189: Introduction to Machine Learning - Discussion 4

1. Norms

   (a) Assuming $x \in \mathbb{R}^n$, define the ℓ_p norm, $\|x_p\|$

   (b) What is the ℓ_0 norm, qualitatively?

   (c) The ℓ_1 norm is often used in sparse machine learning (e.g. bag of words model). Explain with a picture why the ℓ_1 norm often produces sparse results.

Solution:

(a) $\|x\|_p = \sqrt[p]{\sum_i^n |x_i|^p}$

(b) Number of nonzero elements in x.

(c) Taken from the lecture slides:

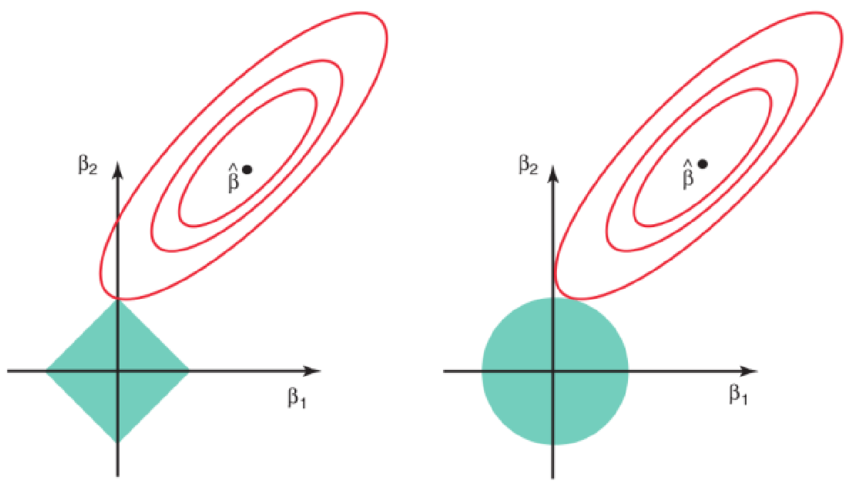

FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \le s$ *and* $\beta_1^2 + \beta_2^2 \le s$*, while the red ellipses are the contours of the RSS.*

2. Ridge Regression with Laplace prior

   As we discussed in class, linear regression is a model of the form $P(y|\mathbf{x}, \sigma^2) \sim \mathcal{N}(\mathbf{w^T x}, \sigma^2)$. The reason that the MLE can overfit is that it is picking the parameter values that are the best for modeling the training data; but if the data is noisy, such parameters often result in complex functions. We can assume some prior distribution on parameters $\mathbf{w}$. Now we assume the prior is Laplace distribution, $w_j \sim Laplace(0, t)$, i.e. $P(w_j) = \frac{1}{2t}e^{-|w_j|/t}$ and $P(\mathbf{w}) = \prod_{j=1}^{D} P(w_j) = (\frac{1}{2t})^D \cdot e^{-\frac{\sum |w_j|}{t}}$

   Show it is equivalent to minimizing the following and find the constant λ. ($\|\mathbf{w}\|_1 = \sum_{j=1}^{D} |w_j|$)

$$J(\mathbf{w}) = \sum_{i=1}^{n} (Y_i - \mathbf{w^T X_i})^2 + \lambda \|\mathbf{w}\|_1$$

Solution: We have to solve the MAP for parameter $\mathbf{w}$ and the posterior of $\mathbf{w}$ is,

$$P(w|\mathbf{X_i}, Y_i) \propto (\prod_{i=1}^{n} \mathcal{N}(Y_i|\mathbf{w^T X_i}, \sigma^2)) \cdot P(\mathbf{w}) = (\prod_{i=1}^{n} \mathcal{N}(Y_i|\mathbf{w^T X_i}, \sigma^2)) \cdot \prod_{j=1}^{D} P(w_j)$$

Taking log and we want to maximize

$$\begin{aligned} l(\mathbf{w}) &= \sum_{i=1}^{n} log\mathcal{N}(Y_i|\mathbf{w^T X_i}, \sigma^2) + \sum_{j=1}^{D} logP(w_j) \\ &= \sum_{i=1}^{n} log(\frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(Y_i - \mathbf{w^T X_i})^2}{2\sigma^2})) + \sum_{j=1}^{D} log(\frac{1}{2t} exp(\frac{-|w_j|}{t})) \\ &= -\sum_{i=1}^{n} \frac{(Y_i - \mathbf{w^T X_i})^2}{2\sigma^2} + \frac{-\sum_{j=1}^{D} |w_j|}{t} + nlog(\frac{1}{\sqrt{2\pi}\sigma}) + Dlog(\frac{1}{2t}) \end{aligned}$$

So it is equivalent to minimize the following function

$$J(\mathbf{w}) = \sum_{i=1}^{n} (Y_i - \mathbf{w^T X_i})^2 + \frac{2\sigma^2}{t} \sum_{j=1}^{D} |w_j| = \sum_{i=1}^{n} (Y_i - \mathbf{w^T X_i})^2 + \lambda \|\mathbf{w}\|_1$$

where $\lambda = \frac{2\sigma^2}{t}$.

3. Weighted Least Squares

In our traditional least squares scenario, we minimize the least squares error, or:

$$L(\beta) = \sum_{i=1}^{n} (y_i - \beta^T \vec{x}_i)^2$$

A generalization of this scenario is one where we minimize a sum of weighted errors, where some training points may have more weight than others. Given some weight vector, $[w_1, w_2, \ldots, w_n]^T$,

$$L(\beta) = \sum_{i=1}^{n} w_i (y_i - \beta^T \vec{x}_i)^2$$

Find the value of β that minimizes the weighted least-squares error. Your answer should be in matrix form.

Solution: We can vectorize this summation and show that

$$L(\beta) = (Y - X\beta)^T W (Y - X\beta)$$

where $Y = [y_1, y_2, \ldots, y_n]^T$, $X = [\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n]^T$, and W is a diagonal matrix of the weights.

Expanding this equation:

$$L(\beta) = (Y^T - \beta^T X^T)(WY - WX\beta)$$

$$L(\beta) = Y^T WY - Y^T WX\beta - \beta^T X^T WY + \beta^T X^T WX\beta$$

Taking the derivative of this quantity, we get:

$$\frac{dL(\beta)}{d\beta} = -Y^T WX - X^T WY + 2X^T WX\beta = 0$$

$$\frac{dL(\beta)}{d\beta} = -2X^T WY + 2X^T WX\beta = 0$$

Solving for β:

$$\hat{\beta} = (X^T WX)^{-1} X^T WY$$