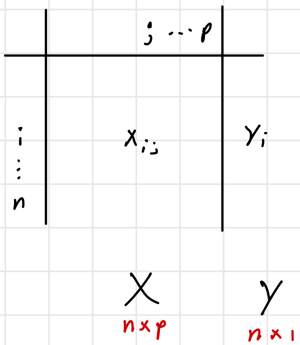
Lecture 9

- Part 3: kernel regression, Gaussian Process, Support Vector Machine

- Recall Ridge Regression



$$L = \text{Loss}(\beta) = |Y - X\beta|^2 + \lambda|\beta|^2$$

$$= (Y - X\beta)^T(Y - X\beta) + \lambda|\beta|^2$$

$$= Y^TY - \beta^TX^TY - Y^TX\beta + \beta^TX^TX^T\beta + \lambda|\beta|^2$$

$$= \text{const} - 2\langle X^TY, \beta\rangle + \beta^TX^TX\beta + \lambda\beta^T\beta$$

$$\frac{\partial L}{\partial \beta} = -\cancel{2}X^TY - \cancel{2}X^TX\beta + \cancel{2}\lambda\beta = 0$$

$$(X^TX + \lambda I_p)\beta = X^TY$$

$$\hat{\beta}_{ridge} = (X^TX + \lambda I_p)^{-1}X^TY$$

Identity: $(X^TX + \lambda I_p)^{-1}X^T = X^T(XX^T + \lambda I_n)^{-1}$

Proof: $X^T(XX^T + \lambda I_n) = (X^TX + \lambda_p)X^T$

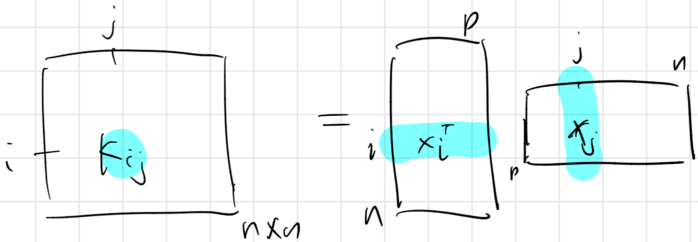
$$\Downarrow$$

$$X^TXX^T + \lambda X^T = X^TXX^T + \lambda X^T \qquad \square$$

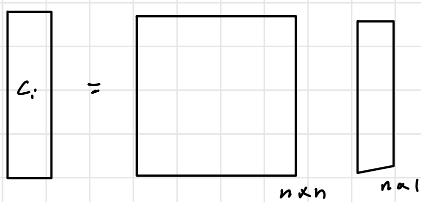
Thus $\hat{\beta} = X^T(XX^T + \lambda I_n)^{-1}Y$

$$\hat{\beta} = X^T (XX^T + \lambda I_n)^{-1} y$$

Let $K = XX^T$ with $k_{ij} = \langle x_i, x_j \rangle$



$$\underset{n\times 1}{c} = \underset{n\times n}{(K + \lambda I_n)^{-1}} \underset{n \times 1}{y}$$



So $\hat{\beta} = X^T c$

$$\hat{\beta} = \underset{p \times n}{\begin{bmatrix} x_1 & \cdots & x_i & \cdots & x_n \end{bmatrix}} \begin{pmatrix} c_1 \\ \vdots \\ c_i \\ \vdots \\ c_n \end{pmatrix} = \sum_{i=1}^{n} \boxed{c_i} \underset{p\times 1}{x_i}$$

Representer Form

Now consider

	$j \cdots p$	
i $\vdots$ n	x_{ij}	y_i
0	x_0^T	$?$

$$s = f(x_0) = x_0^T \hat{\beta}$$

$$= x_0^T \sum_{i=1}^{n} c_i x_i$$

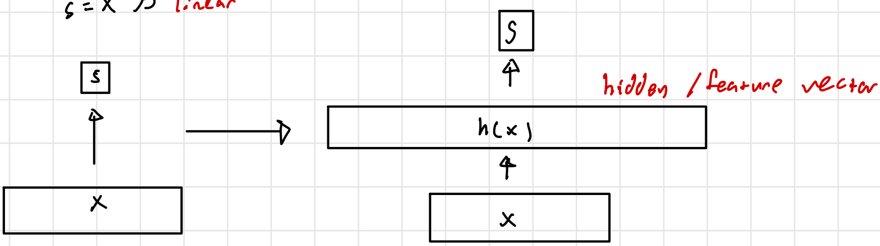
$$= \sum_{i=1}^{n} c_i \langle x_i, x_0 \rangle = \sum_{i=1}^{n} c_i \, k(x_i, x_0)$$

$k(x, x') = \langle x, x' \rangle$ ← Kernel Function

- Kernel Trick

$s = X^T \beta$ linear

$s = h(x)^T \beta$ non-linear



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

$k(x, x') = \langle h(x), h(x') \rangle$

Implicit

Measures similarity

e.g. $\exp(-\gamma |x - x'|^2)$

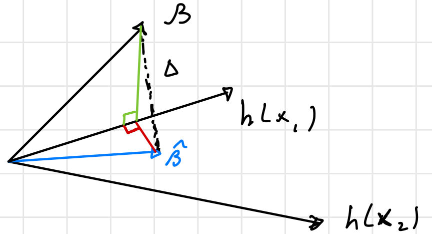
Gaussian Kernel / Radial Basis Function

- More systematic :

$$\text{Loss} = \sum_{i=1}^{n} (y_i - h(x_i)^T \beta)^2 + \lambda |\beta|^2$$

$$s_i = f(x_i) = h(x_i)^T \beta = \langle h(x_i), \beta \rangle$$

- Representer Theorem



$$h(x_i)^T \beta = h(x_i)^T \hat{\beta}$$

$$|\beta|^2 \geq |\hat{\beta}|^2$$

$$\text{So } \hat{\beta} = \sum_{i=1}^{n} c_i h(x_i)$$

$$\text{Loss} = \sum_{i=1}^{n} \Big(y_i - h(x_i)^T \underbrace{\sum_{j=1}^{n} c_j h(x_j)}_{\hat{\beta}} \Big)^2 + \lambda \Big\langle \underbrace{\sum_{i=1}^{n} c_i h(x_i)}_{\hat{\beta}}, \underbrace{\sum_{j=1} c_j h(x_j)}_{\hat{\beta}} \Big\rangle$$

$$= \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{n} c_j k_{ij} \Big)^2 + \lambda \sum_{ij} c_i c_j k_{ij}$$

$$\text{Loss}(c) = |y - Kc|^2 + \lambda c^T K c$$

$$= (y - Kc)^T (y - Kc) + \lambda c^T K c$$

$$= y^T y - c^T K y - y K c - c^T K^2 c + \lambda c^T K c$$

$$\frac{\partial L}{\partial c} = -2Ky + 2K^2 c + 2\lambda K c = 0$$

$$K^{-1}(-\cancel{2}Ky + \cancel{2}K^2 c + \cancel{2}\lambda K c) = 0$$

$$c = (K + \lambda I)^{-1} y \quad \text{Estimation}$$

Thus $\hat{\beta} = \sum_{i=1}^{n} c_i h(x_i)$

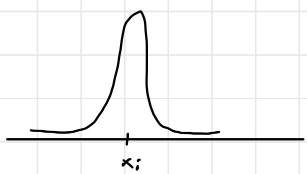
Consider $f(x_0) = h(x_0)^T \hat{\beta}$ Prediction

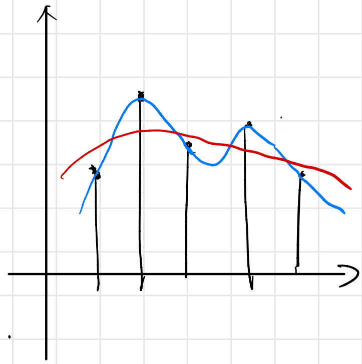
$$= h(x_0)^T \sum_{i=1}^{n} c_i h(x_i)$$

$$= \sum_{i=1}^{n} c_i \, k(x_i, x_0)$$

- $f(x) = \sum_{i=1}^{n} c_i \, k(x_i, x)$

Suppose $k(x_i, x) = \exp(-\gamma |x - x_i|^2)$





let $\lambda \to 0$, $\gamma \to \infty$

$$K(x_i, x_j) = \begin{cases} 1 & \text{if } x_i = x_j \\ \to 0 & \text{if } x_i \neq x_j \end{cases}$$

$K \to I$, $c \to Y$

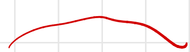
$$f(x) = \sum_{i=1}^{n} y_i \, K(x_i, x) = \begin{cases} y_i & \text{if } x = x_i \\ 0 & \text{otherwise} \end{cases}$$ memorization

for $\gamma < \infty$, $\lambda \to 0$

$C \to K^{-1} Y$ so $KC = Y$

$$f(x) = \begin{cases} y_i & \text{if } x = x_i, \text{ memorization} \\ \text{smooth interpolation} \end{cases}$$

for $\gamma < \infty$, $\lambda > 0$, smooth fitting



Theoretical underpinning

- Reproducing Kernel Hilbert Space: (RKHS)

$$\mathcal{F} = \{ f(x) = h(x)^T \beta \, , \quad \|\beta\|^2_{\ell_2} < \infty \}$$

Suppose
$$f(x) = h(x)^T \beta$$
$$g(x) = h(x)^T \alpha$$

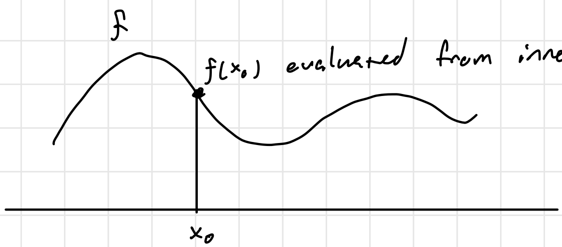
$$\langle f, g \rangle_{\mathcal{F}} = \langle \beta, \alpha \rangle_{\ell_2}$$

$$\|f\|^2_{\mathcal{F}} = \langle f, f \rangle_{\mathcal{F}} = \|\beta\|^2_{\ell_2}$$

- Reproducing Property:

$$\langle f(x) \, , \, K(x_0, x) \rangle_{\mathcal{F}} = h(x_0)^T \beta = f(x_0)$$

$$f(x) \downarrow \; h(x)^T \beta \qquad K(x_0, x) \downarrow \; h^T(x) \underbrace{h(x_0)}_{\alpha}$$



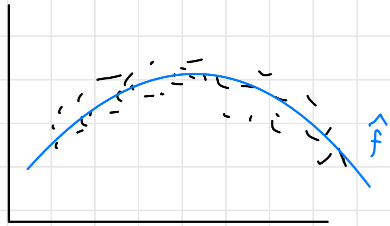
$f(x_0)$ evaluated from inner product

In regular space we need $\langle f, \delta_{x_0} \rangle$

Sweep $h(x)$, β under the rug

non-parametric regression:

$$\text{Loss}(f) = \underbrace{\sum_{i=1}^{n} (y_i - f(x_i))^2}_{\ell(y_i,\, f(x_i))} + \underbrace{\lambda |f|_{\mathcal{F}}^2}_{\text{smoothness}}$$



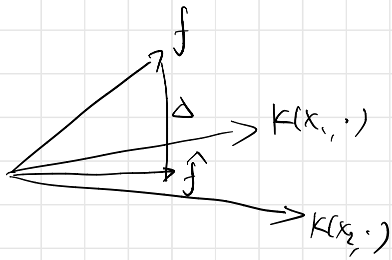
- Representer:

$$\hat{f}(x) = \underbrace{\sum_{i=1}^{n} c_i\, k(x_i, x)}_{\hat{f}(x)} + \Delta(x)$$

Proof:

$$\langle \Delta, k(x_i, \cdot) \rangle_{\mathcal{F}} = 0 \quad \forall i$$

$$\Delta(x_i)$$



$$|f|_{\mathcal{F}}^2 \geq |\hat{f}|_{\mathcal{F}}^2$$

$$f(x_j) = \sum_{i=1}^{n} c_i\, k(x_i, x_j) + \Delta(x_j)$$

$$\Delta(x_j) = \langle \Delta(x), k(x_j, x) \rangle_{\mathcal{F}} = 0$$

$$= \hat{f}(x_j)$$

$$\text{Loss}(f) = \text{Loss}(c) = |y - Kc|^2 + \lambda c^T K c$$

- Mercer Theorem: Condition for K, so that $K(x, x') = \langle h(x), h(x') \rangle$

  - Recall $K_{n \times n} = Q \Lambda Q^T$
    symmetric

    $K \geq 0$ if $\lambda_i \geq 0$, $i = 1, \dots, n$

    equivalently

    $a^T K a \geq 0 \quad \forall a \neq 0$

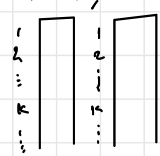
    $\downarrow$

    $b^T \Lambda b$

- kernel $k(x, x')$

Mercer Decomposition $\quad k(x, x') = \sum_k \lambda_k q_k(x) q_k(x') = \sum_k h_k(x) h_k(x') = \langle h(x), h(x') \rangle$

$$k = \sum_k \lambda_k q_k q_k^T$$



$h_k(x) = \sqrt{\lambda_k}\, q_k(x)$

$K \geq 0, \ \lambda_k \geq 0 \ \forall k$