Quiz next Wednesday

Recap: kernels

KRR

Duality and SVM

$k(\underline{x}, \underline{x}')$ kernel function $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

$\underline{\underline{K}}$ kernel matrix $K_{ij} = k(\underline{x}_i, \underline{x}_j)$

Defn: k is a symmetric PSD kernel (SPD)

$\equiv$ $\underline{\underline{K}}$ is a symmetric PSD matrix $\forall x_1, \dots x_n, \forall n$

(PSD function
PSD matrix)

$$\underline{\underline{K}} = \begin{bmatrix} k(x_1,x_1) & k(x_1,x_2) & \cdots \\ \vdots & k(x_2,x_2) & \end{bmatrix}$$

Mercer's Theorem: k is SPD kernel $\Leftrightarrow$ $k(x,x') = \langle \Phi(x), \Phi(x') \rangle$

(outside our scope $\quad k(x,x') = \sum_{i=1}^{\infty} \lambda_i h_i(x) h_i(x')$, $\lambda_i \geq 0$)

$$\Phi(x) = \left[\sqrt{\lambda_1} \cdot h_1(x) \quad \sqrt{\lambda_2} h_2(x) \ \cdots \ \sqrt{\lambda_i} h_i(x) \cdots \right]$$

Ridge regression: $\hat{w} = (\tilde{X}^T \tilde{X} + \lambda I)^{-1} \tilde{X}^T \tilde{y}$ $\qquad \hat{b} = \bar{y} - \hat{w}^T \bar{x}$

$\min_w \| \tilde{X} w - \tilde{y} \|_2^2 + \lambda \cdot \|w\|_2^2$ $\qquad \bar{y} = \frac{1}{n} \sum_i y_i$

$$(A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

$\bar{x} = \frac{1}{n} \sum_i \underline{x}_i \Rightarrow \tilde{\underline{x}}_i = x_i - \bar{x}$

$\tilde{y}_i = y_i - \bar{y}$

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

$$\hat{w} = \frac{1}{\lambda}\left[\tilde{X}^T - \tilde{X}^T(\lambda I + \tilde{X}\tilde{X}^T)^{-1}\underbrace{\tilde{X}\tilde{X}^T}_{\tilde{K}}\right]\tilde{y} \qquad \text{Define } \tilde{K}_{ij} = \langle \tilde{x}_i, \tilde{x}_j \rangle = [\tilde{X}\tilde{X}^T]_{ij}$$

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

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

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

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

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

$$= \bar{y} + \cancel{\frac{1}{\lambda}}\cdot\tilde{y}^T\cdot\cancel{\lambda}\cdot I\cdot(\tilde{K}+\lambda I)^{-1}\cdot\tilde{k}(x)$$

$$= \bar{y} + \tilde{y}^T(\tilde{K}+\lambda I)^{-1}\cdot\tilde{k}(x)$$

Preprocessing: Subtract the mean $\frac{1}{n}\sum y_i = 0$, $\frac{1}{n}\sum \underline{x}_i = \underline{0}$

$\underline{\bar{x}} = \underline{0}$ $\quad \bar{y} = 0$

$$\hat{f}(x) = \hat{\alpha}^T k'(x) = \sum \hat{\alpha}_i \langle x_i, x \rangle$$

where $\hat{\alpha} = (K + \lambda I)^{-1} y$ $\qquad K_{ij} = \langle x_i, x_j \rangle$

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

- Eg. Polynomial kernel

$$K_{ij} = \langle \Phi(x_i), \Phi(x_j) \rangle = (1 + \langle x_i, x_j \rangle)^p$$

$$\hat{f}(x) = \sum \hat{\alpha}_i \cdot \langle \Phi(x_i), \Phi(x) \rangle = \sum_i \hat{\alpha}_i (1 + \langle x_i, x \rangle)^p$$

- Eg. Linear Kernel

Least Squares (Ridge Regression) $\quad K_{ij} = \langle x_i, x_j \rangle$, $K = XX^T$

$\Phi(x_i) = x_i$ $\qquad$ Gram matrix

$n \times n$

Inhomogeneous Polynomial Kernel $K_{ij} = (x_i^T x_j + 1)^4$

$d = 1$

$n = 200$

$K_{ij} = (x_i \cdot x_j + 1)^4$

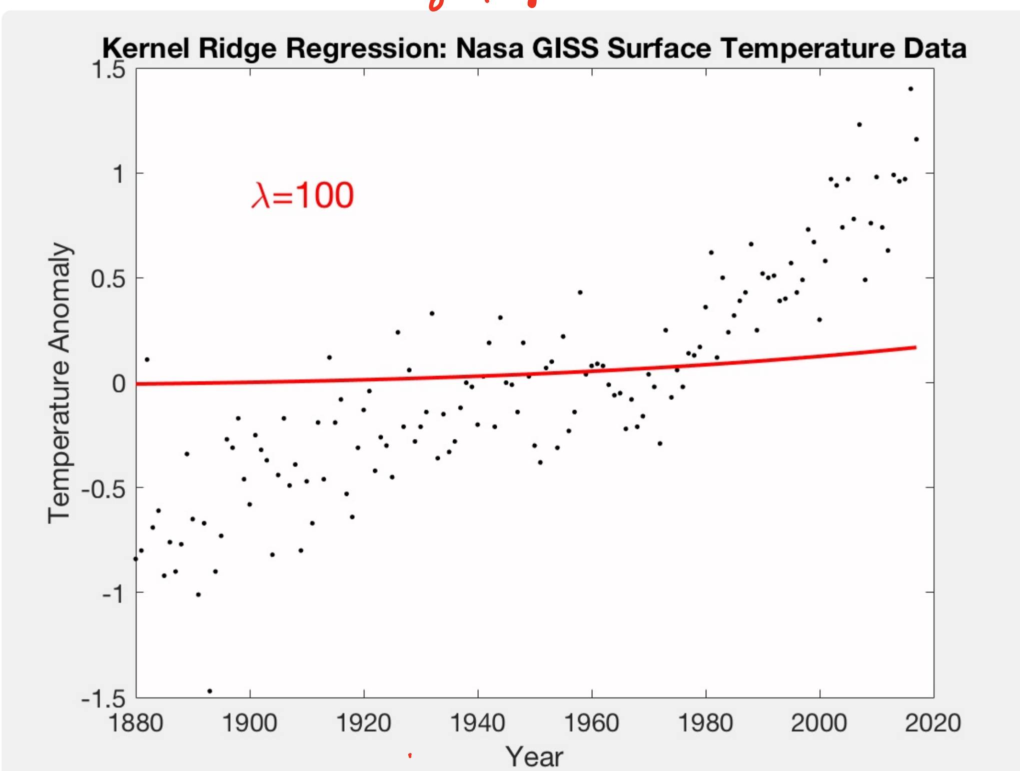
$$\underline{k}'(\underline{x}) = \begin{bmatrix} (x_1 \cdot x + 1)^4 \\ \vdots \\ (x_n \cdot x + 1)^4 \end{bmatrix}$$

$$\hat{f}(x) = \sum^{n} \hat{\alpha}_i \cdot (x_i \cdot x + 1)^4$$

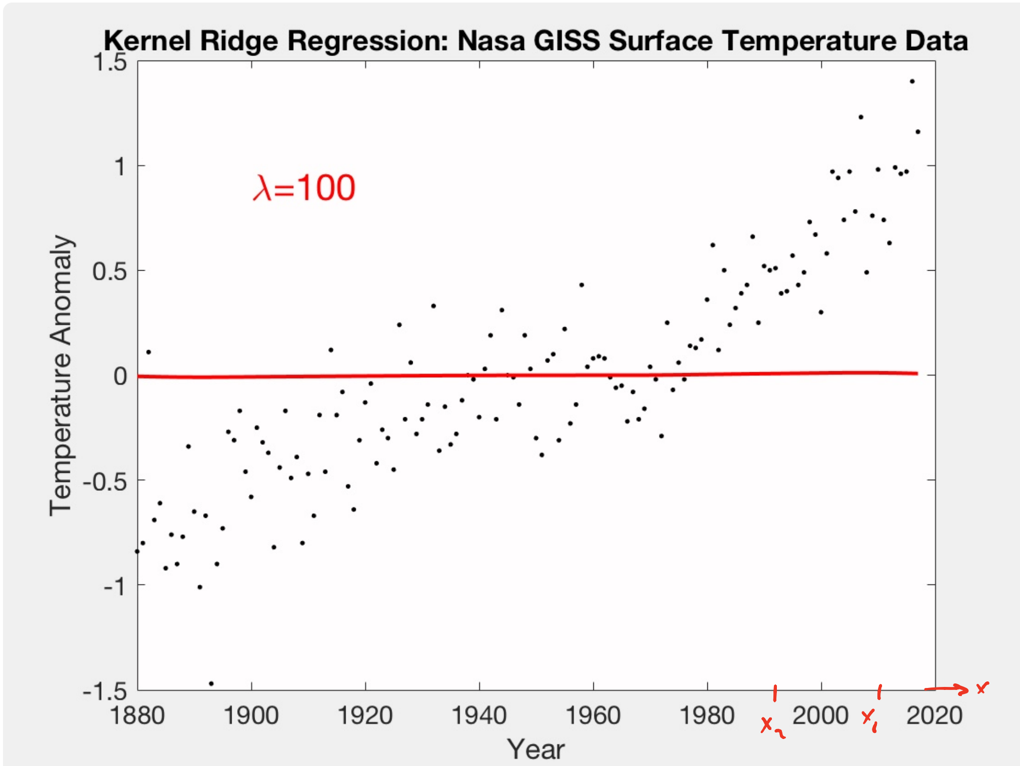
other way

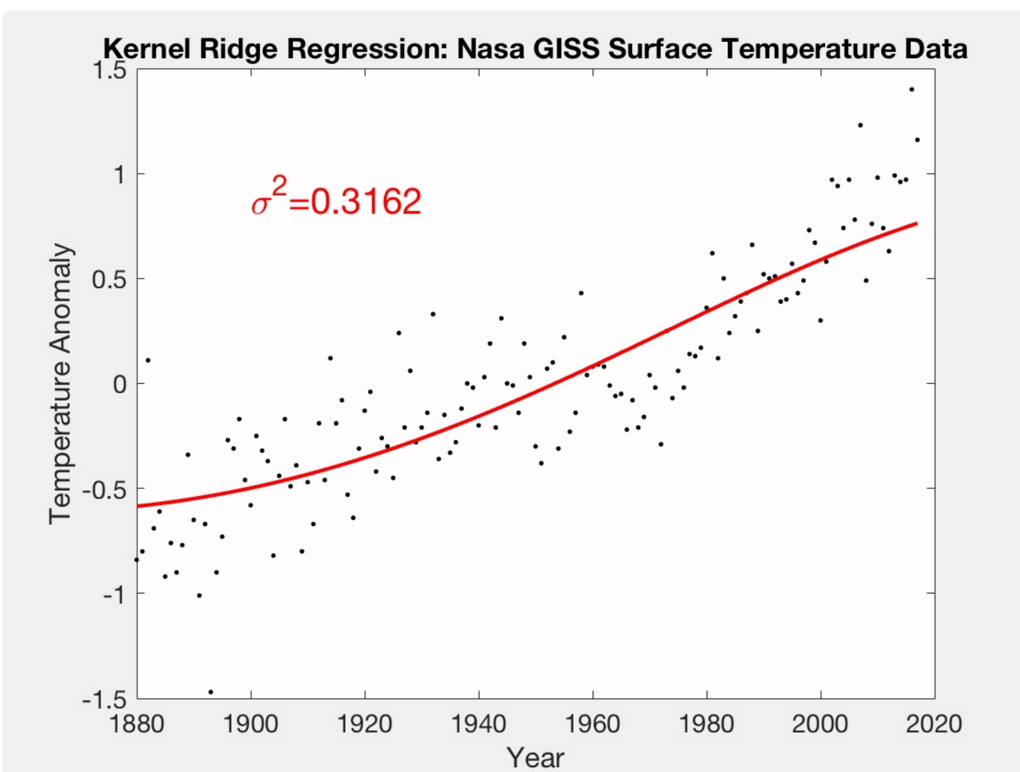
$$\underline{X}' = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^3 & x_1^4 \\ \vdots & & & & & \\ 1 & x_n & x_n^2 & \cdots & & \end{bmatrix}$$

4-deg. poly



Gaussian Kernel $K_{ij} = e^{-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}}$


Kernel Ridge Regression: Nasa GISS Surface Temperature Data
λ=100
Temperature Anomaly
Year
x_2
x_1
x


Kernel Ridge Regression: Nasa GISS Surface Temperature Data
σ²=0.3162
Temperature Anomaly
Year

Analytic

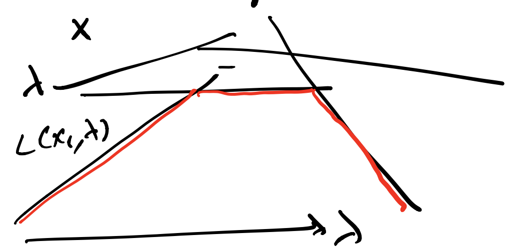
Convex Constrained Problem
$$\min_{x} f(x) \quad \text{s.t.} \quad g_i(x) \le 0 \quad i=1,\dots,m$$

Lagrangian: $L(x,\lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x)$

$\underline{\lambda} = [\lambda_1, \dots, \lambda_m]$ Lagrange multipliers
Dual variables

Lagrange Dual Function $L_D(\lambda) = \min_{x} L(x,\lambda)$

$L_D(\lambda)$ is concave in λ

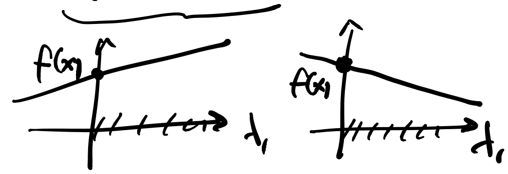


Primal optimization problem

$$\min_{x} \max_{\lambda_i \ge 0\ \forall i} L(x,\lambda) = \min_{x} f(x) \ \text{s.t.}\ g_i(x) \le 0$$

E.g. $d=1$
$g_1(x) = x$

$$\min_{x \le 0} f(x) = \min_{x} \underbrace{\overbrace{\max_{\lambda_1 \ge 0} f(x) + \lambda_1 \cdot x}}$$



$$= \min_{x} \begin{cases} \infty & \text{if } x > 0 \\ f(x) & \text{if } x \le 0 \end{cases}$$

Dual optimization problem

$$\max_{\lambda:\ \lambda_i \geq 0\ \forall i} \ \min_{x} \ L(x,\lambda) = \max_{\lambda_i \geq 0} \ L_D(\lambda)$$

Weak Duality:

$$p^* = \min_{x} \ \max_{\lambda_i \geq 0} \ L(x,\lambda)$$

$$d^* = \max_{\lambda_i \geq 0} \ \min_{x} \ L(x,\lambda)$$

Theorem: $d^* \leq p^*$