Problem Set 3

林小斌

2018年12月20日

1 Regularized Normal Equation for Linear Regression

Consider the cost function

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)} - y^{(i)})^2 + \lambda\sum_{j=1}^{n}\theta_j^2)\right]$$

The normal equation is to find the parameters that minimize the cost function by solving the following equations.

$$\frac{\partial}{\partial\theta_j}J(\theta) = 0$$

Assuming that there are m training examples, each instance has n characteristics, the training example set is

$$X = \begin{bmatrix} x_0^{(1)} & \dots & x_n^{(1)} \\ \vdots & \ddots & \vdots \\ x_0^{(m)} & \dots & x_n^{(m)} \end{bmatrix}$$

where $x_j^{(i)}$ represents the j feature of the i instance.
Consider

$$\theta = \begin{bmatrix}\theta_0 & \theta_1 & \dots & \theta_n\end{bmatrix}^T$$

$$Y = \begin{bmatrix}y^{(1)} & y^{(2)} & \dots & y^{(m)}\end{bmatrix}^T$$

thus

$$
\begin{aligned}
J(\theta) &= \frac{1}{2m}\left[(X\theta - Y)^T(X\theta - Y) + \lambda\theta^2\right] \\
&= \frac{1}{2m}\left[Y^TY - Y^TX\theta - \theta^TX^TY + \theta^TX^TX^T\theta + \lambda L\theta^2\right]
\end{aligned}
$$

where L is $m \times m$ matrix and $L = \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}$

Derivation is equivalent to the following form

$$
\frac{1}{2m}\left(\frac{\partial Y^TY}{\partial\theta} - \frac{\partial Y^TX\theta}{\partial\theta} - \frac{\partial\theta^TX^TY}{\partial\theta} + \frac{\partial\theta^TX^TX^T\theta}{\partial\theta} + \lambda L\frac{\partial\theta^2}{\partial\theta}\right)
$$

(1)For the first item

$$
\frac{\partial Y^TY}{\partial\theta} = 0
$$

(2)For the second item

$$
\begin{aligned}
Y^TX\theta &= \begin{bmatrix} y^{(1)} & y^{(2)} & \dots & y^{(m)} \end{bmatrix} \begin{bmatrix} x_0^{(1)} & \dots & x_n^{(1)} \\ \vdots & \ddots & \vdots \\ x_0^{(m)} & \dots & x_n^{(m)} \end{bmatrix} \begin{bmatrix} \theta_0 & \theta_1 & \dots & \theta_n \end{bmatrix}^T \\
&= \left(x_0^{(1)}y^{(1)} + \dots x_0^{(m)}y^{(m)}\right)\theta_0 + \dots + \left(x_n^{(1)}y^{(1)} + \dots x_n^{(m)}y^{(m)}\right)\theta_n
\end{aligned}
$$

thus

$$
\frac{\partial Y^TX\theta}{\partial\theta} = \begin{bmatrix} \frac{\partial Y^TX\theta}{\partial\theta_0} \\ \frac{\partial Y^TX\theta}{\partial\theta_1} \\ \vdots \\ \frac{\partial Y^TX\theta}{\partial\theta_n} \end{bmatrix} = X^TY
$$

(3)For the third item

$$
\begin{aligned}
\theta^TX^TY &= \begin{bmatrix} \theta_0 & \theta_1 & \dots & \theta_n \end{bmatrix} \begin{bmatrix} x_0^{(1)} & \dots & x_n^{(1)} \\ \vdots & \ddots & \vdots \\ x_0^{(m)} & \dots & x_n^{(m)} \end{bmatrix}^T \begin{bmatrix} y^{(1)} & y^{(2)} & \dots & y^{(m)} \end{bmatrix}^T \\
&= \left(x_0^{(1)}\theta_0 + \dots x_0^{(m)}\theta_n\right)y^{(1)} + \dots + \left(x_n^{(1)}\theta_0 + \dots x_n^{(m)}\theta_n\right)y^{(n)}
\end{aligned}
$$

thus

$$\frac{\partial \theta^T X^T Y}{\partial \theta} = \begin{bmatrix} \frac{\partial \theta^T X^T Y}{\partial \theta_0} \\ \frac{\partial \theta^T X^T Y}{\partial \theta_1} \\ \vdots \\ \frac{\partial \theta^T X^T Y}{\partial \theta_n} \end{bmatrix} = X^T Y$$

(4)For the fourth item

$$\theta^T X^T X \theta = X^T X \left(\theta_0^2 + \theta_1^2 + \cdots + \theta_n^2\right)$$

thus

$$\frac{\partial \theta^T X^T X \theta}{\partial \theta} = \begin{bmatrix} \frac{\partial \theta^T X^T X\theta}{\partial \theta_0} \\ \frac{\partial \theta^T X^T X\theta}{\partial \theta_1} \\ \vdots \\ \frac{\partial \theta^T X^T X\theta}{\partial \theta_n} \end{bmatrix} = 2\left(X^T X\right) \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} = 2X^T X \theta$$

(5)For the fifth item

$$\lambda L \frac{\partial \theta^2}{\theta} = 2\lambda L \theta$$

In summary, the normal equation is:

$$\frac{1}{2m}\left(-2X^T Y + 2X^T X\theta + 2\lambda L\theta\right) = 0$$

thus

$$\theta = \left(X^T X + \lambda L\right)^{-1} X^T Y$$

2 Lagrange Duality

Primal problem formulation

$$\begin{aligned} & min \quad c^T x \\ & s.t \quad Ax \preceq b \end{aligned}$$

where x ∈R is variable, $c \in$ R^n,A ∈ R$^{k\times n}$,b ∈ R.

The Lagrangian

$$\mathcal{L}(x,\alpha) = c^T x + \alpha^T (Ax - b)$$

The Lagrange dual function

$$\begin{aligned}\mathcal{G}(\alpha) &= \inf_x \ \mathcal{L}(x,\alpha)\\ &= \inf_x \ (c^T x + \alpha^T(Ax-b))\\ &= \inf_x \ ((c^T + \alpha^T A)x - \alpha^T b)\end{aligned}$$

To avoid the Lagrange dual function $\mathcal{G}$ be -∞, c$^T + \alpha^T A$ must equal to 0.
Lagrange dual problem

$$\begin{aligned}&\max_\alpha \ G(\alpha) = \max_\alpha \ \inf_x \ \mathcal{L}(x,\alpha) = \max_\alpha \ -\alpha^T b\\ &s.t \quad c^T + \alpha^T A = 0\\ &\qquad \alpha \geq 0\end{aligned}$$

3 SVM

3.1 Convex Functions

Assume

$$w = \begin{bmatrix} x_1 x_2 \ldots x_n \end{bmatrix}^T$$

then we have

$$f(w) = w^T w = \begin{bmatrix} x_1 x_2 \ldots x_n \end{bmatrix}^T \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = x_1^2 + x_2^2 + \cdots + x_n^2 = f(x_1) + f(x_2) + \cdots + f(x_n)$$

where $f(x_i) = x_i^2, \quad i = 1, 2, \ldots, n.$
Since $f(x) = x^2$ is a convex function, for any convex function,

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

For any $i, j \in \{1, 2, \ldots, n\}$, we assume $g(x) = f_i(x) + f_j(x)$, then we have

$$\begin{aligned}g(\lambda x_i + (1-\lambda)x_j) &= f_i(\lambda x_i + (1-\lambda)x_j) + f_j(\lambda x_i + (1-\lambda)x_j)\\ &\leq \lambda f_i(x_i) + (1-\lambda) f_i(x_j) + f_j(x_i) + (1-\lambda) f_j(x_j)\\ &= \lambda(f_i(x_i) + f_j(x_i)) + (1-\lambda)(f_i(x_j) + f_j(x_j))\\ &= \lambda g(x_i) + (1-\lambda) g(x_j)\end{aligned}$$

Then we know $g(x)$ is a convex function too! Finally we can change $f(w)$ into $g(x)$, so $f(w)$ is convex function.

3.2 Soft-Margin for Separable Data

True!
According to the question, we can set the condition that

$$y^{(i)}(\omega^T x^{(i)} + b) \geq 1$$

Lagrangian of soft-margin:

$$L(\omega, b, \xi, \alpha, r) = \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i[y^{(i)}(\omega^T x^{(i)} + b) - 1 + \xi_i] - \sum_{i=1}^{m} r_i\xi_i$$

KKT conditions:
1. $\bigtriangledown_\omega(\omega, b, \xi, \alpha, r) = 0 \Rightarrow \omega = \sum_{i=1}^{m}\alpha_i y^{(i)} x^{(i)}$
2. $\bigtriangledown_b(\omega, b, \xi, \alpha, r) = 0 \Rightarrow \sum_{i=1}^{m}\alpha_i y^{(i)} = 0$
3. $\bigtriangledown_{\xi i}(\omega, b, \xi, \alpha, r) = 0 \Rightarrow \alpha_i + r_i = C$ for $\forall i$
4. $\alpha_i, r_i, \xi_i \geq 0,$ for $\forall i$
5. $y^{(i)}(\omega^T x^{(i)} + b) - 1 + \xi_i = 0,$ for $\forall i$
6. $r_i\xi_i = 0,$ for $\forall i$

If $\alpha_i = 0, y^{(i)}(\omega^T x^{(i)} + b) \geq 1$

$$\begin{aligned}
&\alpha_i = 0,\ \alpha_i + r_i = C \\
&r_i = C \\
&r_i\xi_i = 0,\ \xi_i \geq 0 \\
&\xi_i = 0 \\
&\alpha_i(y^{(i)}(\omega^T x^{(i)} + b) - 1 + \xi_i) \geq 0 \\
&y^{(i)}(\omega^T x^{(i)} + b) - 1 + \xi_i = 0 \\
&y^{(i)}(\omega^T x^{(i)} + b) \geq 1
\end{aligned}$$

If $\alpha \neq C$, it satisfies the condition $y^{(i)}(\omega^T x^{(i)} + b) \geq 1$, then we have $\xi_i = 0$. When we use soft-margin SVM can solve this problem when dataset are linearly separable, it is not necessary to use a hard-margin SVM.

3.3 In-bound Support Vectors in Soft-Margin SVMs

Lagrangian of soft-margin:

$$L(\omega, b, \xi, \alpha, r) = \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i[y^{(i)}(\omega^T x^{(i)} + b) - 1 + \xi_i] - \sum_{i=1}^{m} r_i\xi_i$$

KKT conditions:

1. $\bigtriangledown_\omega(\omega, b, \xi, \alpha, r) = 0 \Rightarrow \omega = \sum_{i=1}^{m}\alpha_i y^{(i)} x^{(i)}$
2. $\bigtriangledown_b(\omega, b, \xi, \alpha, r) = 0 \Rightarrow \sum_{i=1}^{m}\alpha_i y^{(i)} = 0$
3. $\bigtriangledown_{\xi i}(\omega, b, \xi, \alpha, r) = 0 \Rightarrow \alpha_i + r_i = C$ for $\forall i$
4. $\alpha_i, r_i, \xi_i \geq 0,$ for $\forall i$
5. $y^{(i)}(\omega^T x^{(i)} + b) - 1 + \xi_i = 0,$ for $\forall i$
6. $r_i\xi_i = 0,$ for $\forall i$

As for in-bound SVs $0 < \alpha_i < C$

$$\begin{aligned}
&0 < \alpha_i < C,\ \alpha_i + r_i = C \\
&0 < r_i < C \\
&r_i\xi_i = 0,\ \xi_i \geq 0 \\
&\xi_i = 0 \\
&\alpha_i(y^{(i)}(\omega^T x^{(i)} + b) - 1 + \xi_i) = 0 \\
&y^{(i)}(\omega^T x^{(i)} + b) - 1 + \xi_i = 0 \\
&y^{(i)}(\omega^T x^{(i)} + b) = 1
\end{aligned}$$

So the in-bound SVs lie exactly on the margin.

As for bound SVs $\alpha_i = C$

$$\begin{aligned}
&\alpha_i = C,\ \alpha_i + r_i = C \\
&r_i = 0 \\
&r_i\xi_i = 0,\ \xi_i \geq 0 \\
&\xi_i \geq 0 \\
&\alpha_i(y^{(i)}(\omega^T x^{(i)} + b) - 1 + \xi_i) = 0 \\
&y^{(i)}(\omega^T x^{(i)} + b) - 1 + \xi_i = 0 \\
&y^{(i)}(\omega^T x^{(i)} + b) = 1 - \xi \leq 1
\end{aligned}$$

So the bounds SVs can lie both on or in the margin.