Lecture 12 - Random features Continued

In the last lecture we introduced the idea of random (Fourier) features via Bochner's theorem

$$K(\underline{x},\underline{x}') = k(\underline{x}-\underline{x}') = \mathbb{E}_{\omega\sim\Lambda} \exp(i\underline{\omega}^T\underline{x})\exp(-i\underline{\omega}^T\underline{x}')$$

$\varphi_j : \mathbb{R}^d \to \mathbb{C}$

$$\approx \frac{1}{N}\sum_{j=1}^{N} \underbrace{\exp(i\underline{\omega}_j^T\underline{x})}_{\varphi_j(\underline{x})} \underbrace{\exp(-i\underline{\omega}_j^T\underline{x}')}_{\overline{\varphi_j(\underline{x}')}}$$

In fact since both k & Λ are real valued we infer that we can choose our feature maps to be real valued

$$\exp(i\underline{\omega}^T(\underline{x}-\underline{x}')) = \cos(\underline{\omega}^T(\underline{x}-\underline{x}')) + \cancel{i\sin(\underline{\omega}^T(\underline{x}-\underline{x}'))}$$

$$= \cos(\underline{\omega}^T\underline{x})\cos(\underline{\omega}^T\underline{x}') + \sin(\underline{\omega}^T\underline{x})\sin(\underline{\omega}^T\underline{x}')$$

Thus, we can define

$\varphi_j : \mathbb{R}^d \to \mathbb{R}^2$ $\quad \varphi_j(\underline{x}) = \left(\cos(\underline{\omega}_j^T\underline{x}), \sin(\underline{\omega}_j^T\underline{x})\right)$

& take the approx

$$\begin{cases} K(\underline{x},\underline{x}') \approx \frac{1}{N}\sum_{j=1}^{N} \varphi_j^T(\underline{x})\,\varphi_j(\underline{x}') \\ \underline{\omega}_j \overset{iid}{\sim} \Lambda \end{cases}$$

These are not the only possible ways to construct random features. Another option (Preferred by Rahimi & Recht) is to take

$$\varphi_j(\underline{x}) = \cos\left(\sqrt{2}\,\underline{\omega}_j^T \underline{x} + b_j\right),$$

$\varphi_j : \mathbb{R}^d \to \mathbb{R}$

$$\underline{\omega}_j \overset{iid}{\sim} \Lambda \quad \underline{b} \sim U[0, 2\pi]$$

Observe: Here we are exploiting the lack of uniqueness of feature maps! ie, in general there are many maps $\varphi : \mathcal{X} \to \mathcal{F}$ s.t.

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

The question remains, if we use random features then how do we estimate the RKHS norms? There is a simple calculation that reveals this (although it is not a complete proof).

Suppose $N \geq 1$

$$K_N(\underline{x},\underline{y}) = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \varphi_j(\underline{x}) \varphi_j(\underline{x}')$$
$$= \sum_{j=1}^{N} \left(\frac{1}{\sqrt{N}} \varphi_j(\underline{x})\right)\left(\frac{1}{\sqrt{N}} \varphi_j(\underline{x}')\right)$$

Now consider $f(\underline{x}) = \sum_{j=1}^{N} \alpha_j \left(\frac{1}{\sqrt{N}} \varphi_j(\underline{x})\right)$

& try to satisfy the reproducing property

$$\langle f, K_N(\underline{x},\cdot)\rangle = \left\langle \sum_j \alpha_j \left(\frac{1}{\sqrt{N}} \varphi_j\right), \sum_j \underbrace{\left(\frac{1}{\sqrt{N}} \varphi_j(\underline{x})\right)}\left(\frac{1}{\sqrt{N}} \varphi_j\right) \right\rangle$$

$$= f(\underline{x}) = \sum_{j=1}^{N} \alpha_j \underbrace{\left(\frac{1}{\sqrt{N}} \varphi_j(\underline{x})\right)}$$

Same coeff.

from this we see

$$f = \sum_{j=1}^{N} \alpha_j \varphi_j , \quad g = \sum_{j=1}^{N} \beta_j \varphi_j$$

Then $\langle f, g \rangle_N = \underline{\alpha}^T \underline{\beta}$ & so $\|f\|_N = \|\underline{\alpha}\|_2$

Be careful, the $\|\cdot\|_N$ norm is the RKHS norm corresponding to K_N & not $K = \lim_{N\to\infty} K_N$!

An example in Classification

Consider $\mathcal{X} \subseteq \mathbb{R}^d$ & a function $\ell : \mathcal{X} \longrightarrow \{-1, +1\}$. eg $\mathcal{X} \subseteq \mathbb{R}^{784}$ for MNIST! The function ℓ assigns a label to any point $\underline{x} \in \mathcal{X}$. Now suppose we are given a training data $X = \{\underline{x}_1, \dots, \underline{x}_n\} \subset \mathcal{X}$ along with labels $\underline{Y} = \{y_1, \dots, y_n\} \subseteq \{-1, +1\}^n$ & $y_j \approx \ell(\underline{x}_j)$

Goal: given a new point $\underline{x}_{n+1} \in \mathcal{X}$ predict $\ell(\underline{x}_{n+1})$.

we will do this using the so called probit method. Fist, we need a model for the y_j.

We assume

$$(*) \qquad y_j = \mathrm{Sign}\left(f(\underline{x}_j) + \varepsilon_j\right)$$

where $\varepsilon_j \overset{iid}{\sim} \psi$, for eg $\psi(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{t^2}{2\sigma^2}\right)$

& $\mathrm{sign}(t) = \begin{cases} +1 & \text{if } t \geq 0 \\ -1 & \text{if } t < 0 \end{cases}$. The function $f : \mathcal{X} \to \mathbb{R}$ is called a latent variable / function. It is a convenient tool in our model.

Based on $(*)$ we can formulate a log-likelihood / loss function for f.

Suppose ψ is symmetric. Then

$$\mathbb{P}r(y_j = +1 \mid f) = \mathbb{P}r\left(f(\underline{x}_j) + \varepsilon_j \geq 0\right)$$

$$= \mathbb{P}r\left(\varepsilon_j \geq -f(\underline{x}_j)\right) = \int_{-f(x_j)}^{\infty} \psi(t)\, dt$$

$$= \int_{-\infty}^{f(x_j)} \psi(t)\, dt = \Psi\left(f(\underline{x}_j)\right) = \Psi\left(y_j f(\underline{x}_j)\right)$$

where Ψ is the CDF of ψ.

Repeating a similar calculation as above shows

$$\mathbb{P}r(y_j = -1 \mid f) = \mathbb{P}r\left(f(\underline{x}_j) + \varepsilon_j < 0 \right)$$

$$= \Psi(-f(\underline{x}_j)) = \Psi(y_j f(\underline{x}_j))$$

In summary $\mathbb{P}r(y_j \mid f) = \Psi(y_j f(\underline{x}_j))$

Since the ε_j are iid we have

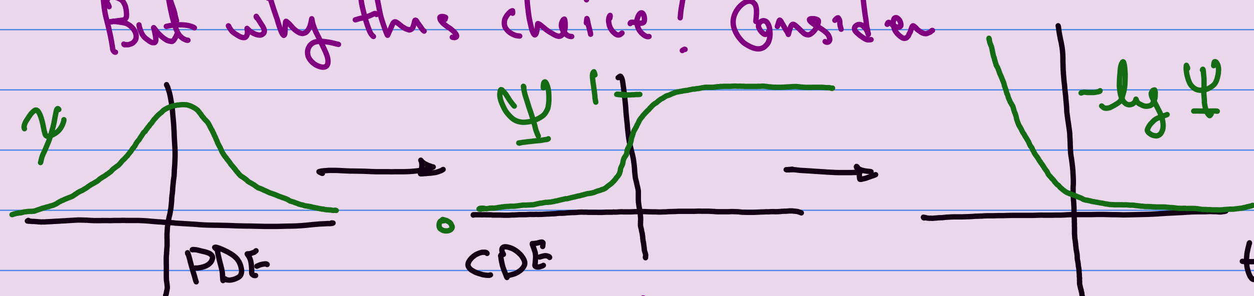
$$\mathbb{P}r(\underline{y} \mid f) = \prod_{j=1}^{n} \Psi(y_j f(\underline{x}_j))$$

This is simply the likelihood of $\underline{y}$ given f. We therefore use the negative log of this function as our misfit term. ie.

$$L(f) = -\log \mathbb{P}r(\underline{y} \mid f) = -\sum_{j=1}^{n} \log \Psi(y_j f(\underline{x}_j))$$

This loss is often referred to as the Probit loss or model.

But why this choice? Consider



This loss is small if $y_j f(\underline{x}_j) \geq 0$, ie, y_j has the same sign as $f(\underline{x}_j)$. Conversely it blows up if the signs disagree!

To this end we can formulate an opt. prob. to find $f : \mathcal{X} \to \mathbb{R}$

$$f^* = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \sum_{j=1}^{n} -\log \Psi(y_j, f(\underline{x}_j)) + \lambda \|f\|_K^2$$

so an application of the rep. thm. yields

$$f^* = K(\cdot, X)\left(K(X,X) + \sigma^2 I\right)^{-1} \underline{z}^*$$

nugget if needed

$$\underline{z}^* = \underset{\underline{z} \in \mathbb{R}^n}{\operatorname{argmin}} \sum_{j=1}^{n} -\log \Psi(y_j, z_j) + \lambda \underline{z}^T \left(K(X,X) + \sigma^2 I\right)^{-1} \underline{z}$$

Alternatively, using random features, ie,

$$K(\underline{x}, \underline{x}') = \sum_{j=1}^{N} \left(\frac{1}{\sqrt{N}} \varphi_j(\underline{x})\right)\left(\frac{1}{\sqrt{N}} \varphi_j(\underline{x}')\right)$$

$$f^* = \sum_{j=1}^{N} w_j^* \left(\frac{1}{\sqrt{N}} \varphi_j\right)$$

$$\underline{w}^* = \underset{\underline{w} \in \mathbb{R}^N}{\operatorname{argmin}} \sum_{j=1}^{n} -\log \Psi\left(y_j, f(\underline{x}_j)\right) + \lambda \|\underline{w}\|_2^2$$

$$\text{s.t. } f(\underline{x}) = \sum_{j=1}^{N} w_j \left(\frac{1}{\sqrt{N}} \varphi_j\right)$$

If ψ is taken to be log concave, then Ψ is also log-concave (eg Gaussian, Logistic, Laplace, etc.) Then both problems become convex! & have unique solutions! They can be solved efficiently!