Lecture VI-2: SVM with Random Fourier Features

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

November, 2021

Reading: Ali Rahimi and Ben Recht "Random features for large-scale Kernel Machine", NIPS 2007. Test of Time Award, NIPS 2017.

Problem: Kernel machines scale with sample size n

- Gram matrix $G = [k(x^i, x^j)]_{i,j=1}^n \in \mathbb{R}^{n \times n}$. Expensive/intractable for n large!
- Want to: benefit from infinite dimensional feature spaces, e.g. Gaussian kernel, AND have constant dimension D for any n
- **Idea** approximate $k(x, x')$ with finite sum.
- Equivalently, approximate feature space $\mathcal{H}$ with D-dimensional feature space. How? Pick D features at random!

Why is this possible? Bochner's Theorem

Let $K(x, x') = K(x - x')$ be a continuous shift invariant kernel.

Theorem [Bochner]

$K(x, x')$ is a positive definite kernel iff $K(\Delta)$ is the Fourier transform of some non-negative measure $p(\omega)$.

$$K(\Delta) \;=\; \int_{\mathbb{R}^d} p(\omega) e^{-i\omega^T \Delta} d\omega \tag{1}$$

$K(\Delta)$	$p(\omega)$	
$e^{-\|\|\Delta\|\|^2/2}$	$(2\pi)^{-d/2} e^{-\|\|\omega\|\|^2/2}$	Gaussian (RBF) kernel
$e^{-\|\|\Delta\|\|_1}$	$(2\pi)^{-d} \prod_{j=1}^d \frac{1}{1+\omega_j^2}$	Laplace kernel
$\prod_{j=1}^d \frac{2\pi}{1+\omega_j^2}$	$e^{-\|\|\Delta\|\|_1}$	product kernel

From Bochner to RFF

- Note that $e^{-i\omega\Delta} = e^{-i\omega^T x}(e^{-i\omega^T x'})^*$ and let $\zeta_\omega(x) = e^{-i\omega^T x}$.
- Then $K(\Delta) = E_{p(\omega)}[\zeta_\omega(x)\zeta_\omega^*(x')] \approx \frac{1}{D}\sum_{j=1}^{D} \zeta_{\omega_j}(x)\zeta_{\omega_j}^*(x')$ with $\omega_{1:D} \sim$ i.i.d. $p(\omega)$
- D is the sample size, must be large enough for good approximation
- $\zeta_{\omega_{1:D}}$ form a **random feature space** of dimension D
- Feature map is $x \to \tilde{\phi}(x) = \frac{1}{\sqrt{D}}[\zeta_{\omega_1} \ldots \zeta_{\omega_D}]$

Fact Because $K()$ is real, the random complex features $\zeta_\omega \leftarrow \sqrt{2}cos(\omega^T x + b)$ with $b \sim uniform[0, 2\pi]$

- **Significance** Infinite dimensional feature vector $\phi(x)$ approximated by D-dimensional feature vector $\tilde{\phi}(x)$. Hence, primal problem of dimension D can be solved instead of dual of dimension n.
- Opens up SVM/kernel machines for large data

Approximation

Theorem [Rahimi and Recht 07]

Assume space $\mathcal{X}$ is compact of diameter $d_{\mathcal{X}}$ and let $\sigma_p^2 = E_p[\omega^T \omega]$ be the standard deviation of $p(\omega)$. Then,

1. $$Pr\left[\sup_{x,x' \in \mathcal{X}} |\tilde{\phi}(x)^T \tilde{\phi}(x') - K(x,x')| \geq \epsilon\right] \leq e^{-\frac{D\epsilon^2}{4(d+2)}} \left(\frac{2^4 \sigma_p d_{\mathcal{X}}}{\epsilon}\right)^2 \tag{2}$$

2. For δ confidence level,
$$D = \Omega\left(\frac{d}{\epsilon^2} \ln \frac{\sigma_p d_{\mathcal{X}}}{\epsilon}\right) \tag{3}$$

Kernel machine with RFF algorithm

In Data $x^{1:n}, y^{1:n}$, kernel K

1. Fourier transform $p(\omega) = \frac{1}{2\pi} \int_{\mathbb{R}^d} e^{-i\omega^T \Delta} K(\Delta) d\Delta$.
2. Choose D.
3. Sample $w_{1:D}$ i.i.d. from p. Sample $b_{1:D}$ uniformly from $[0, 2\pi]$.
4. Map data to features $\tilde{\phi}(x^i) = \sqrt{\frac{2}{D}}[cos(\omega_j^T x^i + b_j)]_{j=1:D}$ for all $i = 1 : n$.
5. Solve SVM Primal problem; obtain $w \in \mathbb{R}^D$ and intercept $b \in \mathbb{R}$. (note that b is not one of $b_{1:D}$).