Ridge regression and its dual problem

Sep 3, 2014

Ridge regression is the name given to least-squares regression with squared Euclidean norm regularisation added. Given n example vectors x_i of dimension m with scalar labels y_i, the problem is expressed as finding the weight vector w and scalar bias b which minimise the objective function

$$f(w, b) = \frac{1}{2} \sum_{i=1}^{n} \left(x_i^T w + b - y_i\right)^2 + \frac{\lambda}{2} \|w\|^2.$$

Eliminating the bias

Setting the derivative of f with respect to b to zero yields

$$\frac{\partial f}{\partial b}(w, b) = \sum_{i=1}^{n} \left(x_i^T w + b - y_i\right) = 0, \qquad b = \bar{y} - \bar{x}^T w$$

and therefore the problem is to find the minimiser of

$$h(w) = f(w, b(w)) = \frac{1}{2} \sum_{i=1}^{n} \left[(x_i - \bar{x})^T w - (y_i - \bar{y})\right]^2 + \frac{\lambda}{2} \|w\|^2.$$

From this point on we will assume that the example vectors and the labels have been pre-processed to have zero-mean, leading to the simplified form

$$h(w) = \frac{1}{2} \sum_{i=1}^{n} \left(x_i^T w - y_i\right)^2 + \frac{\lambda}{2} \|w\|^2.$$

Let us introduce the notation that X is an $m \times n$ matrix whose columns are the example vectors and y is a vector comprising the corresponding labels, writing the objective as $h(w) = \frac{1}{2}\|X^T w - y\|^2 + \frac{1}{2}\lambda\|w\|^2$.

Solving for the weights in the primal

The problem above can be re-written as

$$\arg\min_{w}\left[\frac{1}{2}w^T(S+\lambda I)w - w^T Xy\right]$$

where $S = XX^T$ is the $m \times m$ covariance matrix. The solution to this unconstrained quadratic program is simply $w = (S+\lambda I)^{-1}Xy$.

The dual problem

The problem can be converted into a constrained minimisation problem

$$\arg\min_{w,r}\left[\frac{1}{2}\|r\|^2 + \frac{\lambda}{2}\|w\|^2\right] \text{ subject to } r = X^T w - y$$

whose Lagrangian is

$$L(w,r,\alpha) = \frac{1}{2}\|r\|^2 + \frac{\lambda}{2}\|w\|^2 + \alpha^T(r - X^T w + y).$$

Setting derivatives with respect to the primal variables to zero, we obtain

$$\begin{aligned} \frac{\partial L}{\partial w}(w,r,\alpha) &= \lambda w - X\alpha = 0, & w &= \frac{1}{\lambda}X\alpha \\ \frac{\partial L}{\partial r}(w,r,\alpha) &= r + \alpha = 0, & r &= -\alpha. \end{aligned}$$

Making these substitutions to eliminate r and w gives the dual function

$$\begin{aligned} g(\alpha) &= L(w(\alpha), r(\alpha), \alpha) \\ &= \frac{1}{2}\|\alpha\|^2 + \frac{1}{2\lambda}\|X\alpha\|^2 + \alpha^T\left(-\alpha - \frac{1}{\lambda}X^T X\alpha + y\right) \\ &= -\frac{1}{2}\|\alpha\|^2 - \frac{1}{2\lambda}\|X\alpha\|^2 + \alpha^T y. \end{aligned}$$

and the dual problem is

$$\arg\min_{\alpha}\left[\frac{1}{2}\alpha^T(K+\lambda I)\alpha - \lambda\alpha^T y\right]$$

where $K = X^T X$ is the $n \times n$ kernel matrix. The solution is obtained $\alpha = \lambda(K+\lambda I)^{-1}y$ and then $w = X(K+\lambda I)^{-1}y$.

Primal vs dual

We now have two equivalent solutions, one using the covariance matrix and the other the kernel

$$w = (S+\lambda I)^{-1}Xy = X(K+\lambda I)^{-1}y$$