Lecture 10: Featuremap vs Kernel Perspectives

So far our discussion of kernel methods & in particular, representer theorems has been focused on the so called "kernel perspective"

equiv

$$\min_{f \in \mathcal{H}} \ell(f(x_1), \dots, f(x_n)) + R(\|f\|)$$

$$\min_{z \in \mathbb{R}^n} \ell(z) + R\left(\sqrt{z^T K(X,X)^{-1} z}\right)$$

The second formula involves the kernel matrix $K(X,X)$, hence the name "kernel perspective".

Conceptually this method is very elegant but it has two potential issues.

(i) $K(X,X)$ can be ill-conditioned

(ii) $K(X,X) \in \mathbb{R}^{n \times n}$ & it is typically dense

the loss ℓ involves more points ie $|X| = n$ grows, the cost of computing $K(X,X)^{-1} z$ grows rapidly!

Many potential work arounds exist for this issue some keywords are

- Nystrom or low rank approx
- Sparse GPs
- Sparse Cholesky factorizations

We won't have time to cover these but focus our attention on the so called "feature map perspective" which is both efficient & mathematically enlightening.

10.1 The feature map perspective

Recall that we showed in the RKHS thm. That every PDS kernel K is associated with a feature map $\varphi : \mathcal{X} \to \mathcal{F}$ ← a Hilbert space

$$K(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{F}}$$

We showed (in HW2) that indeed the RKHS $\mathcal{H}$ associated with K is unique! But the feature maps φ are not unique.

To see this consider the Canonical feature map

$$\varphi(x) = K(x, \cdot), \quad \varphi : \mathcal{X} \to \mathcal{H}$$

as one example. At the same time, by Mercer's thm we have that

$$K(x,y) = \sum_{j=1}^{\infty} \lambda_j \psi_j(x) \psi_j(y) \quad \text{ortho sum in } L^2(\mathcal{X}, \mu)$$

$\tilde{\psi}_j = \sqrt{\lambda_j}\,\psi_j$

$$\equiv \sum_{j=1}^{\infty} \tilde{\psi}_j(x) \tilde{\psi}_j(y)$$

$$= \left\langle \{\tilde{\psi}_j(x)\}, \{\tilde{\psi}_j(y)\} \right\rangle_{\ell^2}$$

Thus, we could also pick

$$\varphi : \mathcal{X} \to \ell^2(\mathcal{X}, \mu) \qquad \varphi(x) = \sum_{j=1}^{\infty} \tilde{\psi}_j(x) \psi_j$$

Indeed observe that, by Mercer's thm

$$K(x,x) = \sum_{j=1}^{\infty} \lambda_j \psi_j^2(x) < +\infty \quad \& \text{ so,}$$

$$\left\| \sum_{j=1}^{\infty} \sqrt{\lambda_j}\, \psi_j(x) \psi_j \right\|^2_{L^2(\mathcal{X}, \mu)} = \left\langle \sum_{j=1}^{\infty} \sqrt{\lambda_j}\, \psi_j(x) \psi_j, \sum_{k=1}^{\infty} \sqrt{\lambda_k}\, \psi_k(x) \psi_k \right\rangle_{L^2(\mathcal{X}, \mu)}$$

$$= \sum_{j=1}^{\infty} \lambda_j \psi_j^2(x) < +\infty$$

Thus, $\sum_{j=1}^{\infty} \tilde{\psi}_j(x)\psi_j \in L^2(\chi,\mu)$ Now consider

$$f=\sum_{j=1}^{\infty} \alpha_j \psi_j \,, \quad g=\sum_{j=1}^{\infty} \beta_j \psi_j \,, \quad \{\alpha_j\}_{j=1}^{\infty}, \{\beta_j\}_{j=1}^{\infty} \in \ell^2$$

& inner product $\langle f, g\rangle := \sum_{j=1}^{\infty} \frac{1}{\lambda_j} \alpha_j \beta_j$

Now observe,

$$\langle f, K(x,\cdot)\rangle = \left\langle \sum_{j=1}^{\infty} \alpha_j \psi_j \,, \frac{1}{\lambda_j} \sum_{k=1}^{\infty} \lambda_j \psi_j(x) \psi_j \right\rangle_{L^2(\chi,\mu)}$$

$$= \sum_{j=1}^{\infty} \alpha_j \psi_j(x) = f(x) \quad \text{(Reproducing Property)}$$

$$\langle f, f\rangle = \sum_{j=1}^{\infty} \frac{\alpha_j^2}{\lambda_j} < +\infty \qquad \leftarrow \left\{\frac{\alpha_j}{\sqrt{\lambda_j}}\right\} \in \ell^2$$

Based on these calculations (skipping detailed proof) we can show that

Thm: Suppose Mercer's thm holds. Then the RKHS of a PDS kernel K can be identified as

$$H = \left\{ f : \chi \to \mathbb{R} \;\middle|\; f = \sum_{j=1}^{\infty} \tilde{\alpha}_j \tilde{\psi}_j \,, \sum_{j=1}^{\infty} \tilde{\alpha}_j^2 < +\infty \right\}$$

Note: The $\tilde{\psi}_j$ in our notation are not the eigenfunctions of K & have been scaled by the eigenvalues $\psi_j = \sqrt{\lambda_j}\,\tilde{\psi}_j$

10.2 Linear methods & Kernels

The above, spectral characterization of RKHS's has important implications in practice. Let us demonstrate the ideas with an example in Regression:

Given $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ & data

$$\mathbb{R}^d \ni \underline{y} = f^t(X) + \underline{\varepsilon}, \quad \underline{\varepsilon} \overset{iid}{\sim} N(0, \sigma^2 I)$$

Approximate f^t

The standard kernel approach via rep thm gives

$$f^* = \underset{f \in H}{\operatorname{argmin}} \frac{1}{2\sigma^2}\|f(X) - \underline{y}\|_2^2 + \lambda \|f\|^2$$

$$\Rightarrow f^* = K(\cdot, X) K(X,X)^{-1} \underline{z}^*$$

where $\underline{z}^* = \underset{\underline{z} \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2\sigma^2}\|\underline{z} - \underline{y}\|_2^2 + \lambda \underline{z}^T K(X,X)^{-1} \underline{z}$

Now suppose n is large, then we need to invert $K(X,X) \in \mathbb{R}^{n \times n}$ which is also dense!

Linear solve will cost us $O(n^3)$!

Note: Standard linear solves are inefficient because they do not use the structure of the kernel matrix. There are many ideas for improving this performance & this is an active area of research. eg. see

- Snelson & Ghahramani, "Local & Global Sparse Gaussian Process Approximations" (2007)
- Cao et al, "Variational sparse Cholesky approximation for latent Gaussian processes via double Kullback-Liebler minimization" (2023).

We can also, reformulate the above problem using the spectral approach to RKHSs as

$$f^* = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{2\sigma^2} \|f(X) - Y\|_2^2 + \lambda \sum_{j=1}^{\infty} \tilde{\alpha}_j^2$$

$$\text{s.t. } f = \sum_{j=1}^{\infty} \tilde{\alpha}_j \cdot \tilde{\psi}_j$$

Or equivalently, $f^* = \sum_{j=1}^{\infty} \tilde{\alpha}_j^* \tilde{\psi}_j$ where

$$\tilde{\alpha}^* = \underset{\tilde{\alpha} \in \mathbb{R}^\infty}{\operatorname{argmin}} \frac{1}{2\sigma^2} \left\| \sum_{j=1}^{\infty} \tilde{\alpha}_j \tilde{\psi}_j(X) - Y \right\|_2^2 + \lambda \|\tilde{\alpha}\|_{\ell^2}^2$$

(ℓ^2 norm)

This is an instance of a so-called "feature map" formulation of our regression problem. We can further approximate f^* simply by truncating the expansion:

$$\tilde{f} = \sum_{j=1}^{m} \tilde{\alpha}_j \tilde{\psi}_j \quad \text{& solve}$$

$$\underline{\tilde{\alpha}}^* = \underset{\underline{\tilde{\alpha}} \in \mathbb{R}^m}{\operatorname{argmin}} \frac{1}{2\sigma^2} \left\| \sum_{j=1}^{m} \tilde{\alpha}_j \tilde{\psi}_j(X) - Y \right\|_2^2 + \lambda \|\underline{\tilde{\alpha}}\|_2^2$$

(Euclidean norm)

indeed, define

$$A_{ij} = \tilde{\psi}_j(x_i)$$

then we have

$$\underline{\tilde{\alpha}}^* = \underset{\underline{\tilde{\alpha}}}{\operatorname{argmin}} \frac{1}{2\sigma^2} \left\| A\underline{\tilde{\alpha}} - Y \right\|_2^2 + \lambda \|\underline{\tilde{\alpha}}\|_2^2$$

which is nothing more than a penalized least squares problem!

The feature map perspective allows us to solve a problem for $\tilde{\alpha}^* \in \mathbb{R}^m$ after a small truncation error versus solving for $\underline{z}^* \in \mathbb{R}^n$. If the kernel K is "nice" ie, its eigenvalues λ_j decay rapidly then we can take $m \ll n$.

A few observations:

(i) The above calculations show that all penalized least squares problems, & more broadly, all linear methods with ansatz of the form $f = \sum_j \alpha_j \psi_j$ with ℓ^2-penalty on $\{\alpha_j\}$ coincide with kernel methods!

(ii) The function up to m-terms is the same as a rank m-approx. of the kernel K ie,

$$\tilde{K}(x,y) = \sum_{j=1}^{m} \tilde{\psi}_j(x)\tilde{\psi}_j(y) \approx K(x,y)$$