Random Fourier Features for Kernel Ridge Regression

Approximation Bounds and Statistical Guarantees

Kailong Wang[1]

[1]Ph.D. of ECE
Rutgers University

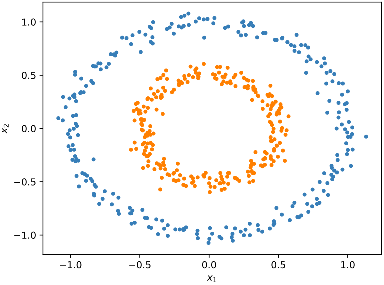
ECE 539 HDP, May 3, 2023

Table of Contents

Linear Classification with Non-linear Input

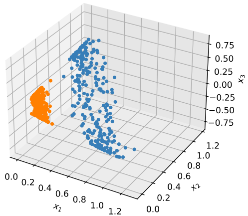
Consider a binary classification problem with non-linear (e.g. polynomial) samples. This is not separable with linear function.

(e.g. $X = \begin{bmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \\ \dots & \\ x_{N,1} & x_{N,2} \end{bmatrix} \in \mathbb{R}^{N\times 2}$.)



Lifting

One idea is to **LIFT** the samples into a higher dimensional space in which the samples are linearly separable.



The Lifting function in this case is $\phi(X) = \begin{bmatrix} x_{1,1}^2 & x_{1,2}^2 & \sqrt{2}x_{1,1}x_{1,2} \\ x_{2,1}^2 & x_{2,2}^2 & \sqrt{2}x_{2,1}x_{2,2} \\ & \cdots & \\ x_{N,1}^2 & x_{N,2}^2 & \sqrt{2}x_{N,1}x_{N,2} \end{bmatrix}$.

Curse of Dimensionality

Consider solving the above problem with *support vector machine* (SVM).

$$\mathcal{L}(\mathbf{w}, \alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n}^{N} \sum_{m}^{N} \alpha_n \alpha_m y_n y_m ({x_n}^\intercal x_m).$$

The $\mathbf{w}$ is the linear decision boundary and α is a vector of Lagrange multipliers.

Curse of Dimensionality

Consider solving the above problem with *support vector machine* (SVM).

$$\mathcal{L}(\mathbf{w}, \alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n}^{N} \sum_{m}^{N} \alpha_n \alpha_m y_n y_m ({x_n}^\intercal x_m).$$

The $\mathbf{w}$ is the linear decision boundary and α is a vector of Lagrange multipliers.

We need to use lifting function $\phi(X)$ to make the samples linearly separable. Specifically, we replace $({x_n}^\intercal x_m)$ with $(\phi(x_n)^\intercal \phi(x_m))$.

$$\begin{aligned}\phi(x_n)^\intercal \phi(x_m) &= \begin{bmatrix} x_{n,1}^2 & x_{n,2}^2 & \sqrt{2} x_{n,1} x_{n,2} \end{bmatrix} \begin{bmatrix} x_{m,1}^2 & x_{m,2}^2 & \sqrt{2} x_{m,1} x_{m,2} \end{bmatrix}^\intercal \\ &= x_{n,1}^2 x_{m,1}^2 + x_{n,2}^2 x_{m,2}^2 + 2 x_{n,1} x_{n,2} x_{m,1} x_{m,2}\end{aligned}$$

Curse of Dimensionality

Consider solving the above problem with *support vector machine* (SVM).

$$\mathcal{L}(\mathbf{w}, \alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n}^{N} \sum_{m}^{N} \alpha_n \alpha_m y_n y_m ({x_n}^\intercal x_m).$$

The $\mathbf{w}$ is the linear decision boundary and α is a vector of Lagrange multipliers.

We need to use lifting function $\phi(X)$ to make the samples linearly separable. Specifically, we replace $({x_n}^\intercal x_m)$ with $(\phi(x_n)^\intercal \phi(x_m))$.

$$\begin{aligned}\phi(x_n)^\intercal \phi(x_m) &= \begin{bmatrix} x_{n,1}^2 & x_{n,2}^2 & \sqrt{2} x_{n,1} x_{n,2} \end{bmatrix} \begin{bmatrix} x_{m,1}^2 & x_{m,2}^2 & \sqrt{2} x_{m,1} x_{m,2} \end{bmatrix}^\intercal \\ &= x_{n,1}^2 x_{m,1}^2 + x_{n,2}^2 x_{m,2}^2 + 2 x_{n,1} x_{n,2} x_{m,1} x_{m,2}\end{aligned}$$

Calculate the inner product in the $\mathbb{R}^3$ across all N pairs of samples is acceptable. However, the lifting function $\phi(X)$ is usually very high dimensional.

Kernel Trick

Consider the following derivation,

$$
\begin{aligned}
({x_n}^\intercal x_m)^2 &= ([x_{n,1} \; x_{n,2}][x_{m,1} \; x_{m,2}]^\intercal)^2 \\
&= (x_{n,1}x_{m,1} + x_{n,2}x_{m,2})^2 \\
&= x_{n,1}^2 x_{m,1}^2 + x_{n,2}^2 x_{m,2}^2 + 2x_{n,1}x_{n,2}x_{m,1}x_{m,2} \\
&= \phi(x_n)^\intercal \phi(x_m)
\end{aligned}
$$

Kernel Trick

Consider the following derivation,

$$
\begin{aligned}
(x_n^\intercal x_m)^2 &= ([x_{n,1}\ x_{n,2}][x_{m,1}\ x_{m,2}]^\intercal)^2 \\
&= (x_{n,1}x_{m,1} + x_{n,2}x_{m,2})^2 \\
&= x_{n,1}^2 x_{m,1}^2 + x_{n,2}^2 x_{m,2}^2 + 2x_{n,1}x_{n,2}x_{m,1}x_{m,2} \\
&= \phi(x_n)^\intercal \phi(x_m)
\end{aligned}
$$

Instead of computing inner product in the high dimensional space, we compute the inner product in the original space.

Kernel Trick

Consider the following derivation,

$$\begin{aligned}({x_n}^\intercal x_m)^2 &= ([x_{n,1}\ x_{n,2}][x_{m,1}\ x_{m,2}]^\intercal)^2 \\ &= (x_{n,1}x_{m,1} + x_{n,2}x_{m,2})^2 \\ &= x_{n,1}^2 x_{m,1}^2 + x_{n,2}^2 x_{m,2}^2 + 2x_{n,1}x_{n,2}x_{m,1}x_{m,2} \\ &= \phi(x_n)^\intercal \phi(x_m)\end{aligned}$$

Instead of computing inner product in the high dimensional space, we compute the inner product in the original space.

The function

$$K(x_n, x_m) = ({x_n}^\intercal x_m)^2 = \phi(x_n)^\intercal \phi(x_m)$$

is called a **kernel function**.

There must be disadvantages...

Given training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$. Consider *Kernel Ridge Regression* (KRR), with $\phi(\mathcal{X}) \subseteq \mathbb{R}^k$, where $k \to \infty$

$$\mathcal{L}(\mathbf{w}, \lambda) = \underset{\mathbf{w}}{\text{argmin}} \sum_{n}^{N} (y_n - \mathbf{w}^\intercal \phi(\mathbf{x}_n))^2 + \lambda \mathbf{w}^\intercal \mathbf{w}.$$

Solving it with Lagrange multipliers α, which is the solution of

$$(\mathbf{K} + \lambda \mathbf{I}_k)\alpha = \mathbf{y},$$

requires $\Theta(k^3)$ time and $\Theta(k^2)$ memory. Here $\mathbf{K} \in \mathbb{R}^{k \times k}$ is the kernel matrix or Gram matrix defined by $\mathbf{K}_{nm} \equiv K(\mathbf{x}_n, \mathbf{x}_m)$.

There must be disadvantages...

Given training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$. Consider *Kernel Ridge Regression* (KRR), with $\phi(\mathcal{X}) \subseteq \mathbb{R}^k$, where $k \to \infty$

$$\mathcal{L}(\mathbf{w}, \lambda) = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{n}^{N} (y_n - \mathbf{w}^\intercal \phi(\mathbf{x}_n))^2 + \lambda \mathbf{w}^\intercal \mathbf{w}.$$

Solving it with Lagrange multipliers α, which is the solution of

$$(\mathbf{K} + \lambda \mathbf{I}_k)\alpha = \mathbf{y},$$

requires $\Theta(k^3)$ time and $\Theta(k^2)$ memory. Here $\mathbf{K} \in \mathbb{R}^{k \times k}$ is the kernel matrix or Gram matrix defined by $\mathbf{K}_{nm} \equiv K(\mathbf{x}_n, \mathbf{x}_m)$.

Intuition: Can we find a kernel function which lifts $\mathcal{X}$ to $\mathbb{R}^s$, where $d < s \ll k$, while not sacrifices model performance?

Some Prerequisites

Shift Invariant Kernel (Radial Basis Function (RBF))

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

Mercer's Theorem

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**.

Bochner's Theorem

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

Random Fourier Features

Conclusion

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 $p(\cdot)$.

$$\phi(\mathbf{x_n})^\intercal \phi(\mathbf{x_m}) = K(\mathbf{x_n}, \mathbf{x_m}) = K(\mathbf{x_n} - \mathbf{x_m}) \tag{1}$$

$$= \int_{\mathbb{R}^d} p(\omega) \exp(i\omega^\intercal(\mathbf{x_n} - \mathbf{x_m}))d\omega \tag{2}$$

$$= \mathbb{E}_\omega[\xi_\omega(\mathbf{x_n})^* \xi_\omega(\mathbf{x_m})] \tag{3}$$

Here $\xi_\omega(\mathbf{x}) = \exp(i\omega^\intercal \mathbf{x}) = \begin{bmatrix} \cos(\omega^\intercal \mathbf{x}) \\ \sin(\omega^\intercal \mathbf{x}) \end{bmatrix}$ and hence $\xi_\omega(\mathbf{x_n})^* \xi_\omega(\mathbf{x_m})$ is an unbiased estimator of $K(\mathbf{x_n}, \mathbf{x_m})$ when ω is drawn from $p(\cdot)$.

Random Fourier Features

Since both the $p(\cdot)$ and $K(\triangle)$ are real-valued, we can replace $\xi(\mathbf{x})$ with $z_\omega(\mathbf{x}) = [\sqrt{2}\cos(\omega^\intercal \mathbf{x} + b)]$ where ω is drawn from $p(\omega)$ and b is uniformly drawn from $[0, 2\pi]$. Then eq. (3) becomes $\mathbb{E}_\omega[z(\mathbf{x_n})^\intercal z(\mathbf{x_m})]$

Random Fourier Features

Since both the $p(\cdot)$ and $K(\triangle)$ are real-valued, we can replace $\xi(\mathbf{x})$ with $z_\omega(\mathbf{x}) = [\sqrt{2}\cos(\omega^\intercal \mathbf{x} + b)]$ where ω is drawn from $p(\omega)$ and b is uniformly drawn from $[0, 2\pi]$. Then eq. (3) becomes $\mathbb{E}_\omega[z(\mathbf{x_n})^\intercal z(\mathbf{x_m})]$

Note: $z(\mathbf{x_n})^\intercal z(\mathbf{x_m})$ is an unbiased estimator of $\phi(\mathbf{x_n})^\intercal \phi(\mathbf{x_m})$. The $z(\mathbf{x})$ is not a lifting function.

Random Fourier Features

Since both the $p(\cdot)$ and $K(\triangle)$ are real-valued, we can replace $\xi(\mathbf{x})$ with $z_\omega(\mathbf{x}) = [\sqrt{2}\cos(\omega^\intercal \mathbf{x} + b)]$ where ω is drawn from $p(\omega)$ and b is uniformly drawn from $[0, 2\pi]$. Then eq. (3) becomes $\mathbb{E}_\omega[z(\mathbf{x_n})^\intercal z(\mathbf{x_m})]$

Note: $z(\mathbf{x_n})^\intercal z(\mathbf{x_m})$ is an unbiased estimator of $\phi(\mathbf{x_n})^\intercal \phi(\mathbf{x_m})$. The $z(\mathbf{x})$ is not a lifting function.

Note: To further reduce the variance of the estimator, we can randomly draw s samples of ω and normalize each corresponding $z(\mathbf{x})$ by $\sqrt{s}$. 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 1 Random Fourier Features

Require: A shift invariant kernel $K(\mathbf{x_n}, \mathbf{x_m}) = K(\mathbf{x_n} - \mathbf{x_m})$.

Ensure: A randomized feature map $z(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^s$ so that $z(\mathbf{x_n})^\intercal z(\mathbf{x_m}) \approx K(\mathbf{x_n}, \mathbf{x_m})$.

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

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

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

Convergence

Bound for a *fixed* pair of points $\mathbf{x_n}$ and $\mathbf{x_m}$

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

$$\mathbb{P}(|z(\mathbf{x_n})^\intercal z(\mathbf{x_m}) - K(\mathbf{x_n}, \mathbf{x_m})| \geq \epsilon) \leq 2\exp\left(-\frac{s\epsilon^2}{4}\right).$$

Convergence

Bound for *all* pair of points $\mathbf{x_n}$ and $\mathbf{x_m}$

Let $\mathcal{M}$ be a compact sunset of $\mathbb{R}^d$ with diameter $\text{diam}(\mathcal{M})$. Then, for the mapping z defined in Algorithm 1, we have

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

Two-column slide

This is a subtitle

This is a text in first column.

$$E = mc^2$$

- First item ℓ
- Second item

This text will be in the second column and on a second thought this is a nice looking layout in some cases.