رگرسیون خطی

Linear Regression

a.fahim@ut.ac.ir

متغیر مستقل تصادفی (Random Independent Variable) که شامل d مشخصه‌ی مشاهده شده در نمونه‌ها می‌شود: $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)^T$
متغیر وابسته (Dependent Variable) و یا پاسخ که هدف از رگرسیون تعیین آن از روی مشاهدات می‌باشد: $\boldsymbol{Y}$
تابع رگرسیون که مدل مورد استفاده برای تعیین متغیر وابسته است: f
جمله‌ی خطای تصادفی (Random Error Term) که از مشخصه‌ها و مشاهدات مستقل هستند. این جمله شامل و اثر تمام عدم قطعیت‌ها و نیز متغیرهای پنهانی است که مشاهده نمی‌شوند.

$$Y = f(X_1, X_2, \cdots, X_d) + \varepsilon = f(\mathbf{X}) + \varepsilon$$

$$f(\mathbf{X}) = \beta + \omega_1 X_1 + \omega_2 X_2 + \cdots + \omega_d X_d = \beta + \sum_{i=1}^{d} \omega_i X_i = \beta + \boldsymbol{\omega}^T \mathbf{X}$$

مقدار پیش‌قدر β و ضرایب رگرسیون $\boldsymbol{\omega}$ واقعی برای ما نامشخص هستند.
با استفاده از مجموعه داده‌ی یادگیری $\boldsymbol{D}$ و مقادیر پاسخ متناظر هر داده، β و $\boldsymbol{\omega}$ را برآورد می‌کنیم. برآوردگرهای متناظر b و $\boldsymbol{w}$ هستند.

$$\hat{y} = b + w_1 x_1 + \cdots + w_d x_d = b + \mathbf{w}^T \mathbf{x}$$

خطای باقیمانده (Residual Error) $\epsilon = y - \hat{y} = y - b - \mathbf{w}^T \mathbf{x}$

خطای باقیمانده ϵ با خطای آماری تصادفی ε متفاوت است. ϵ تفاوت بین پاسخ واقعی (ناشناخته) و مشاهده شده را اندازه گیری می‌کند. خطای باقیمانده ϵ برآوردگری برای خطای تصادفی ε است.

$$SSE = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b - \mathbf{w}^T \mathbf{x}_i)^2$$

رویکرد کلی برای بدست آوردن b و $\boldsymbol{w}$ کمینه کردن جمع مربعات خطا است.

در رگرسیون دو متغیره، فقط یک متغیر مستقل و یک متغیر وابسته (پاسخ) داریم.

$$\hat{y}_i = f(x_i) = b + w \cdot x_i$$

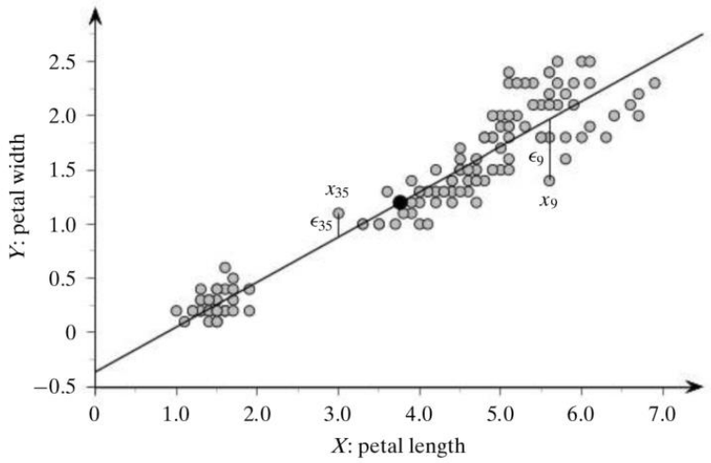
$$\min_{b,w} SSE = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b - w \cdot x_i)^2$$

$$\frac{\partial}{\partial b} SSE = -2 \sum_{i=1}^{n} (y_i - b - w \cdot x_i) = 0$$

$$b = \mu_Y - w \cdot \mu_X$$

$$\frac{\partial}{\partial w} SSE = -2 \sum_{i=1}^{n} x_i (y_i - b - w \cdot x_i) = 0$$

$$w = \frac{\sum_{i=1}^{n} (x_i - \mu_X)(y_i - \mu_Y)}{\sum_{i=1}^{n} (x_i - \mu_X)^2} = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{\text{cov}(X, Y)}{\text{var}(X)}$$



مثال: رگرسیون عرض کاسبرگ‌ها برحسب طول کاسبرگ‌ها در نمونه‌ی گل‌های زنبق

Figure 23.1. Scatterplot: `petal length` (X) versus `petal width` (Y). Solid circle (black) shows the mean point; residual error is shown for two sample points: x_9 and x_{35}.

$$\mu_X = \frac{1}{150} \sum_{i=1}^{150} x_i = \frac{563.8}{150} = 3.7587$$

$$\mu_Y = \frac{1}{150} \sum_{i=1}^{150} y_i = \frac{179.8}{150} = 1.1987$$

$$\sigma_X^2 = \frac{1}{150} \sum_{i=1}^{150} (x_i - \mu_X)^2 = 3.0924$$

$$\sigma_Y^2 = \frac{1}{150} \sum_{i=1}^{150} (y_i - \mu_Y)^2 = 0.5785$$

$$\sigma_{XY} = \frac{1}{150} \sum_{i=1}^{150} (x_i - \mu_X) \cdot (y_i - \mu_Y) = 1.2877$$

$$w = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{1.2877}{3.0924} = 0.4164$$

$$b = \mu_Y - w \cdot \mu_X = 1.1987 - 0.4164 \cdot 3.7587 = -0.3665$$

$$\hat{y} = -0.3665 + 0.4164 \cdot x$$

میانگین‌های μ_x و μ_y روی خط قرار دارند.

در رگرسیون چندگانه برای سادگی پیش‌قدر (Bias) را بردار وزن ادغام می‌کنیم.
برای این منظور به ابعاد مشخصه‌ها یک بعد اضافه می‌کنیم

$$\tilde{\mathbf{x}}_i = (x_{i0}, x_{i1}, x_{i2}, \cdots, x_{id})^T \in \mathbb{R}^{d+1}$$

$\tilde{\mathbf{D}} \in \mathbb{R}^{n \times (d+1)}$ ماتریس داده‌ی افزوده (Augmented Data Matrix)

$$\tilde{\mathbf{w}} = (w_0, w_1, w_2, \cdots, w_d)^T \qquad w_0 = b$$

$$\hat{y}_i = w_0 x_{i0} + w_1 x_{i1} + w_2 x_{i2} + \cdots + w_d x_{id} = \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_i$$

$$\widehat{Y} = \tilde{\mathbf{D}}\tilde{\mathbf{w}} \qquad \widehat{Y} = (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n)^T$$

$$\min_{\tilde{\mathbf{w}}} SSE = \sum_{i=1}^{n} \epsilon_i^2 = \|\boldsymbol{\epsilon}\|^2 = \|Y - \widehat{Y}\|^2$$

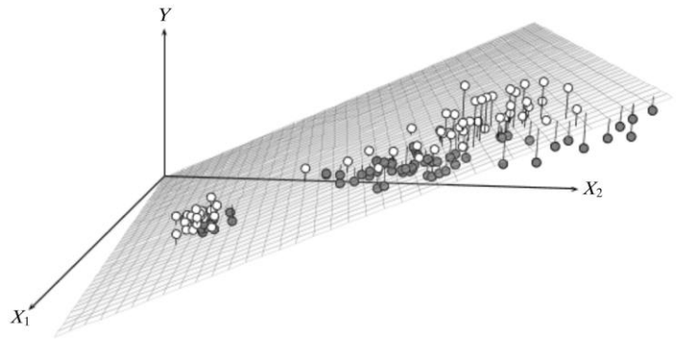
$$= Y^T Y - 2\tilde{\mathbf{w}}^T (\tilde{\mathbf{D}}^T Y) + \tilde{\mathbf{w}}^T (\tilde{\mathbf{D}}^T \tilde{\mathbf{D}}) \tilde{\mathbf{w}}$$

$$\frac{\partial}{\partial \tilde{\mathbf{w}}} SSE = -2\tilde{\mathbf{D}}^T Y + 2(\tilde{\mathbf{D}}^T \tilde{\mathbf{D}})\tilde{\mathbf{w}} = \mathbf{0}$$

$$\tilde{\mathbf{w}} = (\tilde{\mathbf{D}}^T \tilde{\mathbf{D}})^{-1} \tilde{\mathbf{D}}^T Y$$

$$\widehat{Y} = \tilde{\mathbf{D}}\tilde{\mathbf{w}} = \tilde{\mathbf{D}}(\tilde{\mathbf{D}}^T \tilde{\mathbf{D}})^{-1} \tilde{\mathbf{D}}^T Y = \mathbf{H} Y$$

$\mathbf{H} = \tilde{\mathbf{D}}(\tilde{\mathbf{D}}^T \tilde{\mathbf{D}})^{-1} \tilde{\mathbf{D}}^T$ ماتریس کلاه (Hat Matrix)



Example 23.3 (Multiple Regression). Figure 23.5 shows the multiple regression of `sepal length` (X_1) and `petal length` (X_2) on the response attribute `petal width` (Y) for the Iris dataset with $n = 150$ points. We first add an extra attribute $X_0 = \mathbf{1}_{150}$, which is a vector of all ones in $\mathbb{R}^{150}$. The augmented dataset $\tilde{\mathbf{D}} \in \mathbb{R}^{150 \times 3}$ comprises $n = 150$ points along three attributes X_0, X_1, and X_2.

Next, we compute the uncentered 3×3 scatter matrix $\tilde{\mathbf{D}}^T\tilde{\mathbf{D}}$ and its inverse

$$\tilde{\mathbf{D}}^T\tilde{\mathbf{D}} = \begin{pmatrix} 150.0 & 876.50 & 563.80 \\ 876.5 & 5223.85 & 3484.25 \\ 563.8 & 3484.25 & 2583.00 \end{pmatrix} \quad (\tilde{\mathbf{D}}^T\tilde{\mathbf{D}})^{-1} = \begin{pmatrix} 0.793 & -0.176 & 0.064 \\ -0.176 & 0.041 & -0.017 \\ 0.064 & -0.017 & 0.009 \end{pmatrix}$$

We also compute $\tilde{\mathbf{D}}^T Y$, given as

$$\tilde{\mathbf{D}}^T Y = \begin{pmatrix} 179.80 \\ 1127.65 \\ 868.97 \end{pmatrix}$$

The augmented weight vector $\tilde{\mathbf{w}}$ is then given as

$$\tilde{\mathbf{w}} = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix} = (\tilde{\mathbf{D}}^T\tilde{\mathbf{D}})^{-1} \cdot (\tilde{\mathbf{D}}^T Y) = \begin{pmatrix} -0.014 \\ -0.082 \\ 0.45 \end{pmatrix}$$

The bias term is therefore $b = w_0 = -0.014$, and the fitted model is

$$\widehat{Y} = -0.014 - 0.082 \cdot X_1 + 0.45 \cdot X_2$$

Figure 23.5 shows the fitted hyperplane. It also shows the residual error for each point. The white colored points have positive residuals (i.e., $\epsilon_i > 0$ or $\hat{y}_i > y_i$), whereas the gray points have negative residual values (i.e., $\epsilon_i < 0$ or $\hat{y}_i < y$). The SSE value for the model is 6.18.

بردار پیشبینی شده‌ی $\hat{Y}$ در فضای ستونی ماتریس افزوده‌ی $\tilde{\boldsymbol{D}}$ قرار می‌گیرد. در واقع خطای پیشبینی تصویر عمودی Y به زیرفضای $col(\tilde{D})$ می‌باشد و لذا به هر یک از مشخصه‌های مجموعه داده‌ها (بردارهای ستونی) عمود است. از این تعامد $d+1$ معادله بوجود می‌آید.

$$\hat{Y} = \tilde{\mathbf{D}}\tilde{\mathbf{w}} \qquad \boldsymbol{\epsilon} = Y - \hat{Y} \qquad X_i^T \boldsymbol{\epsilon} = 0$$

بردارهای ستونی $\tilde{\boldsymbol{D}}$ حتی اگر فرض کنیم که استقلال خطی هم داشته باشند، لزوما بر هم عمود نیستند. لذا ابتدا با روش گرام-اشمیت (Gram-Schmidt Orthogonalization) از روی آن‌ها یک پایه‌ی متعامد می‌سازیم. این روش منجر به تجذیه‌ی QR ماتریس داده‌های افزوده $\tilde{\boldsymbol{D}}$ می‌شود.

$$\begin{aligned}
U_0 &= X_0 \\
U_1 &= X_1 - p_{10} \cdot U_0 \\
U_2 &= X_2 - p_{20} \cdot U_0 - p_{21} \cdot U_1 \qquad p_{ji} = \text{proj}_{U_i}(X_j) = \frac{X_j^T U_i}{\|U_i\|^2} \\
&\vdots \quad = \qquad\qquad \vdots \\
U_d &= X_d - p_{d0} \cdot U_0 - p_{d1} \cdot U_1 - \cdots - p_{d,d-1} \cdot U_{d-1}
\end{aligned}$$

$$\underbrace{\begin{pmatrix} | & | & | & & | \\ X_0 & X_1 & X_2 & \cdots & X_d \\ | & | & | & & | \end{pmatrix}}_{\tilde{\mathbf{D}}} = \underbrace{\begin{pmatrix} | & | & | & & | \\ U_0 & U_1 & U_2 & \cdots & U_d \\ | & | & | & & | \end{pmatrix}}_{\mathbf{Q}} \cdot \underbrace{\begin{pmatrix} 1 & p_{10} & p_{20} & \cdots & p_{d0} \\ 0 & 1 & p_{21} & \cdots & p_{d1} \\ 0 & 0 & 1 & \cdots & p_{d2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 1 & p_{d,d-1} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}}_{\mathbf{R}}$$

$$\tilde{\mathbf{D}} = \mathbf{QR},$$

بردارهای U_i پایه‌های متعامد زیرفضای ستونی $col(\tilde{\boldsymbol{D}})$ می‌باشد.

$$\hat{Y} = \text{proj}_{U_0}(Y) \cdot U_0 + \text{proj}_{U_1}(Y) \cdot U_1 + \cdots + \text{proj}_{U_d}(Y) \cdot U_d$$

$$\mathbf{Q}^T\mathbf{Q} = \begin{pmatrix} \|U_0\|^2 & 0 & \cdots & 0 \\ 0 & \|U_1\|^2 & \cdots & 0 \\ & & & \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots \| & U_d\|^2 \end{pmatrix} = \boldsymbol{\Delta}$$

Algorithm 23.1: Multiple Regression Algorithm

MULTIPLE-REGRESSION ($\mathbf{D}$, Y):

1 $\tilde{\mathbf{D}} \leftarrow \begin{pmatrix} \mathbf{1} & \mathbf{D} \end{pmatrix}$ // augmented data with $X_0 = \mathbf{1} \in \mathbb{R}^n$

2 $\{\mathbf{Q}, \mathbf{R}\} \leftarrow$ QR-factorization($\tilde{\mathbf{D}}$) // $\mathbf{Q} = \begin{pmatrix} U_0 & U_1 & \cdots & U_d \end{pmatrix}$

3 $\boldsymbol{\Delta}^{-1} \leftarrow \begin{pmatrix} \frac{1}{\|U_0\|^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\|U_1\|^2} & \cdots & 0 \\ & & & \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \frac{1}{\|U_d\|^2} \end{pmatrix}$ // reciprocal squared norms

4 $\mathbf{Rw} \leftarrow \boldsymbol{\Delta}^{-1}\mathbf{Q}^T Y$ // solve for $\mathbf{w}$ by back-substitution

5 $\hat{Y} \leftarrow \mathbf{Q}\boldsymbol{\Delta}^{-1}\mathbf{Q}^T Y$

$$w_0 + 5.843 \cdot w_1 + 3.759 \cdot w_2 = 1.1987$$
$$\Longrightarrow w_0 = 1.1987 + 0.4786 - 1.6911 = -0.0139$$

Thus, the multiple regression model is given as

$$\widehat{Y} = -0.014 \cdot X_0 - 0.082 \cdot X_1 + 0.45 \cdot X_2 \tag{23.27}$$

which matches the model in Example 23.3.

It is also instructive to construct the new basis vectors $U_0, U_1, \cdots, U_d$ in terms of the original attributes $X_0, X_1, \cdots, X_d$. Since $\tilde{\mathbf{D}} = \mathbf{QR}$, we have $\mathbf{Q} = \tilde{\mathbf{D}}\mathbf{R}^{-1}$. The inverse of $\mathbf{R}$ is also upper-triangular, and is given as

$$\mathbf{R}^{-1} = \begin{pmatrix} 1 & -5.843 & 7.095 \\ 0 & 1 & -1.858 \\ 0 & 0 & 1 \end{pmatrix}$$

Therefore, we can write $\mathbf{Q}$ in terms of the original attributes as

$$\underbrace{\begin{pmatrix} | & | & | \\ U_0 & U_1 & U_2 \\ | & | & | \end{pmatrix}}_{\mathbf{Q}} = \underbrace{\begin{pmatrix} | & | & | \\ X_0 & X_1 & X_2 \\ | & | & | \end{pmatrix}}_{\tilde{\mathbf{D}}} \underbrace{\begin{pmatrix} 1 & -5.843 & 7.095 \\ 0 & 1 & -1.858 \\ 0 & 0 & 1 \end{pmatrix}}_{\mathbf{R}^{-1}}$$

which results in

$$U_0 = X_0$$
$$U_1 = -5.843 \cdot X_0 + X_1$$
$$U_2 = 7.095 \cdot X_0 - 1.858 \cdot X_1 + X_2$$

The scalar projection of the response vector Y onto each of the new basis vectors yields:

$$\text{proj}_{U_0}(Y) = 1.199 \qquad \text{proj}_{U_1}(Y) = 0.754 \qquad \text{proj}_{U_2}(Y) = 0.45$$

Finally, the fitted response vector is given as:

$$\begin{aligned} \widehat{Y} &= \text{proj}_{U_0}(Y) \cdot U_0 + \text{proj}_{U_1}(Y) \cdot U_1 + \text{proj}_{U_2}(Y) \cdot U_2 \\ &= 1.199 \cdot X_0 + 0.754 \cdot (-5.843 \cdot X_0 + X_1) + 0.45 \cdot (7.095 \cdot X_0 - 1.858 \cdot X_1 + X_2) \\ &= (1.199 - 4.406 + 3.193) \cdot X_0 + (0.754 - 0.836) \cdot X_1 + 0.45 \cdot X_2 \\ &= -0.014 \cdot X_0 - 0.082 \cdot X_1 + 0.45 \cdot X_2 \end{aligned}$$

which matches Eq. (23.27).

Example 23.4 (Multiple Regression: QR-Factorization and Geometric Approach). Consider the multiple regression of `sepal length` (X_1) and `petal length` (X_2) on the response attribute `petal width` (Y) for the Iris dataset with $n = 150$ points, as shown in Figure 23.5. The augmented dataset $\tilde{\mathbf{D}} \in \mathbb{R}^{150 \times 3}$ comprises $n = 150$ points along three attributes X_0, X_1, and X_2, where $X_0 = \mathbf{1}$. The Gram–Schmidt orthogonalization results in the following QR-factorization:

$$\underbrace{\begin{pmatrix} | & | & | \\ X_0 & X_1 & X_2 \\ | & | & | \end{pmatrix}}_{\tilde{\mathbf{D}}} = \underbrace{\begin{pmatrix} | & | & | \\ U_0 & U_1 & U_2 \\ | & | & | \end{pmatrix}}_{\mathbf{Q}} \cdot \underbrace{\begin{pmatrix} 1 & 5.843 & 3.759 \\ 0 & 1 & 1.858 \\ 0 & 0 & 1 \end{pmatrix}}_{\mathbf{R}}$$

Note that $\mathbf{Q} \in \mathbb{R}^{150 \times 3}$ and therefore we do not show the matrix. The matrix $\mathbf{\Delta}$, which records the squared norms of the basis vectors, and its inverse matrix, is given as

$$\mathbf{\Delta} = \begin{pmatrix} 150 & 0 & 0 \\ 0 & 102.17 & 0 \\ 0 & 0 & 111.35 \end{pmatrix} \qquad \mathbf{\Delta}^{-1} = \begin{pmatrix} 0.00667 & 0 & 0 \\ 0 & 0.00979 & 0 \\ 0 & 0 & 0.00898 \end{pmatrix}$$

We can use back-substitution to solve for $\tilde{\mathbf{w}}$, as follows

$$\mathbf{R}\tilde{\mathbf{w}} = \mathbf{\Delta}^{-1}\mathbf{Q}^T Y$$
$$\begin{pmatrix} 1 & 5.843 & 3.759 \\ 0 & 1 & 1.858 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} 1.1987 \\ 0.7538 \\ 0.4499 \end{pmatrix}$$

In back-substitution, we start with w_2, which is easy to compute from the equation above; it is simply

$$w_2 = 0.4499$$

Next, w_1 is given as:

$$w_1 + 1.858 \cdot w_2 = 0.7538$$
$$\Longrightarrow w_1 = 0.7538 - 0.8358 = -0.082$$

Finally, w_0 can be computed as

نزول شیب تصادفی (Stochastic Gradient Descent)

$$\min_{\tilde{\mathbf{w}}} SSE = \frac{1}{2}\left(Y^T Y - 2\tilde{\mathbf{w}}^T(\tilde{\mathbf{D}}^T Y) + \tilde{\mathbf{w}}^T(\tilde{\mathbf{D}}^T\tilde{\mathbf{D}})\tilde{\mathbf{w}}\right)$$

$$\tilde{\mathbf{w}}^{t+1} = \tilde{\mathbf{w}}^t - \eta \cdot \nabla_{\tilde{\mathbf{w}}} = \tilde{\mathbf{w}}^t + \eta \cdot \tilde{\mathbf{D}}^T(Y - \tilde{\mathbf{D}} \cdot \tilde{\mathbf{w}}^t)$$

در نزول شیب تصادفی (SGD) بردارهای وزن را با استفاده از یک نقطه‌ی تصادفی در هر دفعه به روز می‌کنیم.

$$\nabla_{\tilde{\mathbf{w}}}(\tilde{\mathbf{x}}_k) = -\tilde{\mathbf{x}}_k y_k + \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T \tilde{\mathbf{w}} = -(y_k - \tilde{\mathbf{x}}_k^T \tilde{\mathbf{w}})\tilde{\mathbf{x}}_k$$

$$\begin{aligned}\tilde{\mathbf{w}}^{t+1} &= \tilde{\mathbf{w}}^t - \eta \cdot \nabla_{\tilde{\mathbf{w}}}(\tilde{\mathbf{x}}_k) \\ &= \tilde{\mathbf{w}}^t + \eta \cdot (y_k - \tilde{\mathbf{x}}_k^T \tilde{\mathbf{w}}^t) \cdot \tilde{\mathbf{x}}_k\end{aligned}$$

Algorithm 23.2: Multiple Regression: Stochastic Gradient Descent

MULTIPLE REGRESSION: SGD ($\mathbf{D}, Y, \eta, \epsilon$):

$\tilde{\mathbf{D}} \leftarrow (\mathbf{1} \quad \mathbf{D})$ // augment data
$t \leftarrow 0$ // step/iteration counter
$\tilde{\mathbf{w}}^t \leftarrow$ random vector in $\mathbb{R}^{d+1}$ // initial weight vector
repeat
  foreach $k = 1, 2, \cdots, n$ *(in random order)* **do**
    $\nabla_{\tilde{\mathbf{w}}}(\tilde{\mathbf{x}}_k) \leftarrow -(y_k - \tilde{\mathbf{x}}_k^T \tilde{\mathbf{w}}^t) \cdot \tilde{\mathbf{x}}_k$ // compute gradient at $\tilde{\mathbf{x}}_k$
    $\tilde{\mathbf{w}}^{t+1} \leftarrow \tilde{\mathbf{w}}^t - \eta \cdot \nabla_{\tilde{\mathbf{w}}}(\tilde{\mathbf{x}}_k)$ // update estimate for $\tilde{\mathbf{w}}$
  $t \leftarrow t + 1$
until $\|\mathbf{w}^t - \mathbf{w}^{t-1}\| \leq \epsilon$

Example 23.5 (Multiple Regression: SGD). We continue Example 23.4 for multiple regression of `sepal length` (X_1) and `petal length` (X_2) on the response attribute `petal width` (Y) for the Iris dataset with $n = 150$ points.

Using the exact approach the multiple regression model was given as

$$\widehat{Y} = -0.014 \cdot X_0 - 0.082 \cdot X_1 + 0.45 \cdot X_2$$

Using stochastic gradient descent we obtain the following model with $\eta = 0.001$ and $\epsilon = 0.0001$:

$$\widehat{Y} = -0.031 \cdot X_0 - 0.078 \cdot X_1 + 0.45 \cdot X_2$$

The results from the SGD approach as essentially the same as the exact method, with a slight difference in the bias term. The SSE value for the exact method is 6.179, whereas for SGD it is 6.181.

جمله‌ی تنظیم کننده (Regularization Term) جهت عمومیت بخشیدن به نتایج

$$\min_{\tilde{\mathbf{w}}} J(\tilde{\mathbf{w}}) = \|Y - \widehat{Y}\|^2 + \alpha \cdot \|\tilde{\mathbf{w}}\|^2 = \|Y - \tilde{\mathbf{D}}\tilde{\mathbf{w}}\|^2 + \alpha \cdot \|\tilde{\mathbf{w}}\|^2$$

$$\frac{\partial}{\partial \tilde{\mathbf{w}}} J(\tilde{\mathbf{w}}) = \frac{\partial}{\partial \tilde{\mathbf{w}}} \left\{ \|Y - \tilde{\mathbf{D}}\tilde{\mathbf{w}}\|^2 + \alpha \cdot \|\tilde{\mathbf{w}}\|^2 \right\} = \mathbf{0}$$

$$\tilde{\mathbf{w}} = (\tilde{\mathbf{D}}^T\tilde{\mathbf{D}} + \alpha \cdot \mathbf{I})^{-1}\tilde{\mathbf{D}}^T Y \qquad \mathbf{I} \in \mathbb{R}^{(d+1)\times(d+1)}$$

اگر $\alpha > 0$ باشد، حتی اگر $\tilde{\mathbf{D}}^T\tilde{\mathbf{D}}$ معکوس‌پذیر نباشد، $(\tilde{\boldsymbol{D}}^T\tilde{\boldsymbol{D}} + \alpha.\boldsymbol{I})$ همواره معکوس‌پذیر است. پس وجود یک α هر چند کوچک وجود پاسخ را تضمین می‌کند.

Example 23.6 (Ridge Regression). Figure 23.6 shows the scatterplot between the two attributes `petal length` (X; the predictor variable) and `petal width` (Y; the response variable) in the Iris dataset. There are a total of $n = 150$ data points. The uncentered scatter matrix is given as

$$\tilde{\mathbf{D}}^T\tilde{\mathbf{D}} = \begin{pmatrix} 150.0 & 563.8 \\ 563.8 & 2583.0 \end{pmatrix}$$

Using Eq. (23.32) we obtain different lines of best fit for different values of the regularization constant α:

$$\alpha = 0 : \widehat{Y} = -0.367 + 0.416 \cdot X, \quad \|\tilde{\mathbf{w}}\|^2 = \|(-0.367, 0.416)^T\|^2 = 0.308, \quad SSE = 6.34$$

$$\alpha = 10 : \widehat{Y} = -0.244 + 0.388 \cdot X, \quad \|\tilde{\mathbf{w}}\|^2 = \|(-0.244, 0.388)^T\|^2 = 0.210, \quad SSE = 6.75$$

$$\alpha = 100 : \widehat{Y} = -0.021 + 0.328 \cdot X, \quad \|\tilde{\mathbf{w}}\|^2 = \|(-0.021, 0.328)^T\|^2 = 0.108, \quad SSE = 9.97$$

Figure 23.6 shows these regularized regression lines. We can see that as α increases there is more emphasis on minimizing the squared norm of $\tilde{\mathbf{w}}$. However, since $\|\tilde{\mathbf{w}}\|^2$ is more constrained as α increases, the fit of the model decreases, as seen from the increase in SSE values.

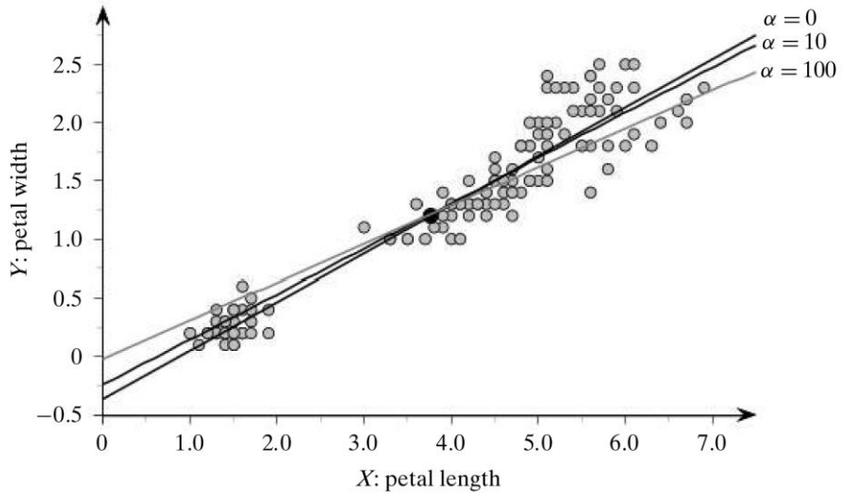


Figure 23.6. Scatterplot: `petal length` (X) versus `petal width` (Y). Ridge regression lines for $\alpha = 0, 10, 100$.

جمله‌ی اریبی بی‌جریمه (Unpenalized Bias Term)

در خیلی از موارد علاقه داریم پارامتر اریبی (Bias) تنظیم نشود و مقدار خود را مستقل از بردار وزن بدست آورد. لذا جمله‌ی $w_0 = b$ را از $\tilde{w}$ جدا می‌کنیم.

$$\min_{\mathbf{w}} \; J(\mathbf{w}) = \|Y - w_0 \cdot \mathbf{1} - \mathbf{D}\mathbf{w}\|^2 + \alpha \cdot \|\mathbf{w}\|^2$$

$$w_0 = b = \mu_Y - \sum_{i=1}^{d} w_i \cdot \mu_{X_i} = \mu_Y - \boldsymbol{\mu}^T \mathbf{w}$$

$$\min_{\mathbf{w}} \; J(\mathbf{w}) = \|\bar{Y} - \bar{\mathbf{D}}\mathbf{w}\|^2 + \alpha \cdot \|\mathbf{w}\|^2$$

$$\bar{\mathbf{D}} = \mathbf{D} - \mathbf{1}\boldsymbol{\mu}^T \qquad \bar{Y} = Y - \mu_Y$$

Example 23.7 (Ridge Regression: Unpenalized Bias). We continue from Example 23.6. When we do not penalize w_0, we obtain the following lines of best fit for different values of the regularization constant α:

$$\alpha = 0 : \widehat{Y} = -0.365 + 0.416 \cdot X \qquad w_0^2 + w_1^2 = 0.307 \qquad SSE = 6.34$$
$$\alpha = 10 : \widehat{Y} = -0.333 + 0.408 \cdot X \qquad w_0^2 + w_1^2 = 0.277 \qquad SSE = 6.38$$
$$\alpha = 100 : \widehat{Y} = -0.089 + 0.343 \cdot X \qquad w_0^2 + w_1^2 = 0.125 \qquad SSE = 8.87$$

From Example 23.6, we observe that for $\alpha = 10$, when we penalize w_0, we obtain the following model:

$$\alpha = 10 : \widehat{Y} = -0.244 + 0.388 \cdot X \qquad w_0^2 + w_1^2 = 0.210 \qquad SSE = 6.75$$

As expected, we obtain a higher bias term when we do not penalize w_0.

رگرسیون خط الرأس: نزول گرادیان تصادفی (Ridge regression: stochastic gradient descent)

بجای معکوس کردن ماتریس $(\tilde{\boldsymbol{D}}^T \tilde{\boldsymbol{D}} + \alpha.\boldsymbol{I})$

$$\tilde{\mathbf{w}}^{t+1} = \tilde{\mathbf{w}}^t - \eta \cdot \nabla_{\tilde{\mathbf{w}}} = (1 - \eta \cdot \alpha)\tilde{\mathbf{w}}^t + \eta \cdot \tilde{\mathbf{D}}^T (Y - \tilde{\mathbf{D}} \cdot \tilde{\mathbf{w}}^t)$$

در هر دفعه برای بروز کردن مقدار بردار وزن تنها از یک نقطه استفاده می‌کنیم. لذا باید ضریب α را با تقسیم کردن بر n تصحیح نماییم.

$$\tilde{\mathbf{w}}^{t+1} = \tilde{\mathbf{w}}^t - \eta \cdot \nabla_{\tilde{\mathbf{w}}}(\tilde{\mathbf{x}}_k) = \left(1 - \frac{\eta \cdot \alpha}{n}\right) \tilde{\mathbf{w}}^t + \eta \cdot (y_k - \tilde{\mathbf{x}}_k^T \tilde{\mathbf{w}}^t) \cdot \tilde{\mathbf{x}}_k$$

Algorithm 23.3: Ridge Regression: Stochastic Gradient Descent

RIDGE REGRESSION: SGD ($\mathbf{D}, Y, \eta, \epsilon$):

$\tilde{\mathbf{D}} \leftarrow (\mathbf{1} \quad \mathbf{D})$ // augment data
$t \leftarrow 0$ // step/iteration counter
$\tilde{\mathbf{w}}^t \leftarrow$ random vector in $\mathbb{R}^{d+1}$ // initial weight vector
repeat
  foreach $k = 1, 2, \cdots, n$ *(in random order)* **do**
    $\nabla_{\tilde{\mathbf{w}}}(\tilde{\mathbf{x}}_k) \leftarrow -(y_k - \tilde{\mathbf{x}}_k^T \tilde{\mathbf{w}}^t) \cdot \tilde{\mathbf{x}}_k + \frac{\alpha}{n} \cdot \tilde{\mathbf{w}}$ // compute gradient at $\tilde{\mathbf{x}}_k$
    $\tilde{\mathbf{w}}^{t+1} \leftarrow \tilde{\mathbf{w}}^t - \eta \cdot \nabla_{\tilde{\mathbf{w}}}(\tilde{\mathbf{x}}_k)$ // update estimate for $\tilde{\mathbf{w}}$
  $t \leftarrow t + 1$
until $\|\mathbf{w}^t - \mathbf{w}^{t-1}\| \leq \epsilon$

نقطه‌ی تبدیل شده‌ی افزوده.

$$\tilde{\phi}(\mathbf{x}_i)^T = \begin{pmatrix} 1 & \phi(\mathbf{x}_i)^T \end{pmatrix}$$

مجموعه‌داده در فضای ویژگی‌ها $\tilde{\boldsymbol{D}}_{\phi}$

تابع هسته‌ی افزوده (Augmented Kernel Matrix)

$$\tilde{K}(\mathbf{x}_i, \mathbf{x}_j) = \tilde{\phi}(\mathbf{x}_i)^T \tilde{\phi}(\mathbf{x}_j) = 1 + \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) = 1 + K(\mathbf{x}_i, \mathbf{x}_j)$$

$$\widehat{Y} = \tilde{\mathbf{D}}_{\phi} \tilde{\mathbf{w}}$$

$$\min_{\tilde{\mathbf{w}}} \; J(\tilde{\mathbf{w}}) = \| Y - \widehat{Y} \|^2 + \alpha \cdot \| \tilde{\mathbf{w}} \|^2 = \| Y - \tilde{\mathbf{D}}_{\phi} \tilde{\mathbf{w}} \|^2 + \alpha \cdot \| \tilde{\mathbf{w}} \|^2$$

$$\frac{\partial}{\partial \tilde{\mathbf{w}}} J(\tilde{\mathbf{w}}) = \frac{\partial}{\partial \tilde{\mathbf{w}}} \left\{ \| Y - \tilde{\mathbf{D}}_{\phi} \tilde{\mathbf{w}} \|^2 + \alpha \cdot \| \tilde{\mathbf{w}} \|^2 \right\} = \mathbf{0}$$

$$\tilde{\mathbf{w}} = \tilde{\mathbf{D}}_{\phi}^T \left(\frac{1}{\alpha} \left(Y - \tilde{\mathbf{D}}_{\phi} \tilde{\mathbf{w}} \right) \right) \qquad \tilde{\mathbf{w}} = \tilde{\mathbf{D}}_{\phi}^T \mathbf{c} = \sum_{i=1}^{n} c_i \cdot \tilde{\phi}(\mathbf{x}_i)$$

$$\mathbf{c} = (c_1, c_2, \cdots, c_n)^T = \tfrac{1}{\alpha}(Y - \tilde{\mathbf{D}}_{\phi} \tilde{\mathbf{w}})$$

$$\mathbf{c} = (\tilde{\mathbf{D}}_{\phi} \tilde{\mathbf{D}}_{\phi}^T + \alpha \cdot \mathbf{I})^{-1} Y$$

$$\mathbf{c} = (\tilde{\mathbf{K}} + \alpha \cdot \mathbf{I})^{-1} Y$$

$$\tilde{\mathbf{D}}_{\phi} \tilde{\mathbf{D}}_{\phi}^T = \left\{ \tilde{\phi}(\mathbf{x}_i)^T \tilde{\phi}(\mathbf{x}_j) \right\}_{i,j=1,2,\cdots,n} = \left\{ \tilde{K}(\mathbf{x}_i, \mathbf{x}_j) \right\}_{i,j=1,2,\cdots,n} = \tilde{\mathbf{K}}$$

$$\begin{aligned} \widehat{Y} &= \tilde{\mathbf{D}}_{\phi} \tilde{\mathbf{w}} \\ &= \tilde{\mathbf{D}}_{\phi} \tilde{\mathbf{D}}_{\phi}^T \mathbf{c} \\ &= \left(\tilde{\mathbf{D}}_{\phi} \tilde{\mathbf{D}}_{\phi}^T \right) \left(\tilde{\mathbf{K}} + \alpha \cdot \mathbf{I} \right)^{-1} Y \\ &= \tilde{\mathbf{K}} \left(\tilde{\mathbf{K}} + \alpha \cdot \mathbf{I} \right)^{-1} Y \end{aligned}$$

ماتریس کلاه هسته‌ای (Kernel Hat Matrix)

$$\tilde{\mathbf{K}}(\tilde{\mathbf{K}} + \alpha \cdot \mathbf{I})^{-1}$$

برای نقطه‌ی آزمون $\boldsymbol{z}$

$$\hat{y} = \tilde{\phi}(\mathbf{z})^T \tilde{\mathbf{w}} = \tilde{\phi}(\mathbf{z})^T \left(\tilde{\mathbf{D}}_{\phi}^T \mathbf{c} \right) = \tilde{\phi}(\mathbf{z})^T \left(\sum_{i=1}^{n} c_i \cdot \tilde{\phi}(\mathbf{x}_i) \right)$$

$$= \sum_{i=1}^{n} c_i \cdot \tilde{\phi}(\mathbf{z})^T \tilde{\phi}(\mathbf{x}_i) = \sum_{i=1}^{n} c_i \cdot \tilde{K}(\mathbf{z}, \mathbf{x}_i) = \mathbf{c}^T \tilde{\mathbf{K}}_{\mathbf{z}}$$

Example 23.9. Consider the nonlinear Iris dataset shown in Figure 23.7, obtained via a nonlinear transformation applied to the centered Iris data. In particular, the `sepal length` (A_1) and `sepal width` attributes (A_2) were transformed as follows:

$$X = A_2$$
$$Y = 0.2A_1^2 + A_2^2 + 0.1A_1A_2$$

We treat Y as the response variable and X is the independent attribute. The points show a clear quadratic (nonlinear) relationship between the two variables.

We find the lines of best fit using both a linear and an inhomogeneous quadratic kernel, with regularization constant $\alpha = 0.1$. The linear kernel yields the following fit

$$\widehat{Y} = 0.168 \cdot X$$

On the other hand, using the quadratic (inhomogeneous) kernel over X comprising constant (1), linear (X), and quadratic terms (X^2), yields the fit

$$\widehat{Y} = -0.086 + 0.026 \cdot X + 0.922 \cdot X^2$$

The linear (in gray) and quadratic (in black) fit are both shown in Figure 23.7. The SSE error, $\|Y - \widehat{Y}\|^2$, is 13.82 for the linear kernel and 4.33 for the quadratic kernel. It is clear that the quadratic kernel (as expected) gives a much better fit to the data.

Algorithm 23.4: Kernel Regression Algorithm

KERNEL-REGRESSION (D, Y, K, α):

1 $\mathbf{K} \leftarrow \{K(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1,\ldots,n}$ // standard kernel matrix

2 $\tilde{\mathbf{K}} \leftarrow \mathbf{K} + 1$ // augmented kernel matrix

3 $\mathbf{c} \leftarrow (\tilde{\mathbf{K}} + \alpha \cdot \mathbf{I})^{-1} Y$ // compute mixture coefficients

4 $\widehat{Y} \leftarrow \tilde{\mathbf{K}}\mathbf{c}$

TESTING (z, D, K, c):

5 $\tilde{\mathbf{K}}_\mathbf{z} \leftarrow \{1 + K(\mathbf{z}, \mathbf{x}_i)\}_{\forall\, \mathbf{x}_i \in \mathbf{D}}$

6 $\hat{y} \leftarrow \mathbf{c}^T \tilde{\mathbf{K}}_\mathbf{z}$

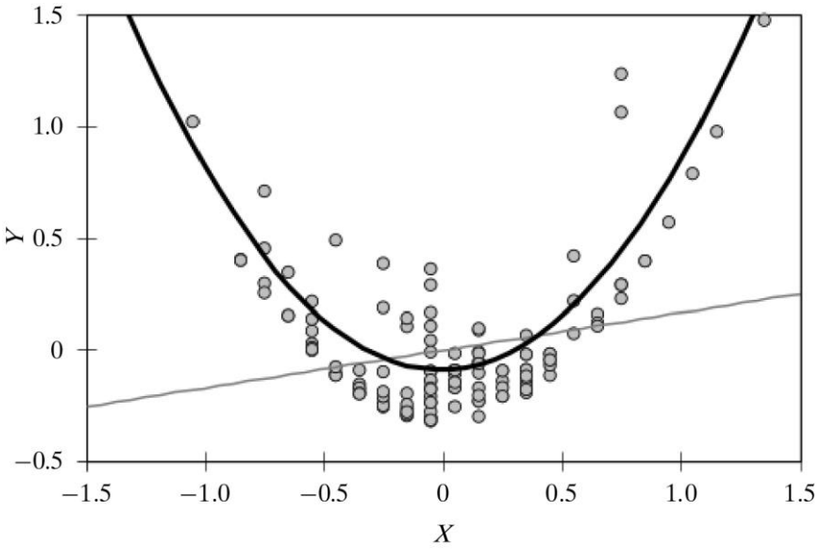
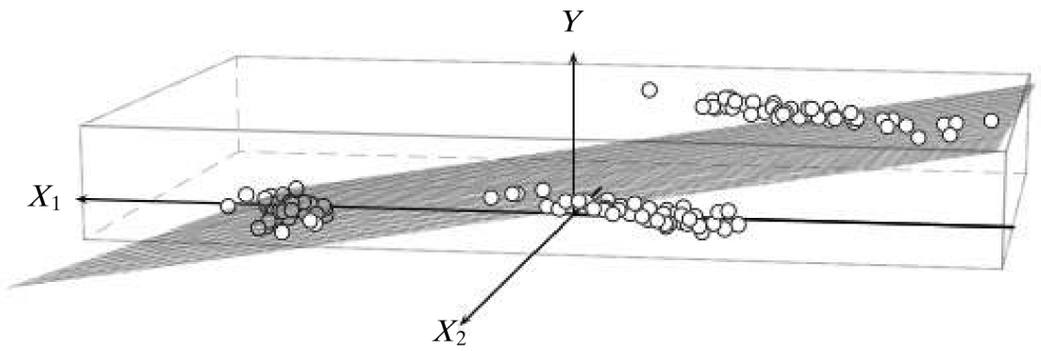
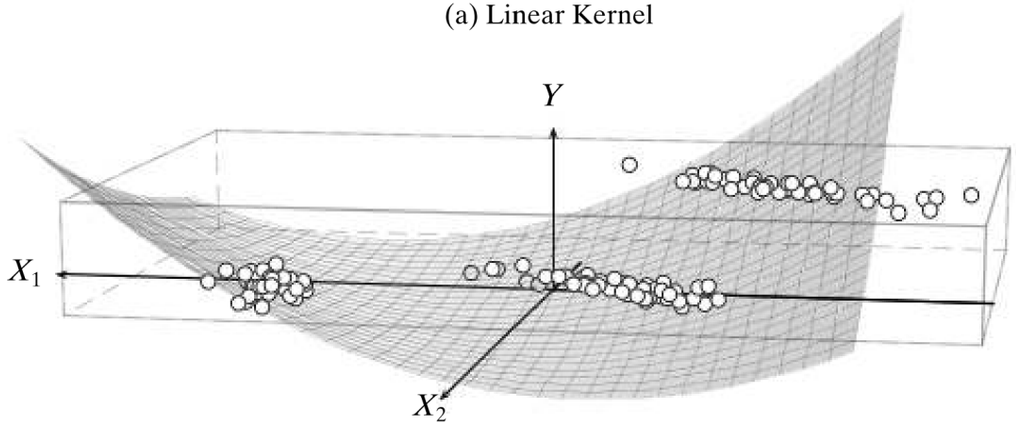


Figure 23.7. Kernel regression on nonlinear Iris dataset.



(a) Linear Kernel



(b) Inhomogeneous Quadratic Kernel

Figure 23.8. Kernel ridge regression: linear and (inhomogeneous) quadratic kernels.

Example 23.10 (Kernel Ridge Regression). Consider the Iris principal components dataset shown in Figure 23.8. Here X_1 and X_2 denote the first two principal components. The response variable Y is binary, with value 1 corresponding to `Iris-virginica` (points on the top right, with Y value 1) and 0 corresponding to `Iris-setosa` and `Iris-versicolor` (other two groups of points, with Y value 0).

Figure 23.8(a) shows the fitted regression plane using a linear kernel with ridge value $\alpha = 0.01$:

$$\widehat{Y} = 0.333 - 0.167 \cdot X_1 + 0.074 \cdot X_2$$

Figure 23.8(b) shows the fitted model when we use an inhomogeneous quadratic kernel with $\alpha = 0.01$:

$$\widehat{Y} = -0.03 - 0.167 \cdot X_1 - 0.186 \cdot X_2 + 0.092 \cdot X_1^2 + 0.1 \cdot X_1 \cdot X_2 + 0.029 \cdot X_2^2$$

The SSE error for the linear model is 15.47, whereas for the quadratic kernel it is 8.44, indicating a better fit for the training data.

مزیت اصلی استفاده از هنجار یا نرم L_1 این است که منجر به تنکی بردار پاسخ می‌شود.

$$\|\mathbf{w}\|_1 = \sum_{i=1}^{d} |w_i| \qquad \boldsymbol{\mu} = (\mu_{X_1}, \mu_{X_2}, \cdots, \mu_{X_d})^T$$

$$\bar{\mathbf{D}} = \mathbf{D} - \mathbf{1} \cdot \boldsymbol{\mu}^T \qquad \mathbf{1} \in \mathbb{R}^n$$

$$\bar{Y} = Y - \mu_Y \cdot \mathbf{1} \qquad b = w_0 = \mu_Y - \sum_{j=1}^{d} w_j \cdot \mu_{X_j}$$

$$\min_{\mathbf{w}} J(\mathbf{w}) = \frac{1}{2} \cdot \left\| \bar{Y} - \bar{\mathbf{D}}\mathbf{w} \right\|^2 + \alpha \cdot \|\mathbf{w}\|_1$$

جمله‌ی جریمه‌ی $\alpha \sum_{i=1}^{d} |w_i|$ در $w_i = 0$ مشتقپذیر نیست. پس به سادگی نمی‌توانیم گرادیان این جمله را محاسبه کنیم. این مشکل را از طریق زیرگرادیان (Sub-gradients) حل خواهیم کرد. هنگامی که مشتق یک تابع وجود ندارد، زیر مشتق تابع معادل یک مجموعه‌ای از زیرگرادیان‌ها خواهد شد.

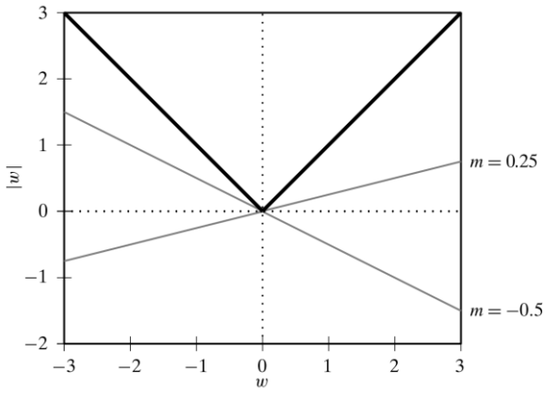


Figure 23.9. Absolute value function (in black) and two of its subgradients (in gray).

$$\partial |w| = \begin{cases} 1 & \text{iff } w > 0 \\ -1 & \text{iff } w < 0 \\ [-1, 1] & \text{iff } w = 0 \end{cases}$$

$$\min_{w} J(w) = \frac{1}{2} \sum_{i=1}^{n} (\bar{y}_i - w \cdot \bar{x}_i)^2 + \alpha \cdot |w|$$

$$\partial J(w) = \frac{1}{2} \cdot \sum_{i=1}^{n} 2 \cdot (\bar{y}_i - w \cdot \bar{x}_i) \cdot (-\bar{x}_i) + \alpha \cdot \partial |w|$$

$$w = \mathcal{S}_{\eta \cdot \alpha}(\eta \cdot \bar{X}^T \bar{Y})$$

$$\eta = 1 / \|\bar{X}\|^2 > 0$$

تابع آستانه‌ی نرم (soft-threshold function)

$$\mathcal{S}_\tau(z) = \text{sign}(z) \cdot \max\left\{0, (|z| - \tau)\right\}$$

رگرسیون L_1 چندگانه

$$\min_{\mathbf{w}} J(\mathbf{w}) = \frac{1}{2} \cdot \left\| \bar{Y} - \sum_{i=1}^{d} w_i \cdot \bar{X}_i \right\|^2 + \alpha \cdot \|\mathbf{w}\|_1$$

$$= \frac{1}{2} \cdot \left(\bar{Y}^T \bar{Y} - 2 \sum_{i=1}^{d} w_i \cdot \bar{X}_i^T \bar{Y} + \sum_{i=1}^{d} \sum_{j=1}^{d} w_i \cdot w_j \cdot \bar{X}_i^T \bar{X}_j \right) + \alpha \cdot \sum_{i=1}^{d} |w_i|$$

$$w_k^{t+1} = \mathcal{S}_{\eta \cdot \alpha}\left(w_k^t + \eta \cdot \bar{X}_k^T \left(\bar{Y} - \bar{\mathbf{D}} \mathbf{w}^t \right) \right)$$

Example 23.11 (L_1 Regression). We apply L_1 regression to the full Iris dataset with $n = 150$ points, and four independent attributes, namely `sepal-width` (X_1), `sepal-length` (X_2), `petal-width` (X_3), and `petal-length` (X_4). The Iris type attribute comprises the response variable Y. There are three Iris types, namely `Iris-setosa`, `Iris-versicolor`, and `Iris-virginica`, which are coded as 0, 1 and 2, respectively.

The L_1 regression estimates for different values of α (with $\eta = 0.0001$) are shown below:

$$\alpha = 0: \ \widehat{Y} = 0.192 - 0.109 \cdot X_1 - 0.045 \cdot X_2 + 0.226 \cdot X_3 + 0.612 \cdot X_4 \qquad SSE = 6.96$$

$$\alpha = 1: \ \widehat{Y} = -0.077 - 0.076 \cdot X_1 - 0.015 \cdot X_2 + 0.253 \cdot X_3 + 0.516 \cdot X_4 \qquad SSE = 7.09$$

$$\alpha = 5: \ \widehat{Y} = -0.553 + 0.0 \cdot X_1 + 0.0 \cdot X_2 + 0.359 \cdot X_3 + 0.170 \cdot X_4 \qquad SSE = 8.82$$

$$\alpha = 10: \ \widehat{Y} = -0.575 + 0.0 \cdot X_1 + 0.0 \cdot X_2 + 0.419 \cdot X_3 + 0.0 \cdot X_4 \qquad SSE = 10.15$$

The L_1 norm values for the weight vectors (excluding the bias term) are 0.992, 0.86, 0.529, and 0.419, respectively. It is interesting to note the sparsity inducing effect of Lasso, as observed for $\alpha = 5$ and $\alpha = 10$, which drives some of the regression coefficients to zero.

We can contrast the coefficients for L_2 (ridge) and L_1 (Lasso) regression by comparing models with the same level of squared error. For example, for $\alpha = 5$, the L_1 model has $SSE = 8.82$. We adjust the ridge value in L_2 regression, with $\alpha = 35$ resulting in a similar SSE value. The two models are given as follows:

$$L_1: \ \widehat{Y} = -0.553 + 0.0 \cdot X_1 + 0.0 \cdot X_2 + 0.359 \cdot X_3 + 0.170 \cdot X_4 \qquad \|\mathbf{w}\|_1 = 0.529$$

$$L_2: \ \widehat{Y} = -0.394 + 0.019 \cdot X_1 - 0.051 \cdot X_2 + 0.316 \cdot X_3 + 0.212 \cdot X_4 \qquad \|\mathbf{w}\|_1 = 0.598$$

where we exclude the bias term when computing the L_1 norm for the weights. We can observe that for L_2 regression the coefficients for X_1 and X_2 are small, and therefore less important, but they are not zero. On the other hand, for L_1 regression, the coefficients for attributes X_1 and X_2 are exactly zero, leaving only X_3 and X_4; Lasso can thus act as an automatic feature selection approach.

Algorithm 23.5: L_1 Regression Algorithm: Lasso

L_1-**REGRESSION** $(\mathbf{D}, Y, \alpha, \eta, \epsilon)$**:**

$\boldsymbol{\mu} \leftarrow \text{mean}(\mathbf{D})$ // compute mean
$\overline{\mathbf{D}} \leftarrow \mathbf{D} - \mathbf{1} \cdot \boldsymbol{\mu}^T$ // center the data
$\overline{Y} \leftarrow Y - \mu_Y \cdot \mathbf{1}$ // center the response
$t \leftarrow 0$ // step/iteration counter
$\mathbf{w}^t \leftarrow$ random vector in $\mathbb{R}^d$ // initial weight vector
repeat
  foreach $k = 1, 2, \cdots, d$ **do**
    $\nabla(w_k^t) \leftarrow -\overline{X}_k^T (Y - \overline{\mathbf{D}} \mathbf{w}^t)$ // compute gradient at w_k
    $w_k^{t+1} \leftarrow w_k^t - \eta \cdot \nabla(w_k^t)$ // update estimate for w_k
    $w_k^{t+1} \leftarrow \mathcal{S}_{\eta \cdot \alpha}\left(w_k^{t+1}\right)$ // apply soft-threshold function
  $t \leftarrow t + 1$
until $\|\mathbf{w}^t - \mathbf{w}^{t-1}\| \leq \epsilon$
$b \leftarrow \mu_Y - \left(\mathbf{w}^t\right)^T \boldsymbol{\mu}$ // compute the bias term