# Nonlinear Function Approximation: Kernel Method

**Qiang Liu**
UT Austin

## Supervised Learning Framework

- Given training data $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$.
- Want to find $f(x)$, such that $y_i \approx f(x_i)$.
- Empirical risk minimization:
  - Decide a function class $\mathcal{F}$.
  - Define an empirical loss function $L(f;\ \mathcal{D})$ for $f \in \mathcal{F}$
  - Solve optimization:

$$\hat{f} = \arg\min_{f \in \mathcal{F}} L(f;\ \mathcal{D}).$$


$$\mathcal{F} \triangleq \left\{ f_\theta(x) = \sum_{\ell=1}^{d} \theta_\ell x_\ell + \theta_0 \;\middle|\; \forall \theta_\ell \in \mathbb{R} \right\}$$

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

$$\min_\theta L(f_\theta, \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} \left( y^{(i)} - f_\theta(x^{(i)}) \right)^2$$

---

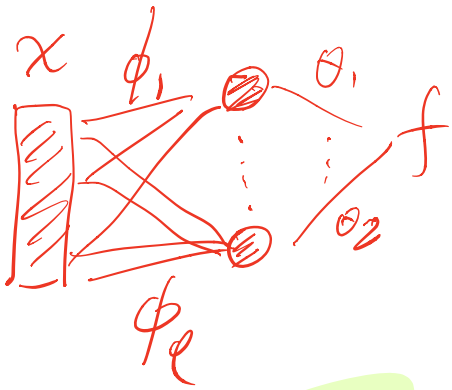
Linear: $f_\theta(x) = \sum_{\ell=1}^{d} \theta_\ell x_\ell$

Nonlinear: $f_\theta(x) = \sum_{\ell=1}^{d'} \theta_\ell \phi_\ell(x)$

$\phi_\ell$: Basis function

$\phi_\ell$ : Basis function

$\phi_\ell(x) = x^\ell \Rightarrow 1, x, x^2, \cdots$



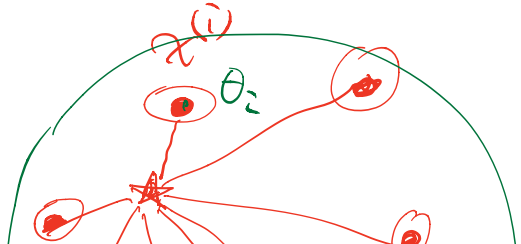

Adaptive Basis function

- Kernel Method
- Neural Networks

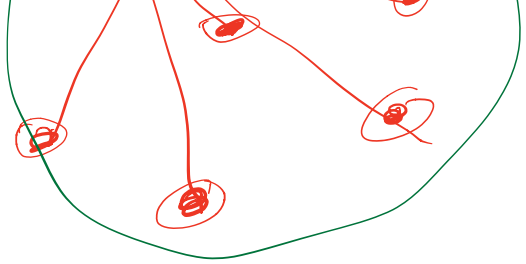
Kernel Method

$D = \{x^{(i)}, y^{(i)}\}$

Similarity function



$K(x, x') : X \times X \mapsto \mathbb{R}$.



Example:

$$K(x, x') = \exp\left(-\frac{1}{2h^2}\|x - x'\|^2\right)$$

Gaussian Radial Basis function (RBF) kernel. h: Bandwidth

Kernel: A two variable that is symmetric

$$K(x, x') = K(x', x)$$

$$K(x, x') = \exp\left(-\frac{1}{h}\|x - x'\|\right).$$

Laplace kernel.

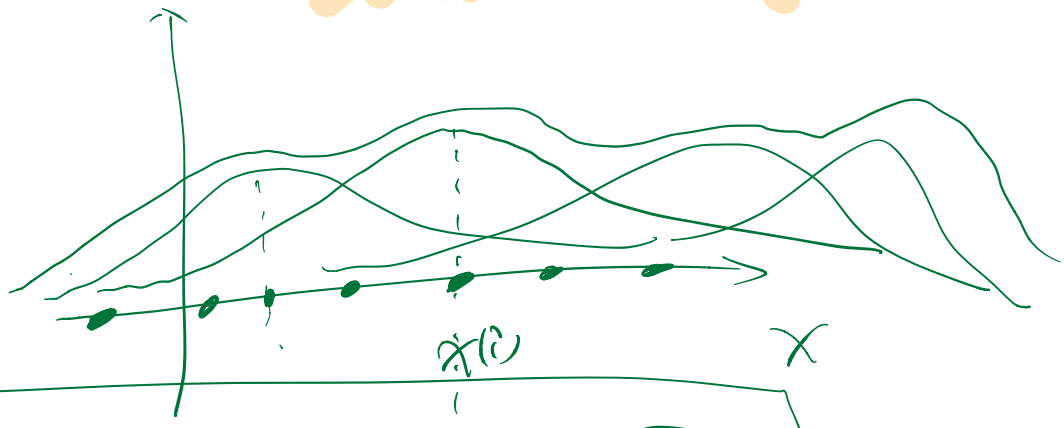
---

$$\phi(x) = \begin{bmatrix} K(x, x_1) \\ \vdots \\ K(x, x_i) \\ \vdots \end{bmatrix} \qquad \theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \vdots \end{bmatrix}$$

$$\begin{bmatrix} k(x, x_n) \end{bmatrix} \qquad \begin{bmatrix} \theta_n \end{bmatrix}$$

$$f_\theta(x) = \underline{\theta}^T \phi(x) = \sum_{i=1}^{n} \theta_i \phi_i(x)$$

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




$\mathcal{F}$ : adaptive with data $D$.

$\dim(\mathcal{F}) = n \longrightarrow +\infty$

Nonparametric Method.

$$\min_{\theta} L(\theta) = \sum_{j=1}^{n} \left( y^{(j)} - \sum_{i=1}^{n} \theta_i K(x^{(j)}, x^{(i)}) \right)$$

$$= \| Y - K\theta \|_2^2$$

$$Y = \underbrace{\begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}}_{n \times 1}, \quad K = \underbrace{\begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & & \\ K(x_n, x_1) & \cdots & K(x_n, x_n) \end{bmatrix}}_{n \times n}.$$

$$\Rightarrow \nabla L(\theta) = 2K^T(K\theta - Y)$$

$$\Rightarrow K^T K \theta = K^T Y$$

$$\Rightarrow \boxed{\theta = K^{-1} Y}$$

(Assume $K$ is invertable)

$$\Rightarrow \min_{\theta} \|Y - K\theta\|_2^2 + \alpha \Phi(\theta)$$

$L_2$: $\|\theta\|_2^2 = \theta^T \theta$

$$\|\theta\|_K^2 = \theta^T K \theta$$

---

Why $\|\theta\|_K^2$ vs. $\|\theta\|_2^2$ ?

$$D = \{x^{(1)}, x^{(2)}, x^{(3)}\}$$

Assume $x^{(1)} = x^{(2)} \neq x^{(3)}$

$x^{(1)} = x^{(2)}$

$$f(x) = \theta_1 K(x, x^{(1)}) + \theta_2 K(x, x^{(2)}) + \theta_3 K(x, x^{(3)})$$

$$= (\theta_1 + \theta_2) K(x, x^{(1)}) + \theta_3 K(x, x^{(3)})$$

$x^{(1)}$ ✓

$$\Rightarrow \|\theta\|_2^2 = \theta_1^2 + \theta_2^2 + \theta_3^2$$

Better: $(\theta_1 + \theta_2)^2 + \theta_3^2 = \|\theta\|_K^2$

$$K(x, x') = \exp\left(-\frac{1}{2h^2}\|x - x'\|^2\right).$$

$$K = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\theta^T K \theta = [\theta_1 \ \theta_2 \ \theta_3] \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

$= (\theta_1 + \theta_2)^2 + \theta_3^2$

---

$$\min_{\theta} \|Y - K\theta\|_2^2 + \lambda \theta^T K \theta.$$

$$\Rightarrow \hat{\theta} = (K + \lambda I)^{-1} Y.$$

---

Classification. $\{x^{(i)}, y^{(i)}\}$

$y^{(i)} \in \{\pm 1\}$.

$$L(\theta) = \sum_{i=1}^{n} \sigma(y^{(i)} f_\theta(x^{(i)}))$$

$$= \sum_{i=1}^{n} \sigma(y^{(i)} \sum_{j=1}^{n} \theta_j k(x^{(i)}, x^{(j)}))$$

$$+ \lambda \theta^T K \theta$$

Kernel = infinite basis function

$$\{\psi_\ell(x) : \ell = 0, 1, 2, \ldots, \infty\}$$

Example: $\psi_\ell(x) = x^\ell$

$$f(x) = \sum_{\ell=0}^{\infty} w_\ell \psi_\ell(x)$$

$$L(\underline{w}) = \hat{E}_D\left[\left(y - \sum_{\ell=0}^{\infty} w_\ell \psi_\ell(x)\right)^2\right] + \alpha \sum_{\ell=0}^{\infty} \frac{w_\ell^2}{\lambda_\ell}$$

$\lambda_\ell \geq 0$ Regularization coeff.

for $w_\ell$.

$\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$

$\lambda_\ell$ small : $w_\ell \to 0$

Define

$$\hat{w} = \arg\min_w \hat{E}_p\left[\left(y - \sum_{\ell=0}^{\infty} w_\ell \psi_\ell(x)\right)^2\right] + \sum_{\ell=0}^{\infty} \frac{w_\ell^2}{\lambda_\ell}.$$

$$\hat{f}(x) = \sum_{\ell=0}^{\infty} \hat{w}_\ell \psi_\ell(x)$$

Then:

$$\hat{f}(x) = \sum_{i=1}^{n} \theta_i K(x, x_i).$$

for $\theta_i$, and

$$K(x,x') = \sum_{i=0}^{\infty} \lambda_i \psi_i(x) \psi_i(x')$$

And.

$$\sum_{i=0}^{\infty} w_i^2/\lambda_i = \theta^T K \theta$$

$$\Longleftrightarrow \hat{\theta} = \text{argmin} \; \hat{E}_D\left[\left(y - \sum_{i=1}^{n} \theta_i K(x, x^{(i)})\right)^2\right] + \theta^T K \theta.$$

Proof:

$$\min \hat{E}_D\left[\left(y - \sum_{i=0}^{\infty} w_i \psi_i(x)\right)^2\right] + \sum_{i=0}^{\infty} \frac{w_i^2}{\lambda_i}$$

$\Rightarrow$ Define $\delta(x) = y - \sum_{\ell=0}^{\infty} \omega_\ell \psi_\ell(x)$

$$\nabla_{\omega_\ell} L(\omega) = 2\hat{E}_D[-\delta(x)\psi_\ell(x)] + \frac{2\hat{\omega}_\ell}{\lambda_\ell} = 0$$

$$\Rightarrow \hat{\omega}_\ell = \lambda_\ell \hat{E}_D[\delta(x)\psi_\ell(x)]$$

$$= \frac{\lambda_\ell}{n} \sum_{i=1}^{n} [\delta(x^{(i)})\psi_\ell(x^{(i)})]$$

$$\hat{f}(x) = \sum_{\ell=0}^{\infty} \hat{\omega}_\ell \psi_\ell(x)$$

$$= \sum_{\ell=0}^{\infty} \frac{\lambda_\ell}{n} \sum_{i=1}^{n} \delta(x^{(i)}) \psi_\ell(x^{(i)}) \psi_\ell(x)$$

$$= \sum_{\ell=0}^{\infty} \sum_{i=1}^{n} \frac{\lambda_\ell}{n} \delta(x^{(i)}) \psi_\ell(x^{(i)}) \psi_\ell(x)$$

$$= \sum_{i=1}^{n} \sum_{\ell=0}^{\infty} \frac{\lambda_\ell}{n} \delta(x^{(i)}) \psi_\ell(x^{(i)}) \psi_\ell(x)$$

$$= \sum_{i=1}^{n} \frac{\delta(x^{(i)})}{n} \boxed{\sum_{\ell=0}^{\infty} \lambda_\ell \psi_\ell(x^{(i)}) \psi_\ell(x)}$$

$$= \sum_{i=1}^{n} \frac{\delta(x^{(i)})}{n} K(x^{(i)}, x)$$

$$= \boxed{\sum_{i=1}^{n} \theta_i K(x^{(i)}, x)}$$

$$\boxed{\theta_i \triangleq \frac{\delta(x^{(i)})}{n}}$$

Two views:

① $\hat{f}(x) = \sum_{i=1}^{n} \hat{\theta}_i K(x^{(i)}, x)$

② $\hat{f}(x) = \sum_{l=0}^{\infty} \omega_l \gamma_l(x)$

$\Leftrightarrow$ $K(x, x') = \sum_{l=0}^{\infty} \lambda_l \gamma_l(x) \gamma_l(x')$

$\lambda_l \geq 0$

① If $K(x, x')$ is positive definite.

$\Leftrightarrow$ $K = [K(x^{(i)}, x^{(j)})]_{i,j=1}^{n} \geq 0$

for any $n$, $\{x^{(i)}\}$

(2) $\Leftrightarrow \exists \psi_\ell, \lambda_\ell \geq 0$

Gaussian RBF Kernel:

$$K(x, x') = \exp(-\gamma \|x - x'\|^2)$$

$$= \exp(-\gamma\|x\|^2 + 2\gamma x^T x' - \gamma\|x'\|^2)$$

$$= \exp(-\gamma\|x\|^2)\exp(-\gamma\|x'\|^2)\exp(2\gamma x^T x')$$

$$\exp(t) = 1 + t + \frac{t^2}{2} + \cdots$$

$$= \sum_{\ell=0}^{\infty} \frac{t^\ell}{\ell!}$$

$$\exp(2\gamma x^T x') = \sum_{\ell=0}^{\infty} \frac{(2\gamma x^T x')^\ell}{\ell!}$$

$\Rightarrow$ Assume $\gamma \in \mathbb{R}$.

$$k(x,x') = \sum_{\ell=0}^{\infty} \lambda_\ell \varphi_\ell(x) \varphi_\ell(x')$$

$$\varphi_\ell(x) = \exp(-\gamma x^2) x^\ell.$$

$$\lambda_\ell = \frac{(2\gamma)^\ell}{\ell!} \geq 0$$