Scalable ML

10605-10805

Random Fourier Features

Barnabás Póczos

Motivation

Kernel Function

$$k(x,z) \doteq \langle \lambda(x), \lambda(z) \rangle_{\mathcal{H}} \; \forall \; x, z \in \mathbb{R}^d$$

Kernel Methods

- Powerful tools in ML
- Can represent complex relations
- Requires inverting an nxn Gram matrix,
  where n is the number of instances in the training set
- Computationally expensive $O(n^3)$
- Pure scalability

Goal: Scale up Kernel Methods for large datasets

- We want $O(n)$ methods instead of $O(n^3)$

Trick: Approximate the kernel with Random Fourier Features motivated by the Bochner's theorem

Bochner's Theorem

Theorem: [Bochner]

Part 1

If ϕ is a **characteristic function** of a **probability distribution** on $\mathbb{R}$, then ϕ is a **positive semidefinite function**.

Part 2

If ϕ is a **positive semidefinite function**, **continuous at 0**, $\phi(0) = 1$, then ϕ is a **characteristic function** of a **probability distribution**.

The Kernel Function

$k(x,z) \doteq \langle \lambda(x), \lambda(z) \rangle_{\mathcal{H}} \; \forall \; x, z \in \mathbb{R}^d$

Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a

- bounded
- continuous
- translation invariant
- PSD kernel
- $k(x,y) \doteq \phi(x-y) \; \forall \; x, y \in \mathbb{R}^d$
- $k(x,x) = \phi(x-x) = \phi(0) = 1 \; \forall \; x \in \mathbb{R}^d$

Now, according to Bochner's theorem

ϕ is a characteristic function of a d-dim probability distribution $\mathbb{P}$.

$$\begin{aligned} k(x,z) = \phi(x-z) &= \int_{\mathbb{R}^d} e^{iw^T(x-z)} d\mathbb{P}(w) \\ &= \mathbb{E}_{w \sim \mathbb{P}} \left[e^{iw^T(x-z)} \right] \end{aligned}$$

From the proof of Bochner's theorem we also know that the density of $\mathbb{P}$ is the inverse Fourier transform of ϕ.

The Kernel Function

$$k(x,z) = \phi(x-z) = \int_{\mathbb{R}^d} e^{iw^T(x-z)} d\mathbb{P}(w)$$
$$= \mathbb{E}_{w\sim\mathbb{P}}\left[e^{iw^T(x-z)}\right]$$

Since $k(x,z) \in \mathbb{R}$, its imaginary part is zero.

$$\Rightarrow k(x,z) = Re\left[\int_{\mathbb{R}^d} e^{iw^T(x-z)} d\mathbb{P}(w)\right]$$

$$= Re\left[\int_{\mathbb{R}^d} \left\{\cos(w^T(x-z)) + i\sin(iw^T(x-z))\right\} d\mathbb{P}(w)\right]$$

$$= \int_{\mathbb{R}^d} \cos(w^T(x-z)) d\mathbb{P}(w)$$

$$= \mathbb{E}_{w\sim\mathbb{P}}\left[\cos(w^T(x-z))\right]$$

Kernel Approximation

We already know that

$$k(x,z) = \mathbb{E}_{w\sim\mathbb{P}}\left[\cos(w^T(x-z))\right]$$

Main idea:

Approximate this expected value with the empirical average of a random sample:

Let $w_1, \dots, w_m \sim \mathbb{P}(w)$ iid, where the density of $\mathbb{P}$ is the inverse Fourier transform of ϕ.

The kernel function can be approximated:

$$\widehat{k}(x,z) = \frac{1}{m}\sum_{i=1}^{m}\cos(w_i^T(x-z))$$

Random Fourier Features

We already know that

$$\widehat{k}(x,z) = \frac{1}{m}\sum_{i=1}^{m}\cos(w_i^T(x-z))$$

approximates $k(x,z) = \mathbb{E}_{w\sim\mathbb{P}}\left[\cos(w^T(x-z))\right]$

Random Fourier Features

Lets calculate the features $\lambda(x)$ and $\lambda(z)$ corresponding to
$\widehat{k}(x,z) = \langle\lambda(x),\lambda(z)\rangle = \lambda^T(x)\lambda(z)$

Since $\cos(a-b) = \cos a\cos b + \sin a \sin b$, we have that

$$\widehat{k}(x,z) = \frac{1}{m}\sum_{i=1}^{m}\cos(w_i^T(x-z))$$

$$\widehat{k}(x,z) = \frac{1}{m}\sum_{i=1}^{m}\left[\cos(w_i^T x)\cos(w_i^T z) + \sin(w_i^T x)\sin(w_i^T z)\right]$$

Random Fourier Features

$$\hat{k}(x,z) = \frac{1}{m}\sum_{i=1}^{m}\left[\cos(w_i^T x)\cos(w_i^T z) + \sin(w_i^T x)\sin(w_i^T z)\right]$$

$$= \langle \frac{1}{\sqrt{m}}\left[\cos(w_1^T x), \ldots, \cos(w_m^T x), \sin(w_1^T x), \ldots, \sin(w_m^T x)\right],$$

$$\frac{1}{\sqrt{m}}\left[\cos(w_1^T z), \ldots, \cos(w_m^T z), \sin(w_1^T z), \ldots, \sin(w_m^T z)\right]\rangle$$

$$= \langle \lambda(x), \lambda(z) \rangle$$

where $w_1, \ldots, w_m \sim \mathbb{P}(w)$ iid.

$\lambda(x), \lambda(z) \in \mathbb{R}^{2m}$

These features are the so-called Random Fourier Features.

Random Fourier Features

Examples $$k(x,z) = \phi(x-z) = \int_{\mathbb{R}^d} e^{iw^T(x-z)} d\mathbb{P}(w)$$
$$= \mathbb{E}_{w\sim\mathbb{P}}\left[e^{iw^T(x-z)}\right]$$

Kernel Name	Kernel	$p(w)$
Gaussian	$\phi(x-z) = \exp\left(\frac{-\|x-z\|_2^2}{2}\right)$	$p(w) = \text{Gauss}$
Laplacian	$\phi(x-z) = \exp\left(-\|x-z\|_1\right)$	$p(w) = \text{Cauchy}$
Cauchy	$\phi(x-z) = \prod_{j=1}^{d} \frac{1}{1+(x_j-z_j)^2}$	$p(w) = \text{Laplace}$

Feature map version 1:

$$\lambda(x) \doteq \frac{1}{\sqrt{m}}\left[\cos(w_1^T x), \ldots, \cos(w_m^T x), \sin(w_1^T x), \ldots, \sin(w_m^T x)\right] \in \mathbb{R}^{2m},$$

where $w_1, \ldots, w_m \sim \mathbb{P}(w)$ iid.

Feature map version 2:

$$\lambda(x) \doteq \sqrt{\frac{2}{m}}\left[\cos(w_1^T x + b_1), \ldots, \cos(w_i^T x + b_i), \ldots \cos(w_m^T x + b_m)\right] \in \mathbb{R}^m$$

where $w_1, \ldots, w_m \sim \mathbb{P}(w)$ iid, and $b_1, \ldots, b_m \sim U[0, 2\pi]$ iid

The Primal Hard SVM with Random Features

- Given $D = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ training data set.
- Assume that D is **linearly separable**.

$$\widehat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^{2m}} \frac{1}{2} \|\mathbf{w}\|^2$$

$$\text{subject to } y_i \langle \lambda(\mathbf{x}_i), \mathbf{w} \rangle \geq 1, \ \forall i = 1, \ldots, n$$

Prediction: $f_{\widehat{\mathbf{w}}}(\mathbf{x}) = \text{sign}(\langle \widehat{\mathbf{w}}, \lambda(\mathbf{x}) \rangle)$

Here $\lambda(x)$ is a random Fourier feature map:

$$\lambda(x) \doteq \frac{1}{\sqrt{m}} \left[\cos(w_1^T x), \ldots, \cos(w_m^T x), \sin(w_1^T x), \ldots, \sin(w_m^T x) \right] \in \mathbb{R}^{2m}$$

The Primal Soft SVM problem

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^{2m}, \boldsymbol{\xi} \in \mathbb{R}^{n}} C \sum_{i=1}^{n} \xi_i + \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{subject to } y_i \langle \lambda(\mathbf{x}_i), \mathbf{w} \rangle \geq 1 - \xi_i, \ \forall i = 1, \ldots, n$$

$$\xi_i \geq 0, \ \forall i = 1, \ldots, n$$

Prediction: $f_{\hat{\mathbf{w}}}(\mathbf{x}) = \text{sign}(\langle \hat{\mathbf{w}}, \lambda(\mathbf{x}) \rangle)$

Here $\lambda(x)$ is a random Fourier feature map:

$$\lambda(x) \doteq \frac{1}{\sqrt{m}} \left[\cos(w_1^T x), \ldots, \cos(w_m^T x), \sin(w_1^T x), \ldots, \sin(w_m^T x) \right] \in \mathbb{R}^{2m}$$

Thanks for your Attention! ☺