KERNEL RIDGE REGRESSION

Recall ridge regression: Given $(x_i, y_i)_{i=1}^n$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

$$(\hat{\beta}, \hat{\beta}_0) \leftarrow \min_{\beta, \beta_0} \sum_{i=1}^n (y_i - \beta^T x_i - \beta_0)^2 + \lambda \|\beta\|^2$$

Solution:

$$\left.\begin{aligned} \hat{\beta} &= (A^T A + \lambda I)^{-1} A^T \tilde{y} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}^T \bar{x} \end{aligned}\right\} \Rightarrow \hat{f}(x) = \hat{\beta}^T x + \beta_0$$

where

$$A = \begin{bmatrix} \tilde{x}_1^T \\ \vdots \\ \tilde{x}_n^T \end{bmatrix}, \quad \begin{aligned} \tilde{x}_i &= x_i - \bar{x} \\ \tilde{y}_i &= y_i - \bar{y} \end{aligned}$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \qquad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

Can we express RR in terms of inner products $\langle x_i, x_j \rangle$ and $\langle x_i, x \rangle$?

Not immediately. Note A^TA is <u>not</u> $\left[\langle \tilde{x}_i, \tilde{x}_j \rangle\right]_{i,j=1}^n$

Let's apply the <u>matrix inversion lemma</u>, also known as the Woodbury matrix identity:

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

where

$$P = \lambda I \;,\quad Q = A^T \;,\quad R = I \;,\quad S = A \;.$$

We have

$$(\lambda I + A^TA)^{-1} = \frac{1}{\lambda}I - \frac{1}{\lambda}I \cdot A^T\left(I + \frac{1}{\lambda}AA^T\right)^{-1}A \cdot \frac{1}{\lambda}$$

$$= \frac{1}{\lambda}\left[I - A^T(\lambda I + AA^T)^{-1}A\right]$$

$$\Rightarrow \quad (A^TA + \lambda I)^{-1}A^T\underline{y} = \frac{1}{\lambda}\left[A^T - A^T(AA^T + \lambda I)^{-1}AA^T\right]y$$

$$= \frac{1}{\lambda}\left[A^T - A^T(K + \lambda I)^{-1}K\right]y$$

where $K = \left[\langle \tilde{x}_i, \tilde{x}_j \rangle\right]_{i,j=1}^n$. Note $\langle \tilde{x}_i, \tilde{x}_j \rangle$

$$= \langle x_i, x_j \rangle - \frac{1}{n}\sum_{r=1}^n \langle x_i, x_r \rangle - \frac{1}{n}\sum_{s=1}^n \langle x_s, x_j \rangle + \frac{1}{n^2}\sum_{r,s=1}^n \langle x_r, x_s \rangle.$$

What about the remaining A^T? It is handled when we evaluate the estimate:

$$\hat{\beta}^T x = \frac{1}{\lambda} y^T \left[A - K(K+\lambda I)^{-1} A \right] x$$

$$= \frac{1}{\lambda} y^T \left[I - K(K+\lambda I)^{-1} \right] \underline{k}(x)$$

where

$$\underline{k}(x) = \begin{bmatrix} \langle \tilde{x}_1, x \rangle \\ \vdots \\ \langle \tilde{x}_n, x \rangle \end{bmatrix} = \begin{bmatrix} \langle x_1, x \rangle - \frac{1}{n} \Sigma \langle x_r, x \rangle \\ \vdots \\ \langle x_n, x \rangle - \frac{1}{n} \Sigma \langle x_r, x \rangle \end{bmatrix}$$

We can simplify further:

(A) $$I - K(K + \lambda I)^{-1} =$$

$$\Longrightarrow \hat{\beta}^T x =$$

$$\Rightarrow \quad \hat{\beta}^T x = \tilde{y}^T (K + \lambda I)^{-1} \underline{k}(x)$$

What about $\hat{\beta}_0 = \bar{y} - \hat{\beta}^T \bar{x}$?

$$\hat{\beta}_0 = \bar{y} - \frac{1}{n} \sum_{i=1}^{n} \hat{\beta}^T x_i$$

$$= \bar{y} - \frac{1}{n} \sum_{i=1}^{n} \tilde{y}^T (K + \lambda I)^{-1} \underline{k}(x_i)$$

$$= \bar{y} - \tilde{y}^T (K + \lambda I)^{-1} \cdot \frac{1}{n} \sum_{i=1}^{n} \underline{k}(x_i)$$

Note | For some kernels, $k(x, x') = \langle \underline{\Phi}(x), \underline{\Phi}(x') \rangle$,

$\underline{\Phi}(x)$ already contains a constant component, in which case β_0 is not needed. Examples include the inhomogeneous polynomial kernels.

Example | Gaussian kernel

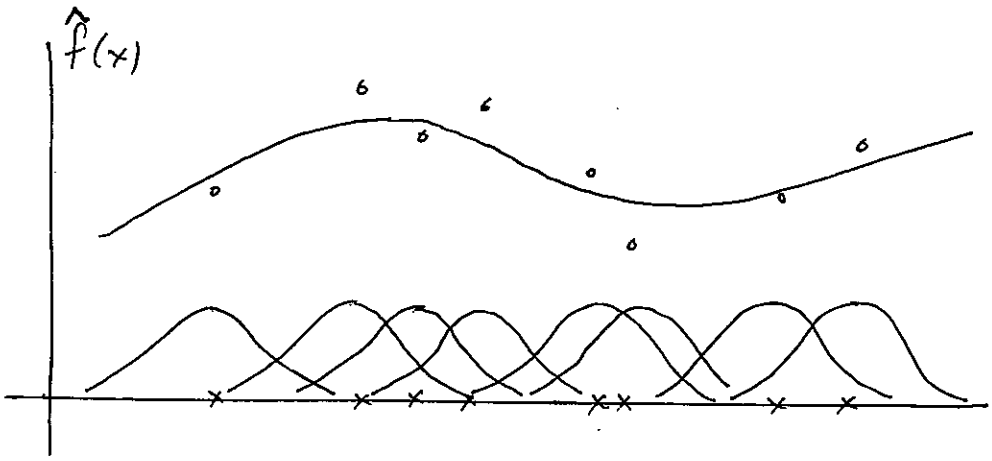
$$k(x, x') = \exp\left\{-\frac{\|x - x'\|^2}{2\sigma^2}\right\}$$

Then $\hat{\beta}^T x = \underline{y}^T (K + \lambda I)^{-1} \underline{k}(x)$

$= \underline{\alpha}^T \underline{k}(x)$

$= \sum \alpha_i k(x, x_i)$

$\underline{\alpha}$ independent of x



Key A. $I - K(K + \lambda I)^{-1}$

$= (K + \lambda I - K)(K + \lambda I)^{-1}$

$= \lambda \cdot (K + \lambda I)^{-1}$

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