Kernel Methods Notes

Part I: Kernels basics, kernel PCA & Ridge regression

- A kernel is a function $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that there exists a Hilbert space $\mathcal{H}$ and mapping $\phi: \mathcal{X} \to \mathcal{H}$ where $K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$

- A Hilbert Space is a vector space on which an inner product ~~...~~ $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is defined, this having the following properties:
  - $\langle af_1 + bf_2, g \rangle_{\mathcal{H}} = a \langle f_1, g \rangle_{\mathcal{H}} + b \langle f_2, g \rangle_{\mathcal{H}}$ (linear)
  - $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$ (symmetric)
  - $\langle f, f \rangle_{\mathcal{H}} \geq 0$, $= 0$ only when $f = 0$

- All kernels $K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$ are positive definite functions:
  - given arbitrary $a_1, \ldots, a_n \in \mathbb{R}$, $x_1, \ldots, x_n \in \mathcal{X}$

$$\sum_i \sum_j a_i a_j K(x_i, x_j) = \sum_i \sum_j \langle a_i \phi(x_i), a_j \phi(x_j) \rangle_{\mathcal{H}}$$

$$= \left\langle \sum_i a_i \phi(x_i), \sum_j a_j \phi(x_j) \right\rangle_{\mathcal{H}}$$

$$= \left\| \sum_i a_i \phi(x_i) \right\|^2_{\mathcal{H}} \geq 0 \quad \dashv$$

- It turns out that the opposite direction holds as well: all positive definite functions are kernels!
- Therefore, all sums of kernels $K(x, x') = K_1(x, x') + K_2(x, x')$ are kernels: for arbitrary $a_1, \dots, a_n \in \mathbb{R}$, $x_1, \dots, x_n \in \mathcal{X}$

$$\sum_i \sum_j a_i a_j K(x_i, x_j) = \sum_i \sum_j a_i a_j \left(K_1(x_i, x_j) + K_2(x_i, x_j)\right)$$

$$= \left\| \sum_i a_i \phi_1(x_i) \right\|^2_{\mathcal{H}_1} + \left\| \sum_i a_i \phi_2(x_i) \right\|^2_{\mathcal{H}_2} \geq$$

$\Rightarrow$ positive-definite $\therefore$ a kernel

- All products of kernels $K(x, x') = K_1(x, x') K_2(x, x')$ are kernels:

$$K_1(x, x') K_2(x, x') = \langle \phi_1(x), \phi_1(x') \rangle_{\mathcal{H}_1} \langle \phi_2(x), \phi_2(x') \rangle_{\mathcal{H}_2}$$

$$= \phi_1(x')^T \phi_1(x^*) \, \phi_2(x)^T \phi_2(x')$$

can always take trace of a scalar

$$= \phi_1(x')^T \phi_1(x^*) \, \mathrm{Trace}\left[\phi_2(x') \phi_2(x)^T\right]$$

can move a scalar into a trace

$$= \mathrm{Tr}\Big[\underbrace{\phi_2(x') \phi_1(x')^T}_{A^T} \underbrace{\phi_1(x^*) \phi_2(x)^T}_{B}\Big]$$

$$= \mathrm{Tr}\left[A^T B\right]$$

Frobenius product

$$= \mathrm{vec}(A)^T \mathrm{vec}(B)$$

$$= \left\langle \mathrm{vec}\left(\phi_1(x') \phi_2(x')^T\right), \mathrm{vec}\left(\phi_1(x^*) \phi_2(x)^T\right) \right\rangle_{\mathcal{H}}$$

$$= \langle \psi(x'), \psi(x) \rangle_{\mathcal{H}} = K(x, x') \checkmark$$

- Every kernel is associated with a unique RKHS $\mathcal{H}$, which has the following properties:
  - $\forall x \in \mathcal{X}$, $K(\cdot, x) \in \mathcal{H}$
  - $\forall x \in \mathcal{X}$, $\forall f \in \mathcal{H}$, $\underbrace{\langle f, K(\cdot, x)\rangle = f(x)}_{\text{reproducing property}}$

- Ex. RKHS defined by a Fourier Series

Consider the space of all periodic functions on $[-\pi, \pi]$:

$$f(x) = \sum_{l=-\infty}^{\infty} \hat{f}_l e^{ilx}$$

We can then define the ∞-D vector space spanned by the orthonormal basis $\{e^{ilx}\}_{l=-\infty}^{\infty}$, $x \in \mathbb{R}$ together with the standard L2 dot product $\langle \cdot, \cdot \rangle$, to give us a Hilbert space $\mathcal{H}$, where $\langle f, g\rangle_{L_2} = \sum_{l=-\infty}^{\infty} \hat{f}_l \bar{\hat{g}}_l$.

Is $\mathcal{H}$ an RKHS? Let $K(x,y) = K(x-y)$

$$= \sum_{l=-\infty}^{\infty} \hat{K}_l e^{il(x-y)}$$

$$= \sum_{l=-\infty}^{\infty} \hat{K}_l e^{ily} e^{-ilx}$$

We check for the reproducing property:

$$\langle f, K(\cdot, x)\rangle_{L_2} = \sum_{l=-\infty}^{\infty} \hat{f}_l \hat{K}_l e^{ilx}$$

$$= \sum_{l=-\infty}^{\infty} \hat{K}_l \hat{f}_l e^{ilx} \neq f(x)$$

Given this kernel, what is the dot product of the associated RKHS?

So $\mathcal{H}$ is not an RKHS. But we can easily modify it so that it is: $\mathcal{H}^*$ with $\langle f, g\rangle_{\mathcal{H}^*} = \sum_{l=-\infty}^{\infty} \frac{\hat{f}_l \bar{\hat{g}}_l}{\hat{K}_l}$

Now, $\langle f, K(\cdot,x)\rangle_{\mathcal{H}^*} = \sum_{l=-\infty}^{\infty} \frac{\hat{f}_l \cancel{\hat{K}_l} e^{ilx}}{\cancel{\hat{K}_l}} = \sum_{l=-\infty}^{\infty} \hat{f}_l e^{ilx} = f(x)$

$$\langle K(\cdot,x), K(\cdot,y)\rangle_{\mathcal{H}^*} = \sum_{l=-\infty}^{\infty} \frac{\hat{K}_l e^{-ilx} \cancel{\hat{K}_l} e^{ily}}{\cancel{\hat{K}_l}} = \sum_{l=-\infty} \hat{K}_l e^{il(y-x)} = K(y-x)$$

Importantly, $\langle f, f\rangle_{\mathcal{H}^*} = \|f\|^2_{\mathcal{H}^*} = \sum_{l=-\infty}^{\infty} \frac{|\hat{f}_l|^2}{\hat{K}_l}$, so ~~the~~ kernel enforces smoothness since any $f \in \mathcal{H}^*$ must have $\hat{f}_l$ that decay faster than $\hat{K}_l$ for $\|f\|^2_{\mathcal{H}^*} < \infty$, i.e. $f(\cdot)$ must be at least as smooth (low amplitudes at higher frequencies) as $K(\cdot)$.

– Kernel PCA: just like normal PCA but performed in feature space, via the reproducing property:

$f^* = \underset{\substack{f\in\mathcal{H} \\ \|f\|_{\mathcal{H}}=1}}{\operatorname{argmax}}$ variance of data projected into $\mathcal{H}$ via feature map $\phi(x) = K(x,\cdot)$ along unit vector f

$$= \cancel{\underset{\|f\|_{\mathcal{H}}=1}{\operatorname{argmax}} \langle f, C \rangle_{\mathcal{H}}} \quad \cancel{C = \frac{1}{N}\sum_i \left(\phi(x_i) - \left(\frac{1}{N}\sum_j \phi(x_j)\right)\right)^2}$$

$$= \underset{\|f\|_{\mathcal{H}}=1}{\operatorname{argmax}} \frac{1}{N}\sum_{i=1}^{N} \langle f, \phi(x_i) - \bar{\phi}\rangle^2_{\mathcal{H}} \qquad \cancel{= \frac{1}{N}\sum_i \tilde{\phi}(x_i)^2}$$

$$\bar{\phi} = \frac{1}{N}\sum_i \phi(x_i)$$

$$\tilde{\phi}(x_i) = \phi(x_i) - \bar{\phi}$$

$$\frac{1}{N}\sum_i \langle f, \tilde{\phi}(x_i)\rangle \langle f, \tilde{\phi}(x_i)\rangle$$

$$\frac{1}{N}\sum_i \langle f, \tilde{\phi}(x_i) \otimes \tilde{\phi}(x_i) f\rangle$$

$$= \underset{\|f\|_{\mathcal{H}}}{\operatorname{argmax}} \langle f, Cf\rangle, \quad C = \frac{1}{N}\sum_{i=1}^{N} \tilde{\phi}(x_i) \otimes \tilde{\phi}(x_i)$$

$$\Rightarrow \frac{\partial}{\partial f}\left[\langle f, Cf\rangle_{\mathcal{H}} + \lambda\left(\langle f,f\rangle_{\mathcal{H}} - 1\right)\right] = 0$$

$$\Leftrightarrow \quad Cf = \lambda f$$

$\Rightarrow f^* =$ largest e-vector C

↳ but this requires computing C, which lives [redacted] in $\mathbb{R}^{\infty \times \infty}$

→ How can we avoid feature space?

$\Rightarrow$ We can always express f as a linear combination of data points, without loss of generality, since any dimensions orthogonal to the space spanned by $\{\tilde{\phi}(x_i)\}_{i=1}^{N}$ will disappear in the first line $\langle f, \tilde{\phi}(x_i)\rangle_{\mathcal{H}}$, thus rendering them irrelevant to the optimization:

$$f = \sum_{i=1}^{N} \alpha_i \tilde{\phi}(x_i)$$

$$\tilde{K}(x, x') = \langle \tilde{\phi}(x), \tilde{\phi}(x')\rangle_{\mathcal{H}}$$

$$\Leftrightarrow f(\cdot) = \sum_{i=1}^{N} \alpha_i \tilde{K}(x_i, \cdot) \quad \text{(by reproducing property)}$$

Thus we need only solve for the α's:

$$Cf = \frac{1}{N}\sum_{i=1}^{N} \tilde{\phi}(x_i) \sum_{j=1}^{N} \alpha_j \langle \tilde{\phi}(x_i), \tilde{\phi}(x_j)\rangle_{\mathcal{H}}$$

$\alpha_j \tilde{K}(x$

$$= \frac{1}{N}\sum_{i=1}^{N} \tilde{\phi}(x_i) \sum_{j=1}^{N} \alpha_j \tilde{K}(x_i, x_j) \Rightarrow \langle \tilde{\phi}(x_k), Cf\rangle = \frac{1}{N}\sum_i \tilde{K}(x_k, x_i) \sum_j$$

$$\langle \tilde{\phi}(x_k), \lambda f\rangle_{\mathcal{H}} = \lambda \sum_i \alpha_i \tilde{K}(x_k, x_i) \quad \Rightarrow \frac{1}{N}\tilde{K}\tilde{K}\alpha = \lambda \tilde{K}\alpha$$

where $\tilde{K}_{ij} = \tilde{K}(x_i, x_j)$. Since this matrix is symmetric and positive semidefinite, its inverse exists, so we get the following eigenvalue equation:

$$\tilde{K}\alpha = N\lambda\alpha$$

So we can solve for α by constructing the Gram matrix $\tilde{K}$ and solving the e-value equation, giving us the directions ~~of~~ f of greatest variance, without having to work at all in feature space. (ie. biggest λ

Importantly, f is a function, so kernel PCA, as opposed to regular PCA, can give us ~~[illegible]~~ non-linear principal subspaces rather than just ~~[illegible]~~ hyperplanes (depending on the kernel).

- <u>Kernel Ridge Regression:</u> ridge regression in feature space

$$y = w^T \phi(x) + \epsilon, \quad \phi(x) \in \mathcal{H}$$

$$\Rightarrow w^* = \underset{w \in \mathcal{H}}{\arg\min} \left[\sum_{i=1}^{N} \left(y_i - \langle w, \phi(x_i) \rangle_{\mathcal{H}} \right)^2 + \lambda \|w\|_{\mathcal{H}}^2 \right]$$

$$= \underset{w \in \mathcal{H}}{\arg\min} \left[\|Y - X^T w\|_{\mathcal{H}}^2 + \lambda \|w\|_{\mathcal{H}}^2 \right], \quad X = \begin{bmatrix} \phi(x_1) & \cdots & \phi(x_N) \end{bmatrix}$$

$$= \underset{w \in \mathcal{H}}{\arg\min} \left[Y^T Y - 2 Y^T X^T w + w^T (X X^T + \lambda I) w \right]$$

completing the square

$$= \underset{w \in \mathcal{H}}{\arg\min} \left[Y^T Y + \| (XX^T + \lambda I)^{\frac{1}{2}} w - (XX^T + \lambda I)^{-\frac{1}{2}} X Y \|_{\mathcal{H}}^2 - \| (XX^T + \lambda I)^{-\frac{1}{2}} X$$

$$= (XX^T + \lambda I)^{-1} XY$$

(we could've done this by taking derivatives, but derivatives don't necessarily exist for discrete x_i, y_i)

To avoid having to do anything in feature space, we rewrite this in terms of the Gram matrix $K = X^T X$; $K_{ij} = K(x_i, x_j)$

① Via SVD:

$$\underset{D \times N}{X} = \underset{D \times D}{\left[\tilde{U}\right]} \underset{D \times N}{\begin{bmatrix} \tilde{S} \\ 0 \end{bmatrix}} \underset{N \times N}{\left[\tilde{V}\right]}$$

(orthogonal) (diagonal) (orthogonal)

Let
$$U = \tilde{U} \qquad D \times D$$
$$S = \begin{bmatrix} \tilde{S} & 0 \\ 0 & 0 \end{bmatrix} \qquad D \times D$$
$$V = \begin{bmatrix} \tilde{V} & 0 \end{bmatrix} \qquad N \times D$$

such that $X = USV^T$

We then have:

$$w^* = (US^2U^T + \lambda I)^{-1} USV^T Y$$
$$= U(S^2 + \lambda I)^{-1} U^T U S V^T Y$$
$$= US(S^2 + \lambda I)^{-1} V^T Y$$
$$= USV^T V (S^2 + \lambda I)^{-1} V^T Y$$
$$= USV^T (V^T S^2 V + \lambda I)^{-1} Y$$
$$= X(X^T X + \lambda I)^{-1} Y$$
$$= X(K + \lambda I)^{-1} Y$$

(can do this since S is diagonal and square (hence the change from the usual SVD))

② Via Woodbury Identity:

$$w^* = (XX^T + \lambda I)^{-1} XY$$

$$= \left(\lambda^{-1} I - \lambda^{-1} X (\lambda^{-1} X^T X + I)^{-1} X^T \lambda^{-1}\right) XY$$

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

$$= \Big[\lambda^{-1} X + \lambda^{-1} X (\lambda^{-1} X^T X + I)^{-1}$$
$$- \lambda^{-1} X (\lambda^{-1} X^T X + I)^{-1}$$
$$- \lambda^{-1} X (\lambda^{-1} X^T X + I)^{-1} \lambda^{-1} X^T X\Big] Y$$

$$= \Big[\cancel{\lambda^{-1} X} + \lambda^{-1} X (\lambda^{-1} X^T X + I)^{-1}$$
$$\cancel{- \lambda^{-1} X (\lambda^{-1} X^T X + I)^{-1} (\lambda^{-1} X^T X + I)}\Big] Y$$

$$= \lambda^{-1} X (\lambda^{-1} X^T X + I)^{-1} Y$$

$$= X(\underbrace{X^T X}_{K} + \lambda I)^{-1} Y$$

Thus, the optimal weights are a weighted sum of the data points: $w^* = \sum_i \alpha_i \phi(x_i)$, $\underline{\alpha} = (K + \lambda I)^{-1} Y$

Note that w^* is a function in $\mathcal{H}$, such that its smoothness is constrained by the kernel since $\|w^*\|_{\mathcal{H}}^2 < \infty$. The larger our regularizing ~~the~~ constant λ, the smoother our resulting regression function $\langle w^*, \phi(x) \rangle_{\mathcal{H}} = w^*(x)$ will be