承接上一篇推送，今天继续来看看论文 *Random Features for Large-Scale Kernel Machines* 中提出的第二种随机特征构造方法，姑且叫做随机装箱特征（Random Binnin Features）吧。

Random Binning Features

第二种特征特征提取方法，有着非常有趣的 Idea。用随机的分辨率和平移量，将数据所在的空间等分成小块，然后记录数据点在哪些小块当中。重复这个操作若干次，看看 2 个数据点被划分到同一个小块区域的频率是多少，用这个频率来近似这 2 个数据点的核函数值（核内积）。直观的来说，当 2 个数据点靠的越近的时候，它们被分到同一个小块区域的频率会越大，这样按上面的 Idea 所逼近的核函数值也应该越大。这是符合许多反应亲密度的核函数的特点。

这个想法也可以用映射的观点来刻画。 令 $z(x)$ 是数据点 x 所落区域的二进制编号（比如 01011 这样），这样就定义了一个映射 $z: R^d \to \{0,1\}^D$ ，其中 D 是编号的位数。 那么逻辑与运算 $z(x)\&z(y) := z(x)z(y)$ 的结果为 1 则表示数据点 x 和 y 落在了同一个区域中，为 0 则表示不在一个区域中。比方说，我们用不同的分辨率和平移量对空间做了 P 次分割，对应的有编号映射 $z_1, \cdots, z_P$ 。这样，数据点 x 和 y 落在同一个区域中的频率就是：

$$\frac{1}{P}\sum_{p=1}^{P} z_p(x)z_p(y) := z(x)^T z(y) \approx k(x,y)$$

其中 $z(x) = \frac{1}{\sqrt{P}}[z_1(x)\cdots z_P(x)]$ ，就是我们要找的特征映射。

带着这个 Idea，问题的重心就落在了如何随机的选取空间分割的分辨率和平移量，使得上面的近似能够尽可能精确。

首先我们要利用概率论知识来对整个分割空间的操作进行刻画，然后考察上述近似的精确度，并设法提高。一般思路是，确定分割区域的分辨率和平移量应该服从什么分布，才能使得频率 $z(x)^T z(y)$ 是 $k(x,y)$ 的无偏估计，然后刻画分割次数 P 对近似的精确度有何影响，比如估计随着 P 增大，$z(x)^T z(y)$ 收敛到 $k(x,y)$ 的速度（如果收敛的话）。

先考虑 1 维的情形。假设有一个核函数 $k(x, y)$ 。给定任意 2 个实数轴上的点 $x, y \in R$ 。我们把实数轴用随机选取的间隔 δ 等分成一系列区间，设 $p(\delta), \delta > 0$ 是 δ 服从的分布。然后再从 $[0, \delta]$ 的均匀分布中随机取 u 作为分割区间的偏移量，最后将整条实数轴均分成形如 $[u + k\delta, u + (k+1)\delta), n \in Z$ 的一系列区间。现在，为了让 $z(x)z(y) \approx k(x, y)$ ，当然首先希望 $z(x)z(y)$ 是 $k(x, y)$ 的无偏估计，就是说，我们希望：

$$k(x, y) = E_{\delta,u}[z(x)z(y)]$$

所以问题就集中在，怎么确定分布 $p(\delta)$ 使得上式成立。考虑到在分割中，我们是先取定 δ ，再取定 u 的，于是想到把 δ 作为条件，利用条件期望定义，得到：

$$E_{\delta,u}[z(x)z(y)] = E_{\delta}[E_u[z(x)z(y)|\delta]] = \int_0^{\infty} E_u[z(x)z(y)|\delta]p(\delta)d\delta$$

回忆 $z(x)z(y)$ 的含义：

$$z(x)z(y) = \begin{cases} 1 & \text{如果}x, y\text{落在同一个区间,} \\ 0 & \text{否则} \end{cases}$$

于是可以计算：

$$E_u[z(x)z(y)|\delta] = Pr_u[z(x)z(y) = 1|\delta]$$

接下来再计算上式右边，也就是 x, y 两个点落在同一个区间的概率。当 $|x - y| > \delta$ 的时候，2 个点无论如何都不可能在同一个区间内，因此这时它们落在同一区间内的概率是 0；而当 $|x - y| \leq \delta$ 时，由几何知识知道，给定的 2 点落在同一个区间的概率是 $1 - \frac{|x-y|}{\delta}$ 。因此，综合起来有：

$$Pr_u[z(x)z(y) = 1|\delta] = max\left(0, 1 - \frac{|x - y|}{\delta}\right) := \hat{k}(x, y; \delta)$$

这样，我们就得到了确定分布 $p(\delta)$ 的一个积分方程：

$$k(x, y) = \int_0^{\infty} \hat{k}(x, y; \delta)p(\delta)d\delta = \int_{|x-y|}^{\infty} \left(1 - \frac{|x - y|}{\delta}\right) p(\delta)d\delta$$

这时，就要对核函数的形状做一些约束了。假设核函数只和数据点的 L_1 距离有关，即有这样的形状：

$$k(x,y)=k(|x-y|)$$

这样，如果记 $\Delta=|x-y|$ ，上述方程改写成：

$$k(\Delta)=\int_{\Delta}^{\infty}\left(1-\frac{\Delta}{\delta}\right)p(\delta)d\delta$$

两边对 Δ 求 2 次导数，就可以得到：

$$\Delta\frac{\partial^2 k}{\partial\Delta^2}=p(\Delta)$$

至此，就得到了确定分布 p 的公式。并且，由于 p 是一个分布函数，上式成立，自然要求核函数是凸的，这样它的二阶导数才会大于 0。比如 Gauss 核函数 $e^{-|x-y|^2}$ 就不是这样的函数，也就是说，这次讨论的随机装箱特征不可能使用在 Gauss 核函数上面。但是 Laplace 核函数 $e^{-|x-y|}$ 就完全符合上面所有的要求，可以说随机装箱特征完全就是为 Laplace 核函数量身定做的。比如，Laplace 核函数对应的分布 p 恰好是 Gamma 分布函数 $\delta e^{-\delta}$ 。

接下来，就是重复做 P 次上面的分割，每次都随机的从分布 p 取不同的分辨率 δ ，从区间 $[0,\delta$ 随机的取偏移量 u ，得到一系列编码映射 $z_1,\cdots,z_P$ 。因为每个 $z_p(x)z_p(y)$ 都是核函数 $k(x,y)$ 的无偏估计，所以统计任意 2 点落在同一区间的频率：

$$\frac{1}{P}\sum_{p=1}^{P}z_p(x)z_p(y):=z(x)^Tz(y)\approx k(x,y)$$

也是核函数的一个无偏估计，而且方差更小。

到这里，我们就已经得到了 1 维情形的随机装箱特征算法，更高维的讨论是类似的，论文里面有相关讨论，这里就不费口舌了。我们把 1 维情形的算法整理如下：

算法 随机装箱特征

前提：数据空间 1 维。核函数 $k(x,y)$ 有形状 $k(|x-y|)=k(\delta)$ ，而且用下式构造的函数

$$p(\delta)=\delta\frac{\partial^2 k}{\partial\delta^2},\delta>0$$

是一个概率密度函数。

效果：得到随机特征映射 $z(x)$ 可以使得 $z(x)^T z(y)\approx k(|x-y|)$ 。

for $m=1,\cdots,P$

从分布 p 中随机选取分辨率 δ_m ，从区间 $[0,\delta_m]$ 内随机选取偏移量 u_m ，把实数轴等分成一系列区间。

把有数据点下落的区间用二进制编码，用 $z_p(x)$ 表示数据 x 下落的区间编号。

end for

令 $z(x)=\frac{1}{\sqrt{P}}[z_1(x)\cdots z_P(x)]$ 得所求。

可能要提下的是，论文里面没有提到用随机装箱特征的话，evaluation 里面 $w^T z(x)=\frac{1}{P}\sum_p w_p z_p(x)$ 里面权重的每一个分量是什么，那么为了统一运算，可以姑且认为也是一个二进制串。

最后，论文还讨论了随机装箱特征逼近核函数的收敛速度，这一段是很体现作者数学功力的。它的思路是从概率测度的意义上探讨算法随着分割次数 P 增大，逼近过程的收敛速度。结论是，逼近达到指定精确度的概率随着 P 增大，成指数增长到 1。有需要的话，笔者可能会专门花一篇文章来学习作者的这些技巧。

后记

总体而言，整篇论文的奇思妙想非常多，阅读过程也很愉快。但是可以看到，随机装箱特征适用的核函数是有限的，相比较起来，随机 Fourier 特征的适用范围更广一些。但是随机装箱特征也是有用武之地的，比如论文的实证部分提到的，一些分类问题数据集的分割平面高度不光滑，这时候随机 Fourier 特征的效果就远不如随机装箱特征。

这篇论文给我们的启示是，可以多用概率分布来刻画带有随机性的操作，然后借用概率论和数理统计的知识对问题进行建模和解决。另外，论文在推导随机 Fourier 特征时提到的那个调和分析的定理，也启发我们，看到一些概率密度或者测度的相关定理，也应该反方向的思考是否可以由此开发出对应的随机操作。