（翻译）大规模核机器的随机特性

原文:

Ali Rahimi,Benjamin Recht.Random Features for farge-ScaleKernel

Machines.Intel Research Seattle,Caltech IST,2019

摘要

为了加速核机器的训练，我们提出将输入数据映射到随机低维特征空间，然后应用现有的快速线性方法。这些特征的设计使得转换后的数据的内积近似等于用户指定的移位不变核的特征空间中的内积。我们探索了两组随机特征，给出了它们逼近各种径向基核的能力的收敛界，并证明在大规模分类和回归任务中，应用于这些特征的线性机器学习算法优于目前最先进的大规模核机器。

介绍

支持向量机的核机器具有很强的吸引力，因为它们可以用足够的训练数据任意逼近任意函数或决策边界。不幸的是，对数据的核矩阵（Gram 矩阵）进行操作的方法与训练数据集的大小相比伸缩性很差。例如，即使使用最强大的工作站，在一个有 50 万个训练示例的数据集上训练非线性支持向量机也可能需要几天的时间。另一方面，当数据的维数很小时，线性机器可以很快地在大数据集上进行训练[1，2，3]。利用这些线性训练算法训练非线性机器的一种方法是近似地对核矩阵进行因子化，并将因子矩阵的列视为线性机器中的特征（参见示例[4]）。相反，我们建议将核函数本身考虑在内。这种分解不依赖于数据，通过将数据映射到一个相对低维的随机特征空间，可以将核机器的训练和评估转化为线性机器的相应操作。我们的实验表明，这些随机特征与非常简单的线性学习技术相结合，在速度和精度上与最先进的基于核的分类和回归算法（包括那些影响核矩阵的算法）形成了良好的竞争。

内核技巧是一种简单的方法，可以为算法生成仅依赖于输入点对之间内积的特征。它依赖于任何正定函数，$k(x,y)\ x,y \in \mathcal{R}^d$定义了内积和一个上升函数$\phi$，因此提升数据点之间的内积可以快速计算为$\langle \phi(x),\phi(y) \rangle = k(x,y)$。这种方便的代价是，算法只通过对 $k(x,y)$的求值，或通过对所有数据点对应用 k 的核矩阵来访问数据。因此，大型训练集会产生大量计算和存储成本。

我们不依赖核技巧提供的隐式提升，而是使用随机化特征映射 z 将数据映射到低维欧氏内积空间：$\mathcal{R}^d \to \mathcal{R}^D$，使一对变换点之间的内积逼近来估计它们的核的值：

$$k(x,y) = \langle \phi(x),\phi(y) \rangle \approx z(x,y) \tag{1}$$

与核的提升函数ϕ不同，z 是低维的。因此，我们可以简单地用 z 来对输入进行变换，然后应用快速线性学习方法逼近相应的非线性核机器的解。在下面的内容中，我们将展示如何构造一致的特征空间，使得一致逼近的平移不变核$k(x-y)$仅用在$D=O(d\epsilon^{-2}\log\frac{1}{\epsilon^2})$的维度内将一致逼近的平移不变核$k(x-y)$近似，经验表明，优秀的回归和分类性能可以在甚至更小的 D 维上实现。

除了给我们提供极快的学习算法，这些随机特征映射还提供了一种快速评估机器的方法。使用内核技巧，评估机器在测试点x需要计算$f(x)=\sum_{i=1}^{N}c_i\,k(x_i,x)$，这需要在$O(Nd)$内进行操作计算并要求保留大部分数据集，除非机器非常稀疏。这通常是不接受大量的数据集的。另一方面，在学习了超平面 w 后，线性机器可以通过简单地计算$f(x)=w'z(x)$来评估，在这里呈现的随机特征图中，只需要$\mathrm{O(D+d)}$步操作和存储量。

我们展示了两个随机特征映射来逼近移位不变核。我们的第一个随机图在第 3 节中提出，由随机抽取的正弦函数是从核函数的傅立叶变换中寻找的。因为这个变换是平滑的，它非常适合进行插值计算。我们的第二个随机图在第 4 节中，使用随机移动的网格以随机选择的分辨率划分输入空间。这个映射不平滑，但利用了输入点之间的接近性，非常适合依赖于数据点之间定义在$L1$距离上对核进行估计。我们在第 5 节中的实验表明，将这些随机映射与简单的线性学习算法相结合，可以在各种回归和分类场景中与最新的训练算法竞争。

相关工作

大规模核机器最常用的方法是求解支持向量机（SVM）的分解方法。这些方法使用坐标上升迭代更新核机器系数的子集，直到 KKT 条件满足在差误差范围内为止[5，6]。虽然这些方法是多用途的工作，但它们并不总是针对非线性问题扩展到具有数十万个数据点的数据集。为了将核机器学习扩展到这些尺度，人们提出了几种近似方案来加速涉及核矩阵的运算。

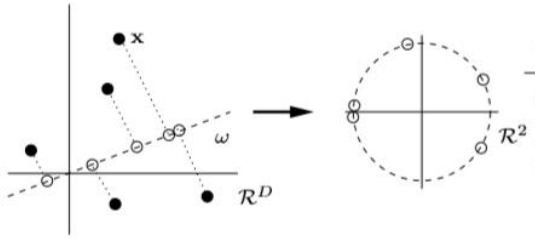
利用线性随机投影可以加快核函数的计算速度[7]。丢弃核矩阵的单个条目[7]或整行[8、9、10]，降低了操作核矩阵的存储和计算成本。这些近似要么保持数据的可分性[8]，要么产生真核矩阵的良好低秩或稀疏近似[7，9]。为此，人们也提出了快速多极和多重网格方法，但是，尽管它们在处理小维和低维问题时似乎是有效的，但在大型数据集上却没有得到证明。此外，这些方法所依赖的 Hermite 或 Taylor 近似的质量随着数据集的维数呈指数级下降[11]。利用 KD 树进行快速最近邻查找，可以近似地估算与核矩阵的乘法，进而进行其他各种运算[12]。我们在第 4 节中给出的特征图让人想起 KD 树，因为它使用多分辨率轴对齐网格划分输入空间，类似于[13]中为嵌入线性分配问题而开发的网格。

随机傅里叶特征

我们的第一组随机特征将数据点投射到随机选择的直线上，然后将得到的标量通过正弦曲线（参见图 1 和算法 1）。为了保证两个变换点的内积逼近期望的平移不变核，从分布图中画出随机线。

以下谐波分析的经典定理提供了这种转换背后的关键见解：

定理 1（Bochner[15]）当且仅当 $k(\delta)$是非负测度的 Fourier 变换时，$\mathcal{R}^d$上的连续核 $k(x,y) = k(x-y)$是正定的。



Kernel Name	$k(\Delta)$	$p(\omega)$
Gaussian	$e^{-\frac{\|\Delta\|_2^2}{2}}$	$(2\pi)^{-\frac{D}{2}} e^{-\frac{\|\omega\|_2^2}{2}}$
Laplacian	$e^{-\|\Delta\|_1}$	$\prod_d \frac{1}{\pi(1+\omega_d^2)}$
Cauchy	$\prod_d \frac{2}{1+\Delta_d^2}$	$e^{-\|\Delta\|_1}$

图 1：随机傅里叶特征。特征映射 $z(x)$的每个分量将 x 投影到从 $k(\Delta)$的 Fourier 变换 $p(\omega)$中提取的随机方向ω上，并将这条线包裹到$\mathcal{R}^2$上的单位圆上。用这种方法变换两个点 x 和 y 后，它们的内积是 $k(x,y)$的无偏估计。该表列出了一些常用的移位不变核及其傅里叶变换。要处理非各向同性核，在应用这些内核之前，数据可能会清空。

如果核 $k(\delta)$具有适当的标度，Bochner 定理保证其 Fourier 变换 $p(\omega)$是一个适当的概率分布。定义$\zeta_\omega(x) = e^{j\omega' x}$我们有

$$k(x-y) = \int_{\mathcal{R}^d} p(\omega)\, e^{j\omega'(x-y)}\, d\omega = E_\omega[\zeta_\omega(x)\zeta_\omega(y)^*] \qquad (2)$$

因此，当从 p 中取ω时，$\zeta_\omega(x)\zeta_\omega(y)^*$是 $k(x,y)$的无偏估计。

要获得 k 的实值随机特征，注意概率分布 $p(\omega)$和核 $k(\Delta)$都是实的，因此被积函数$e^{j\omega'(x-y)}$可以替换为 $\cos\omega'(x-y)$。定义$z_\omega(x) = [\cos(x)\sin(x)]'$给出满足条件 $\mathrm{E}\left[z_\omega(x)' z_\omega(y)\right] = k(x,y)$的实值映射，因为$z_\omega(x)' z_\omega(y) = \cos\omega'(x-y)$。其他映射如$z_\omega(x) = \sqrt{2}cos(\omega' x + b)$，其中$\omega$从 $p(\omega)$中取出，b 从[0，2π]中均匀地取出，同时满足条件 $\mathrm{E}\left[z_\omega(x)' z_\omega(y)\right] = k(x,y)$。

我们可以通过将随机选择的 D 个z_ω串接成为一个向量 z，并对 z 的每一个分量通过$\sqrt{D}$进行标准化来降低$z_\omega(x)' z_\omega(y)$的方差。由

由 2D 维随机特征 z 表征的点的内积，$\mathrm{z(x)'z(y)} = \frac{1}{D} z_\omega(x)' z_\omega(y)$的样本平均值，因此是对期望值（2）的较低方差近似。

由于$z_\omega(x)'z_\omega(y)$在-1 和 1 之间有界，对于一对固定的点 x 和 y，Hoeffding 不等式保证了 z(x)'z(y)和 $k(x,y)$之间 D 维的指数快速收敛：$Pr\left[\left|z(x)'z(y) - k(x,y)\right| \geq \epsilon\right] \leq 2exp\,(-D\epsilon^2/2)$。基于这一观察结果，可以同时为输入空间中的每一对点证明一个更有力的断言：

结论 1：（Fourier 特征的一致收敛性） 让 M 定义为$\mathcal{R}^4$上的紧致子集，定义其直径为 $diam(M)$。然后，对于算法 1 中定义的映射，我们有

$$Pr\left[\sup_{x,y\in M}\left|z(x)'z(y) - k(x,y)\right| \geq \epsilon\right] \leq 2^8\left(\frac{\sigma_p diam(M)}{\epsilon}\right)^2 \exp\left(-\frac{D\epsilon^2}{d(d+2)}\right)$$

这里的$\sigma_p \equiv E_p\left[\omega'\omega\right]$ 是 k 的 Fourier 变换的第二阶矩。另外，$\sup_{x,y\in M}\left|z(x)'z(y) - k(x,y)\right| \leq \epsilon$是一个恒定的概率当 $\mathrm{D} = \Omega\left(\frac{d}{\epsilon^2}\log\frac{\sigma_p diam(M)}{\epsilon}\right)$。

这个断言的证明首先保证 $z(x)'z(y)$很接近 $k(x,y)$在 M*M 的ϵ-网的中心。然后，使用以下事实将此结果扩展到整个空间，也就是地图平滑的概率很高高。详见附件。

通过标准的 Fourier 恒等式，标量σ_p^2等于 k 的 Hessian 矩阵在 0 处的迹。它量化了核在原点的曲率。对于球面高斯核，$k(x,y) = exp\,(-\gamma\| x - y \|^2)$，在这里我们有$\sigma_p^2 = 2d\gamma$。

Algorithm 1 Random Fourier Features.

Require: A positive definite shift-invariant kernel $k(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y})$.
Ensure: A randomized feature map $\mathbf{z}(\mathbf{x}) : \mathcal{R}^d \to \mathcal{R}^{2D}$ so that $\mathbf{z}(\mathbf{x})'\mathbf{z}(\mathbf{y}) \approx k(\mathbf{x} - \mathbf{y})$.
Compute the Fourier transform p of the kernel k: $p(\omega) = \frac{1}{2\pi}\int e^{-j\omega'\Delta}k(\Delta)\, d\Delta$.
Draw D iid samples $\omega_1, \cdots, \omega_D \in \mathcal{R}^d$ from p.
Let $\mathbf{z}(\mathbf{x}) \equiv \sqrt{\frac{1}{D}}\left[\cos(\omega_1'\mathbf{x}) \cdots \cos(\omega_D'\mathbf{x}) \sin(\omega_1'\mathbf{x}) \cdots \sin(\omega_D'\mathbf{x})\right]'$.

结论

我们提出了随机特征，其内积一致地逼近许多流行的核函数。我们的经验表明，将这些特征作为标准线性学习算法的输入，可以产生在精度、训练时间和评估时间方面与最先进的大型核机器相媲美的结果。

值得注意的是，傅立叶特征和二值化特征的混合可以通过结合这些特征来构造。当我们专注于回归和分类时，我们的特性可以应用于加速其他核心方法，包括半监督和非监督学习算法。在所有这些情况下，首先计算随机特征，然后应用相关的线性技术，可以显著加快计算速度。