Lecture 11 - Random feature methods

Let us review what we have learned about the kernel vs feature map perspective so far:

In both cases we consider problems of the form

$$f^* = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \; \ell(f(x_1), \dots, f(x_n)) + R(\|f\|)$$

where $\mathcal{H}$ is the RKHS of a kernel K.

If we employ the so called "Kernel Perspective" then we use the representer formula

$$\begin{cases} f^* = K(\cdot, X) K(X,X)^{-1} \underline{z}^* \\ \underline{z}^* = \underset{\underline{z} \in \mathbb{R}^n}{\operatorname{argmin}} \; \ell(\underline{z}) + R\left(\sqrt{\underline{z}^T K(X,X)^{-1} \underline{z}}\right) \end{cases}$$

If we employ the so called "feature map perspective" then, we may use Mercer's theorem to write

$$K(x,y) = \sum_{j=1}^{\infty} \lambda_j \psi_j(x) \psi_j(y) \qquad f^* = \sum_{j=1}^{\infty} \alpha_j^* \psi_j$$

& solve the problem

$$\alpha^* = \underset{\alpha^* \in \ell^2}{\operatorname{argmin}} \; L\left(\sum_{j=1}^{\infty} \alpha_j^* \psi_j(X)\right) + R\left(\left(\sum_{j=1}^{\infty} \frac{\alpha_j^2}{\lambda_j}\right)^{1/2}\right)$$

The Kernel vs Feature map perspective often comes down to flexibility vs computational cost!

- Kernel perspective allows us to use "all of the features" but needs us to deal with the $K(X,X)$ matrix which is $n \times n$. $O(n^2)$ to construct!
- The feature map perspective allows us to conveniently truncate the sum $f^* = \sum_{j=1} \alpha_j^* \psi_j$ to as many terms as we would like.

However, there is still another hidden cost to the feature map perspective & that is the cost of finding the λ_j & ψ_j. Recall these are the eigen pairs of the operator

$$T_K f(x) = \int_{\mathcal{X}} K(x,y) f(y) \, d\mu(y)$$

ie,

$$T_K \psi_j(x) = \int_{\mathcal{X}} K(x,y) \psi_j(y) \, d\mu(y) = \lambda_j \psi_j(x)$$

So one needs to solve these eigenproblems.

This is potentially costly, for example if $\mathcal{X}$ is a non-trivial set.

For example let $\mathcal{X} \subset \mathbb{R}^d$ be a smooth domain & let $K(x,y)$ be the Green's function of an elliptic operator, eg. $(-\Delta + \tau^2 I)^\alpha$ where $\alpha > d/2$. Take μ to be the Lebesgue measure on $\mathcal{X}$. Then the eigen functions ψ_j coincide with those of $-\Delta$ on $\mathcal{X}$ & the λ_j are given by $\lambda_j = \frac{1}{(\sigma_j + \tau^2)^\alpha}$ where σ_j are eigen values of $-\Delta$.

so, we need PDE solvers to approximate these! & the prob is potentially high dim!

so this is not quite how we use feature maps in practice. At least not in modern applications.

One solution is to explicitely prescribe the features. That is pick functions

$$\psi_j : \mathcal{X} \to \mathbb{R} \qquad \text{for } j = 1, \dots, m$$

& implicitely define

$$K(x,y) = \sum_{j=1}^{m} \psi_j(x)\,\psi_j(y)$$

(we could also incorporate the eigenvalues λ_j).

An example of this would be a fourier method

$$\psi_j(x) = \exp(i\pi k x)$$

or polynomials

$$\psi_j(x) = x^j$$ – or any orth. polynomial.

This is very useful of course but it often defines the kernel implicitely & will limit our ability to "learn" or "adapt" kernels.

11.1 Bochner's thm & Random features

A popular & elegant approach to overcome the above challenges, & to sort of have the best of both worlds, is to use the so called Random feature approach that is based on the celebrated theorem of Bochner.

Def^n: We say that a kernel $K(x,y)$ is translation invariant / stationary if

$$K(x,x') = \kappa(x-x') \quad \forall x,x' \in \mathcal{X}$$

\kappa

Thm (Bochner)

A shift invariant kernel $K(\underline{x},\underline{x}') = \kappa(\underline{x}-\underline{x}')$ on $\mathbb{R}^d$ is positive definite iff there exists a finite, non-negative Borel measure Λ on $\mathbb{R}^d$ such that

$$\kappa(\underline{x}-\underline{x}') = \int_{\mathbb{R}^d} \exp(i\underline{\omega}^T(\underline{x}-\underline{x}'))\, d\Lambda(\underline{\omega}) \qquad (*)$$

Some Observations:

(i) Observe that ⊛ is nothing but the inverse Fourier transform of Λ, this is called the Characteristic function of Λ.

(ii) Since Λ is finite we can always renormalize K such that $K(0) = 1$ & so Λ becomes a probability measure.

(iii) Thus, Bochner's thm says that:

"All pos. def. shift invariant kernels are proportional to the characteristic function of a probability measure"

Observation (iii) is the key to the so called Random feature approach introduced in the Seminal paper

Rahimi & Recht, "Random features for Large Scale kernel machines" (2007).

Suppose K is normalized so that Λ is a prob. meas. Then we can write

$$
\begin{aligned}
K(\underline{x},\underline{x}') &= K(\underline{x}-\underline{x}') \\
&= \int_{\mathbb{R}^d} \exp\left(i\,\underline{w}^T(\underline{x}-\underline{x}')\right) d\Lambda(\underline{w}) \\
&= \mathbb{E}_{\underline{w}\sim\Lambda} \exp\left(i\,\underline{w}^T\underline{x}\right)\exp\left(-i\,\underline{w}^T\underline{x}'\right)
\end{aligned}
$$

φ is complex valued here

$$
=: \left\langle \varphi(\underline{x};\underline{w}),\ \varphi(\underline{x}';\underline{w}) \right\rangle_{L^2(\mathbb{R}^d,\Lambda)}
$$

Monte Carlo Approx

$$
\approx \frac{1}{N}\sum_{j=1}^{N} \exp\left(i\,\underline{w}_j^T\underline{x}\right)\exp\left(-\underline{w}_j^T\underline{x}'\right)
$$

where $\underline{w}_j \overset{iid}{\sim} \Lambda$

$$
\begin{aligned}
&= \sum_{j=1}^{N} \frac{1}{\sqrt{N}}\exp\left(i\,\underline{w}_j^T\underline{x}\right)\frac{1}{\sqrt{N}}\exp\left(i\,\underline{w}_j^T\underline{x}'\right) \\
&= \varphi^N(\underline{x})^T\ \overline{\varphi^N(\underline{x}')} \\
&\equiv K^N(\underline{x},\underline{x}')
\end{aligned}
$$

Comp. Conjugate

Thus, to approx K with K^N, we only need to be able to sample from $\Lambda(\underline{w})$. Luckily this measure is known for many commonly used kernels. For example:

Table 2.1: Well-known kernel functions and their corresponding spectral densities. More examples can be found in Rasmussen and Williams (2005), Rahimi and Recht (2007), Fukumizu et al. (2009b), Kar and Karnick (2012), Pham and Pagh (2013). K_ν is the modified Bessel function of the second kind of order ν and Γ is the Gamma function. We also define $h(\nu, d\sigma) = \frac{2^d \pi^{d/2}\Gamma(\nu+d/2)(2\nu)^\nu}{\Gamma(\nu)\sigma^{2\nu}}$.

Kernel	$k(\mathbf{x}, \mathbf{x}')$	$\Lambda(\boldsymbol{\omega})$
Gaussian	$\exp\left(-\frac{\lVert\mathbf{x}-\mathbf{x}'\rVert_2^2}{2\sigma^2}\right),\ \sigma > 0$	$\frac{1}{(2\pi/\sigma^2)^{d/2}}\exp\left(-\frac{-\sigma^2\lVert\boldsymbol{\omega}\rVert_2^2}{2}\right)$
Laplace	$\exp\left(-\frac{\lVert\mathbf{x}-\mathbf{x}'\rVert_1}{\sigma}\right),\ \sigma > 0$	$\prod_{i=1}^d \frac{\sigma}{\pi(1+\omega_i^2)}$
Cauchy	$\prod_{i=1}^d \frac{2}{1+(x_i-x_i')^2}$	$\exp\left(-\lVert\boldsymbol{\omega}\rVert_1\right)$
Matérn	$\frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\lVert\mathbf{x}-\mathbf{x}'\rVert_2}{\sigma}\right)K_\nu\left(\frac{\sqrt{2\nu}\lVert\mathbf{x}-\mathbf{x}'\rVert_2}{\sigma}\right)$ $\sigma > 0,\ \nu > 0$	$h(\nu, d\sigma)\left(\frac{2\nu}{\sigma^2}+4\pi^2\lVert\boldsymbol{\omega}\rVert_2^2\right)^{\nu+d/2}$

- Table from "Kernel Mean Embedding of distributions: A review & Beyond" by Muandet, Fukumizu, Sriperumbudur, & Schölkopf, (2020).