# Notes for Fundations of Data Science 1

//** 笔记基本根据上课的顺序完成，但是对一些问题做了注释和补充 **//

## 1 概率统计

### 大数定理

(1) 切比雪夫大数定理:

设 $x_1, x_2, \cdots, x_n, \cdots$ 是一系列相互独立的随机变量，都分别有期望 $\mathbb{E}(x_i)$ 和方差 $Var(x_i)$ 并且数列 $\{Var(x-1,), Var(x_2), \cdots\}$ 有上界，则对 $\forall \varepsilon > 0$，有:

$$\lim_{n \longrightarrow \infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} x_k - \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}x_k\right| < \varepsilon\right) = 1$$

(2) 伯努利大数定理:

设 $\mu$ 是 n 次独立事件中 A 发生的次数，且 A 事件在每次独立实验中发生的概率皆为 p，则对 $\forall \varepsilon > 0$，有:

$$\lim_{n \longrightarrow \infty} \mathbb{P}\left(\left|\frac{\mu_n}{n} - p\right| < \varepsilon\right) = 1$$

(3) 辛钦大数定律:

设 $\{a_i, i > lgeq1\}$ 是独立同分布的随机变量序列，若 $a_i$ 的数学期望存在，$\mathbb{E}(a_i 0 = \mu$，则对 $\forall \varepsilon > 0$，有:

$$\lim_{n \longrightarrow \infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} x_k - \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}x_k\right| < \varepsilon\right) = 1$$

直观上，辛钦大数定律还可以被表示成:

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^{n} x_i}{n} - \mathbb{E}(x)\right| \geq \varepsilon\right) \leq \frac{Var(x)}{n\varepsilon^2}$$

我们之后会谈到它的证明过程。

**【Markov's inequality】** 设 x 是一个非负随机变量，则对于 $a > 0$，有:

$$\mathbb{P}(x \geq a) \leq \frac{\mathbb{E}(x)}{a}$$

**证明:** 对于一个连续的非负随机变量 x 具有概率密度 p:

$$\begin{aligned}
\mathbb{E}(x) &= \int_0^{\infty} xp(x)dx && (1)\\
&= \int_0^{a} xp(x)dx + \int_a^{\infty} xp(x)dx && (2)\\
&\geq \int_a^{\infty} xp(x)dx && (3)\\
&\geq a\int_a^{\infty} p(x)dx && (4)\\
&= a \cdot \mathbb{P}(x \geq a) && (5)
\end{aligned}$$

即证；对于离散型随机变量 x，同理即可。

**推论：**

$$\mathbb{P}(x \geq b\mathbb{E}(x)) \leq \frac{1}{b}$$

**【Chebyshev's inequality】** 设 x 是一个随机变量，则对于 $c > 0$，有：

$$\mathbb{P}\left(|x - \mathbb{E}(x)| \geq c\right) \leq \frac{Var(x)}{c^2}$$

**证明：** 由于：

$$\mathbb{P}\left(|x - \mathbb{E}(x)| \geq c\right) = \mathbb{P}\left(|x - \mathbb{E}(x)|^2 \geq c^2\right)$$

令 $y = |x - \mathbb{E}(x)|^2$ 是非负随机变量，且 $\mathbb{E}(y0 = Var(x)$，利用马尔可夫不等式得：

$$\mathbb{P}(|x - \mathbb{E}(x)| \geq c) \leq \frac{\mathbb{E}\left(|x - \mathbb{E}(x)|^2\right)}{c^2} = \frac{Var(x)}{c^2}$$

利用切比雪夫不等式以及性质 $Var(x) = Var(x_i)$ 可以证明辛钦大数定律的推导形式。

**【Master Tail Bounds Theorem】** 设 $x = x_1 + x_2 + \cdots, +x_n$，其中 $x_1, x_2, \cdots, x_n$ 是相互独立的随机变量，且均值为零，方差不超过 $\sigma^2$。假设 $0 \leq a \leq \sqrt{2n\sigma^2}$，假设对于 $s = 3, 4, \cdots, \frac{a^2}{4n\sigma^2}$ 有：$|\mathbb{E}(x_i^s)| \leq \sigma^2 s!$，则有：

$$\mathbb{P}(|x| \geq a) \leq 3e^{-\frac{a^2}{12n\sigma^2}}$$

**lemma** 设 $x = x_1 + x_2 + \cdots + x_n$，其中 $x_1, x_2 \cdots, x_n$ 是相互独立的随机变量，各自的均值为 0，方差至多为 $\sigma^2$。假设 $a \in [0, \sqrt{2n\sigma^2}]$ 并且 $s < \frac{n\sigma^2}{2}$ 是一个正偶数且 $|\mathbb{E}(x_i^r)| \leq \sigma^2 r!$，对于 $r = 3, 4, \cdots, s$，有：

$$\mathbb{P}(|x_1 + x_2 + \cdots + x_n| \geq a) \leq \left(\frac{2sn\sigma^2}{a^2}\right)^{s/2}$$

进一步要求 $s \geq \frac{a^2}{4n\sigma^2}$ 就可以得到 Master Tail Bounds Theorem。

**证明：** 先给出一个适合任意正偶数 r 的上界。由于：

$$(x_1 + x_2 + \cdots + x_n)^r = \sum_{\sum r_i = r} \binom{r}{r_1, r_2, \cdots, r_n} x_1^{r_1} x_2^{r_2} \cdots x_n^{r_n}$$

根据独立性：

$$\mathbb{E}(x^r) = \sum \frac{r!}{r_1! r_2! \cdots r_n!} \mathbb{E}x_1^{r_1} \mathbb{E}x_2^{r_2} \cdots \mathbb{E}x_n^{r_n}$$

对于任意一项，由于 $\mathbb{E}(x_i) = 0$ 可知只要 $r_i = 1$，该项值为 0。此后取 $(x_1, x_2, \cdots, x_n)$ 非 0 且满足 $x_i \geq 2$ 的集合，每个集合中至多有 $\frac{r}{2}$ 个非 $0 r_i$，由于 $|\mathbb{E}(x_i^{r_i})| \leq \sigma^2 r_i!$，有：

$$\mathbb{E}(x^r) \leq r! \sum_{(r_1, r_2, \cdots, r_n)} \sigma^2 (number\ of\ non - zero\ r_i\ in\ set)$$

对 $t = 1, 2, \cdots, r/2$，将求和的项按照 t 个非 0 的 $r_i$ 分组，对于每个固定的 t，这样的组内有 $\binom{n}{t}$ 个子集；对这样的确定的子集，设 $r_i \geq 2$，为每个 $r_i$ 分配 2，然后随机分配剩余的 $r - 2t$，这样的方法数为 $\binom{r-2t+t-1}{t-1} = \binom{r-t-1}{t-1}$，从而：

$$\mathbb{E}(x^r) \leq r! \sum_{t=1}^{\frac{r}{2}} f(t), \qquad where \quad f(t) = \binom{n}{t} \binom{r-t-1}{t-1} \sigma^{2t}$$

假设 $h(t)=\frac{(n\sigma^2)^t}{t!}2^{r-t-1}$，有 $f(t)<h(t)$，由于 $t\le\frac{r}{2}\le\frac{n\sigma^2}{4}$，我们有：

$$\frac{h(t)}{h(t-1)}=\frac{n\sigma^2}{2t}\ge 2$$

于是：

$$\mathbb{E}(x^r)=r!\sum_{t=1}^{\frac{R}{2}}f(t)\le r!h(r/2)\left(1+\frac{1}{2}+\frac{1}{4}+\cdots\right)\le\frac{r!}{\frac{r}{2}!}2^{r/2}(n\sigma^2)^{r/2}$$

应用马尔可夫不等式：

$$\mathbb{P}(|x|>a)=\mathbb{P}(|x|^r>a^r)\le\frac{r!(n\sigma^2)^{r/2}2^{r.2}}{(r/2)!a^r}=g(r)\le\left(\frac{2rn\sigma^2}{a^2}\right)^{r/2}$$

对于偶数 $r\le s$ 皆成立。

对偶数 r，$\frac{g(r)}{g(r-2)}=\frac{4(r-1)n\sigma^2}{a^2}$，因此，只要 $r-1\le\frac{a^2}{4n\sigma^2}\mathrm{g(r)}$ 就会减小。令 r 为小于等于 $\frac{a^2}{6n\sigma^2}$ 的最大偶数，尾部概率至大为 $e^{-r/2}$，最多为 $e\cdot e^{-a^2/(12n\sigma^2)}\le 3\cdot e^{-a^2/(12n\sigma^2)}$

**尾界定理的应用**

设 $y_1,y_2,\cdots,y_n$ 是 n 个独立的 0-1 随机变量，且对于所有的 i 有 $\mathbb{E}(y_i)=p$，设 $y=\sum_{i=1}^n y_i$，对于任意 $c\in[0,1]$ 有：

$$\mathbb{P}(|y-\mathbb{E}(y)|\ge cnp)\le 3e^{-npc^2/8}$$

令 $x_i=y_i-p$，于是 $\mathbb{E}(x_i)=0$ 以及 $\mathbb{E}(x_i^2)=\mathbb{E}(y-p)^2=p$，对于 $s>3$：

$$\begin{aligned}|\mathbb{E}(x_i^s)|&=|\mathbb{E}(y_i-p)^s| &(6)\\ &=|p(1-p)^s+(1-p)(0-p)^s| &(7)\\ &=|p(1-p)\left((1-p)^{s-1}+(-p)^{s-1}\right)| &(8)\\ &\le p. &(9)\end{aligned}$$

对上一个 lemma 应用 $a=cnp$，且注意到 $a<\sqrt{2}np$ 立证。

**幂律分布**

k 阶的幂律分布如下定义（这里 k 是一个正整数）：

$$f(x)=\frac{k-1}{x^k}\qquad for\qquad x\ge 1$$

对阶数至少为 4 的幂律分布，有：

$$\mu=\mathbb{E}(x)=\frac{k-1}{k-2}\qquad and\qquad Var(x)=\frac{k-1}{(k-2)^2(k-3)}$$

◦ 设 $x_1,x_2,\cdots,x_n$ 分别是依不少于 4 阶的幂律分布独立同分布的（$n>10k^2$），则对于 $x=\sum_{i=1}^n x_i$ 及 $\forall\varepsilon\in(1/(2\sqrt{nk}),1/k^2)$，有：

$$\mathbb{P}\left(|x-\mathbb{E}(x)|\ge\varepsilon\mathbb{E}(x)\right)\le\left(\frac{4}{\varepsilon^2(k-1)n}\right)^{(k-3)/2}$$

对于整数 s，变量 $x_i$ 的 s 阶中心矩矩存在当且仅当 $s < k-2$，对 $s < k-2$，有：

$$\mathbb{E}((x_i - \mu)^s) = (k-1)\int_1^\infty \frac{y-\mu)^s}{y^k} dy$$

替换变量 $z = \frac{\mu}{y}$：

$$\frac{(y-\mu)^s}{y^k} = y^{s-k}(1-z)^s = \frac{z^{k-s}}{\mu^{k-s}}(1-z)^s$$

当 y 从 1 增长到 $\infty$ 时，z 从 $\mu$ 变化到 0，并且 $dz = -\frac{\mu}{y^2}dy$，从而：

$$\mathbb{E}((x_i - \mu)^s) = (k-1)\int_1^\infty \frac{(y-\mu)^s}{y^k} dy \tag{10}$$

$$= \frac{k-1}{\mu^{k-s-1}}\int_0^1 (1-z)^s z^{k-s-2} dz + \frac{k-1}{\mu^{k-s-1}}\int_1^\mu (1-z)^s z^{k-s-2} dz \tag{11}$$

积分的前半部分是一个标准的 beta 积分，它的值为 $\frac{s!(k-s-2)!}{(k-1)!}$，为了计算后半部分积分的界，注意到 $z \in [1, \mu], |z-1| \le \frac{1}{k-2}$ 并且：

$$z^{k-s-2} \le (1 + (1/(k-2)))^{k-s-2} \le e^{(k-s-2)/(k-2)} \le e$$

从而：

$$|\mathbb{E}((x_i - \mu)^s)| \le \frac{(k-1)s!(k-2-s)!}{(k-10)!} + \frac{e(k-1)}{(k-2)^{s+1}} \tag{12}$$

$$\le s! Var(y)\left(\frac{!}{k-4} + \frac{e}{3!}\right) \tag{13}$$

$$\le s! Var(x) \tag{14}$$

根据尾界定理中的第一个不等式，我们需要将参数 s 设为 $k-2$ 或 $k-3$，具体值的选择基于哪个值是偶数，此外，结合题目中提到的 $a = \varepsilon\mathbb{E}(x) \le \sqrt{2n\sigma^2}$ 和 $\varepsilon \le \frac{!}{k^2}$ 可证。

## 2 高维几何学

高维对象的一个重要特性是它们的大部分体积集中在表面附近。考虑任意 d-维空间中的对象 A，现将 A 缩小一个很小的比例 $\varepsilon$，产生一个新对象 $(1-\varepsilon)A = \{(1-\varepsilon)x | x \in A\}$，有下式成立：

$$V((1-\varepsilon)A) = (1-\varepsilon)^d \cdot V(A)$$

这里 $V(A)$ 表示 A 的体积。利用不等式 $1 - x \le e^{-x}$ 可以更好得说明。

**d-维球体积与表面积**

在直角坐标系下，d-维球的体积可以由下面的公式给出：

$$V(d) = \int_{x_1=-1}^{x_1=1} \int_{x_2=-\sqrt{1-{x_1}^2}}^{x_2=\sqrt{1-{x_1}^2}} \cdots \int_{x_d=-\sqrt{1-x_1^2-\cdots-x_{d-1}^2}}^{x_d=\sqrt{1-x_1^2-\cdots-x_{d-1}^2}} dx_d \cdots dx_2 dx_1$$

如果使用极坐标系，我们可以得到：

$$V(d) = \int_{S^d}\int_{r=0}^1 r^{d-1} dr d\Omega$$

由于变量 Ω和$r$ 不相关，因此有：

$$V(d) = \int_{S^d} d\Omega \int_{r=0}^{1} r^{d-1} dr = \frac{1}{d} \int_{S^d} d\Omega = \frac{A(d)}{d}$$

其中 $A(d)$ 是 d-维单位球的表面积。考察另一个积分：

$$I(d) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-(x_1^2+x_2^2+\cdots+x_d^2)} dx_d \cdots dx_2 dx_1$$

事后考察两种积分的关系即可。

利用直角坐标系计算 $I(d)$：

$$I(d) = \left( \int_{-\infty}^{\infty} e^{-x^2} dx \right)^d = (\sqrt{\pi})^d = \pi^{\frac{d}{2}}$$

再利用极坐标系计算 $I(d)$：

$$I(d) = \int_{S^d} d\Omega \int_0^{\infty} e^{-r^2} r^{d-1} dr$$

积分的前半部分是对整个立体角的积分，相当于 d-单位球的表面积 $A(d)$，对于剩余部分，利用 gamma 函数：

$$\int_0^{\infty} e^{-r^2} r^{d-1} dr = \int_0^{\infty} e^{-t} t^{(d-1)/2} \frac{1}{2} t^{-1/2} dt = \frac{1}{2} \int_0^{\infty} e^{-t} t^{d/2-1} dt = \frac{1}{2} \Gamma\left(\frac{d}{2}\right)$$

结合 $I(d) = \pi^{\frac{d}{2}}$ 和 $I(d) = A(d) = A(d) \frac{1}{2} \Gamma\left(\frac{d}{2}\right)$ 可知：

$$A(d) = \frac{2\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)}$$

从而可以得到下面的结论：

$$A(d) = \frac{2\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)}$$

$$V(d) = \frac{2\pi^{d/2}}{d\Gamma\left(\frac{d}{2}\right)}$$

利用已有的知识也可以简单验证低维下公式的正确性。

下面还需要到前文论证中的两个事实进行说明，其一是积分 $\int_{-\infty}^{\infty} e^{-x^2} dx$ 的计算：

取独立于 x 分布的变量 y，有：

$$\int_{-\infty}^{\infty} e^{-x^2} dx \int_{-\infty}^{\infty} e^{-y^2} dy = \int_{-\infty}^{\infty} e^{-x^2-y^2} dx dy = \left( \int_{-\infty}^{\infty} e^{-x^2} dx \right)^2$$

不妨：

$$x = r \cdot cos\theta \qquad y = r \cdot sin\theta$$

从而上式转化为：

$$\left( \int_{-\infty}^{\infty} e^{-x^2} dx \right)^2 = \int_0^{\infty} \int_0^{2\pi} e^{-r^2} r dr d\theta \tag{15}$$

$$= \int_0^{\infty} e^{-r^2} r dr \int_0^{2\pi} d\theta \tag{16}$$

$$= -2\pi \left[ \frac{e^{-r^2}}{2} \right]_0^{\infty} \tag{17}$$

$$= \pi \tag{18}$$

即证。另一方面，Gamma 函数是对阶乘函数在复数域上的拓展，有余元公式：

$$\Gamma(x)\Gamma(1-x)=\frac{\pi}{\sin\pi x}\ (0<x<1)$$

代入 $x=\frac{1}{2}$ 即可得到 $\Gamma\left(\frac{1}{2}\right)$ 的结果。

**高维球在赤道附近的体积性质**

**Theorem** 对于 $c\geq 1$ 和 $d\geq 3$，至少有 $1-\frac{2}{c}e^{-c^2/2}$ 的单位球体积满足 $|x_1|\leq\frac{c}{\sqrt{d-1}}$.

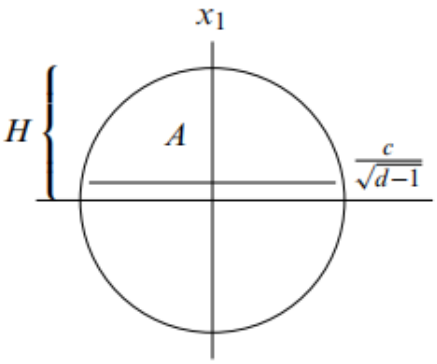


图 1: d-维上半球体积的绝大部分在超平面$x_1=\frac{c}{\sqrt{d-1}}$的下方

**证明** 只需证明定理的另一个侧面。不妨设 $A$ 是球中满足 $x_1\geq\frac{c}{\sqrt{d-1}}$ 的部分，$H$ 代表上半球，如图 1 所示。我们下面证明 A 和 H 的体积比将趋于 0：

$$\frac{V(A)}{V(H)}\leq\frac{min\{V(A)\}}{min\{V(H)\}}=\frac{2}{e}e^{-\frac{c^2}{2}}$$

考察 A 的体积：将其表示为厚度为 $dx_1$ 的超圆盘的增量体积，其截面是一个半径为 $\sqrt{1-x_1^2}$ 的 $d-1$ 维球，于是：

$$V(A)=\int_{\frac{c}{\sqrt{d-1}}}^{1}(1-x_1^2)^{\frac{d-1}{2}}V(d-1)dx_1$$

利用 $1-x\leq e^{-x}$ 有：

$$V(A)\leq\int_{\frac{c}{\sqrt{d-1}}}^{\infty}\frac{x_1\sqrt{d-1}}{c}e^{-\frac{d-1}{2}x_1^2}V(d-1)dx_1 \tag{19}$$

$$=V(d-1)\frac{\sqrt{d-1}}{c}\int_{\frac{c}{\sqrt{d-1}}}^{\infty}x_1e^{-\frac{d-1}{2}x_1^2}dx_1 \tag{20}$$

$$=V(d-1)\frac{\sqrt{d-1}}{c}\frac{1}{d-1}e^{-\frac{c^2}{2}} \tag{21}$$

$$=\frac{V(d-1)}{c\sqrt{d-1}}e^{-\frac{c^2}{2}} \tag{22}$$

另一方面，对于 $H$ 的体积，尽管它的确切体积显然可求，但希望可以在化简比例的时候消去 $V(d-1)$，因此，考察 H 在超平面 $x_1=\frac{1}{\sqrt{d-1}}$ 下方的近似柱体，其高度为 $x_1$，底面半

径为 $\sqrt{1-\frac{1}{d-1}}$，体积为 $V(d-1)(1-\frac{1}{d-1})^{\frac{d-1}{2}}\frac{1}{\frac{1}{d-1}}$，利用伯努利可知该部分的体积总不会小于 $\frac{V(d-1)}{2\sqrt{d-1}}(d\geq 3)$，于是：

$$ratio \leq \frac{\frac{V(d-1)}{c\sqrt{d-1}}e^{-\frac{c^2}{2}}}{\frac{V(d-1)}{2\sqrt{d-1}}} = \frac{2}{c}e^{-\frac{c^2}{2}}$$

上述分析的一个直接结论是，如果我们从单位球中随机取出两个点，他们几乎正交，即他们的夹角几乎为 $\frac{\pi}{2}\pm O(\frac{1}{\sqrt{d}})$，从而引出下面的定理：

考虑从单位球中随机抽取的 n 个点：$x_1, x_2, \cdots, x_n$，以概率 $1-O(\frac{1}{n})$：

- 对所有 i，有 $|x_i| \geq 1-\frac{2\ln n}{d}$；
- 对所有 $i\neq j$，有 $|x_i \cdot x_j| \leq \frac{\sqrt{6\ln n}}{\sqrt{d-1}}$。

利用大部分体积集中在表面的分析，$\mathbb{P}(|x_i|<1-\varepsilon)<e^{-\varepsilon d}$，从而：

$$\mathbb{P}\left(|x_i|<1-\frac{2\ln n}{d}\right)\leq e^{-(\frac{2\ln n}{d})d}=\frac{1}{n^2}$$

利用联合界性质，存在某个 $i$ 使得 $|x_i|<1-\frac{2\ln n}{d}$ 的概率至多为 $\frac{1}{n}$。

另一方面，利用高维球在赤道附近的体积性质，某个向量满足 $|x_i|>\frac{c}{\sqrt{d-1}}$ 的概率至多为 $\frac{2}{c}e^{-\frac{c^2}{2}}$，有 $\binom{n}{2}$ 对 i 和 j，对每对向量，定义 $x_i$ 为“北”，则 $x_j$ 在“北”方向的投影超过 $\frac{\sqrt{6\ln n}}{\sqrt{d-1}}$ 的概率至多为 $O(e^{-\frac{6\ln n}{2}})=O(n^{-3})$，从而，$|x_i\cdot x_j|>\frac{\sqrt{6\ln n}}{\sqrt{d-1}}$ 的概率至多为 $O\left(\binom{n}{2}n^{-3}\right)=O(\frac{1}{n})$。

除此之外，考虑中心在原点，边长为 $\frac{2c}{\sqrt{d-1}}$ 的小盒子，注意到当 c 取 $2\sqrt{\ln d}$ 时，盒子包含了超过一半的单位球的体积，并且显然盒子的体积趋于 0，从而单位球的体积也趋于 0；并且在这个 c 的取值下，单位球中 $|x_1|\geq\frac{c}{d-1}$ 的部分至多为占：

$$\frac{2}{c}e^{-\frac{c^2}{2}}=\frac{1}{\sqrt{\ln d}}e^{-2\ln d}=\frac{1}{d^2\sqrt{\ln d}}<\frac{1}{d^2}$$

利用加法原理，最多有 $d\cdot\frac{1}{d^2}=\frac{1}{d}$ 的体积位于立方体之外，即证。

**随机生成点的相关结论**

考虑在单位球表面均匀随机生成点。对于二维情况下，在单位圆周上生成点，可以独立地从区间 $[0,1]$ 中均匀随机生成每个坐标。这会产生一个足够大的正方形内的点，该正方形完全包含单位圆。将每个点投影到单位圆上。然而，由于从原点到正方形顶点的线段比从原点到正方形边中点的线段长，因此这种分布不是均匀的。为了解决这个问题，可以丢弃所有位于单位圆外的点，并将剩余的点投影到圆上。在更高维度中，这种方法不再适用，因为落在球内的点的比例会趋近于零，导致几乎所有点都会被丢弃。解决方法是生成每个坐标都是独立高斯变量的点。具体来说，生成 $x_1, x_2, \cdots, x_d$，使用均值为 0、方差为 1 的高斯分布，即：

$$p(x)=\frac{1}{(2\pi)^{\frac{d}{2}}}e^{-\frac{x_1^2+x_2^2+\cdots+x_d^2}{2}}$$

上述等式 RHS 的幂在球中恰好是一个常量，因此这个分布关于单位球是均匀分布（球对称），为了得到在球面上均匀分布的点，只需将每一个生成向量归一化即可，但这也使得他们的坐标不再是统计独立的。

为了在单位球（包括表面和内部）内均匀生成点，可以将生成的表面点 $\frac{x}{|x|}$ 按标量 $\rho \in [0,1]$ 放缩，注意到 $d-$ 维半径为 $r$ 的球的体积是 $r^d V(d)$，因此在半径为 $r$ 处密度恰好是 $\frac{d}{dr}(r^d V(d)) = dr^{d-1}V(d)$，因此，取 $dr^{d-1}$ 为 $\rho$ 才是妥当的。

**Box-Muller** 要得到服从正态分布的随机数，基本思想是先得到服从均匀分布的随机数再将服从均匀分布的随机数转变为服从正态分布。

**d-维球形高斯分布** 对于 d-维球形高斯分布，其均值为 0，每个坐标上的方差为 $\sigma^2$，其概率密度函数为：

$$p(x) = \frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d} e^{-\frac{|x|^2}{2\sigma^2}}$$

尽管密度函数在原点处达到最大值，但在原点附近的体积非常小。当 $\sigma^2 = 1$ 时，在原点为中心的单位球内积分概率密度几乎为零，因为这样的球体体积可以忽略不计。实际上，需要将球的半径增加到接近 $\sqrt{d}$ 才会有显著的体积和概率质量。如果进一步增加半径，即使体积增加，积分也几乎不会增加，因为概率密度下降的速度更快。以下定理正式表明，几乎所有概率都集中在半径为 $\sqrt{d}$ 的薄环（annulus）中。

**Gaussian annulus theorem** 对于 d-维球形高斯分布，每个方向上的方差为 1，对于任意 $\beta \le \sqrt{d}$ 最多有 $3e^{-c\beta^2}$ 的概率质量集中在环 $\sqrt{d} - \beta \le |x| \le \sqrt{d} + \beta$ 内，这里 c 为一个确定的常量。

**证明：** 设 $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ 是从原点为中心、单位方差的高斯分布中选取的点，令 $r = |\mathbf{x}|$。条件 $d - \beta \le |\mathbf{x}| \le \sqrt{d} + \beta$ 等价于 $|r - \sqrt{d}| \ge \beta$.

如果 $|r - \sqrt{d}| \ge \beta$，则两边乘以 $r + \sqrt{d}$ 得到：$|r^2 - d| \ge \beta(r + \sqrt{d}) \ge \beta\sqrt{d}$. 因此，我们希望界定 $|r^2 - d| \ge \beta\sqrt{d}$ 的概率。

重写 $r^2 - d$ 为：$r^2 - d = (x_1^2 + \cdots + x_d^2) - d = (x_1^2 - 1) + \cdots + (x_d^2 - 1)$, 并进行变量替换：$y_i = x_i^2 - 1$。我们需要考察 $|y_1 + \cdots + y_d| \ge \beta\sqrt{d}$ 的概率。注意 $E(y_i) = E(x_i^2) - 1 = 0$。考察 $y_i$ 的 $s$-阶矩。

对于 $|x_i| \le 1$，有 $|y_i|^s \le 1$；对于 $|x_i| \ge 1$，有 $|y_i|^s \le |x_i|^{2s}$。因此：

$$|E(y_i^s)| = E(|y_i|^s) \le E(1 + x_i^{2s}) = 1 + E(x_i^{2s}) = 1 + \sqrt{\frac{2}{\pi}} \int_0^\infty x^{2s} e^{-x^2/2} dx$$

使用替换 $2z = x_i^2$，得到：

$$|E(y_i^s)| = 1 + \frac{1}{\pi} \int_0^\infty 2^s z^{s-1/2} e^{-z}\, dz \le 2^s s!.$$

最后一个不等式来自伽马积分。

由于 $E(y_i) = 0$，因此 $\mathrm{Var}(y_i) = E(y_i^2) \le 2^2 \cdot 2 = 8$. 不幸的是，这并不满足要求的 $|E(y_i^s)| \le 8s!$。为了解决这个问题，我们再次进行变量替换，使用 $w_i = y_i/2$。此时，$\mathrm{Var}(w_i) \le 2$ 且 $|E(w_i^s)| \le s!$，我们的目标是界定了 $|w_1 + \cdots + w_d| \ge \frac{\beta}{2}\sqrt{d}$ 的概率。

根据主尾界定理，其中 $\sigma^2 = 2$ 且 $n = d$，这种情况发生的概率小于或等于 $3e^{-\frac{\beta^2}{96}}$.

在处理高维数据的任务中，最近邻搜索是最常用的子程序之一。在最近邻搜索问题中，我们有一个包含 $n$ 个点的数据库，这些点位于 $\mathbb{R}^d$ 空间中，其 n 和 d 通常较大。数据库可以预先处理并存储在一个高效的数据结构中。之后，我们会收到查询点，任务是找到与查询点最接近或近似最接近的数据库点。由于查询的数量往往很大，每个查询的回答时间应该非常短，理想情况下是 $\log n$和 $\log d$ 的小函数，而预处理时间可以更大，即 n 和 d 的多项式函数。为了提高效率，降维技术（将数据库点投影到 $k$ $d$ 维空间）是非常有用的，只要点之间的相对距离能够大致保持不变。我们将使用高斯环定理来证明这样的投影确实存在且简单。

**随机投影方法**

考虑以下投影 $f : \mathbb{R}^d \to \mathbb{R}^k$（许多相关的投影也被证明有效）。选择 $k$ 个 $\mathbb{R}^d$ 中的单位方差向量 (满足高斯分布) $u_1, u_2, \ldots, u_k$。对于任何向量 $v$，定义投影 $f(v)$ 为：

$$f(v) = (u_1 \cdot v, u_2 \cdot v, \ldots, u_k \cdot v)$$

这里 $f(v)$ 是 $v$ 与 $u_i$ 的点积组成的向量。我们将证明，以高概率 $|f(v)| \approx \sqrt{k}|v|$。对于任意两个向量 $v_1$ 和 $v_2$，有 $f(v_1 - v_2) = f(v_1) - f(v_2)$。因此，要估计 $\mathbb{R}^d$ 空间中两个向量 $v_1$ 和 $v_2$ 之间的距离 $|v_1 - v_2|$，只需计算 $k$-维空间中的 $|f(v_1) - f(v_2)| = |f(v_1 - v_2)|$，因为 $\sqrt{k}$ 是已知的，可以除以它。需要注意的是，投影后距离增加是因为向量 $u_i$ 不是单位长度，而且它们不是正交的。如果要求它们正交，就会失去统计独立性。

**Theorem-随机投影定理** 设 $v$ 是 $\mathbb{R}^d$ 中的一个固定向量，且 $f$ 如上定义。存在常数 $c > 0$，对于 $\epsilon \in (0, 1)$，有：

$$\mathrm{Prob}\left(\left||f(v)| - \sqrt{k}|v|\right| \ge \epsilon\sqrt{k}|v|\right) \le 3e^{-ck\epsilon^2}$$

这里的概率基于用于构造 $f$ 的随机向量 $u_i$ 。

**证明：** 通过将内不等式的两边都缩放 $|v|$，我们可以假设 $|v| = 1$。独立正态分布实变量的和也是正态分布的，其均值和方差分别是各个变量的均值和方差之和。由于 $u_i \cdot v = \sum_{j=1}^d u_{ij}v_j$，随机变量 $u_i \cdot v$ 具有零均值和单位方差的高斯密度，特别地，

$$\mathrm{Var}(u_i \cdot v) = \mathrm{Var}\left(\sum_{j=1}^d u_{ij}v_j\right) = \sum_{j=1}^d v_j^2 \mathrm{Var}(u_{ij}) = \sum_{j=1}^d v_j^2 = 1$$

由于 $u_1 \cdot v, u_2 \cdot v, \ldots, u_k \cdot v$ 是满足高斯分布的独立随机变量，$f(v)$ 是一个 $k$-维球形高斯分布的随机向量，每个坐标上的方差为 1。因此，根据高斯环定理（定理 2.9），当 $d$ 替换为 $k$ 时，定理成立。

**Johnson-Lindenstrauss Lemma** 对于任意 $0 < \epsilon < 1$ 和整数 $n$，令 $k \geq \frac{3}{c\epsilon^2} \ln n$，其中 $c$ 如定理 2.9 所示。对于 $\mathbb{R}^d$ 空间中的任意 $n$ 个点集合，上述定义的随机投影 $f : \mathbb{R}^d \to \mathbb{R}^k$ 具有以下性质：对于所有点对 $v_i$ 和 $v_j$，以至少 $1 - \frac{3}{2n}$ 的概率，

$$(1-\epsilon)\sqrt{k}\|v_i - v_j\| \leq \|f(v_i) - f(v_j)\| \leq (1+\epsilon)\sqrt{k}\|v_i - v_j\|$$

**证明：** 应用随机投影定理，对于任意固定的 $v_i$ 和 $v_j$，$|f(v_i - v_j)|$ 落在范围

$$[(1-\epsilon)\sqrt{k}\|v_i - v_j\|, (1+\epsilon)\sqrt{k}\|v_i - v_j\|]$$

之外的概率至多为 $3e^{-ck\epsilon^2} \leq \frac{3}{n^3}$（当 $k \geq \frac{3}{c\epsilon^2} \ln n$ 时）。由于点对的数量 $\binom{n}{2} < \frac{n^2}{2}$，通过联合界，任何一对点具有大失真的概率小于 $\frac{3}{2n}$。

注意，上述定理的结论断言对于所有的 $v_i$ 和 $v_j$，而不是大多数。较弱的断言对于大多数 $v_i$ 和 $v_j$ 通常不太有用，因为我们的算法（如最近邻搜索）可能会返回一些“坏”点对。引理的一个显著方面是投影维度 $k$ 仅依赖于 $n$ 的对数。由于 $k$ 通常远小于 $d$，这被称为降维技术。在应用中，主导项通常是 $\frac{1}{\epsilon^2}$ 项。

对于最近邻问题，如果数据库中有 $n_1$ 个点，并且预期在算法生命周期内会有 $n_2$ 个查询，取 $n = n_1 + n_2$ 并将数据库投影到随机 $k$-维空间，其中 $k$ 如定理 2.11 所述。在接收到查询时，将其投影到同一子空间，并计算附近的数据库点。Johnson-Lindenstrauss 引理表明，以高概率，这将给出正确的答案，无论查询是什么。注意到 $k$ 依赖于 $\ln n$ 而不是 $n$，这是由于 $k$ 在 $\epsilon$ 下指数级小的概率使得 $k$ 只依赖于 $\ln n$。

### 混合高斯模型及其应用

假设我们在记录某城市 20-30 岁人群的身高。我们知道，平均而言，男性比女性更高，因此一个自然的模型是高斯混合模型：

$$p(x) = w_1 p_1(x) + w_2 p_2(x),$$

其中 $p_1(x)$ 表示女性身高的高斯密度，$p_2(x)$ 表示男性身高的高斯密度，而 $w_1$ 和 $w_2$ 是表示城市中女性和男性比例的混合权重。

对于混合模型的参数估计问题，给定从总体密度 $p$ 中抽取的样本（例如，城市中人的身高，但不告知该身高对应的是男性还是女性），我们需要重建分布的参数（例如，对 $p_1$ 和 $p_2$ 的均值和方差以及混合权重的良好近似）。

即使解决了高度参数估计问题，给定一个数据点，我们也不一定能确定它来自哪个群体。也就是说，给定一个身高，我们无法确定它是来自男性还是女性。在下文中，我们将探讨一个在某些方面更简单、而在其他方面更困难的问题。它将更困难，因为我们关注的是高维空间中的两个高斯分布的混合，而不是一维的高度情况。但它将更简单，因为我们将假设均值之间的距离相对于方差来说非常大。具体来说，我们的重点将是两个球形单位方差高斯分布的混合，其均值之间的距离为 $\Theta(d^{1/4})$。我们将证明，在这种分离水平下，我们可以以高概率唯一确定每个数据点来自哪个高斯分布。为此所需的算法实际上相当简单：计算所有点对之间的距离。距离较近的点来自同一个高斯分布，而距离较远的点则来自不同的高斯分布。稍后我们将看到，使用更复杂的算法，即使分离距离为 $\Theta(1)$ 也足够了。

首先考虑仅有一个以原点为中心的球形单位方差高斯分布。根据高斯环定理，其大部分概率质量集中在半径为 $\sqrt{d}$ 的薄环内，宽度为 $O(1)$。此外，几乎所有的质量都在厚度为 $O(1)$ 的板条内，即 $\{x \mid -c \le x_1 \le c\}$，其中 $c \in O(1)$。

选择一个点 $x$ 来自这个高斯分布；然后旋转变换坐标系使得坐标系的 $x$ 轴与向量 x 重合，此后独立地选择第二个点 $y$。由于高斯分布的大部分概率质量位于赤道附近的板条内，$y$ 在 $x$ 方向上的分量以高概率为 $O(1)$。因此，$y$ 几乎与 $x$ 垂直。所以，

$$|x-y| \approx \sqrt{|x|^2+|y|^2}.$$

见图 2(a)。更精确地说，由于坐标系已旋转使得 $x$ 位于北极，即 $x=(\sqrt{d}\pm O(1),0,\dots,0)$。因为 $y$ 几乎位于赤道上，进一步旋转坐标系，使得 $y$ 垂直于北极轴的分量位于第二坐标。那么

$$y=(O(1),\sqrt{d}\pm O(1),0,\dots,0).$$

因此，

$$(x-y)^2 = d\pm O(\sqrt{d}) + d \pm O(\sqrt{d}) = 2d \pm O(\sqrt{d}),$$

并且

$$|x-y| = \sqrt{2d}\pm O(1) \quad \text{以高概率成立}$$

考虑两个球形单位方差高斯分布，中心分别为 $p$ 和 $q$，且它们之间的距离为 $\Delta$。随机选择第一个高斯分布中的一个点 $x$ 和第二个高斯分布中的一个点 $y$，这两个点之间的距离接近 $\sqrt{\Delta^2+2d}$，因为 $x-p$、$p-q$ 和 $q-y$ 几乎相互垂直。选择 $x$ 并旋转坐标系使 $x$ 位于北极。设 $z$ 是近似第二个高斯分布球体的北极。现在选择 $y$。第二个高斯分布的大部分质量位于与 $z-q$ 垂直的赤道附近 $O(1)$ 内。同样，每个高斯分布的大部分质量位于与 $q-p$ 垂直的相应赤道附近 $O(1)$ 内。见图 2(b)。因此，

$$|x-y|^2 \approx \Delta^2 + |z-q|^2 + |q-y|^2 = \Delta^2 + 2d \pm O(\sqrt{d}).$$

为确保来自同一高斯分布的两点之间的距离小于来自不同高斯分布的两点之间的距离，需要满足同一高斯分布的点对之间的最大距离不超过不同高斯分布的点对之间的最小距离，即：

$$2d+O(1) \le \sqrt{2d+\Delta^2} - O(1) \quad \text{或} \quad 2d+O(\sqrt{d}) \le 2d+\Delta^2,$$

当 $\Delta \in \omega(d^{1/4})$ 时成立。因此，只要两个高斯分布的中心之间距离为 $\omega(d^{1/4})$，就可以用这种方法分离它们。如果我们有 $n$ 个点，并希望以高概率正确分离所有这些点，我们需要个体高概率陈述以概率 $1-1/\mathrm{poly}(n)^3$ 成立，这意味着定理 2.9 中的 $O(1)$ 项变为 $O(\sqrt{\log n})$。因此，我们需要在分离距离中包含额外的 $O(\sqrt{\log n})$ 项。

分离来自两个高斯分布的点的算法如下：

1. 计算所有点对之间的距离。
2. 找到最小成对距离的簇：这些点必须来自同一个高斯分布。
3. 移除这些点：剩余的点来自第二个高斯分布。

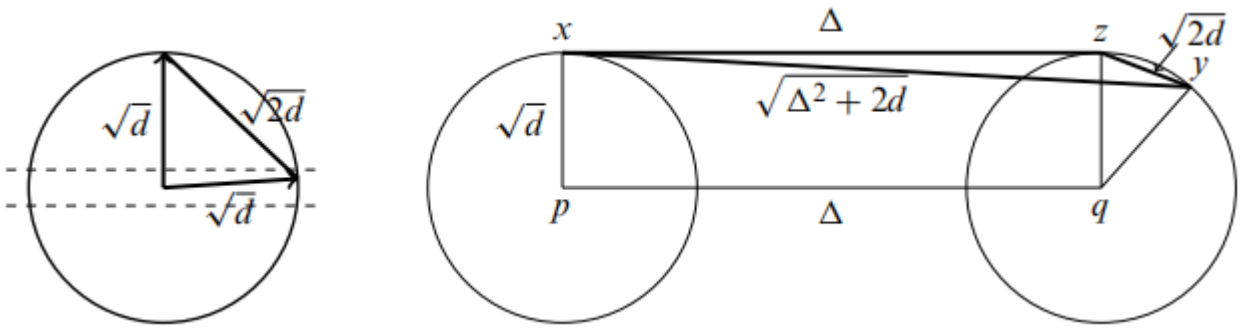


图 2: f2

实际上，可以分离中心距离更近的高斯分布。以后我们将使用奇异值分解来分离两个高斯分布的混合，即使它们的中心距离为 $O(1)$ 时也能成功分离。

给定一组 $d$-维空间中的样本点 $x_1, x_2, \ldots, x_n$，我们希望找到最能拟合这些点的球形高斯分布。设 $f$ 是未知的高斯分布，其均值为 $\mu$，每个方向上的方差为 $\sigma^2$。根据 $f$ 抽样得到这些点的概率密度为：

$$c \exp\left(-\frac{(x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_n-\mu)^2}{2\sigma^2}\right)$$

其中归一化常数 $c$ 是 $\left(\int e^{-\frac{|x-\mu|^2}{2\sigma^2}} dx\right)^n$ 的倒数。在从 $-\infty$ 到 $\infty$ 积分时，可以将原点移动到 $\mu$，因此 $c = \left(\int e^{-\frac{|x|^2}{2\sigma^2}} dx\right)^{-n} = \frac{1}{(2\pi)^{n/2}}$，并且它与 $\mu$ 无关。

最大似然估计（MLE）是指在给定样本 $x_1, x_2, \ldots, x_n$ 的情况下，最大化上述概率密度的 $f$。

对于一组 $n$ 个 $d$-维点 $\{x_1, x_2, \ldots, x_n\}$，$(x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_n-\mu)^2$ 在 $\mu$ 为这些点的质心时最小，即 $\mu = \frac{1}{n}(x_1+x_2+\cdots+x_n)$。

**证明：** 对 $(x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_n-\mu)^2$ 关于 $\mu$ 求梯度并令其为零，得到

$$-2(x_1-\mu)-2(x_2-\mu)-\cdots-2(x_n-\mu)=0.$$

解得 $\mu = \frac{1}{n}(x_1+x_2+\cdots+x_n)$。

为了确定 $\sigma^2$ 的最大似然估计，首先将 $\mu$ 设为真实的质心。接下来，往证 $\sigma$ 应该设置为样本的标准差。将 $\nu = \frac{1}{2\sigma^2}$ 和 $a = (x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_n-\mu)^2$ 代入选择这些点的概率公式中，得到

$$\frac{e^{-a\nu}}{\left(\int e^{-x^2\nu} dx\right)^n}$$

现在 $a$ 是固定的，而 $\nu$ 需要确定。取对数后，需要最大化的表达式是

$$-a\nu - n\ln\left[\int_x e^{-\nu x^2}dx\right].$$

为了找到最大值，对 $\nu$ 求导，令导数为零，并解出 $\sigma$。导数为

$$-a + n\frac{\int_x |x|^2 e^{-\nu x^2}dx}{\int_x e^{-\nu x^2}dx}.$$

令 $y = |\sqrt{\nu}x|$ 在导数中，得到

$$-a + \frac{n}{\nu}\frac{\int_y y^2 e^{-y^2}dy}{\int_y e^{-y^2}dy}.$$

由于两个积分的比值是 $d$-维球形高斯分布（标准差为 $\sqrt{\frac{1}{2}}$）到中心的期望平方距离，已知这个值为 $\frac{d}{2}$，因此我们得到

$$-a + \frac{nd}{2\nu}.$$

将 $\sigma^2$ 代入 $\frac{1}{2\nu}$，得到

$$-a + \frac{nd\sigma^2}{2}.$$

令 $-a + nd\sigma^2 = 0$，表明最大值出现在

$$\sigma = \sqrt{\frac{a}{nd}}.$$

注意，这个量是样本到其均值的平均坐标距离的平方根，即样本的标准差。因此，我们得到以下结论：

> 一组样本的最大似然球形高斯分布是均值等于样本均值, 标准差等于样本与真实均值的标准差的高斯分布。

假设 $x_1, x_2, \ldots, x_n$ 是由高斯分布生成的样本点。那么 $\mu = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$ 是分布期望值的无偏估计。然而，如果在估计方差时使用样本均值而不是真实期望值，则不会得到方差的无偏估计，因为样本均值不是独立于样本集的。应该使用

$$\tilde{\mu} = \frac{1}{n-1}(x_1 + x_2 + \cdots + x_n)$$

来估计方差。