KERNEL RIDGE REGRESSION

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

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

The solution is

$$\hat{\beta} = (A^T A + \lambda I)^{-1} A^T \underline{y}, \qquad A = \begin{bmatrix} x_1^{(1)} & \cdots & x_1^{(d)} \\ \vdots & & \\ x_n^{(1)} & \cdots & x_n^{(d)} \end{bmatrix}$$

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

and the final estimator is

$$\hat{f}(x) = \hat{\beta}^T x + \hat{\beta}$$

Can these operations be expressed entirely in terms of $\langle x_i, x_j \rangle$ and $\langle x_i, x \rangle$?

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

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

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

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

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

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

where $K = \left[\langle x_i, x_j\rangle\right]_{i,j=1}^n$

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

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

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

where

$$\underline{k}(x) \begin{bmatrix} \langle x_1, x \rangle \\ \vdots \\ \langle x_n, x \rangle \end{bmatrix}$$

we can simplify further:

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

(A)

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

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

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

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

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

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 & Gaussian kernels.

Example | Gaussian kernel

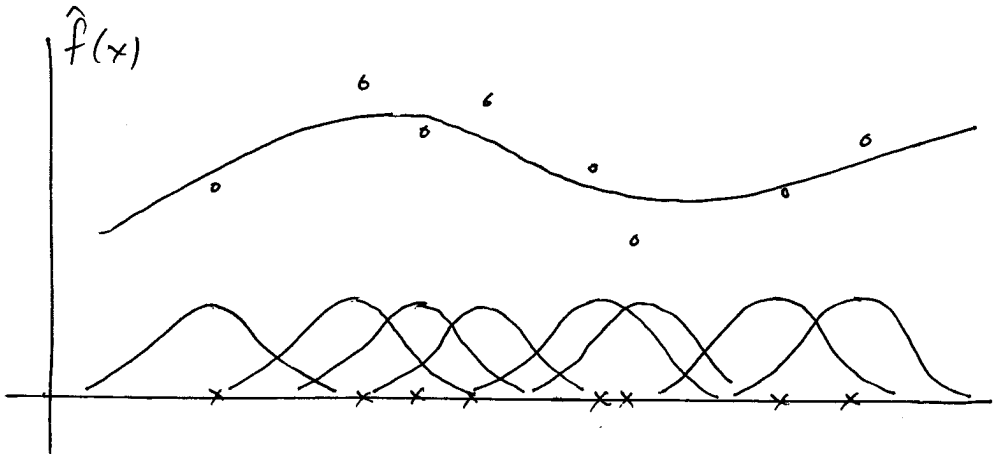
$$k_\sigma(x, x_i) = \exp\left\{-\frac{\|x - x_i\|^2}{2\sigma^2}\right\}$$

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

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

$$= \sum \alpha_i k_\sigma(x, x_i)$$

α independent of x



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

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

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