

# Random Features for Large-Scale Kernel Machines

## 用于大型核机器的随机特征

Ali Rahimi

Intel Research Seattle

Seattle, WA 98105

ali.rahimi@intel.com

Benjamin Recht

Caltech IST

Pasadena, CA 91125

brecht@ist.caltech.edu

### 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化的低维特征空间，并随后应用已有的快速线性方法。我们通过构造随机化特征以使得转换后的数据的内积近似于特征空间中自定义的平移不变核。我们探索了两类随机特征，提供了他们估计不同径向基核的收敛界，并且证明了在大规模分类和回归任务中，使用这些特征的线性机器学习算法更优于目前的新大型核机器。

## 1 引言

核机器，诸如支持向量机等，因其能在充足的训练数据下以任意精度逼近任何函数或决策边界而引人注目。不幸的是，对数据的核矩阵（格拉姆矩阵）进行操作的方法，根据训练数据集的大小表现得不尽如人意。例如，对于一个有 50 万训练样本的数据集来说，即使是在算力最强的工作站上，仍可能需要花费数天来训练非线性 SVM。另一方面，当数据集很大而维数较小时，线性机器能够很快被训练出来 [1, 2, 3]。一种能够利用线性训练算法来训练非线性机器的方法是近似分解核矩阵，并将因子矩阵的列作为线性机器中的特征（如 [4] 中所述）。不同的是，我们提出直接分解核函数本身。这样的分解不依赖于数据，并且通过将数据映射到相对低维的随机特征空间中使得对核机器的训练和评估转换为对线性机器的相应操作。我们的实验表明，随机特征与非常简单的线性学习技术相结合，可以有力地与基于核方法的最新分类和回归算法在速度和精确性上竞争，包括那些分解核矩阵的方法。

对于只依赖于输入点对之间内积的算法，核技巧是一种能够为其生成特征的简单方法。它依赖于以下结论：任何正定函数  $k(\mathbf{x}, \mathbf{y})$ ,  $\mathbf{x}, \mathbf{y} \in \mathcal{R}^d$ ，定义了一个内积和一个映射  $\phi$ ，以使得映射后的数据点能被快速计算为  $\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle = k(\mathbf{x}, \mathbf{y})$ 。这种便利的代价是算法仅通过评估  $k(\mathbf{x}, \mathbf{y})$  或通过将  $k$  作用于所有数据而组成的核矩阵来处理数据。最终，若训练集很大会导致计算和存储成本增大。

相较于依赖核技巧提供的隐式映射，我们提出用随机化的特征映射  $z: \mathcal{R}^d \rightarrow \mathcal{R}^D$  来显式地将数据映射到一个低维欧几里得内积空间，以使得变换后的点对之间的内积近似于它们的核估计：

$$k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle \approx z(\mathbf{x})'z(\mathbf{y}). \quad (1)$$

不像核中的映射  $\phi$ ,  $z$  相对而言是低维度的。因此, 我们可以简单地用  $z$  来转换输入, 然后应用快速的线性学习方法来近似相应的非线性核机器的结果。在下文中, 我们展示了如何构造能够一致逼近熟知的平移不变核  $k(\mathbf{x} - \mathbf{y})$  且在给定  $\epsilon$  下仅有  $D = O(d\epsilon^{-2} \log \frac{1}{\epsilon^2})$  维的特征空间, 并经验上证明了甚至在更小的  $D$  下也能有良好的回归和分类结果。

除了使我们能够使用极快速的学习算法之外, 这些随机化的特征映射也提供了一个快速评估机器的方法。使用核技巧时, 在测试点  $\mathbf{x}$  处评估机器需要计算  $f(\mathbf{x}) = \sum_{i=1}^N c_i k(\mathbf{x}_i, \mathbf{x})$ , 除非机器是非常稀疏的, 这要求  $O(Nd)$  的计算复杂度以及存储数据集中大部分的点。通常来说, 这对于大的数据集而言是不可接受的。另一方面, 在学习了一个超平面  $\mathbf{w}$  之后, 一个线性机器可以通过简单地计算  $f(\mathbf{x}) = \mathbf{w}'z(\mathbf{x})$  来评估, 基于这里的随机化特征映射, 这仅仅要求  $O(D + d)$  的复杂度。

我们展示了两种近似的平移不变核的随机化特征映射。我们的第一种随机化映射, 呈现在第三节, 由从我们想近似的核函数的傅立叶变换中随即采样的正弦函数组成。因为这个映射是光滑的, 它非常适合于插值问题。我们的第二种随机化映射, 呈现在第四节, 用随机选择的分辨率下的随机平移网格来划分输入空间。这个映射不是光滑的, 但是衡量了输入点之间的接近度, 非常适合用来估计依赖于点对间  $L_1$  距离的核。我们在第五节中的实验证明了结合这些随机映射和线性学习算法能够有力地与最新的各种回归和分类情形的训练算法相竞争。

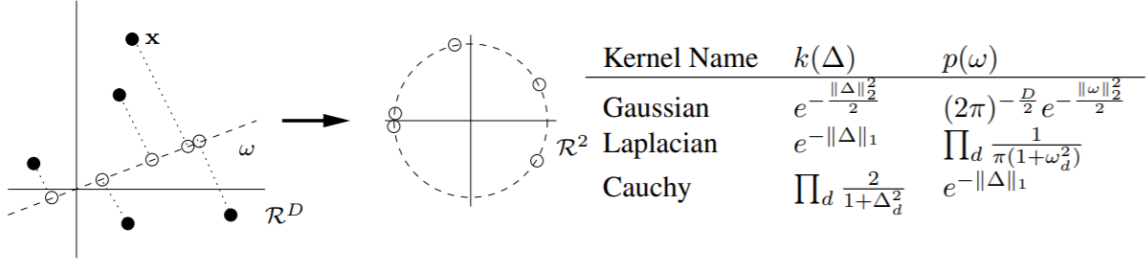
## 2 相关工作

大型核机器最流行的方法是用于求解支持向量机 (SVM) 的分解方法。这些方法使用坐标上升迭代地更新核机器的系数的子集, 直到满足 KKT 条件以使其在误差范围内 [5, 6]。尽管此类方法是通用的主力, 但它们并不总是能够扩展到具有超过数十万个数据点的非线性问题的数据集。为了将利用核机器的学习扩展到这些规模, 几种近似方案已经被提出以加速涉及核矩阵的运算。

核函数的评估可以使用线性随机投影来加速 [7]。丢弃核矩阵的各个元素 [7] 或整个行 [8, 9, 10] 会降低对核矩阵进行操作的存储和计算成本。这些近似值可以保留数据的可分离性 [8], 也可以生成实际核矩阵的良好低秩或稀疏近似值 [7, 9]。快速的多极和多网格方法也因此而被提出, 但是, 尽管它们似乎在解决小的和低维问题上是有用的, 但尚未在大型数据集上得到证明。此外, 这些方法所依赖的 Hermite 或 Taylor 近似的效果随着数据集的维数呈指数下降 [11]。使用 KD 树进行快速最近邻居查找已被用于近似核矩阵的乘法运算, 进而进行多种其他运算 [12]。我们在第 4 节中介绍的特征映射联系到 KD 树, 因为它使用多解析度的轴对齐网格来划分输入空间, 类似于 [13] 中的嵌入线性分配问题。

## 3 随机傅立叶特征

我们的第一类随机特征将数据点投影到随机选择的直线, 然后通过正弦曲线来传递得到的标量, 即投影点的值 (见图 1 和算法 1)。上述随机选择的直线从一个分布中采样以保证变换后的点之间的内积能近似所需的平移不变核。



**图 1:** 随机傅立叶特征。特征映射  $z(\mathbf{x})$  的每一个分量将  $\mathbf{x}$  投影到一个从  $k(\Delta)$  的傅立叶变换  $p(w)$  采样得到的随机方向  $w$ ，并将这条直线对应到  $\mathcal{R}^2$  中的单位圆上。在将两个点  $\mathbf{x}$  和  $\mathbf{y}$  按这种方式变换后，它们的内积是  $k(\mathbf{x}, \mathbf{y})$  的无偏估计。表格中列出了一些常用的平移不变核和他们的傅立叶变换。为了处理一些各向异性的核，在应用这些核之前应先对数据进行上述变换。

以下出自调和的经典理论是这一变换背后的关键：

**定理 3.1.** (Bochner [15])  $\mathcal{R}^d$  上的连续核  $k(x, y) = k(x - y)$  是正定的当且仅当  $k(\delta)$  是一个非负测度的傅立叶变换。

如果核  $k(\delta)$  被适当缩放，Bochner 定理保证了它的傅立叶变换  $p(w)$  是恰当的概率分布。定义  $\zeta_\omega(\mathbf{x}) = e^{j\omega' \mathbf{x}}$ ，我们有

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

因此当  $\omega$  从  $p$  采样时， $\zeta_\omega(\mathbf{x}) \zeta_\omega(\mathbf{y})^*$  是  $k(\mathbf{x}, \mathbf{y})$  的无偏估计。

为了获取  $k$  的实值随机特征，注意到概率分布  $p(w)$  和核  $k(\Delta)$  都是实值的，因此可以用  $\cos \omega'(\mathbf{x} - \mathbf{y})$  来替换被积函数  $e^{j\omega'(\mathbf{x} - \mathbf{y})}$ 。定义  $z_\omega(\mathbf{x}) = [\cos(\mathbf{x}) \sin(\mathbf{x})]'$  给出了一个满足条件  $E[z_\omega(\mathbf{x})' z_\omega(\mathbf{y})] = k(\mathbf{x}, \mathbf{y})$  的实值映射，这是因为  $z_\omega(\mathbf{x})' z_\omega(\mathbf{y}) = \cos \omega'(\mathbf{x} - \mathbf{y})$ 。其他的映射比如  $z_\omega(\mathbf{x}) = \sqrt{2} \cos(\omega' \mathbf{x} + b)$ ，其中  $\omega$  从  $p(\omega)$  采样以及  $b$  均匀采样自  $[0, 2\pi]$ ，也满足条件  $E[z_\omega(\mathbf{x})' z_\omega(\mathbf{y})] = k(\mathbf{x}, \mathbf{y})$ 。

我们可以通过连接  $D$  个随机选择的  $z_\omega$  为一个列向量  $\mathbf{z}$  并用  $\sqrt{D}$  规范化每一个分量来降低  $z_\omega(\mathbf{x})' z_\omega(\mathbf{y})$  的方差。由  $2D$  维随机特征  $\mathbf{z}$  刻画的点的内积， $z(\mathbf{x})' z(\mathbf{y}) = \frac{1}{D} \sum_{j=1}^D z_{\omega_j}(\mathbf{x})' z_{\omega_j}(\mathbf{y})$  是  $z_{\omega_j}(\mathbf{x})' z_{\omega_j}(\mathbf{y})$  的样本均值，从而是期望 (2) 的有更小方差的近似。

由于  $z_\omega(\mathbf{x})' z_\omega(\mathbf{y})$  的上下界为 1 和 -1，对于一对固定的点  $\mathbf{x}$  和  $\mathbf{y}$ ，Hoeffding 不等式保证了  $z(\mathbf{x})' z(\mathbf{y})$  指数快速收敛于  $k(\mathbf{x}, \mathbf{y})$ ： $\Pr[|z(\mathbf{x})' z(\mathbf{y}) - k(\mathbf{x}, \mathbf{y})| \geq \epsilon] \leq 2 \exp(-D\epsilon^2/2)$ 。基于这个结论，一个关于输入空间中每一对点的更强的论断可以同时被证明：

**论断 1.** (傅立叶特征的一致收敛). 令  $\mathcal{M}$  为  $\mathcal{R}^d$  的一个直径为  $\text{diam}(\mathcal{M})$  的紧子集。对于定义在算法 1 中的映射  $\mathbf{z}$ ，我们有：

$$\Pr \left[ \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{M}} |z(\mathbf{x})' z(\mathbf{y}) - k(\mathbf{x}, \mathbf{y})| \geq \epsilon \right] \leq 2^8 \left( \frac{\sigma_p \text{diam}(\mathcal{M})}{\epsilon} \right)^2 \exp \left( -\frac{D\epsilon^2}{4(d+2)} \right),$$

其中  $\sigma_p^2 \equiv E_p[\omega' \omega]$  是  $k$  的傅立叶变换的二阶矩。当  $D = \Omega \left( \frac{d}{\epsilon^2} \log \frac{\sigma_p \text{diam}(\mathcal{M})}{\epsilon} \right)$  时，对于任意确定的概率有  $\sup_{\mathbf{x}, \mathbf{y} \in \mathcal{M}} |z(\mathbf{x})' z(\mathbf{y}) - k(\mathbf{y}, \mathbf{x})| \leq \epsilon$ 。

该断言的证明首先保证了对于  $\mathcal{M} \times \mathcal{M}$  上的  $\epsilon$ -网的中心， $z(\mathbf{x})' z(\mathbf{y})$  接近于  $k(\mathbf{x} - \mathbf{y})$ 。基于在很高的概率下特征映射是光滑的，这一结果很快能被拓展到整个空间。细节见附录。

根据标准傅立叶等式，标量  $\sigma_p^2$  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， $k(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2)$ ，我们有  $\sigma_p^2 = 2d\gamma$ 。

---

**Algorithm 1** 随机傅立叶特征

---

**Require:** 正定的平移不变核  $k(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y})$ .

**Ensure:** 随机化特征映射  $\mathbf{z}(\mathbf{x}) : \mathcal{R}^d \rightarrow \mathcal{R}^{2D}$ ，满足  $\mathbf{z}(\mathbf{x})' \mathbf{z}(\mathbf{y}) \approx k(\mathbf{x} - \mathbf{y})$ .

计算核  $k$  的傅立叶变换  $p$ :  $p(\omega) = \frac{1}{2\pi} \int e^{-j\omega' \Delta} k(\Delta) d\Delta$ .

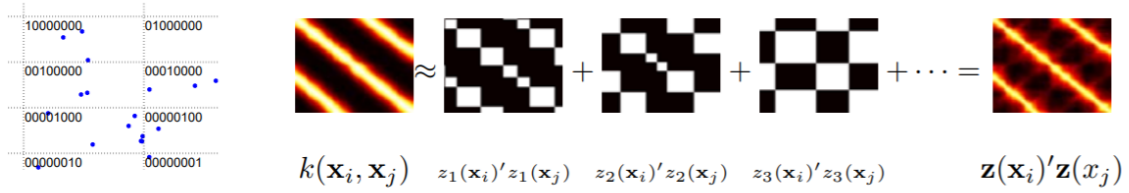
从  $p$  采样  $D$  个独立同分布的样本  $\omega_1, \dots, \omega_D \in \mathcal{R}^D$ .

令  $\mathbf{z}(\mathbf{x}) \equiv \sqrt{\frac{1}{D}} [\cos(\omega_1' \mathbf{x}) \cdots \cos(\omega_D' \mathbf{x}) \sin(\omega_1' \mathbf{x}) \cdots \sin(\omega_D' \mathbf{x})]'$ .

---

## 4 随机分箱特征

我们的第二个随机映射使用随机选择的分辨率下的随机平移网格来划分输入空间并赋予输入点一个相应于其所在箱子的二进制位串（见图2和算法2）。通过使得两个点被划分到同一个箱子的概率与  $k(\mathbf{x}, \mathbf{y})$  成比例来构造网格。变换后的点对之间的内积与它们被分箱在一起的次数成比例，因此是  $k(\mathbf{x}, \mathbf{y})$  的无偏估计。



**图 2:** 随机分箱特征。（左）该算法用随机选择的分辨率下的随机平移网格不断得划分输入空间，并赋予每一个输入点相应于其所在箱子的位串  $\mathbf{z}(\mathbf{x})$ 。（右）二进制邻接矩阵，描述了这个划分以  $\mathbf{z}(\mathbf{x}_i)' \mathbf{z}(\mathbf{x}_j)$  作为其第  $ij$  个元素，是一个对核矩阵的无偏估计。

我们首先在  $\mathcal{R} \times \mathcal{R}$  的一个紧子集上描述了用一个随机映射来近似“帽子”核  $k_{hat}(x, y; \delta) = \max(0, 1 - \frac{|x-y|}{\delta})$ ，然后展示如何为更多一般的可分多维核构造映射。以间距  $\delta$  的网格来划分实直线，并从  $[0, \delta]$  随机均匀采样出总量为  $u$  的网格。对所有整数  $n$ ，这样的网格把实直线划分成区间  $[u + n\delta, u + (n+1)\delta]$ 。两个点  $x$  和  $y$  落进这个网格中的同一个箱子的概率是  $\max(0, 1 - \frac{|x-y|}{\delta})$  [13]。换句话说，如果我们给网格中的箱子标号使得  $x$  落进箱子  $\hat{x} = \lfloor \frac{x-u}{\delta} \rfloor$  和  $y$  落进箱子  $\hat{y} = \lfloor \frac{y-u}{\delta} \rfloor$ ，则  $\Pr_u[\hat{x} = \hat{y} | \delta] = k_{hat}(x, y; \delta)$ 。如果我们把  $\hat{x}$  编译为箱子上的二进制指示向量  $\mathbf{z}(x)$ ，则当  $x$  和  $y$  落进同一个箱子时  $\mathbf{z}(x)' \mathbf{z}(y)$  的值为 1，不然为 0，如此  $\Pr_u[\mathbf{z}(x)' \mathbf{z}(y) = 1 | \delta] = E_u[\mathbf{z}(x)' \mathbf{z}(y) | \delta] = k_{hat}(x, y; \delta)$ 。从而  $\mathbf{z}$  是  $k_{hat}$  的一个随机映射。

在  $\mathcal{R} \times \mathcal{R}$  的一个紧子集上，考虑能够表示成上述帽子核的凸组合的平移不变核： $k(x, y) = \int_0^\infty k_{hat}(x, y; \delta) p(\delta) d\delta$ 。若网格间距  $\delta$  采样自  $p$ ， $\mathbf{z}$  再一次给出  $k$  的一个随机映射，这是因为  $E_{\delta, u}[\mathbf{z}(x)' \mathbf{z}(y)] = E_\delta[E_u[\mathbf{z}(x)' \mathbf{z}(y) | \delta]] = E_\delta[k_{hat}(x, y; \delta)] = k(x, y)$ 。也就是说，如果网格间距  $\delta$  采样自  $p$  以及  $u$  均匀采样自  $[0, \delta]$ ，那么  $x$  和  $y$  分箱在一起的概率是  $k(x, y)$ 。附录中引理??指出令  $p(\delta) = \delta \ddot{k}(\delta)$  可以轻易地通过  $k$  重新获取  $p$ 。例如，在拉普拉斯核情形， $k_{Laplacian}(x, y) =$

$\exp(-|x-y|)$ ,  $p(\delta)$  是伽玛分布  $\delta \exp(-\delta)$ 。对于高斯核,  $\ddot{k}$  不是处处正定, 所以相应的过程不会产生随机映射。

对于形如  $k(\mathbf{x}-\mathbf{y}) = \prod_{m=1}^d k_m(|x^m - y^m|)$  的可分多元平移不变核 (比如多元拉普拉斯核), 若每一个  $k_m$  都能被表示成帽子核的凸组合, 则对应的随机映射可以通过相似的方法来构造。我们独立地应用上面的分箱过程于  $\mathcal{R}^d$  的每一维,  $x^m$  和  $y^m$  在  $m$  维中被分箱在一起的概率是  $k_m(|x^m - y^m|)$ 。因为分箱过程是独立于维数的,  $\mathbf{x}$  和  $\mathbf{y}$  在所有维数中被分箱在一起的概率是  $\prod_{m=1}^d k_m(|x^m - y^m|) = k(\mathbf{x}-\mathbf{y})$ 。在这个多元情形,  $z(\mathbf{x})$  将与  $d$  维网格对应的每个箱子的整数向量  $[\hat{\mathbf{x}}^1, \dots, \hat{\mathbf{x}}^d]$  编译为二进制指示向量。事实上, 当  $d$  很大时, 为防止计算  $z(\mathbf{x})$  导致溢出, 我们的实现消除了式子中的空箱。因为箱子的数量总是不超过训练点的数量, 这保证了不溢出是可能的。

我们可以再一次通过连接  $P$  个随机分箱函数  $z$  为一个更大的特征列表  $\mathbf{z}$  并用  $\sqrt{\frac{1}{P}}$  来规范化它以降低估计量  $z(\mathbf{x})'z(\mathbf{y})$  的方差。内积  $\mathbf{z}(\mathbf{x})'z(\mathbf{y}) = \frac{1}{P} \sum_{p=1}^P z_p(\mathbf{x})'z_p(\mathbf{y})$  是  $P$  个独立的  $z(\mathbf{x})'z(\mathbf{y})$  的均值因此有更小的方差。

因为  $z(\mathbf{x})'z(\mathbf{y})$  是二进制的, Hoeffding 不等式保证了对固定的点对  $\mathbf{x}$  和  $\mathbf{y}$ ,  $\mathbf{z}(\mathbf{x})'z(\mathbf{y})$  作为  $P$  维函数指数快速收敛到  $k(\mathbf{x}, \mathbf{y})$ 。又一次, 一个更强的论断是这个收敛性对所有的点都成立:

**论断 2.** 令  $\mathcal{M}$  为  $\mathcal{R}^d$  的一个直径为  $\text{diam}(\mathcal{M})$  的紧子集。令  $\alpha = E[\frac{1}{\delta}]$  以及令  $L_k$  表示  $k$  相对于  $L_1$  范数的李普希兹常数。对于上面的  $\mathbf{z}$ , 我们有:

$$\Pr \left[ \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{M}} |\mathbf{z}(\mathbf{x})'z(\mathbf{y}) - k(\mathbf{x}, \mathbf{y})| \leq \epsilon \right] \geq 1 - 36dP\alpha \text{diam}(\mathcal{M}) \exp \left( \frac{-\left(\frac{P\epsilon^2}{8} + \ln \frac{\epsilon}{L_k}\right)}{d+1} \right),$$

注意到  $\alpha = \int_0^\infty \frac{1}{\delta} p(\delta) d\delta = \int_0^\infty \ddot{k}(\delta) d\delta$  为 1, 以及  $L_k = 1$  对于拉普拉斯核。该断言的证明 (见附录) 将  $\mathcal{M} \times \mathcal{M}$  划分成很小的矩形块, 在其上  $k(\mathbf{x}, \mathbf{y})$  因  $\mathbf{z}(\mathbf{x})$  和  $\mathbf{z}(\mathbf{y})$  为常数而保持不变。在很高的概率下, 这些块的中心  $\mathbf{z}(\mathbf{x})^{prime}z(\mathbf{y})$  接近于  $k(\mathbf{x}, \mathbf{y})$ , 这保证了  $k(\mathbf{x}, \mathbf{y})$  和  $\mathbf{z}(\mathbf{x})^{prime}z(\mathbf{y})$  在  $\mathcal{M} \times \mathcal{M}$  上接近。

---

#### Algorithm 2 随机分箱特征

---

**Require:** 点  $\mathbf{x} \in \mathcal{R}^d$ 。核函数  $k(\mathbf{x}, \mathbf{y}) = \prod_{m=1}^d k_m(|x^m - y^m|)$ , 因此  $p_m(\Delta) \equiv \Delta \ddot{k}_m(\Delta)$  是  $\Delta > 0$  上的概率分布。

**Ensure:** 随机化特征映射  $\mathbf{z}(\mathbf{x})$ , 满足  $\mathbf{z}(\mathbf{x})'z(\mathbf{y}) \approx k(\mathbf{x}-\mathbf{y})$ .

**for**  $p = 1 \dots P$  **do**

用间隔  $\delta^m \sim p_m$  采样网格参数  $\delta, \mathbf{u} \in \mathcal{R}^d$ , 并从  $[0, \delta^m]$  上的均匀分布中采集  $u^m$ 。

令  $z$  返回包含  $\mathbf{x}$  作为二进制指示向量  $z_p(\mathbf{x}) \equiv \text{hash}(\lceil \frac{x^1 - u^1}{\delta^1} \rceil, \dots, \lceil \frac{x^d - u^d}{\delta^d} \rceil)$  的箱子的坐标。

**end for**

$\mathbf{z}(\mathbf{x}) \equiv \sqrt{\frac{1}{P}} [z_1(\mathbf{x}) \dots z_P(\mathbf{x})]'$ .

---

## 5 实验

总结在表1中的实验表明使用我们的随机特征的岭回归能够很快地完成监督核机器的训练。我们将比较的重心放在核心向量机 [14] 上，因为在 [14] 中证明它比训练核机器的其他已知方法（包括在大多数情况下，随机采样数据点 [8]）更快、更准确。实验是在 [14] 中评估的 5 个大型数据集上进行的，不包括合成数据集。我们通过使用有关 *CVM*，*SVM<sup>light</sup>* 和 *libSVM* 的文献的相应作者提供的二进制文件，复现了其结果。<sup>1</sup>对于随机特征实验，我们通过求解岭回归问题  $\min_{\mathbf{w}} \|\mathbf{Z}'\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_2^2$  来训练回归器和分类器，其中  $\mathbf{y}$  表示期望的输出向量以及  $\mathbf{Z}$  表示随机特征矩阵。为了评估数据点  $\mathbf{x}$  上的训练好的机器，我们可以简单地计算  $\mathbf{w}^{prime}\mathbf{z}(\mathbf{x})$ 。尽管简单，但具有随机特征的岭回归比其他方法要快，并且在精确性上具有竞争优势。它也产生非常紧凑的函数，因为仅需要保留  $\mathbf{w}$  和一组  $O(D)$  随机向量或划分的哈希表。随机傅立叶特征在特别依赖插值的任务上表现更好。另一方面，随机分箱特征在记忆任务（如需要许多支持向量的标准 SVM）上表现得更好，因为它们明确保留了输入空间的局部性。在 *Forest* 数据集中，这种差异最为明显。

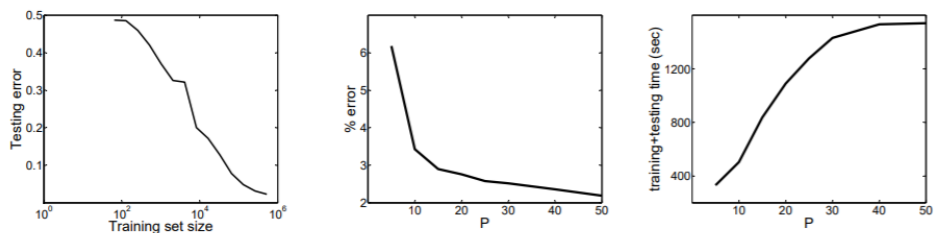
图3（左）说明了在大型数据集上训练分类器的好处，即随着越来越多的数据用于训练，精确度不断提高。图3（中）和（右）表明即使只有少量特征也能获得良好的表现。

数据集	Fourier+LS	Binning+LS	CVM	Exact SVM
CPU 回归 6500 实例 21 维	3.6% 20 secs D = 300	5.3% 3 mins P = 350	5.5% 3 mins 51 secs	11% 31 secs ASVM
Census 回归 18,000 实例 119 维	5% 36 secs D = 500	7.5% 19 mins P = 30	8.8% 7.5 mins	9% 13 mins SVM Torch
Adult 分类 32,000 实例 123 维	14.9% 9 secs D = 500	15.3% 1.5 mins P = 30	14.8% 73 mins	15.1% 7 mins SVMlight
Forest Cover 分类 522,000 实例 54 维	11.6% 71 mins D = 5000	2.2% 25 mins P = 50	2.3% 7.5 hrs	2.2% 44 hrs libSVM
KDDCUP99 (见脚注) 分类 4,900,000 实例 127 维	7.3% 1.5 min D = 50	7.3% 35 mins P = 10	6.2% (18%) 1.4 secs (20 secs)	8.3% <1 s SVM+sampling

**表 1:** 比较具有随机特征的岭回归，核心支持向量机和文献中报道的各种最新方法之间的测试误差和训练时间。对于分类任务，报告了预测错误的测试点的百分比；对于回归任务，报告了根据实际情况规范化的均方根误差。

<sup>1</sup>我们为了完整起见将 *KDDCUP99* 结果包括在内，但是请注意，此数据集本质上是过采样的：对 50 个训练示例（训练数据集的 0.001%）进行随机抽样训练 SVM（或具有随机特征的最小二乘法）足以持续生成测试误差约为 8%。同样，虽然我们能够使用作者提供的参数来复制 CVM 的 6.2% 的错误率，但是在随机改组训练集后进行再训练会导致 18% 的错误，并将计算时间增加一个数量级。即使按照原始顺序，将 CVM 的正则化参数仅扰动 15% 也会在测试集上产生 49% 的错误率 [16]





**图 3:** 随着训练集的增大，测试数据的准确性持续提高。在 Forest 数据集上，使用随机装箱，将数据集大小加倍可将测试错误最多减少 40% (左)。随着 P 的增加 (中)，误差迅速衰减。训练时间随着 P 的增长而缓慢增长 (右)。

## 6 结论

我们介绍了两种随机特征，其内积一致地近似于许多常用的核。我们凭经验证明，将这些特征作为标准线性学习算法的输入，所产生的结果在准确性，训练时间和评估时间上均能与最新的大型核机器相竞争。

值得注意的是，可以通过串联这些特征来构造傅立叶特征和分箱特征的混合体。虽然我们专注于回归和分类，但是我们的功能可以应用于加速其他核方法，包括半监督和无监督学习算法。在所有这些情况下，通过首先计算随机特征，然后应用相关的线性技术，可以显著提高计算速度。

## 7 致谢

我们感谢 Eric Garcia 对这些特征的早期版本的帮助，感谢 Sameer Agarwal 和 James R. Lee 的有用的讨论，以及 Erik Learned-Miller 和 Andres Corrada-Emmanuel 的有用的修正。