Kernel methods: ... "Using the kernel trick" ...

(related to: kernel regression, SVM, kernel PCA, kernel discriminant analysis.)

- $\underline{\text{Kernel regression}}$ < feature embedding into RKHS; optimization perspective (representer thm.)

Nonparametric regression: $(Y_i, X_i)_{i=1}^n$ (Y_i: response, X_i: covariates.) $x^* \xrightarrow{\text{predict}} Y^*$.

Learning the regression fn:

$$f(x) = E[Y \mid X = x]$$

Recall: The prediction at x^* using linear ridge regression is $x^{*T}\hat{\beta}_R = x^{*T}(X^TX + \lambda I_d)^{-1}(X^TY)$

$$= (X \cdot x^*)^T (\cancel{X}X' + \lambda I_n)^{-1} Y$$

$$X x^* = \begin{pmatrix} \langle x_1, x^* \rangle \\ \vdots \\ \langle x_n, x^* \rangle \end{pmatrix}$$

$$(X'X)_{jk} = \langle x_j, x_k \rangle$$

- $\langle x_j, x_k \rangle$ is a (bilinear) measure of similarity b/w x_j & $x_k \in \mathbb{R}^d$. $\cos\theta = \frac{\langle x_i, x_j \rangle}{\|x_i\| \|x_j\|}$

(Idea) - Replace $\langle x_j, x_k \rangle$ with $k(x_j, x_k)$: nonlinear measure of similarity.

Define $K^* = \begin{pmatrix} k(x_1, x^*) \\ \vdots \\ k(x_n, x^*) \end{pmatrix}$ $\quad (K)_{jk} = k(x_j, x_k)$

$K^* \in \mathbb{R}^n$, $K \in \mathbb{R}^{n \times n}$

$$\boxed{\hat{f}(x^*) = K^{*T}(K + \lambda I_n)^{-1} Y}$$ Kernel Ridge Regression (KRR) estimator.

$k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

- Conditions on k: (pd (nnd) kernel)
  - i) symmetric: $k(x, x') = k(x', x) \quad \forall x, x'$
  - ii) $\forall n \in \mathbb{N}$, $\forall x_1, \dots, x_n \in \mathcal{X}$, $K_n = (k(x_i, x_j))_{ij}$ p.d. (n.n.d at least)

$K_n = X'X$ satisfies i) & ii)

Hadamard product: $(A \circ B)_{ij} = A_{ij} \cdot B_{ij}$

Claim: $k(x, x') = \langle x, x' \rangle^2$ is pd (nnd) $(J_n)_{ij} = \langle x_i, x_j \rangle^2$

Fact: if A and B are nnd, then so is $A \circ B$ (Schur's theorem)

$J_n = K_n \circ K_n$

Let $K_m(x_i, x_j) = \langle x_i, x_j \rangle^m$. $m \in \mathbb{N}_+$

Then K_m is a pd kernel $\Rightarrow$ polynomial kernel.

$x_i \in \mathbb{R}^2$

$x_i = (x_{i1}, x_{i2})$

$$\langle x_i, x_j \rangle^2 = (x_{i1} x_{j1} + x_{i2} x_{j2})^2$$
$$= x_{i1}^2 x_{j1}^2 + x_{i2}^2 x_{j2}^2 + 2 x_{i1} x_{j1} x_{i2} x_{j2}$$
$$= \langle \psi(x_i), \psi(x_j) \rangle$$

where $\psi: \mathbb{R}^2 \to \mathbb{R}^3$

$$x \mapsto (x_1^2, x_2^2, \sqrt{2}\, x_1 x_2)$$

$(J_n)_{ij} = \langle \psi(x_i), \psi(x_j) \rangle \Rightarrow J_n$ is nnd.

KRR $\Leftrightarrow$ ordinary ridge regression in the transformed feature space produced by the feature map ψ

"Kernel Engineering"

Creating new kernels out of simple ones.

Exs: if k_1 and k_2 are kernels, then so is any of the following:

i) $k(x, x') = k_1(x, x') + k_2(x, x')$

ii) $k(x, x') = f(x) \cdot k_1(x, x') f(x')$

iii) $k(x, x') = k_1(x, x') \cdot k_2(x, x')$

iv) $k(x, x') = \exp\{k_1(x, x')\}$ (Taylor expansion)

Gaussian kernel

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

Ex: Show (using i) ~ iv)) that k is pd kernel.

If $k(x,x') = \phi(\|x-x'\|)$ then k is called an isotropic kernel.

$k(x,x') = \phi(x-x')$ stationary kernel

⇓

Gaussian

Matern

Mercer's theorem (general way of getting a feature map)

If $k: \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is a p.d. fn and $\mathcal{X} \subseteq \mathbb{R}^d$ is compact, then there is an orthonormal set of functions $\{\psi_j\}$ s.t.

$$k(x,x') = \sum_{j=1}^{\infty} \lambda_j \psi_j(\cdot)\cdot\psi_j(x')$$

i.e. $x \xmapsto{\psi} (\sqrt{\lambda_1}\psi_1(x), \sqrt{\lambda_2}\psi_2(x), \dots)$

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

~~Bochners~~ Bochners Thm (complete characterization of stationary kernels)

Let $k(x,x') = \phi(x-x')$. Then k is p.d. $\Leftrightarrow$ $\hat{\phi}(\omega) \geq 0$ $\forall \omega$.

where $\hat{\phi}(\omega) = (2\pi)^{-d/2}\int e^{i\langle\omega,t\rangle}\phi(t)\,dt$.

proof: $\hat{\phi}(\omega) \geq 0\ \forall\omega \Rightarrow k$ is p.d.

Inverse-Fourier Transform: $\phi(t) = (2\pi)^{-d/2}\cdot\int e^{-i\langle\omega,t\rangle}\hat{\phi}(\omega)\,d\omega$

(For any $x_1,\dots,x_n$, and any $a_1,\dots,a_n \in \mathbb{R}$, we want to show: $\sum_{i=1}^n\sum_{j=1}^n a_i a_j k(x_i,x_j) \geq 0$)

$\sum\sum a_i a_j k(x_i,x_j)$

$= (2\pi)^{-d/2}\cdot\sum\sum a_i a_j \int e^{-i\langle\omega, x_i - x_j\rangle}\hat{\phi}(\omega)\,d\omega$

$= (2\pi)^{-d/2}\cdot\int \sum\sum a_i a_j \cdot e^{-i\omega x_i}\cdot\overline{e^{-i\omega x_j}}\cdot\hat{\phi}(\omega)\,d\omega$

$= (2\pi)^{-d/2}\int |\sum a_i e^{-i\omega x_i}|^2\,\hat{\phi}(\omega)\,d\omega$

≥ 0

This gives a recipe for constructing kernel fns

~~taking~~ Take any non-negative fun. (prob. density) and