Random Features for Large-Scale Kernel Machines[1]

Kailong Wang[1]

[1]Rutgers University

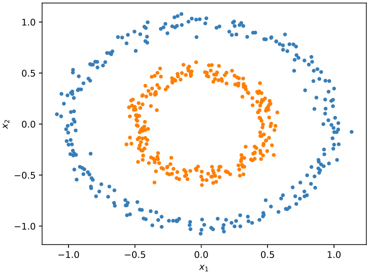
September 20, 2023

[1]Ali Rahimi and Benjamin Recht. "Random Features for Large-Scale Kernel Machines". In: 20 (2007). Ed. by J. Platt et al. URL: https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf.

Table of Contents

Linear Non-separable Problem

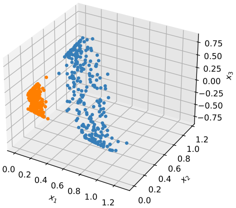
Consider a binary classification problem with non-linear samples.



e.g. For the above dataset $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$ where $\mathbf{x}_i \in \mathbb{R}^2$, a linear decision boundary does not exist.

Lifting

One idea is **LIFTING** samples into a high dimensional space in which the samples are linearly separable.



In this case, the function $\phi(\mathbf{x}_i) = \left[x_{i,1}^2, x_{i,2}^2, \sqrt{2}x_{i,1}x_{i,2}\right]$, lifts the samples into $\mathbb{R}^3$ and the samples are linearly separable.

SVM[2]

The idea of lifting has been implemented in many classification algorithms such as *support vector machine* (SVM).

Dual Problem of SVM

$$\max_{\alpha} \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \langle \mathbf{x}_n, \mathbf{x}_m \rangle$$

Dual Problem with Lifting

$$\max_{\alpha} \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \langle \phi(\mathbf{x}_n), \phi(\mathbf{x}_m) \rangle$$

This is the hard-margin SVM. The soft-margin SVM is similar.

[2] Stephen Boyd and Lieven Vandenberghe. "Convex optimization". In: (2004).

Curse of Dimensionality–Type I

$$\begin{aligned}\langle \phi(\mathbf{x}_n), \phi(\mathbf{x}_m)\rangle &= \left[x_{n,1}^2, x_{n,2}^2, \sqrt{2}x_{n,1}x_{n,2}\right]^\mathsf{T}\left[x_{m,1}^2, x_{m,2}^2, \sqrt{2}x_{m,1}x_{m,2}\right] \\ &= x_{n,1}^2x_{m,1}^2 + x_{n,2}^2x_{m,2}^2 + 2x_{n,1}x_{n,2}x_{m,1}x_{m,2}\end{aligned}$$

As shown in the given example, the inner product (a constant value) of the lifted vector has computational complexity depends on lifted dimension. For a function lifts the original vector space to a much higher dimension, such a calculation can be computationally thirsty. Alternatively, this can be done as follows, whose computational complexity only depends on the dimension of the original vector space.

$$\begin{aligned}(\langle \mathbf{x}_n, \mathbf{x}_m\rangle)^2 &= \left([x_{n,1}, x_{n,2}]^\mathsf{T}[x_{m,1}, x_{m,2}]\right)^2 \\ &= (x_{n,1}x_{m,1} + x_{n,2}x_{m,2})^2 \\ &= x_{n,1}^2x_{m,1}^2 + x_{n,2}^2x_{m,2}^2 + 2x_{n,1}x_{n,2}x_{m,1}x_{m,2} \\ &= \langle \phi(\mathbf{x}_n), \phi(\mathbf{x}_m)\rangle\end{aligned}$$

Kernel Trick

The type of function, such as $(\langle \cdot, \cdot \rangle)^2$, that provides a computationally efficient way to compute the inner product in the high dimensional space is called a **Kernel Function**.

$$K(\cdot, \cdot) = \langle \phi(\cdot), \phi(\cdot) \rangle$$

The matrix that is formed by stacking the kernel function for all samples is called the **Kernel Matrix $\mathbf{K}$**, which is a **Gram Matrix**,

$$\mathbf{K}_{nm} \equiv K(\mathbf{x}_n, \mathbf{x}_m).$$

Some kernel functions can lift the original vector space to an infinite dimensional space. The algorithms involve kernel trick is called **Kernel Machines**.

Curse of Dimensionality–Type II

Another famous kernel machine is kernel ridge regression (KRR). With $\mathbf{y} \in R^N$, $\mathbf{X} \in \mathbb{R}^{N \times d}$, and $\phi_{d \to k}(\cdot) : \mathbb{R}^d \to \mathbb{R}^k$, the normal equation of KRR is (using matrix inversion lemma)

$$\mathbf{w} = (\phi(\mathbf{X})^\intercal \phi(\mathbf{X}) + \lambda \mathbf{I}_k)^{-1} \phi(\mathbf{X})^\intercal \mathbf{y} \tag{1}$$

$$= \phi(\mathbf{X})^\intercal (\lambda \mathbf{I}_N + \phi(\mathbf{X})\phi(\mathbf{X})^\intercal)^{-1} \mathbf{y} \tag{2}$$

$$= \phi(\mathbf{X})^\intercal (\lambda \mathbf{I}_N + \mathbf{K})^{-1} \mathbf{y}. \tag{3}$$

For any input $\mathbf{x}^*$, the prediction is (let $\alpha = (\lambda \mathbf{I}_N + \mathbf{K})^{-1}\mathbf{y}$)

$$\phi(\mathbf{x}^*)^\intercal \mathbf{w} = \sum_{i=1}^{N} K(\phi(\mathbf{x}^*), \phi(\mathbf{x}_i))\alpha_i. \tag{4}$$

Eq. (**??**) is problematic since $k \to \infty$. Common approach solves eq. (**??**) with $O(N^3)$ time and $O(N^2)$ memory. This is not scalable in modern big data era, where $N \to \infty$.

Motivation

Can we find a **Kernel Function** $K(\cdot, \cdot)$, which is equivalent to lifting $\mathbf{X}$ to $\mathbb{R}^s$ with $d < s \ll k$, while not sacrifices model performance?
In such a case, we can solve eq. (**??**) instead of eq. (**??**), where the algorithm complexity dependents on s, where $s \ll N$.

Some Preliminaries

A kernel function $K(\mathbf{x}_n, \mathbf{x}_m)$ is called **shift invariant** if it can be written as $K(\mathbf{x}_n, \mathbf{x}_m) = k(\mathbf{x}_n - \mathbf{x}_m)$ for some function $k(\cdot)$ (*e.g.* $K_{Gaussian}(\mathbf{x}_n, \mathbf{x}_m) = \exp(-\gamma\|\mathbf{x}_n - \mathbf{x}_m\|_2^2)$).

A continuous function $K(\mathbf{x}_n, \mathbf{x}_m)$ is a valid kernel function if and only if the kernel matrix $\mathbf{K}$ is **positive semi-definite**.

A continuous function $k(\cdot)$ is **positive semi-definite** if and only if it is the Fourier transform of a non-negative measure.

Random Fourier Features

A continuous **shift invariant** kernel $K(\mathbf{x}_n, \mathbf{x}_m)$, which is **positive semi-definite** (Mercer's Theorem), is the Fourier transform of a non-negative measure $p(\cdot)$.

$$\phi(\mathbf{x}_n)^{\mathsf{T}} \phi(\mathbf{x}_m) = K(\mathbf{x}_n, \mathbf{x}_m) = k(\mathbf{x}_n - \mathbf{x}_m) \tag{5}$$

$$= \int_{\mathbb{R}^d} p(\omega) \exp(i\omega^{\mathsf{T}} (\mathbf{x}_n - \mathbf{x}_m)) \, \mathrm{d}\omega \tag{6}$$

$$= \mathbb{E}_\omega \left[\xi_\omega(\mathbf{x}_n)^{\mathsf{H}} \xi_\omega(\mathbf{x}_m) \right] \tag{7}$$

Here $\xi_\omega(\mathbf{x}_i) = \exp(i\omega^{\mathsf{T}} (\mathbf{x}_i))$.

Random Fourier Features

Since both the $p(\cdot)$ and $k(\triangle)$ are real-valued, we can replace $\xi_\omega(\mathbf{x}_i) = \exp(i\omega^\intercal(\mathbf{x}_i))$ with $z_\omega(\mathbf{x}_i) = \sqrt{2}\cos(\omega^\intercal(\mathbf{x}_i) + b)$. Then eq. (**??**) becomes $\mathbb{E}_\omega[z_\omega(\mathbf{x}_n)^\intercal z_\omega(\mathbf{x}_m)]$, which means $z_\omega(\mathbf{x}_n)^\intercal z_\omega(\mathbf{x}_m)$ is an unbiased estimator of $\phi(\mathbf{x}_n)^\intercal \phi(\mathbf{x}_m)$.

Let ω been drawn from $p(\omega)$ and b been uniformly drawn from $[0, 2\pi]$, the calculation of $\phi(\mathbf{x}_n)^\intercal \phi(\mathbf{x}_m)$ becomes a sampling problem.

To further reduce the variance of the estimator, we can randomly draw s samples of ω and b, normalize each corresponding $z_\omega(\mathbf{x}_i)$ by $\sqrt{s}$, and concatenate them into one vector. Then the inner product $z(\mathbf{x}_n)^\intercal z(\mathbf{x}_m) = \frac{1}{s}\sum_{j=1}^{s} z_{\omega_j}(\mathbf{x}_n)^\intercal z_{\omega_j}(\mathbf{x}_m)$

Algorithm

Algorithm Random Fourier Features

Require: A shift invariant kernel $K(\mathbf{x}_n, \mathbf{x}_m) = k(\mathbf{x}_n - \mathbf{x}_m) = k(\triangle)$.

Ensure: A randomized feature map $z(\cdot) : \mathbb{R}^d \to \mathbb{R}^s$ so that $z(\mathbf{x}_n)^\mathsf{T} z(\mathbf{x}_m) \approx K(\mathbf{x}_n, \mathbf{x}_m)$.

Compute the Fourier transform $p(\cdot)$ of the kernel K : $p(\omega) = \frac{1}{2\pi} \int \exp(-i\omega^\mathsf{T} \triangle) k(\triangle) \, \mathrm{d}\triangle$

Draw s i.i.d. samples $\omega_1, \omega_2, \ldots, \omega_s \in \mathbb{R}^d$ from $p(\omega)$ and s i.i.d. samples $b_1, b_2, \ldots, b_s \in [0, 2\pi]$.

Let $z(\mathbf{x}_i) \equiv \sqrt{\frac{2}{s}} [\cos(\omega_1{}^\mathsf{T} \mathbf{x}_i + b_1), \cos(\omega_2{}^\mathsf{T} \mathbf{x}_i + b_2), \ldots, \cos(\omega_s{}^\mathsf{T} \mathbf{x}_i + b_s)]$

Common RFF

Kernel	$K(\triangle)$	$p(\omega)$
Gaussian	$\exp(-\gamma\|\triangle\|_2^2)$	$(2\pi)^{-\frac{s}{2}} \exp -\gamma\|\omega\|_2^2$
Laplacian	$\exp(-\|\triangle\|_1)$	$\prod_d (\pi(1+\omega_d^2))^{-1}$
Cauchy	$\prod_d 2(1+\triangle_d^2)^{-1}$	$\exp(-\|\omega\|_1)$

Convergence with Hoeffding's Inequality[3]

Let $X_1, X_2, \cdots, X_N$ be independent random variables. Assume that $X_i \in [m_i, M_i]$ for every i. Then, for any $\epsilon > 0$, we have

$$\mathbb{P}\left(\left|\sum_{i=i}^{N}(X_i - \mathbb{E}[X_i])\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{N}(M_i - m_i)^2}\right).$$

Given z_ω is bounded random variable between $[-\sqrt{2/s}, \sqrt{2/s}]$, with Hoeffding's Inequality, we have

$$\mathbb{P}(|z(\mathbf{x}_n)^\mathsf{T} z(\mathbf{x}_m) - K(\mathbf{x}_n, \mathbf{x}_m)| \geq \epsilon) \leq 2\exp\left(-\frac{s\epsilon^2}{4}\right).$$

[3] Roman Vershynin. "High-Dimensional Probability: An Introduction with Applications in Data Science". In: (2018).

Convergence

Let $\mathcal{M}$ be a compact subset of $\mathbb{R}^d$ with diameter diam$(\mathcal{M})$. Then, for the mapping z defined in Algorithm **??**, we have

$$\mathbb{P}\left(\sup_{x,y\in\mathcal{M}} |z(\mathbf{x}_n)^\mathsf{T} z(\mathbf{x}_m) - K(\mathbf{x}_n, \mathbf{x}_m)| \geq \epsilon\right) \leq 2^8 \left(\frac{\sigma_{p(\cdot)} \text{diam}(\mathcal{M})}{\epsilon}\right)^2 \exp\left(-\frac{s\epsilon^2}{4(d+2)}\right).$$

The $\sigma_{p(\cdot)}^2 = \mathbb{E}_{p(\cdot)}\left[\omega^\mathsf{T}\omega\right]$ is the second moment of the Fourier transform of the $K(\cdot,\cdot)$.

The proof of this bound uses the knowledge of ϵ-net and ϵ-covering number.

Experiment

Datasets	Fourier + LS	SVM
CPU regression 6,500 instances; 21 dims	3.6% 20 secs s = 300	5.5% 51 secs
Census regression 18,000 instances; 119 dims	5% 36 secs s = 500	8.8% 7.5 mins
Adult classification 32,000 instances; 123 dims	14.9% 9 secs s = 500	14.8% 73 mins
KDDCUP99 classification 4,900,000 instances; 127 dims	7.3% 1.5 mins s = 50	6.2% (18%) 1.4 secs (20 secs)

Table: Comparison of testing error and training time between ridge regression with random features and Support Vector Machine. For classification tasks, the percent of testing points incorrectly predicted is reported. For regression tasks, the RMS error normalized by the norm of the ground truth is reported.