STAT561 Lab 3

February 4, 2019

1 Sparse Regression

We call $b_d(x, \epsilon)$ the ϵ-ball around x with the distance metric d. That is,

$$b_d(x, \epsilon) = \{y : d(x, y) \leq \epsilon\}$$

1. Let d be the l_1 metric. Draw the boundary of the ball around $(0, 0)$ with $\epsilon = 1$.

2. Now let d be the l_2 metric. Draw the boundary of the ball around $(0, 0)$ with $\epsilon = 1$.

3. Which of the following is equivalent to $\arg\min_\beta \frac{1}{n}\sum_{i=1}^n (y_i - \beta^T x_i)^2 + \lambda||\beta||_1$?

   a) $\arg\min_\beta \frac{1}{n}\sum_{i=1}^n (y_i - \beta^T x_i)^2$ s.t. $||\beta||_1 \geq \tau$

   b) $\arg\min_\beta \frac{1}{n}\sum_{i=1}^n (y_i - \beta^T x_i)^2$ s.t. $||\beta||_1 \leq \tau$

Note that as the dimensionality of the cube $b_{l_1}(0, \epsilon)$ increases, its volume becomes increasingly concentrated in the corners. This means that in high dimensional settings, the cube becomes quite pointy or spindly, with its corners protruding along the coordinate axes. To see why, consider the case of the cube circumscribing a sphere and determine how p varies as a function of the dimensionality, where p is the probability that a point uniformly drawn from the cube will fall within the sphere.

Figure from *Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman

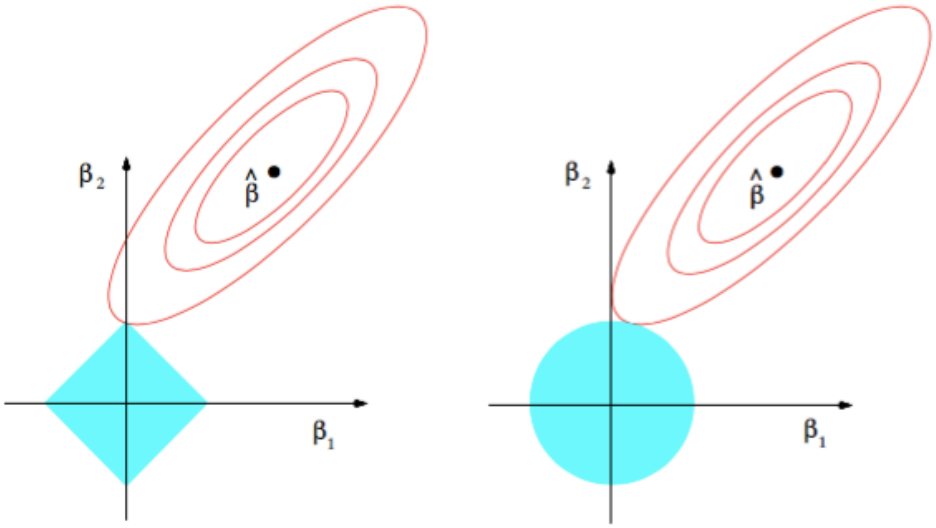

FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

2 Sparsity Computational Example

Regression with polynomial basis:
`https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-ridge-lasso-regression-python/`

Regression with Gaussian basis: `https://github.com/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.06-Linear-Regression.ipynb`

3 Gaussian Process Regression

Recall the definition for a Gaussian process.

Definition. *A stochastic process over domain $\mathcal{X}$ with mean function μ and covariance kernel K is a Gaussian process if and only if for any $\{x_1, ..., x_n\} \in \mathcal{X}$ and $n \in \mathbb{N}$ the distribution of $\mathbf{f} = \{f(x_1), ..., f(x_n)\}^T$ is*

$$\mathbf{f} = \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \sim N\left(\begin{bmatrix} \mu(x_1) \\ \vdots \\ \mu(x_n) \end{bmatrix}, \begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & \ddots & \vdots \\ K(x_1, x_n) & \cdots & K(x_n, x_n) \end{bmatrix} \right).$$

Suppose we have $D = \{(x_i, y_i)\}_{i=1}^n$. Here we assume

$$y_i = f(x_i) + \varepsilon, \varepsilon \sim N(0, \sigma^2)$$

and we place a prior on f in the form of

$$f \sim \mathcal{G}(\mu(\cdot), K(\cdot, \cdot))$$

To do inference, we only need to specifiy the mean function $\mu(\cdot)$ and the covariance kernel $K(\cdot, \cdot)$. Once we specify these, we know our distribution over $\mathbf{f}$.

1. What should we set $\mu(\cdot)$ to if $\frac{1}{n}\sum_{i=1}^n y_i = 5$?

2. We get K_{ij} from $k(x_i, x_j)$, with the requirement that K is positive definite. What does it mean for K to be positive definite? Why must K be positive definite?

3. k must represent a valid inner product in some space $\mathcal{X}_\phi$ where the vectors $\phi(x)$ live. A common choice of k is

   $$k(x_i, x_j) = \nu \exp\left(\frac{-||x_i - x_j||^2}{2l^2}\right)$$

   where ν and l are hyperparameters. How might we interpret ν and l?

4. Would obtaining K as a diagonal matrix be a good or bad idea?

5. Consider the joint Gaussian distribution,

$$p(\mathbf{f},\mathbf{g}) = \mathcal{N}\left(\begin{bmatrix}\mathbf{f}\\ \mathbf{g}\end{bmatrix};\ \begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix},\ \begin{bmatrix}A & C\\ C^\top & B\end{bmatrix}\right)$$

We can directly obtain the conditional distribution as

$$p(\mathbf{f}\,|\,\mathbf{g}) = \mathcal{N}(\mathbf{f};\ \mathbf{a}+CB^{-1}(\mathbf{g}-\mathbf{b}),\ A-CB^{-1}C^\top)$$

In the context of our regression, we have the joint distribution

$$\begin{bmatrix}\mathbf{Y}\\ \mathbf{Y}^*\end{bmatrix}\Big|\,\mathbf{X}^*,\mathbf{X} = \begin{bmatrix}\mathbf{f}\\ \mathbf{f}^*\end{bmatrix}+\begin{bmatrix}\varepsilon\\ \varepsilon^*\end{bmatrix} \sim \mathrm{N}\left(\mathbf{0},\begin{bmatrix}K(\mathbf{X},\mathbf{X})+\sigma^2\mathbf{I} & K(\mathbf{X},\mathbf{X}^*)\\ K(\mathbf{X}^*,\mathbf{X}) & K(\mathbf{X}^*,\mathbf{X}^*)+\sigma^2\mathbf{I}\end{bmatrix}\right)$$

Determine the predictive distribution over some test points $\mathbf{Y}^*$, given data $\mathbf{X}^*, \mathbf{Y}, \mathbf{X}$.

4 Gaussian Process Computational Example

`http://shogun-toolbox.org/notebook/latest/gaussian_processes.html`