

Consider the linear model with additive noise,

$$y_i = w^T x_i + \epsilon_i,$$

where ϵ_i is the noise term. We will assume that w is normally distributed,

$$Pr(W = w) = \frac{1}{(2\pi)^{d/2} |\Sigma_W|^{1/2}} \exp \left(-\frac{1}{2} (w - \mu_W)^T \Sigma_W^{-1} (w - \mu_W) \right),$$

where Σ_W is the covariance matrix and μ_W is the mean. We will also assume that the noise is independent and normally identically distributed with zero mean and covariance σ_ϵ^2 ,

$$Pr(\epsilon = \epsilon) = \frac{1}{(2\pi)^{d/2} |\sigma_\epsilon|} \exp \left(-\frac{\epsilon^T \epsilon}{2\sigma_\epsilon^2} \right).$$

Let X and Y be stacked versions of x_i and $y_i \forall i$ in our training set. Let's first find the mode of the posterior distribution of w , given our training data X and Y :

$$\begin{aligned} w^* &= \arg \max_w Pr(w|X, Y) \\ &= \arg \max_w \frac{Pr(X, Y|w)Pr(w)}{Pr(X, Y)} \\ &= \arg \max_w \frac{Pr(Y - w^T X|w)Pr(w)}{Pr(X, Y)}. \end{aligned}$$

There are two things to note here before we continue. First the denominator does not depend on w , so we can drop it, and, second, that $Y - w^T X = \epsilon$, for which we know the distribution. We will also, for simplicity, assume that $\mu_W = 0$ and that $\Sigma_W = \kappa I$, where I is the identity matrix. So we now have:

$$\begin{aligned} w^* &= \arg \max_w \frac{1}{(2\pi)^d |\sigma_\epsilon| |k|^{1/2}} \exp \left(-\frac{(Y - w^T X)^T (Y - w^T X)}{2\sigma_\epsilon^2} + -\frac{1}{2k} w^T w \right) \\ &= \arg \min_w -\log \left((2\pi)^d |\sigma_\epsilon| |k|^{1/2} \right) + \frac{(Y - w^T X)^T (Y - w^T X)}{2\sigma_\epsilon^2} + \frac{1}{2k} w^T w \\ &= \arg \min_w (Y - w^T X)^T (Y - w^T X) + \frac{\sigma_\epsilon^2}{k} w^T w \\ &= \arg \min_w (Y - w^T X)^T (Y - w^T X) + \lambda w^T w. \end{aligned}$$

Which is the familiar form of ridge regression.