

KERNEL RIDGE REGRESSION

In these notes we will show how to incorporate kernels into ridge regression yielding a non-linear regression method.

Ridge Regression

Ridge regression solves

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

The solution is

$$\hat{w} = (X^T X + n\lambda I)^{-1} X^T \hat{y}$$

$$\hat{b} = \bar{y} - \hat{w}^T \bar{x}$$

where

$$\hat{y} = \begin{bmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_n \end{bmatrix}, \quad X = \begin{bmatrix} \tilde{x}_1^T \\ \vdots \\ \tilde{x}_n^T \end{bmatrix},$$

$$\tilde{x}_i = x_i - \bar{x} \in \mathbb{R}^d$$

$$\tilde{y}_i = y_i - \bar{y} \in \mathbb{R}$$

$$\bar{x} = \frac{1}{n} \sum x_i$$

$$\bar{y} = \frac{1}{n} \sum y_i$$

$$\begin{bmatrix} \tilde{y}_n \end{bmatrix} \quad \begin{bmatrix} \tilde{x}_n^T \end{bmatrix} \quad \bar{y} = \frac{1}{n} \sum y_i$$

The regression function estimate is

$$\hat{f}(x) = \hat{w}^T x + \hat{b} = \bar{y} + \hat{w}(x - \bar{x})$$

To apply kernels, we must show that $\hat{f}(x)$ depends on x, x_1, \dots, x_n only in terms of inner products, such as $\langle x_i, x_j \rangle$ or $\langle x_i, x \rangle$.

Kernelizing Ridge Regression

Note that $X^T X$ is not the Gram matrix of $\tilde{x}_1, \dots, \tilde{x}_n$. Let's apply the matrix inversion lemma:

$$(P + QRS)^{-1} = P^{-1} - P^{-1}Q(R^{-1} + SP^{-1}Q)^{-1}SP^{-1}$$

with

$$P = \mu I, \quad Q = X^T, \quad R = I, \quad S = X$$

$\mu = nd$ for brevity

This yields

$$\begin{aligned}
 (\mu \mathbf{I} + \mathbf{X}^T \mathbf{X})^{-1} &= \frac{1}{\mu} \mathbf{I} - \frac{1}{\mu} \mathbf{I} \cdot \mathbf{X}^T (\mathbf{I} + \frac{1}{\mu} \mathbf{X} \mathbf{X}^T)^{-1} \mathbf{X} \cdot \frac{1}{\mu} \mathbf{I} \\
 &= \frac{1}{\mu} [\mathbf{I} - \mathbf{X}^T (\mu \mathbf{I} + \mathbf{X} \mathbf{X}^T)^{-1} \mathbf{X}]
 \end{aligned}$$

Therefore

$$\begin{aligned}
 (\mathbf{X}^T \mathbf{X} + \mu \mathbf{I})^{-1} \mathbf{X}^T \tilde{\mathbf{y}} &= \frac{1}{\mu} [\mathbf{X} - \mathbf{X}^T (\mathbf{X} \mathbf{X}^T + \mu \mathbf{I})^{-1} \mathbf{X} \mathbf{X}^T] \tilde{\mathbf{y}} \\
 &= \frac{1}{\mu} [\mathbf{X}^T - \mathbf{X}^T (\tilde{\mathbf{K}} + \mu \mathbf{I})^{-1} \tilde{\mathbf{K}}] \tilde{\mathbf{y}}
 \end{aligned}$$

where

$$\tilde{\mathbf{K}} = [\langle \tilde{x}_i, \tilde{x}_j \rangle]_{i,j=1}^n \quad (n \times n)$$

Note

$$\begin{aligned}
 \langle \tilde{x}_i, \tilde{x}_j \rangle &= \langle x_i - \bar{x}, x_j - \bar{x} \rangle \\
 &= \langle x_i, x_j \rangle - \frac{1}{n} \sum_{r=1}^n \langle x_i, x_r \rangle \\
 &\quad - \frac{1}{n} \sum_{s=1}^n \langle x_s, x_j \rangle + \frac{1}{n^2} \sum_{r=1}^n \sum_{s=1}^n \langle x_r, x_s \rangle
 \end{aligned}$$

In addition,

$$\begin{aligned}
\hat{f}(x) &= \bar{y} + \hat{w}^T (x - \bar{x}) \\
&= \bar{y} + \frac{1}{\mu} \tilde{y}^T [X - \tilde{K} (\tilde{K} + \mu I)^{-1} X] (x - \bar{x}) \\
&= \bar{y} + \frac{1}{\mu} \tilde{y}^T [I - \tilde{K} (\tilde{K} + \mu I)^{-1}] \tilde{k}(x)
\end{aligned}$$

where

$$\tilde{k}(x) = \begin{bmatrix} \langle \tilde{x}_1, x - \bar{x} \rangle \\ \vdots \\ \langle \tilde{x}_n, x - \bar{x} \rangle \end{bmatrix}.$$

The entries of this vector are

$$\begin{aligned}
\langle \tilde{x}_i, x - \bar{x} \rangle &= \langle x_i - \bar{x}, x - \bar{x} \rangle \\
&= \langle x_i, x \rangle - \frac{1}{n} \sum_r \langle x_i, x_r \rangle \\
&\quad - \frac{1}{n} \sum_s \langle x_s, x \rangle + \frac{1}{n^2} \sum_r \sum_s \langle x_r, x_s \rangle.
\end{aligned}$$

Finally, observe

$$\begin{aligned}
I - \tilde{K} (\tilde{K} + \mu I)^{-1} \\
= [\tilde{K} + \mu I - \tilde{K}] (\tilde{K} + \mu I)^{-1}
\end{aligned}$$

$$= \mu (\bar{K} + \mu \mathbf{I})^{-1}$$

which yields the estimate

$$\hat{f}(x) = \bar{y} + \tilde{y}^T (\bar{K} + n\lambda \mathbf{I})^{-1} \tilde{k}(x).$$

The point of these manipulations is that \hat{f} depends on elements of the feature space only in terms of inner products. If k is an inner product kernel with feature map Φ , and we substitute

$$\langle x, x' \rangle \rightarrow k(x, x')$$

whenever inner products occur, then \hat{f} corresponds to ridge regression after the original feature space is transformed via Φ .

No Offset

For some kernels $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$, $\Phi(x)$ already contains a constant component, in which case the offset/bias b is not necessary. The

inhomogeneous polynomial kernels are an example.

The Gaussian kernel also does not require an offset, even though its feature space does not contain a constant term (Steinwart and Christmann, 2008).

If b is omitted, the KRR estimator is

$$\hat{f}(x) = y^T (K + n\lambda I)^{-1} k(x)$$

where

$$K = [k(x_i, x_j)]_{i,j=1}^n \quad (n \times n)$$

and

$$k(x) = \begin{bmatrix} k(x_1, x) \\ \vdots \\ k(x_n, x) \end{bmatrix}$$

Example

Gaussian kernel

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

Then

$$\begin{aligned}\hat{f}(x) &= y^T (K + n\lambda I)^{-1} k(x) \\ &= \alpha^T k(x) \\ &= \sum_{i=1}^n \alpha_i k(x, x_i)\end{aligned}$$

where

$$\alpha = (K + n\lambda I)^{-1} y \in \mathbb{R}^n$$

is independent of x .

