信息科学中的数学

林晓疏

March 25, 2022

Contents

1 **高维空间** 4

1.1 大数定律 4

1.2 高维空间中的几何 5

1.2.1 赤道附近的体积 6

1.3 随机投影和 Johnson-Lindenstrauss 引理 7

2 **奇异值分解** 9

2.1 奇异值分解简介 9

2.2 奇异向量 10

2.3 奇异值分解 11

2.4 最佳 k 维近似 12

3 **机器学习** 14

3.1 误差估计 14

3.2 Online Learning 14

3.2.1 逻辑或的学习 15

3.2.2 折半算法 Halving Algorithm 15

3.2.3 感知机算法 15

3.2.4 12.10 Random Weighted Majority 15

3.2.5 Boosting 提升方法 16

4 **针对大量数据的算法：** Streaming, Sketching and Sampling 18

4.1 导引 18

4.2 数据流中不同元素的个数 18

4.3 Majority Vote 20

4.4 Sketching 20

4.4.1 查重 20

4.5 矩阵的 sketching 21

5 **随机图** 23

6 **附录一：Γ 函数以及部分性质** 25

7 **附录二：** Johnson-Lindenstrauss Lemma **的另一种证明** 27

前言

这本简短的笔记是笔者于 2020 年秋季学期在孔雨晴老师主讲的《信息科学中的数学》课程学习期间整理而成。选课同学来自各个年级。由于课程为英文授课，第一节又从不太直观的高维空间讲起，在课程初期可能会感到有些不适应。笔者当时恰好接触过相关内容，方而度过了课程的前几个课时。如今又到了这门课的开课学期，笔者重新翻阅了年初做的笔记，对一些错误之处进行了修改补充，并增添了两章附录。希望能对选课同学有所帮助。

限于时间和笔者水平，疏漏之处在所难免。

内容如有任何问题可联系笔者：wangyuanqing@pku.edu.cn

1 高维空间

1.1 大数定律

我们直觉上能够感觉到，对于某项数据，当样本数量足够大的时候，其平均值就会越来越接近这项数据的期望。在统计学上用**大数定理**来描述，即

$$\text{Prob}\left(\left|\frac{x_1+x_2+\cdots+x_n}{n}-E(x)\right|\geq\epsilon\right)\leq\frac{Var(x)}{n\epsilon^2} \tag{1}$$

其中 $E(x)$ 和 $Var(x)$ 分别代表了 x 的数学期望和方差。为证明这一定律，我们使用两个不等式作为引理。

引理 1.1 (Markov's inequality). *设 x 是一个非负随机变量。对于 $\forall a>0$，有*

$$P(x\geq a)\leq\frac{E(x)}{a} \tag{2}$$

证明 1.1. *根据随机变量期望的定义可得*

$$\begin{aligned}E(x)&=\int_0^{+\infty}xp(x)\mathrm{d}x\int_0^{a}xp(x)\mathrm{d}x+\int_a^{+\infty}xp(x)\mathrm{d}x\\&\geq\int_a^{+\infty}xp(x)\mathrm{d}x\geq a\int_a^{+\infty}p(x)\mathrm{d}x=aP(x\geq a)\end{aligned} \tag{3}$$

因此有 $P(x\geq a)\leq\frac{E(x)}{a}$。

推论 1. $P(x\geq bE(x))\leq\frac{1}{b}$

马尔可夫不等式仅仅使用了数据的期望就控制住了数据“尾部”的概率。下面的切比雪夫不等式利用数据的方差给出了更强的约束。

引理 1.2 (Chebyshev's inequality). *设 x 是一个非负随机变量。对于 $\forall c>0$，有*

$$P(|x-E(x)|\geq c)\leq\frac{Var(x)}{c^2} \tag{4}$$

证明 1.2. *我们知道 $P(|x-E(x)|\geq c)=P(|x-E(x)|^2\geq c^2)$。不妨记 $y=|x-E(x)|^2$，显然有 $E(y)=Var(x)$，运用公式 (2) 即得：*

$$P(|x-E(x)|\geq c)=P(|x-E(x)|^2\geq c^2)\leq\frac{E(y)}{c^2}=\frac{Var(x)}{c^2} \tag{5}$$

此外，还有一些计算公式：

$$\begin{aligned}E(x+y)&=E(x)+E(y)\\Var(x-c)&=Var(x)\\Var(cx)&=c^2Var(x)\end{aligned} \tag{6}$$

此外，如果 x,y 相互独立，那么有 $E(xy)=E(x)E(y)$。因此对于两个独立的随机变量，有

$$Var(x+y)=E(x+y)^2-E^2(x+y)=Var(x)+Var(y) \tag{7}$$

定理 1.3 (大数定理). *设 $x_1,x_2,\cdots,x_n$ 是随机变量 x 的 n 个独立样本。那么有*

$$\text{Prob}\left(\left|\frac{x_1+x_2+\cdots+x_n}{n}-E(x)\right|\geq\epsilon\right)\leq\frac{Var(x)}{n\epsilon^2} \tag{8}$$

Proof. 由于 $E(\frac{x_1+x_2+\cdots+x_n}{n}) = E(x)$，因此

$$\begin{aligned} P\left(\left|\frac{x_1+x_2+\cdots+x_n}{n} - E(x)\right| \geq \epsilon\right) &= P\left(\left|\frac{x_1+x_2+\cdots+x_n}{n} - E\left(\frac{x_1+x_2+\cdots+x_n}{n}\right)\right| \geq \epsilon\right) \\ &\leq \frac{Var\left(\frac{x_1+x_2+\cdots+x_n}{n}\right)}{\epsilon^2} \\ &= \frac{1}{n^2\epsilon^2} Var(x_1+x_2+\cdots+x_n) \\ &= \frac{1}{n^2\epsilon^2}\left(Var(x_1)+Var(x_2)+\cdots+Var(x_n)\right) \\ &= \frac{Var(x)}{n\epsilon^2} \end{aligned} \tag{9}$$

■

1.2 高维空间中的几何

现在我们来考虑 $\mathbb{R}^d$ 中的几何。高维空间中的几何与二维、三维有许多截然不同的性质。比如，设 A 是 $\mathbb{R}^d$ 中的一个单位球体，很容易得出 d 维空间中 $1-\epsilon \leq r \leq 1$ 描述的单位球体表面厚度为 ϵ 的球壳的体积是 $V_\epsilon = [1-(1-\epsilon)^d]\text{volume}(A)$，显然有

$$\lim_{d\to\infty} V_\epsilon = \text{volume}(A) \tag{10}$$

也就是说，高维球体的体积几乎全部集中在球的表面。由不等式 $1-tx \leq (1-x)^t \leq e^{-tx}$，可以得出

$$\frac{V_{1-\epsilon}}{V} = (1-\epsilon)^d \leq e^{-\epsilon d} \tag{11}$$

下文的符号规定：记 $V(d,r)$ 和 $A(d,r)$ 分别为 $\mathbb{R}^d$ 中半径为 r 的球体的“体积”和“表面积”，并在 $r=1$ 时简写为 $V(d)$ 和 $A(d)$. 从量纲中我们不难得出 $V(d,r) = V(d)r^d, A(d,r) = A(d)r^{d-1}$.

那么单位球的体积显然应该定义为

$$V(d) = \int_{x_1=-1}^{x_1=1}\int_{x_2=-\sqrt{1-x_1^2}}^{x_2=\sqrt{1-x_1^2}}\cdots\int_{-\sqrt{1-\sum_{i=1}^{d-1}x_i}}^{\sqrt{1-\sum_{i=1}^{d-1}x_i}} \mathrm{d}x_d\cdots\mathrm{d}x_2\mathrm{d}x_1 \tag{12}$$

在直角坐标系下积分形式较为复杂[1]。用类比的方法，求 $\mathbb{R}^3$ 中球体的体积，我们可以认为是用球面面积对半径进行积分，即

$$V(3,R) = \int_{r=0}^{r=R}(4\pi r^2)\mathrm{d}r = \frac{4\pi}{3}R^3 = \int_{r=0}^{r=R} A(3,r)\mathrm{d}r$$

类比到高维情形，则有

$$V(d,R) = \int_{r=0}^{r=R} A(d,R)\mathrm{d}r = A(d)\int_{r=0}^{r=R} r^{d-1}\mathrm{d}r = \frac{A(d)}{d}R^d \tag{13}$$

因此，要求出 d 维单位球体的体积，就要计算其“面积”。利用正态分布中的结果 $\int_{-\infty}^{+\infty} e^{-x^2}\mathrm{d}x = \sqrt{\pi}$，我们构造积分

[1]或许读者能够记得在数学分析第三册 P297 给出了从直角坐标到球坐标的积分过程。

$$I = \int \cdots \int_{\boldsymbol{x}\in\mathbb{R}^d} e^{-\sum x_i^2} \mathrm{d}x_1 \cdots \mathrm{d}x_d = (\sqrt{\pi})^d \tag{14}$$

另一方面，如果我们作球坐标变换，因为 $e^{-\sum x_i^2} = e^{-r^2}$，

$$I = \int_0^{+\infty} e^{-r^2} A(d,r)\mathrm{d}r = A(d) \int_0^{+\infty} e^{-r^2} r^{d-1} = \frac{1}{2} A(d)\Gamma(\frac{d}{2}) \tag{15}$$

这样就确定出[2]

$$\begin{aligned} A(d) &= \frac{2(\sqrt{\pi})^d}{\Gamma(\frac{d}{2})} \\ V(d) &= \frac{2(\sqrt{\pi})^d}{d\Gamma(\frac{d}{2})} \end{aligned} \tag{16}$$

1.2.1 赤道附近的体积

另一个非常有趣的结果是，高维球体几乎所有的体积都集中在“赤道”附近，换句话说，如果指定向量 $\boldsymbol{v}$ 的方向为“北方”，那么绝大多数单位向量 $\boldsymbol{u}$ 满足 $\boldsymbol{u}\cdot\boldsymbol{v} = \mathcal{O}(1/\sqrt{d})$。换句话说，绝大多数的单位向量满足 $|x_1| = \mathcal{O}(1/\sqrt{d})$。

定理 1.4. 对于 $\forall c \geq 1, d \geq 3$，满足 $|x_1| \leq \frac{c}{\sqrt{d-1}}$ 的点集的体积至少占据球体的 $1 - \frac{2}{c}e^{-\frac{c^2}{2}}$。

Proof. 根据对称性，我们只需要证明对于 $x_1 \geq 0$ 的上半球体，至多有 $\frac{2}{c}e^{-\frac{c^2}{2}}$ 的体积满足 $x_1 \geq \frac{c}{\sqrt{d-1}}$。我们将半球的上半部分记为 A，整个半球记为 H，也就是说我们需要证明

$$\frac{V(A)}{V(H)} \leq \frac{\text{upper bound of V(A)}}{\text{lower bound of V(A)}} = \frac{2}{c}e^{-\frac{c^2}{2}} \tag{17}$$

$V(A)$ 的精确表达式为

$$V(A) = \int_{\frac{c}{\sqrt{d-1}}}^1 V(d-1)(1-x_1^2)^{\frac{d-1}{2}} \mathrm{d}x_1 \tag{18}$$

利用不等式 $1 - x \leq e^x$，我们有

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

这个积分不容易计算，由于在 $x \geq \frac{c}{\sqrt{d-1}}$ 时有 $\frac{x\sqrt{d-1}}{c} \geq 1$，因此插入这一项，使得积分容易算出：

$$V(A) \leq V(d-1)\frac{\sqrt{d-1}}{c} \int_{\frac{c}{\sqrt{d-1}}}^{\infty} x_1 e^{-\frac{d-1}{2}x_1^2} \mathrm{d}x_1 = \frac{V(d-1)}{c\sqrt{d-1}} e^{-\frac{c^2}{2}} \tag{20}$$

虽然我们已经有了 H 的表达式，但是其表达式较为复杂，因此我们估计一个下界。为了能消去上面得出的 $\sqrt{d-1}$，我们选取上半球中高度为 $\frac{1}{\sqrt{d-1}}$ 的圆柱，从而有

$$V(H) \geq V(d-1)(1 - \frac{1}{d-1})^{\frac{d-1}{2}} \cdot \frac{1}{\sqrt{d-1}} \geq \frac{V(d-1)}{2\sqrt{d-1}} \tag{21}$$

[2]关于 Γ 函数的定义和部分性质可参见附录一

综上就可以得出

$$\frac{V(A)}{V(H)} \leq \frac{\frac{V(d-1)}{c\sqrt{d-1}}e^{-\frac{c^2}{2}}}{\frac{V(d-1)}{2\sqrt{d-1}}} = \frac{2}{c}e^{-\frac{c^2}{2}} \quad \blacksquare \tag{22}$$

■

从上面的分析中可以得知，任意选取两个点它们的内积大概率接近于 0，也就是任选两个向量，它们都非常接近垂直。下面的定理可以更精确地告诉我们对于 n 个点的情形。

定理 1.5. *对于在单位球面上 n 个随机选取的点 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$，有 $1 - \mathcal{O}(1/n)$ 的可能性*

1. *对于所有的 i 都有 $|\boldsymbol{x}_i| \geq 1 - \frac{2\log n}{d}$*
2. *对于所有的 $i \neq j$ 都有 $|\boldsymbol{x}_i \cdot \boldsymbol{x}_j| \leq \frac{\sqrt{6\log n}}{\sqrt{d-1}}$*

Proof. 第一条定理根据前面我们的结果，满足 $|x_i| \leq 1 - \epsilon$ 的点的比例小于 $e^{-\epsilon d}$。因此有

$$P(|\boldsymbol{x}_i| \leq 1 - \frac{2\log n}{d}) \leq e^{-\left(\frac{2\log n}{d}\right)d} = \frac{1}{n^2} \tag{23}$$

根据概率公式 $P(\bigvee_{i=1}^{n} x_i) \leq \sum_{i=1}^{n} P(x_i)$，我们有

$$P\left(\neg \bigvee_{i=1}^{n}\left(|\boldsymbol{x}_i| \geq 1 - \frac{2\log n}{d}\right)\right) \geq 1 - \frac{1}{n} \tag{24}$$

对于第二条定理，由于共有 $C_n^2 = \frac{n(n-1)}{2}$ 种组合，我们前面已经证明一个单位向量与另一个单位向量内积大于 $\frac{c}{\sqrt{d-1}}$ 的概率至多是 $\frac{2}{c}e^{-\frac{c^2}{2}}$，因此对于 $c = \sqrt{6\log n}$，其每一对的概率为 $\wr(e^{-\frac{6\log n}{2}}) = \mathcal{O}(\frac{1}{n^3})$，因此 C_n^2 对中内积大于给定值的概率为 $\mathcal{O}(1/n)$。 ■

1.3 随机投影和 Johnson-Lindenstrauss 引理

在涉及高维数据的处理时，我们经常需要找到一个高维点的最近邻点。由于我们需要处理 n 个 d 维数据，而 n, d 通常都非常大，因此难以做到直接在原数据上进行搜索。一般来说，我们希望能够在 $O(\log n), O(\log d)$ 的多项式时间内完成搜索，而在此之前的预处理可以花费 n, d 的多项式时间。通常的操作是将这些 d 维数据在尽量保证距离变化不大的情形下映射到维度较小的 k 维空间。使用 Gaussian Annulus Theorem(高斯环定理)，我们发现这样的映射确实存在，而且并不复杂。

定理 1.6 (Gaussian Annulus Theorem). *对于一个每个维度都是一个方差为 1 的 d 维球面高斯向量，对任意 $\beta \leq \sqrt{d}$，至多 $3e^{-c\beta^2}$ 的概率其位置位于环 $\sqrt{d} - \beta \leq |\mathbf{x}| \leq \sqrt{d} + \beta$ 之外，这里 c 是一个正常数.*

要使用的映射 $f : \mathbb{R}^d \to \mathbb{R}^k$ 定义如下：随机选取 $\mathbb{R}^d$ 中 k 个坐标服从方差为 1 的高斯分布的 Gaussian Vector $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_k$. 对于 $\forall \boldsymbol{v} \in \mathbb{R}^d$，定义

$$f(\boldsymbol{v}) = (\boldsymbol{u}_1 \cdot \boldsymbol{v},\ \boldsymbol{u}_2 \cdot \boldsymbol{v},\ \cdots,\ \boldsymbol{u}_k \cdot \boldsymbol{v})$$

我们将证明，会以高概率满足 $|f(\boldsymbol{v})| \approx \sqrt{k}\,|\boldsymbol{v}|$. 因为映射满足 $f(\boldsymbol{v}_2 - \boldsymbol{v}_1) = f(\boldsymbol{v}_2) - f(\boldsymbol{v}_1)$，故如果要估计 $|\boldsymbol{v}_2 - \boldsymbol{v}_1|$，只需要计算映射后的点在 $\mathbb{R}^k$ 中的距离即可，因为 $\sqrt{k}$ 是一个已知的常数因子。距离会增大的原因是我们的 $\boldsymbol{u}_i$ 并不是单位向量。另外注意 $\boldsymbol{u}_i$ 并不是相互正交的。如果我们要求正交性，$\boldsymbol{u}_i$ 之间将会失去独立性。

定理 1.7. 设 $\boldsymbol{v}$ 是 $\mathbb{R}^d$ 中的任意一个向量，f 的定义如上所述。则存在 $c > 0$ 使得 $\forall \varepsilon \in (0,1)$，有

$$\text{Prob}\left(\left||f(\boldsymbol{v})| - \sqrt{k}|\boldsymbol{v}|\right| \geq \varepsilon\sqrt{k}|\mathbf{v}|\right) \leq 3e^{-ck\varepsilon^2} \tag{25}$$

随机性来源于用于构造映射 f 的随机向量 $\boldsymbol{u}_i$ 的随机性。

Proof. 由于 $\boldsymbol{v}$ 可被单位化，不妨认为 $|\boldsymbol{v}| = 1$. 因此映射后的每个坐标 $\boldsymbol{u}_i \cdot \boldsymbol{v}$ 实际上是 d 个 Gaussian Vector 的线性组合，均值为 0，方差为

$$Var(\boldsymbol{u}_i \cdot \boldsymbol{v}) = \sum_{j=1}^{d} v_j^2 Var(u_{ij}) = \sum_{j=1}^{d} v_j^2 = 1$$

因此映射后的坐标就是在 $\mathbb{R}^k$ 中随机选取的单位 Gaussian 向量。再使用 Gaussian Annulus Theorem 即可证明。 ■

课本上对于 Gaussian Annulus Theorem 的证明使用了一些结论来 bound 随机变量的矩，对 Chernoff Bound 及其证明过程较为熟悉的读者可以参照附录二对于 Johnson-Lindenstrauss 引理的证明[3]。

[3]笔者按：附录二的过程是孔老师在随机算法课上教授的，她认为两个证明过程我们会更喜欢用 Gaussian Annulus Thm 的方法，但笔者觉得哪个方法都不是正常人能想出来的

2 奇异值分解

2.1 奇异值分解简介

对于矩阵 $\boldsymbol{A}_{n\times d}$,我们将其每一行视为一个 d 维的向量,奇异值分解的目标是寻找矩阵 $\boldsymbol{U}_{n\times r}, \boldsymbol{D}_{r\times r}, \boldsymbol{V}_{r\times d}$ 使得

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T \tag{26}$$

其中 $\boldsymbol{U}, \boldsymbol{V}$ 是正交矩阵，$\boldsymbol{D}$ 是对角元全为正数对角矩阵。其中 $\boldsymbol{V}$ 的列向量（也就是 $\boldsymbol{V}^T$ 的行向量）即为“最佳近似子空间”的标准正交基向量。而 $\boldsymbol{U}$ 的元素就是 $\boldsymbol{A}$ 在相应基向量的投影值。也就是

$$\boldsymbol{A} = \begin{bmatrix}\boldsymbol{u}_1 & \boldsymbol{u}_2 \cdots & \boldsymbol{u}_r\end{bmatrix} \operatorname{diag}\{\sigma_1, \sigma_2, \cdots, \sigma_r\} \begin{bmatrix}\boldsymbol{v}_1^T \\ \boldsymbol{v}_2^T \\ \vdots \\ \boldsymbol{v}_r^T\end{bmatrix} \tag{27}$$

将上述写法展开，就是

$$\boldsymbol{A} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T \tag{28}$$

其中的 $\boldsymbol{u}_i$ 是 $n\times 1$ 的向量，$\boldsymbol{v}_i$ 是 $d\times 1$ 的向量。

就像将 $\boldsymbol{x}$ 投影到一组基向量 $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_d$ 中一样，对于 SVD，前 k 个向量保证 $\boldsymbol{A}$ 的行向量在 k 维子空间中的投影平方和最大（距离平方和最小）。

我们已经了解过特征值分解。对于方阵 $\boldsymbol{A}$，满足 $\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$ 的向量 $\boldsymbol{x}$ 称为特征向量，对应的 λ 称为特征值。如果 $\boldsymbol{A}$ 实对称，那么 $\boldsymbol{A}$ 必然有 n 个特征值，从而 $\boldsymbol{A}$ 就可以分解为

$$\boldsymbol{A} = \boldsymbol{P}^{\boldsymbol{T}}\boldsymbol{B}\boldsymbol{P} \tag{29}$$

其中 $\boldsymbol{P}$ 是正交矩阵，$\boldsymbol{B} = \operatorname{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$，$\boldsymbol{P}$ 的行向量维对应的特征向量。

特征值分解对于 $\boldsymbol{A}$ 有着要求。首先必须是方阵，其次必须有 n 个特征向量。但是 SVD 对于矩阵没有要求，我们总可以找到这样的两个由单位正交向量组成的矩阵 $\boldsymbol{U}, \boldsymbol{V}$ 和对角矩阵 $\boldsymbol{D}$。其中 $\boldsymbol{V}$ 的列向量称为右奇异向量，$\boldsymbol{U}$ 的列向量称为左奇异向量。如果 $\boldsymbol{A}$ 是可逆的，那么

$$\boldsymbol{A}^{-1} = \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^T \tag{30}$$

我们接下来将会看到，有下面的关系式成立：

$$\boldsymbol{A}\boldsymbol{v}_i = \sigma_i \boldsymbol{u}_i \quad \boldsymbol{A}^T\boldsymbol{u}_i = \sigma_i \boldsymbol{v}_i \tag{31}$$

这一点不难验证，将公式 27 右乘 $\begin{bmatrix}\boldsymbol{v}_1 & \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_r\end{bmatrix}$ 可知

$$\boldsymbol{A}\begin{bmatrix}\boldsymbol{v}_1 & \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_r\end{bmatrix} = \begin{bmatrix}\sigma_1\boldsymbol{u}_1 & \sigma_2\boldsymbol{u}_2 & \cdots & \sigma_r\boldsymbol{u}_r\end{bmatrix} \tag{32}$$

左乘 $\begin{bmatrix}\boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_r\end{bmatrix}^T$ 可得

$$\boldsymbol{A}^T\begin{bmatrix}\boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_r\end{bmatrix} = \begin{bmatrix}\sigma_1\boldsymbol{v}_1 & \sigma_2\boldsymbol{v}_2 & \cdots & \sigma_r\boldsymbol{v}_r\end{bmatrix} \tag{33}$$

换言之，$\boldsymbol{A}$ 作用于 $\boldsymbol{v}_i$ 会将其变换为 $\boldsymbol{u}_i$ 的 σ_i 倍，$\boldsymbol{A}^T$ 作用于 $\boldsymbol{u}_i$ 会将其变换为 $\boldsymbol{v}_i$ 的 σ_i 倍。注意到 $\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{v}_i = d_i^2\boldsymbol{v}_i$，也就是说第 i 个奇异向量 $\boldsymbol{v}_i$ 是方阵 $\boldsymbol{A}^T\boldsymbol{A}$ 的第 i 个特征向量。

我们选取奇异向量的标准是：其生成的子空间能够让矩阵 $\boldsymbol{A}$ 在其上的投影平方和最大。我们不妨考虑一维的情形。对于一个向量 $\boldsymbol{a}_i$ 和一条过原点的单位方向向量为 $\boldsymbol{v}$ 直线，用 dist i 和 proj i 分别代表向量 $\boldsymbol{a}_i$ 到直线的距离和投影，我们有

$$(\text{dist } i)^2 + (\text{proj } i)^2 = \|\boldsymbol{a}_1\| \tag{34}$$

那么最小化距离平方和实际上就是最大化投影平方和。前者的表述类似于最小二乘法，后者代表在同维度的子空间中最大限度地保留了原数据的信息。可能你会觉得选取投影平方和作为数据留存度的指标略显武断，但是我们随后就会看到，“平方”拥有非常良好的性质，式 34 就是其中之一。

2.2 奇异向量

将 $n \times d$ 的矩阵 $\boldsymbol{A}$ 的行向量视为 d 维空间的点，考虑过原点的直线，其单位向量为 $\boldsymbol{v}$，对于 $\forall \boldsymbol{a} \in \mathbb{R}^d$，$\boldsymbol{v} \cdot \boldsymbol{a}$ 就是 $\boldsymbol{a}$ 在 $\boldsymbol{v}$ 上的投影，因此 $\|\boldsymbol{A}\boldsymbol{v}\|^2$ 即为 $\boldsymbol{A}$ 中 n 个行向量在 $\boldsymbol{v}$ 上的投影平方和。我们的目的是使得 $\|\boldsymbol{A}\boldsymbol{v}\|$ 最小，也就是找到

$$\boldsymbol{v}_1 = \arg\max_{|v|=1} \|\boldsymbol{A}\boldsymbol{v}\| \tag{35}$$

当然 $\boldsymbol{v}_1$ 可能不止一个（或者说必然不止一个），如果 $\boldsymbol{v}_1$ 是奇异向量，那么 $-\boldsymbol{v}_1$ 也是。随便选择一个即可，后面我们都这样处理。$\boldsymbol{v}_1$ 称为第一个奇异向量。而 $\sigma_1(\boldsymbol{A}) = \|\boldsymbol{A}\boldsymbol{v}_1\|$ 称为 $\boldsymbol{A}$ 的第一奇异值。因此有

$$\sigma_1^2(\boldsymbol{A}) = \sum_{i=1}^{n} (\boldsymbol{a}_i \cdot \boldsymbol{v}_1)^2 \tag{36}$$

即为 $\boldsymbol{A}$ 的所有行向量在 $\boldsymbol{v}_1$ 上的投影平方和。

当然，数据未必会集中在一条线附近，可能是一个平面，换句话说，如果我们已经有了一个寻找 $\boldsymbol{v}_1$ 的算法，怎么去寻找到更高维的空间？

一种贪心的做法是，在 $\boldsymbol{v}_1$ 的正交补空间用同样的方法寻找第二个奇异向量和相应的奇异值。也就是

$$\boldsymbol{v}_2 = \arg\max_{\substack{|v|=1 \\ \boldsymbol{v}\cdot\boldsymbol{v}_1=0}} \|\boldsymbol{A}\boldsymbol{v}\| \tag{37}$$

同样的方法我们可以定义下面的奇异向量，直到

$$\max_{\substack{v \perp \boldsymbol{v}_1, \cdots, \boldsymbol{v}_r \\ |v|=1}} \|\boldsymbol{A}\boldsymbol{v}\| = 0 \tag{38}$$

此时 $r = \text{rank}(\boldsymbol{A})$。

如何保证这种贪心法的正确性呢？我们给出下面的定理。

定理 2.1. *设 $\boldsymbol{A}$ 是一个 $n \times d$ 的矩阵，其奇异向量为 $\boldsymbol{v}_1, \boldsymbol{v}_2 \cdots, \boldsymbol{v}_r$。对于 $1 \le k \le r$，令 $V_k = L(\boldsymbol{v}_1, \boldsymbol{v}_2 \cdots, \boldsymbol{v}_k)$，那么 V_k 就是 $\boldsymbol{A}$ 的 k 维最佳近似子空间。*

Proof. 当 $k = 1$ 时显然成立。对于 $k = 2$ 时，假设 $W = L(\boldsymbol{w}_1, \boldsymbol{w}_2)$ 是一个二维的子空间，我们选取 $\boldsymbol{w}_2 \perp \boldsymbol{v}_1$。方法是：如果 $\boldsymbol{v}_1 \perp W$，则任选一个即可，否则，选取 W 中垂直于 $\boldsymbol{v}_1$ 投影的向量即可。这样，由 $\boldsymbol{v}_1, \boldsymbol{v}_2$ 的定义可知 $\|\boldsymbol{A}\boldsymbol{w}_1\|^2 \le \|\boldsymbol{A}\boldsymbol{v}_1\|^2$，$\|\boldsymbol{A}\boldsymbol{w}_2\|^2 \le \|\boldsymbol{A}\boldsymbol{v}_2\|^2$，从而有

$$\|A\boldsymbol{w_1}\|^2 + \|A\boldsymbol{w_2}\|^2 \leq \|A\boldsymbol{v_1}\|^2 + \|A\boldsymbol{v_1}\|^2 \tag{39}$$

对于 k 维情形也可以用类似方法证明。 ■

我们注意到 n 维向量 $A\boldsymbol{v_i}$ 是一张记录了 $\boldsymbol{A}$ 的 n 个行向量在 $\boldsymbol{v}_i$ 的投影（带符号）的表格。假如我们认为 $\sigma_1(\boldsymbol{A}) = \|A\boldsymbol{v_1}\|$ 就是矩阵 $\boldsymbol{A}$ 在 $\boldsymbol{v_1}$ 方向上的一部分。要让这个认识符合我们的一般认识，那么所有这样部分的平方和应该等于 “$\boldsymbol{A}$ 的全部”。设 $\boldsymbol{a}_j$ 是 $\boldsymbol{A}$ 的第 j 个行向量，那么有

$$\sum_{j=1}^{n} \left|\boldsymbol{a}_j\right|^2 = \sum_{j=1}^{n}\sum_{i=1}^{r}(\boldsymbol{a}_j \cdot \boldsymbol{v}_i)^2 = \sum_{i=1}^{r}\sum_{j=1}^{n}(\boldsymbol{a}_j \cdot \boldsymbol{v}_i)^2 = \sum_{i=1}^{r}\sigma_i^2(\boldsymbol{A}) \tag{40}$$

而所有行向量模方之和实际就是 $\boldsymbol{A}$ 中所有元素的平方和。这就是我们前面提到的 “$\boldsymbol{A}$ 的全部”，称为 Frobenius 范数，即

$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{j,k} a_{jk}^2} \tag{41}$$

我们前面已经证明了

$$\sum_{i=1}^{r}\sigma_i^2(\boldsymbol{A}) = \|\boldsymbol{A}\|_F \tag{42}$$

向量 $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_r$ 被称为右奇异向量。用 $\boldsymbol{A}$ 与其作用后正交化，即

$$\boldsymbol{u}_i = \frac{1}{\sigma_i(\boldsymbol{A})} A\boldsymbol{v_i} \tag{43}$$

事实上,$\boldsymbol{u}_i$ 就是最大化 $\left\|\boldsymbol{u}^T\boldsymbol{A}\right\|$ 的向量，且彼此正交，称为左奇异向量。

2.3 奇异值分解

设 $\boldsymbol{A}$ 是一个 $n \times d$ 的矩阵，其 r 个奇异向量为 $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_n$，对应的奇异值为 $\sigma_1, \sigma_2, \cdots, \sigma_r$. 我们定义其左奇异向量为 $\boldsymbol{u}_i = \frac{1}{\sigma_i} A\boldsymbol{v}_i$。那么 $\sigma_i \boldsymbol{u}_i \boldsymbol{v}^T$ 是一个秩为 1 的矩阵，其行向量就是矩阵 $\boldsymbol{A}$ 的行向量 “在 $\boldsymbol{v}_i$ 方向的部分”，也就是 $\boldsymbol{A}$ 的行向量在 $\boldsymbol{v}_i$ 方向的分量的向量坐标表示。我们将要证明 $\boldsymbol{A}$ 可以表示为一系列秩为 1 的矩阵之和，形如

$$\boldsymbol{A} = \sum_{i=1}^{r}\sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T \tag{44}$$

从几何上来说，每个 $\boldsymbol{A}$ 中行向量代表的点被分解为它在 $\boldsymbol{v}_i$ 方向的分量之和。我们也会从代数角度证明这个结论。我们先从下面的引理出发。

引理 2.2. *如果 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 对于任意向量 $\boldsymbol{v}$ 都有 $\boldsymbol{Av} = \boldsymbol{Bv}$，那么 $\boldsymbol{A} = \boldsymbol{B}$.*

证明 2.1. *选取标准正交基即可。*

定理 2.3. *设 $\boldsymbol{A}$ 是一个 $n \times d$ 的矩阵，其 r 个右奇异向量为 $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_n$，对应的奇异值为 $\sigma_1, \sigma_2, \cdots, \sigma_r$ 其左奇异向量为 $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_r$。那么*

$$\boldsymbol{A} = \sum_{i=1}^{r}\sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T \tag{45}$$

Proof. 对于任意一个 $\boldsymbol{v}_j$，我们分别用 $\boldsymbol{A}$ 和 $\sum_{i=1}^{r}\sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^T$ 左乘，有

$$\sum_{i=1}^{r}\sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^T\boldsymbol{v}_j=\sigma_j\boldsymbol{u}_j=\boldsymbol{A}\boldsymbol{v}_j \tag{46}$$

由于任意一个向量 $\boldsymbol{v}$ 都可以分解为一个与所有右奇异向量都垂直的向量和右奇异向量的线性组合，所以对于任意 $\boldsymbol{v}$ 都有 $\sum_{i=1}^{r}\sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^T\boldsymbol{v}=\sigma_j\boldsymbol{u}_j=\boldsymbol{A}\boldsymbol{v}$，由上面的引理即可得出结论 ■

这样，我们就完成了对矩阵 $\boldsymbol{A}$ 的奇异值分解。如果奇异值互异，只需要规定其序关系即得到唯一的分解。如果有相同的奇异值，那么对应的奇异向量会张成子空间，此子空间的任意一组正交基都可以作为奇异向量。

2.4 最佳 k 维近似

本节的标题是英语奇差的笔者直译的结果。

设 $\boldsymbol{A}$ 是一个 $n\times d$ 的矩阵，其奇异值分解为

$$\boldsymbol{A}=\sum_{i=1}^{r}\sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^T \tag{47}$$

对于任意 $1\le k\le r$，令

$$\boldsymbol{A}_k=\sum_{i=1}^{k}\sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^T \tag{48}$$

显然 $\mathrm{rank}(\boldsymbol{A}_k)=k$。我们随后会看到，$\boldsymbol{A}_k$ 是对 $\boldsymbol{A}$ 的最佳的 k 维近似，且其误差可用 F-范数估计。换言之，$\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_k$ 张成的 k 维空间是所有 k 维子空间中，使 $\boldsymbol{A}$ 的行向量距离平方和最小的那一个。为了证明这个结论，我们先给出下面的引理：

引理 2.4. *$\boldsymbol{A}_k$ 的行向量是对应的 $\boldsymbol{A}$ 的行向量在 V_k 的投影。*

证明 2.2. *设 a 是任意一个 $\boldsymbol{A}$ 的行向量，由于 $\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_k$ 彼此正交，因此它在 V_k 的投影就是 $\sum_{i=1}^{k}(\boldsymbol{a}\cdot\boldsymbol{v}_i)\boldsymbol{v}_i^T$。因此整个矩阵 $\boldsymbol{A}$ 的投影就是 $\sum_{i=1}^{k}\boldsymbol{A}\boldsymbol{v}_i\boldsymbol{v}_i^T$. 于是有*

$$\sum_{i=1}^{k}\boldsymbol{A}\boldsymbol{v}_i\boldsymbol{v}_i^T=\sum_{i=1}^{k}\sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^T=\boldsymbol{A}_k \tag{49}$$

定理 2.5. *对于任意秩不大于 k 的矩阵 $\boldsymbol{B}$，有*

$$\|\boldsymbol{A}-\boldsymbol{A}_k\|\le\|\boldsymbol{A}-\boldsymbol{B}\| \tag{50}$$

Proof. 设 $\boldsymbol{B}$ 是所有秩不大于 k 的矩阵中使得 $\|\boldsymbol{A}-\boldsymbol{B}\|^2$ 最小的那一个。设 V 是 $\boldsymbol{B}$ 的行向量张成的子空间。那么 $\dim V\le k$。由于 $\boldsymbol{B}$ 是使得 $\|\boldsymbol{A}-\boldsymbol{B}\|$ 最小的矩阵，因此，$\boldsymbol{B}$ 的行向量就是 $\boldsymbol{A}$ 相应行向量在 V 的投影。否则，用对应的投影代替 $\boldsymbol{B}$ 的行，这仍然保证了 $\boldsymbol{B}$ 的行向量包含在 V 中，但是这使得 $\|\boldsymbol{A}-\boldsymbol{B}\|$ 减小了，与 $\boldsymbol{B}$ 的定义不符。

现在由于 $\boldsymbol{B}$ 的行向量就是 $\boldsymbol{A}$ 相应行向量在 V 的投影，而能使得距离平方和最小的子空间的投影构成的矩阵就是 $\boldsymbol{A}_k$，因此有 $\|\boldsymbol{A}-\boldsymbol{A}_k\|\le\|\boldsymbol{A}-\boldsymbol{B}\|$.■

■

如果我们要对若干数据 $\boldsymbol{x}$ 计算 $\boldsymbol{A}\boldsymbol{x}$，那么每次计算要进行 nd 次乘法，加上加法的时间，时间复杂度为 $\mathcal{O}(nd)$。但是如果我们用 $\boldsymbol{A}_k$ 来近似代替 $\boldsymbol{A}$，那么计算 $\boldsymbol{A}_k\boldsymbol{x}$ 时，只需计算 $k(n+d)$ 次乘法，时间复杂度为 $\mathcal{O}(k(n+d))$。这在 $k \ll \min\{n,d\}$ 的时候能够显著提高计算效率。那么其误差是多少呢？由于 $\boldsymbol{x}$ 未知，我们应该给出一个对所有 $\boldsymbol{x}$ 都符合的估计，所以我们选取最大的 $|(\boldsymbol{A}_k - \boldsymbol{A})\boldsymbol{x}|$。当然，在 $|\boldsymbol{x}|$ 没有约束的话，最大值也没有界限。因此我们限定 $|\boldsymbol{x}| \leq 1$。这样，我们就定义了 $\boldsymbol{A}$ 的一种新的范数，即

$$\|\boldsymbol{A}\|_2 = \max_{|\boldsymbol{x}|\leq 1} |\boldsymbol{A}\boldsymbol{x}| \tag{51}$$

即 2-范数。注意到它就是 $\sigma_1(\boldsymbol{A})$。[4]

[4]本节的内容还有部分遗漏，比如为何左奇异向量相互正交，以及幂法 (Power Method) 求特征值等，这些在课本上都有较为完整的证明。

3 机器学习

3.1 误差估计

如果 $\mathcal{H}$ 是一个规则集，$\varepsilon, \delta > 0$ 如果一个大小为

$$n \geq \frac{1}{\varepsilon}\left(\log|\mathcal{H}| + \log\left(\frac{1}{\delta}\right)\right)$$

的训练集是依据分布 D 挑选的，那么有大于 $1-\delta$ 的概率，任何 $h \in \mathcal{H}$ 且真实错误 $err_D(h) > \varepsilon$ 都会使得训练错误 $err_S(h) > 0$. 换句话说，有大于 $1-\delta$ 的概率，任何 $h \in \mathcal{H}$ 且训练错误 $err_S(h) = 0$ 都会使真实错误 $err_D(h) < \varepsilon$

证明过程如下：设 $h_1, h_2, \cdots$ 是 $\mathcal{H}$ 中真实错误率大于等于 ε 的概念，那么对于一个大小为 n 的训练集 S，设 A_i 是事件 “h_i 的训练集错误率为 0”，那么

$$P(A_i) \leq (1-\varepsilon)^n$$

那么根据联合概率的不等式，至少一个发生的概率

$$P\left(\bigcup_i A_i\right) \leq |\mathcal{H}|(1-\varepsilon)^n \leq |\mathcal{H}|e^{-n\varepsilon} \leq |\mathcal{H}|e^{-\log|\mathcal{H}|-\log(1/\delta)} = \delta$$

上述估计是建立在概念集中有训练错误率为 0 的概念的基础上的。如果表现最好的概念的错误率也只有 5% 呢？我们能否在足量的样本下，使得真实错误率也接近训练错误率呢？答案是可以的，要证明这个结论，首先需要一个二项分布的结论：

若随机变量 $X \sim B(n,p)$，那么对于任意 $\alpha \in [0,1]$ 有

$$P(X/n > p + \alpha) \leq e^{-2n\alpha^2}$$
$$P(X/n < p + \alpha) \leq e^{-2n\alpha^2}$$

利用上述结论很容易证明，如果

$$n \geq \frac{1}{2\varepsilon^2}(\log|\mathcal{H}| + \log(2/\delta))$$

就能保证以大于等于 $1-\delta$ 的概率，所有 $h \in \mathcal{H}$ 都满足 $|err_S(h) - err_D(h)| \leq \varepsilon$.

3.2 Online Learning

到目前为止我们考虑的都是批处理学习（batch learning）。也就是说，给定一批（batch）数据，以及一个训练样本 S，你的目标是通过这个训练集，生成一个概念 h 使它在新的数据上也极少犯错。

我们现在将目光转向更困难的 Online Learning 的情形，这里我们不能假定数据是根据某一概率分布或是某一概率性的过程来选定的。

OL 的过程是这样的：在每个时刻 $t = 1, 2, \cdots$，有两件事情发生：

首先，算法会获得一个随机的样例 $x_t \in \mathcal{X}$，并被要求给出一个预测 l_t 作为它的标签。

然后，算法将被告知这个样例的真实标签 $c^*(x_t)$，如果 $c^*(x_t) \neq l_t$，那么算法将会获得一个 debuff。

3.2.1 **逻辑或的学习**

这个学习算法的目标是尽可能少犯错误。我们举一个学习“逻辑或”的学习算法的例子。比如在垃圾邮件的判定中，假定“重要邮件”是指满足一系列“重要”要求中的至少一个的邮件，比如一共有 d 个条件，我们想知道这些条件中哪些是“重要”的。每个到来的邮件将被标记为 $\{0,1\}^d$ 中的一个标记。从 $h_0 = x_1 \vee x_2 \vee \cdots \vee x_d$ 开始，每次如果犯错，就将犯错标签从“重要标签”中去掉，犯错至多不超过 d 次。

实际上，d 次同时也是一个下界，我们可以证明，没有任何一种算法能够保证对任何一种样例输入的序列，犯错次数都小于 d。只需要每次置第 i 个特征为 1，然后总是告诉算法预测是错误的即可。

3.2.2 **折半算法** Halving Algorithm

如果我们不关心算法的运行时间，那么又一种简单的 OL 能够保证犯错次数不超过 $\log_2(|\mathcal{H}|)$。对每个给定样例，我们都询问所有 $\mathcal{H}$ 中的规则，并将更多人的选择作为输出，这样如果犯错一次，我们就能至少减少一半的规则。

当然这个算法在 $\mathcal{H}$ 中没有完美规则的时候无法给出出错率最低的那个，它可能一开始就把它剔除了。

3.2.3 **感知机算法**

这一部分因为 12 月 3 日的课程回放只有一节，所以不知道讲了多少

3.2.4 12.10 Random Weighted Majority

如果我们有一个大小为 n 的 $\mathcal{H}$，并且确定有一个完美规则，那么我们有 $\log_2(n)$ 的折半算法。

如果并不存在完美规则呢？如何找到出错最少的规则（optimal expert）？经过 T 轮之后，我们选择出错最少的专家。

我们将会给每个专家一个初始权重为 1，每次有一个样例输入，我们将会按照权重选择一个专家的回答作为这一轮的回答（但此时我们仍然知道其他专家的回答）。当我们知道正确答案之后，我们将把所有回答错误的专家的权重乘以 $1-\varepsilon$.

由于我们是按照权重选择的，如果令 $\sum_{wrong} w_i^t$ 代表所有在第 t 轮中犯错的的专家的权重的和，令 $\sum_{right} w_i^t$ 代表所有在第 t 轮中犯错的的专家的权重的和，$W_t = \sum w_i^t$ 代表这一轮所有专家的权重和，那么我们在这一轮选择错误的概率是

$$F_t = \frac{\sum_{wrong} w_i^t}{\sum w_i^t}$$

t 轮过后，出错的次数的期望是

$$E(wrong\ times) = \sum_{i=1}^{t} F_i$$

每次更新的时候，我们将对犯错的权重更新，那么

$$W_{t+1} = \left(\sum_{right} w_i^t + (1-\varepsilon) \sum_{wrong} w_i^t \right) = W_t(1 - F_t\varepsilon)$$

那么经过 T 轮之后，假设最优的专家只犯错了 N_{opt} 次，我们可以知道

$$W_T \geq (1-\varepsilon)^{N_{opt}}$$

（为什么采取如此粗略的估计，大概其他的如果次次犯错，权重下降得更快?）

初始权重为 n，那么我们就有

$$W_T = n\prod_{i=1}^{T}(1 - F_i\varepsilon) \geq (1-\varepsilon)^{N_{opt}}$$

两侧取对数，并利用 $\log(1-x) \leq -x$，因此有

$$E(Total\ num\ of\ mistakes) = \sum_{i=1}^{T} F_i \leq \frac{1}{\varepsilon}\left(\log n - N_{opt}\log(1-\varepsilon)\right)$$

Batch Learning 比 Online Learning 更简单。

3.2.5 Boosting **提升方法**

提升方法是一种可以用来减小监督式学习中偏差的机器学习算法。面对的问题是迈可 · 肯斯（Michael Kearns）提出的：一组“弱学习者”的集合能否生成一个“强学习者”? 弱学习者一般是指一个分类器，它的结果只比随机分类好一点点；强学习者指分类器的结果非常接近真值。

也就是说，弱学习者对于任何分布的样本，其给出的专家的错误率

$$P(wrong) \leq \frac{1}{2} - \gamma$$

其中 $0 < \gamma \leq \frac{1}{2}$. 一个强学习者则可以在足量样本（比如说，是 $\frac{1}{\varepsilon}$ 的多项式）下以极大概率到达任意小的错误率。以下我们将说明一个事实，即通过提升方法，一个在任何样本分布下都能工作的弱学习者可以提升为一个强学习者。简单地说，我们会给训练集的样本加权，每轮运行结束后我们将提升回答错误的样本的权重（乘以 $\alpha > 1$），注意与 RWM 的区别是，这里是提升错误样本的权重，而非降低错误专家的权重。在足够多的轮数之后，我们将所有得到的专家 $h_1, h_2, \cdots, h_t$ 投票决定作为最终的专家。

假如在第 t 轮运行结束之后，样本的总权值为 W_t，那么在第 $t+1$ 轮中，回答错误的样本所占的权重小于等于 $\frac{1}{2} - \gamma$，从而更新后就有

$$W_{t+1} \leq W_t\left(\left(\frac{1}{2} - \gamma\right)\alpha + \frac{1}{2} + \gamma\right)$$

考虑 t 轮运行之后，共得到 $h_1, h_2, \cdots, h_t$ 这 t 个专家。设 m 是使得这 t 个专家投票之后仍然判断错误的样本个数，那么可以知道这 m 个样本至少令 $\frac{t}{2}$ 个专家判断错误，取定 $\alpha = \frac{\frac{1}{2}+\gamma}{\frac{1}{2}-\gamma}$，那么对于 t 轮运行后的总权值，我们有如下估计：

$$m\alpha^{\frac{t}{2}} \leq W_t \leq n(1+2\gamma)^t$$

解出

$$m \leq (1-4\gamma^2)^{\frac{t}{2}} \leq ne^{-2t\gamma^2}$$

上式用到了 $1-x \leq e^{-x}$. 当 $t > \frac{\log n}{2\gamma^2}$ 时，$m < 1$.

本章的一个重要概念——VC dimension 因时间原因未能整理.

4 针对大量数据的算法：Streaming, Sketching and Sampling

4.1 导引

如果你有 n 个正数 $a_1, a_2, \cdots, a_n$，并且都不超过 m，如果让你以其大小为权重，随机选择一个数字作为输出，你会如何操作？直观的方法是，将这 n 个数字都存储起来，这样你需要 $O(n \log m)$ 大小的空间。另一种方法是，存储两个值，sum 代表当前输入的所有数字之和，num 代表选择结果。初始时，sum = num = a_1. 当输入 a_j 时，更新 sum←sum + a_j，并以 $\frac{a_j}{sum}$ 的概率将选择结果变为 a_j。容易证明算法结束后，a_j 被选中的概率是

$$P(a_j) = \frac{a_j}{\sum_{i=1}^{n} a_i}$$

但是这个算法只需要 $O(\log m + \log n)$ 的空间。

4.2 数据流中不同元素的个数

对于大量数据的流,统计其中不同元素的个数是很有意义的。考虑 n 个整数组成的序列 $a_1, a_2, \cdots, a_n$，$1 \le a_n \le m$，并且 n, m 都非常大。如何统计其中不同元素的个数？你可以用桶的方法在 $O(m)$ 的空间下完成这一目标，也可以用全部存储的方法在 $O(n \log m)$ 的空间下完成。但是无论如何，只要你想要精确求解其中的相异元素个数，你所需要的空间至少是 $O(m)$（证明见 188 页）。但是我们可以用随机和近似的方法，在更小的空间代价之下获得一个较好的近似答案。

要引出我们的估计方法，我们不妨考虑在 $[0,1]$ 内独立随机挑选 s 个实数。我们考虑这 s 个实数最小值的数学期望。由几何概型，最小值的概率分布函数为

$$F(x) = P(min \le x) = 1 - (1-x)^s$$

求导得到其概率密度为 $f(x) = s(1-x)^{s-1}$. 因此其数学期望

$$E(min) = \int_0^1 x f(x) \mathrm{d}x = \frac{1}{s+1}$$

类似的思想，在 n, m 非常大的时候，在 $\{1, 2, 3, \cdots, m\}$ 中独立随机挑选一些数字构成集合 S，作出估计 $min \approx \frac{m}{|S|+1}$，所以 $|S| = \frac{m}{min} - 1$。

当然这样的估计是建立在数据相互独立的情况下。如果数据不相互独立，比如说，就挑选了 $\{1, 2, 3, \cdots, m\}$ 中最小的 $|S|$ 个数，那么算法会给出非常糟糕的结果。将不相互独立的数据进行完全的随机映射需要大量的空间，但是我们将会看到，通过 2-universal 哈希函数可以使得映射后的数据两两独立，这足以让我们完成精确度估计的关键步骤，而且只需要花费 $O(\log m)$ 的空间。

定义一个哈希函数集

$$H = \{h | h : \{1, 2, 3, \cdots, m\} \to \{0, 1, 2, 3, \cdots, M-1\}\}$$

我们说 H 是 2-universal 的或者两两独立（pairwise independent）的，是指如果 $\forall x, y \in \{1, 2, 3, \cdots, m\}$ 并且 $x \ne y$，$h(x)$ 和 $h(y)$ 能够等概率地取到 $\{0, 1, 2, 3, \cdots, M-1\}$ 中的值并且两两独立。换句话说，对 $\forall w, z \in \{0, 1, 2, 3, \cdots, M-1\}$，有

$$P\left(h(x) = w \text{ and } h(y) = z\right) = \frac{1}{M^2}$$

我们给出一个 2-universal 的例子。令 M 是一个大于 m 的素数，对任意整数对 $(a,b) \in [0, M-1]^2$，定义函数

$$h_{ab}(x) = ax + b \mod M$$

存储这个函数只需要记录两个数字即可，空间为 $O(\log M)$。那么 $h(x) = w$ and $h(y) = z$ 当且仅当

$$\begin{pmatrix} x & 1 \\ y & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} w \\ z \end{pmatrix} \mod M$$

当 $x \neq y$ 时，有

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} x & 1 \\ y & 1 \end{pmatrix}^{-1} \begin{pmatrix} w \\ z \end{pmatrix} \mod M$$

因此 (w,z) 和 (a,b) 一一对应，也就是

$$P\left(h(x) = w \text{ and } h(y) = z\right) = \frac{1}{M^2}$$

下面我们证明，有极大概率（大于 2/3）使得

$$\frac{m}{6s} \leq min \leq \frac{6m}{s}$$

成立。

首先我们估计 $P(min < \frac{m}{6s})$。这部分不需要两两独立，只需要联合概率公式就有

$$P(min < \frac{m}{6s}) \leq s \times \frac{m}{6s} \times \frac{1}{m} = \frac{1}{6}$$

另一个方向，估计 $P(min > \frac{6m}{s})$. 此时我们不能用独立事件相乘的公式，因为我们无法保证全部独立，只能通过 Hash 方法保证两两独立。我们定义随机变量

$$I_i = \begin{cases} 0, & h(a_i) \geq \frac{6m}{s} \\ 1, & else \end{cases}$$

令 $Y = \sum I_i$，那么 $Y = 0$ 就代表所有数字都大于 $\frac{6m}{s}$，我们希望 $P(Y = 0)$ 越小越好。根据切比雪夫不等式，有

$$P(Y = 0) \leq P(|Y - E(Y)| \geq E(Y)) \leq \frac{Var(Y)}{E^2(Y)}$$

这里 $E(Y) = \sum E(I_i) = s \times \frac{6m}{s} \times \frac{1}{m} = 6$，重点在于 $Var(Y)$ 的计算。

我们知道 $Var(X_1 \pm X_2) = Var(X_1) + Var(X_2) \pm 2Cov(X_1, X_2)$，其中

$$Cov(X_1, X_2) = E\left[(X_1 - E(X_1))(X_2 - E(X_2))\right] = E(X_1X_2) - E(X_1)E(X_2)$$

是双线性函数。且按照定义有 $Cov(X, X) = Var(X)$.

那么

$$Var\left(\sum_{i=1}^{s} I_i\right) = \sum_{i=1}^{s} Var(I_i) + \sum_{i \neq j} Cov(I_i, I_j)$$

由于两两独立，因此协方差部分全部为 0，也就是说此处仍然成立

$$Var\left(\sum_{i=1}^{s} I_i\right) = \sum_{i=1}^{s} Var(I_i)$$

而根据伯努利分布的方差有

$$Var\left(\sum_{i=1}^{s} I_i\right) \leq s \times \frac{6m}{sm}\left(1 - \frac{6m}{sm}\right) \leq 6$$

所以有

$$P(Y = 0) \leq \frac{1}{6}$$

4.3 Majority Vote

如果有一串超长的字符序列（长度为 n），其中有一个字符出现频率超过的一半，每次读取一个字符，如何确定这个超过一半的字符?

方法是类似 “擂台” 的方式，设置一个位置记录当前的胜者及其出现次数，每次到来一个字符，如果位置为空，则记下该字符并将次数设置为 1；如果位置不为空且两字符相同，则次数 $+1$；如果两字符不同，则次数 -1，若减到 0 则位置变为空。最后剩下的字符即为答案。此方法的正确性可由反证法证明。若最后留下的不是所求字符，由于每次 “擂台” 都会消耗两个字符，那么可得总字符数大于 n，矛盾。

如果想求出字符频率超过 $\frac{n}{k+1}$ 的所有字符，那么只需设置 k 个位置，下一个字符到来时，若还有空位，则放入；若没有空位且与所有位置内字符不同，则所有位置的次数 -1。此算法正确性也可同样证明。

4.4 Sketching

在这一部分我们讨论两件事：查重与矩阵近似。本节标题译为 “画素描，概述”，研究的问题就是如何提取数据的特征。在查重的时候，我们不可能将待查重文章与数据库内所有论文逐一比对，而是用一些各自的关键字句来代替待查重文章和数据库内文章，先进行简单的比对，若重复率高，再进行详细比对。同样，在处理超大型矩阵的乘法运算时，为了节省空间，我们需要牺牲一些精确度，用某些数字或向量代替原矩阵进行运算。（这里 SVD 是不行的，因为求大型矩阵的奇异值太慢了）

4.4.1 查重

如何提取特征？我们最先想到的是，在两个部分里各自随机取样，但是这个方法相当糟糕，因为假如有一个部分是重复的，你必须在两个部分的取样中都抽到这一部分，这会使得概率变成原来的平方的量级。

孔老师在此用了两个班 A, B 的学生来举例。这两个班的学生可能有所交叉，我们如何估计两个班学生的重合度？比如用

$$\frac{|A \cap B|}{|A \cup B|}$$

来估计。这在两个班人数都不大的时候很容易做到，但是如果人数非常庞大呢？我们的方法是，在 $A \cup B$ 上进行随机排序（Random rank），获得一个排名。我们分别取出 A 中的后 10 名、B 中的后 10 名、总体中的后 10 名构成集合 $F(A), F(B), C$。那么我们用

$$\frac{|F(A) \cap F(B) \cap C|}{|C|}$$

来作为度量指标。这里的思想是："如果一个人在总体排名后 10 名，那么如果它在一个集合中，大概率也会是后 10 名"。当然 10 这个数字可以进行调整。

4.5 矩阵的 sketching

这一节的任务是估计两个超大型低秩矩阵的乘积 $\boldsymbol{AB}$，其中 $\boldsymbol{A}, \boldsymbol{B}$ 的形状分别为 $n \times k, k \times m$. 若 $\boldsymbol{A} = \begin{pmatrix} \boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2 & \cdots & \boldsymbol{\alpha}_k \end{pmatrix}, \boldsymbol{B} = \begin{pmatrix} \boldsymbol{\beta}_1^T & \boldsymbol{\beta}_2^T & \cdots & \boldsymbol{\beta}_k^T \end{pmatrix}^T$，那么

$$\boldsymbol{AB} = \sum_{i=1}^{k} \boldsymbol{\alpha}_i \boldsymbol{\beta}_i$$

很自然地，我们会想到选择一些指标，用对应的列向量、行向量相乘再相加作为近似。那么应该如何选择呢？首先我们考虑选择一个指标的情形，那么我们就需要给出一个概率向量 $\boldsymbol{p} = \begin{pmatrix} p_1 & p_2 & \cdots & p_k \end{pmatrix}$，分别对应选择第 k 个指标的概率，如果选择的结果为 i，那么我们将 $\boldsymbol{X} = \frac{1}{p_i} \boldsymbol{\alpha}_i \boldsymbol{\beta}_i$ 作为估计，其期望值为

$$E(\boldsymbol{X}) = \sum_{j=1}^{k} p_j \times \frac{1}{p_j} \boldsymbol{\alpha}_j \boldsymbol{\beta}_j = \boldsymbol{AB}$$

虽然上述估计方法保证了期望是正确的，但是不同的概率选择会影响方差，其结果的好坏也千差万别。一个可行的方案是等概率选择，即令 $p_i = \frac{1}{k}, i = 1, 2, \cdots, k$. 但是直觉告诉我们这并不可行，因为很显然不同的指标对于结果的贡献应该是不同的，似乎模长更大的贡献更大一些。我们用估计值与真实值的差的 F-范数平方的期望来度量性能好坏，即

$$E(\|\boldsymbol{X} - \boldsymbol{AB}\|_F^2)$$

令 $\boldsymbol{X} = (x_{ij})_{n \times m}, \boldsymbol{AB} = (c_{ij})_{n \times m}$，则有 $E(x_{ij}) = c_{ij}$，因此

$$\begin{aligned} E\left(\|\boldsymbol{X} - \boldsymbol{AB}\|_2^2\right) &= E\left(\sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - c_{ij})^2\right) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} E\left(x_{ij}^2 - 2 x_{ij} c_{ij} + c_{ij}^2\right) \\ &= E(\|\boldsymbol{X}\|_F^2) - E(\|\boldsymbol{AB}\|_F^2) \end{aligned}$$

也就是说要让 $E(\|\boldsymbol{X}\|_F^2)$ 最小。

$$E(\|\boldsymbol{X}\|_F^2) = \sum_{i=1}^{k} p_i E\left(\left\|\frac{\boldsymbol{\alpha}_i \boldsymbol{\beta}_i}{p_i}\right\|_F^2\right) = \sum_{i=1}^{k} \frac{1}{p_i} |\boldsymbol{\alpha}_i|^2 |\boldsymbol{\beta}_i|^2$$

由于 $\sum p_i = 1$，根据 Cauchy 不等式，当

$$p_i = \frac{|\boldsymbol{\alpha}_i|\,|\boldsymbol{\beta}_i|}{\sum_{j=1}^{k}|\boldsymbol{\alpha}_j|\left|\boldsymbol{\beta}_j\right|}$$

时，$E(\|\boldsymbol{X}\|_2^2)$ 最小，即概率正比于两个模长乘积。（插一句：孔老师在问大家应该怎么得到最小值的时候说了一句："You are very good at it, right? Otherwise you won't be here."）

如果可以选择多组那么按照上述概率选择即可。

这是两个矩阵乘法的情形。如果我们有一个高阶低秩的矩阵 $\boldsymbol{A}$，想对其自身进行估计，应该怎么做？根据我们对乘法的经验，由 $\boldsymbol{A} = \boldsymbol{AI}$ 可以对 $\boldsymbol{AI}$ 进行上述估计，然而此处并不太可行，因为 $\boldsymbol{I}$ 的秩太高了，这样做会丢失大量信息。

另一个想法是，找一个矩阵 $\boldsymbol{P}$ 使得 $\boldsymbol{A} \approx \boldsymbol{AP}$，再进行估计。我们想这样操作：从 $\boldsymbol{A}$ 中独立地选择一些行、列（无需一一对应），构成矩阵 $\boldsymbol{R}, \boldsymbol{C}$（$\boldsymbol{R}$ 是 $\boldsymbol{A}$ 中线性无关的 r 个行向量组成的矩阵），寻找矩阵 $\boldsymbol{U}$ 使得 $\boldsymbol{CUR}$ 是 $\boldsymbol{A}$ 的一个估计。

将 $\boldsymbol{A}_{n\times m}$ 看作是对 m 维向量的一个算子，则我们希望如果 $\boldsymbol{x} \in \mathbb{R}^m$ 是重要的，则 $\boldsymbol{APx} = \boldsymbol{Ax}$。若 $\boldsymbol{x}$ 不重要，那么 $\boldsymbol{APx} = \boldsymbol{0}$。

如何定义"重要"和"不重要"？我们将属于 $\boldsymbol{R}$ 的行向量空间 S 中的向量称为重要，若与之正交（$\boldsymbol{Rx} = \boldsymbol{0}$）则称为不重要。那么令

$$\boldsymbol{P} = \boldsymbol{R}^T(\boldsymbol{RR}^T)^{-1}\boldsymbol{R}$$

容易看出如果 $\boldsymbol{x} \in S^\perp$，那么 $\boldsymbol{APx} = \boldsymbol{0}$。如果 $\boldsymbol{x} \in S$，则 $\boldsymbol{x}$ 可被 $\boldsymbol{R}$ 中行向量线性表出，令 $\boldsymbol{x} = \boldsymbol{R}^T\boldsymbol{y}$，那么

$$\boldsymbol{APx} = \boldsymbol{AR}^T(\boldsymbol{RR}^T)^{-1}\boldsymbol{RR}^T\boldsymbol{y} = \boldsymbol{AR}^T\boldsymbol{y} = \boldsymbol{Ax}$$

我们考察一下二者的相似度。这次采用 2-范数，回忆

$$\|\boldsymbol{A} - \boldsymbol{AP}\|_2 = \max_{|\boldsymbol{x}|=1}\|(\boldsymbol{A} - \boldsymbol{AP})\boldsymbol{x}\|_2$$

对于 $\boldsymbol{x} \in S$，上式恒等于 0，因此只需考虑 $S^\perp$ 中的向量。因此

$$\max_{|\boldsymbol{x}|=1}\|(\boldsymbol{A} - \boldsymbol{AP})\boldsymbol{x}\|_2 = \max_{|\boldsymbol{x}|=1}\|\boldsymbol{Ax}\|_2$$

范数最大与范数平方最大等价，又由 $\boldsymbol{Rx} = \boldsymbol{0}$，且 $\boldsymbol{R}^T\boldsymbol{R}$ 可作为 $\boldsymbol{A}^T\boldsymbol{A}$ 的近似，那么

$$\|\boldsymbol{Ax}\|_2^2 = \left\|\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{Ax}\right\|_2^2 = \left\|\boldsymbol{x}^T(\boldsymbol{A}^T\boldsymbol{A} - \boldsymbol{R}^T\boldsymbol{R})\boldsymbol{x}\right\|_2^2 \approx 0$$

5 随机图

所谓随机图，就是边的存在是有一定概率的。本节主要讨论的是 $G(n,p)$，也就是 n 个顶点、每条边出现的概率是 p 的随机图。当 $p=0$ 时为零图，当 $p=1$ 时为 n 阶完全图。

随机图的一个特点就是，图的很多性质（如连通性）都是在 p 大于某一阈值之后突然出现的，而在此阈值之前出现概率几乎为零。这被称为 Threshold Property。比如，当 p 大于某个阈值之后，随机图几乎必然联通，等等。

另外需要注意的是，对于确定的 n，其某种性质出现的概率当然可以计算，我们这里说的“突然出现”，是指在 $n\to\infty$ 的意义下的。后面我们会看到这一点。

我们先考虑一个简单的问题：对于随机图 $G(n,p)$，其中构成三角形的个数的期望是多少？这个问题并不难回答，首先取定三个点，然后乘以相应概率即可。对每组三个点我们都定义一个随机变量 $I_\triangle$，当这三个点之间三条边都存在时为 1，反之为 0。即

$$E\left(\sum_{\text{所有}\triangle} I_\triangle\right)=\binom{n}{3}p^3\sim n^3p^3$$

从直观上我们可以看出，当 $p=o(\frac{1}{n})$ 时 $E\to 0(n\to\infty)$. 当 $p\gg O(\frac{1}{n})$ 时 E 几乎不可能为 0。下面我们来证明这一点，分别需要使用马尔可夫不等式和切比雪夫不等式。

低于阈值时，三角形几乎不会出现，这很容易证明。令 $X=\sum I_\triangle$ 代表三角形个数，由马尔可夫不等式

$$P\left(X>a\right)<\frac{E(X)}{a}$$

由于 $E(X)\to 0$，取一个小于 1 的 a 即可。

但是当 p 大于阈值的时候却并不能单纯地用期望来证明。孔老师举例如下：假如班级的平均分是 99 分，但是实际上这个班级的同学可以得到任意高的分数，那么并不能保证随便抽取一个同学，他的得分在 99 分附近。相反，完全有可能是一个同学得到了几千分，其他同学都很低。不过，如果还能给定这个班级的分数方差很小，那么这种断言就可信了。我们下面证明就是从方差入手。这里给出了一种感性的认识。

用我们前面用过的技术，有

$$P(X=0)\le P(|X-E(X)|\le E(X))\le\frac{Var(X)}{E^2(X)}$$

以及

$$Var(X)=\sum_{\triangle}Var(I_\triangle)+\sum_{\triangle\neq\triangle'}Cov(I_\triangle,I_{\triangle'})$$

对于方差，有

$$Var(I_\triangle)=E(I_\triangle^2)-E^2(I_\triangle)\le E(I_\triangle)$$

对于协方差，有

$$Cov(I_\triangle,I_{\triangle'})=E(I_\triangle I_{\triangle'})-E(I_\triangle)E(I_{\triangle'})\le E(I_\triangle I_{\triangle'})$$

将不独立的形状画出来，是四个顶点五个边构成了两个相邻的三角形，出现概率约为 n^4p^5。因此 $Var(X) \sim n^3p^3 + n^4p^5$。所以

$$P(X=0) \leq \frac{Var(X)}{E^2(X)} \sim \frac{n^3p^3 + n^4p^5}{(n^3p^3)^2} \to 0(n \to \infty)$$

我们用相同的思想处理一下图中出现 $4-clique$（即 4 阶完全子图）的阈值。首先其数量的期望为 $\binom{n}{4}p^6$，方差的阶为 $n^4p^6 + n^5p^9 + n^6p^{11}$，同样的方法可以证明。

上述讨论可以推广到任意给定的图 H 上。假定图形 H 有 v 个定点，e 条边，那么其出现个数的均值为 $\Theta(n^vp^e)$，我们自然会认为 $p = n^{-v/e}$ 是一个可能的阈值。（注意到这里 v/e 是图 H 的平均度数倒数的两倍）

那么什么时候 $p = n^{-v/e}$ 确实是图 H 出现的阈值呢？我们称一个图是平衡的，当且仅当这个图的平均度数大于等于任何其子图的平均度数。由此我们可以证明：

定理 5.1. *如果 H 是平衡的，那么 $p = n^{-v/e}$ 是它出现的阈值。*

对于一般的图 H，定理仍然成立，但是此时的 v/e 必须是 H 所有子图的最小值，换言之取决于最“稠密”的子图。

最后我们来研究随机图直径问题。直径是指图中顶点最小路径长度的最大值。如果图很稀疏，那么直径会很大，反之很小。对于完全图，$d = 1$。那么如何度量 $d > 2$ 的概率临界值呢?

容易知道，对于顶点 u, v，如果 $\exists w$ 使得 $(u,w),(w,v)$ 是边，那么 $d(u,v) \leq 2$。对于每对顶点，有另外 $n-2$ 个顶点可作为桥梁，而且它们自身不能直接相连，那么不能在两条边内到达的顶点数的期望是

$$E\left(\sum_{pairs} I_{pairs}\right) = \binom{n}{2}(1-p)(1-p^2)^{n-2}$$

6 附录一：Γ 函数以及部分性质

$$\Gamma(z)=\int_0^{+\infty}t^{z-1}e^{-t}\mathrm{d}t\ \ (\mathrm{Re}(z)>0)$$

Γ 函数有如下性质:

(1) $\Gamma(1)=1$

(2) $\Gamma(z+1)=z\Gamma(z)$

(3) $\Gamma(n)=(n-1)!$

(4) $\Gamma(z)\Gamma(1-z)=\dfrac{\pi}{\sin\pi z}$

性质 4 的证明:

$$\Gamma(z)\Gamma(1-z)=\iint_{\substack{s>0\\t>0}}e^{-(t+s)}\left(\frac{t}{s}\right)^x\frac{1}{t}\mathrm{d}s\mathrm{d}t$$

换元令 $\xi=t+s,\eta=t/s$，得到

$$\Gamma(z)\Gamma(1-z)=\int_0^{+\infty}\frac{\eta^{z-1}}{1+\eta}\mathrm{d}\eta=\frac{\pi}{\sin\pi z}$$

此积分可用复变函数求解。

在性质 4 中取 $z=\dfrac{1}{2}$ 就有 $\Gamma(\dfrac{1}{2})=\dfrac{\sqrt{\pi}}{2}$，这个结果也就是高斯积分

$$\int_0^{+\infty}e^{-x^2}\mathrm{d}x=\frac{\sqrt{\pi}}{2}$$

从这个结果出发，结合递推公式，我们就有

$$\Gamma(\frac{2n+1}{2})=\frac{(2n-1)!!}{2^n}\sqrt{\pi}$$

在函数定义式中换元，令 $t=r^2$ 可得

$$\int_0^{\infty}r^pe^{-r^2}\mathrm{d}r=\frac{1}{2}\Gamma(\frac{p+1}{2})$$

我们在正文中求出了高维单位球体的表面积、体积表达式

$$A(n)=\frac{2(\sqrt{\pi})^n}{\Gamma(\frac{n}{2})}$$
$$V(n)=\frac{2(\sqrt{\pi})^n}{n\Gamma(\frac{n}{2})}$$

容易看出当 $n\to\infty$ 时，n 维球体的体积趋于零。画出函数图像如下:

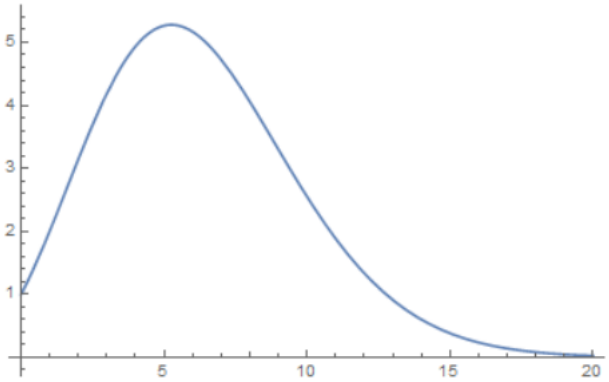

Figure 1: 函数 $\frac{2(\sqrt{\pi})^x}{x\Gamma(\frac{x}{2})}$ 的图像

可以看出五维单位球体体积最大为 $\frac{8\pi^2}{15} \approx 5.264$.

7 附录二：Johnson-Lindenstrauss Lemma 的另一种证明

定理 7.1. *对于任意的 $\mathbb{R}^d$ 中的 n 个点构成的集合 X 和 $\forall \varepsilon \in (0,1)$，存在一个映射 $\varphi : \mathbb{R}^d \to \mathbb{R}^k$，这里*

$$k = \left\lceil \frac{4\ln n}{\varepsilon^2/2 - \varepsilon^2/3} \right\rceil \leq \left\lceil \frac{24}{\varepsilon^2} \ln n \right\rceil$$

使得 $\forall u, v \in X$，

$$(1-\varepsilon)\|u-v\|_2^2 \leq \|\varphi(u) - \varphi(v)\|_2^2 \leq (1+\varepsilon)\|u-v\|_2^2 .$$

Proof. 对于这个结果有多种证明方法，我们采用 [DG99] 的证明方法。证明采用了概率方法，其构造的嵌入方式非常简单：仅仅将每个点映射到一个随机的 k 维超平面上。更准确地说，对于一个原空间中的点 v，我们令它的象 $\varphi(v) = \sqrt{\frac{d}{k}} v'$，这里 v' 代表 v 在这个超平面上的投影。

为了分析这个映射的性能，我们需要考虑当映射 φ 随机时，随机变量

$$\frac{\|\varphi(u) - \varphi(v)\|_2^2}{\|u-v\|_2^2}$$

的分布。我们不妨假定 $\|u-v\|_2^2 = 1$，也就是 $u-v$ 是单位向量。注意到 $\|\varphi(u) - \varphi(v)\|_2^2$ 就是一个固定单位向量映射到一个随机超平面之后的模的平方，等价于一个随机单位向量映射到一个固定的超平面。因此，我们可以通过在 d 维单位球面上随机选取一个点，将其投影到前 k 个基向量构成的超平面上的方法来研究。首先生成一个随机向量 $X = (X_1, X_2, \cdots, X_d), X_i \sim N(0,1)$ i.i.d. 随后将其单位化形成单位向量 $Z = \frac{1}{\|X\|_2}(X_1, X_2, \cdots, X_d)$. 随后将其投影，我们得到了向量 $Y = \frac{1}{\|X\|_2}(X_1, X_2, \cdots, X_k)$.

我们的目标就是分析随机变量

$$L = \|Y\|_2^2 = \frac{X_1^2 + \cdots + X_k^2}{X_1^2 + \cdots + X_d^2}$$

的分布。注意，根据对称性我们有 $\mu = \mathbb{E}[L] = \frac{k}{d}$. 这也正是我们要在映射前乘以系数 $\sqrt{\frac{d}{k}}$ 的原因。

这里，我们要用到的关键事实是一个如下的 Chernoff 型的边界估计：

命题 7.2. *对于如上定义的 L 和 μ，我们有*

1. $\Pr[L < (1-\varepsilon)\mu] \leq \exp(-\frac{\varepsilon^2 k}{4})$
2. $\Pr[L \geq (1+\varepsilon)\mu] \leq \exp(-\frac{k}{2}(\frac{\varepsilon^2}{2} - \frac{\varepsilon^2}{3}))$

这个命题我们将稍后证明。我们首先完成原定理的证明。

根据命题，对于 $k \geq \left\lceil \frac{4\ln n}{\varepsilon^2/2 - \varepsilon^2/3} \right\rceil$，我们有

$$\Pr[|L-\mu| \geq \varepsilon\mu] \leq 2\exp(-2\ln n) = \frac{2}{n^2}$$

因此这个映射“不好”的概率，也就是

$$\Pr[|L-\mu| \geq \varepsilon\mu \text{ for any pair } u, v] \leq \frac{2}{n^2}\binom{n}{2} = \left(1 - \frac{1}{n}\right)$$

因此映射是满足要求的概率大于 $\frac{1}{n}$. 所以必然存在满足要求的映射。我们也可以知道大约运行 $O(n)$ 轮次我们就可以得到一个满足要求的映射。 ■

下面我们要证明前面的命题。

Proof. 同样使用证明 Chernoff Bound 的方法，我们有

$$\begin{aligned}
\Pr[L \le (1-\varepsilon)\mu] &= \Pr\left[\left(\frac{X_1^2+\ldots+X_k^2}{X_1^2+\ldots+X_d^2}\right) \le (1-\varepsilon)\frac{k}{d}\right] \\
&= \Pr\left[k(1-\varepsilon)\left(X_1^2+\ldots+X_d^2\right) - d\left(X_1^2+\ldots+X_k^2\right)\right) \ge 0\Big] \\
&= \Pr\left[\exp\left\{t\left[k(1-\varepsilon)\left(X_1^2+\ldots+X_d^2\right) - d\left(X_1^2+\ldots+X_k^2\right)\right]\right\} \ge 1\right] \\
&\le \mathbb{E}\left[\exp\left\{t\left[k(1-\varepsilon)\left(X_1^2+\ldots+X_d^2\right) - d\left(X_1^2+\ldots+X_k^2\right)\right]\right\}\right] \\
&= \mathbb{E}\left[\exp\left\{tk(1-\varepsilon)X_1^2\right\}\right]^{(d-k)} \mathbb{E}\left[\exp\left\{t(k(1-\varepsilon)-d)X_1^2\right\}\right]^k \\
&= (1-2tk(1-\varepsilon))^{-(d-k)/2}(1-2t(k(1-\varepsilon)-d))^{-k/2}
\end{aligned}$$

最后一步的等式使用了如下的事实：如果 $X \sim N(0,1)$, 那么当 $s < \frac{1}{2}$ 时有 $\mathbb{E}[e^{sX^2}] = (1-2s)^{-\frac{1}{2}}$. 这是显然的，因为

$$\mathbb{E}[e^{sX^2}] = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-(\frac{1}{2}-s)x^2} \mathrm{d}x = (1-2s)^{-\frac{1}{2}}$$

此广义积分当且仅当 $s < \frac{1}{2}$ 时收敛。

余下要做的有两点：第一, 取到上式的最小值; 第二, 证明取到最小值的 t 能够满足 $tk(1-\varepsilon) < \frac{1}{2}$ 和 $t(k(1-\varepsilon)-d) < \frac{1}{2}$, 以便保证上式这一放缩是合理的。求导数可知极值点为

$$t = \frac{\varepsilon}{2(1-\varepsilon)(d-k(1-\varepsilon))}$$

并且 t 满足上述两个条件。

$$\begin{aligned}
\Pr[L \le (1-\varepsilon)\mu] &\le \left(1-\frac{k\varepsilon}{d-k(1-\varepsilon)}\right)^{-(d-k)/2}\left(1+\frac{\varepsilon}{1-\varepsilon}\right)^{-k/2} \\
&= \left(1+\frac{k\varepsilon}{d-k}\right)^{(d-k)/2}(1-\varepsilon)^{k/2} \\
&< \exp\left(\frac{k\varepsilon}{2}\right)(1-\varepsilon)^{k/2} \quad \left[\text{ since } \left(1+\frac{x}{y}\right)^y < e^x\right] \\
&= \exp\left(\frac{k\varepsilon}{2}+\frac{k}{2}\ln(1-\varepsilon)\right) \\
&< \exp\left(-\frac{\varepsilon^2 k}{4}\right) \quad \left[\text{ by Taylor expansion: } \ln(1-\varepsilon) < \left(-\varepsilon-\frac{\varepsilon^2}{2}\right)\right]
\end{aligned}$$

另一个不等式可以用同样的方法进行证明。唯一的不同在于需要使用 $\ln(1+\varepsilon) < \varepsilon - \frac{\varepsilon^2}{2} + \frac{\varepsilon^3}{3}$. ■

参考文献

[DG99] S. DASGUPTA and A. GuPTA, "An elementary proof of the Johnson-Lindenstrauss Lemma," Technical Report TR-99-006, International Computer Science Institute, Berkeley, CA, 1999 .