Math in Information Age

Course Review

Cheng Chen

EECS, PKU

Outline

1. 作业讲评
2. 课程回顾

作业讲评

课程回顾

高维空间几何体的特征

单位球体积集中在赤道附近：

基本证明思路是分析以下比值：

$$r = \frac{Volume(A: \text{高于 } x_1 = \epsilon \text{ 部分的体积})}{Volume(H: \text{球的上半部分体积})}$$

- **一个简单的界**
  $x = \epsilon$ 处的横截面是一个半径为 $\sqrt{(1-\epsilon^2)}$ 的 $d-1$ 维球。
  所以 $V(A) \le (1-\epsilon)(1-\epsilon^2)^{\frac{d-1}{2}} V(d-1) < (1-\epsilon^2)^{\frac{d-1}{2}} V(d-1)$
  而 $V(H) = \frac{1}{2} V(d)$
  由此得到
  $\frac{V(A)}{V(H)} = 2(1-\epsilon^2)^{d-1} \frac{V(d-1)}{V(d)} = 2(1-\epsilon^2)^{d-1} \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}(1-\frac{1}{d})\Gamma(\frac{d-1}{2})} \le \frac{(1-\epsilon^2)^{\frac{d-1}{2}}(d-1)}{\sqrt{\pi}(1-\frac{1}{d})} = \frac{(1-\epsilon^2)^{\frac{d-1}{2}} d}{\sqrt{\pi}}$

高维单位球的特征

书本的证明:

$$
\begin{aligned}
V(A) &= \int_{\frac{c}{\sqrt{d-1}}}^{1} (1-x^2)^{\frac{d-1}{2}} V(d-1)dx \\
&\leq \int_{\frac{c}{\sqrt{d-1}}}^{1} e^{\frac{d-1}{2}x^2} V(d-1)dx \\
&\leq \int_{\frac{c}{\sqrt{d-1}}}^{\infty} \frac{x}{\frac{c}{\sqrt{d-1}}} e^{\frac{d-1}{2}x^2} V(d-1)dx \\
&= \frac{\sqrt{d-1}V(d-1)}{c} \int_{\frac{c}{\sqrt{d-1}}}^{\infty} xe^{\frac{d-1}{2}x^2} dx \\
&= \frac{V(d-1)}{c\sqrt{d-1}} e^{-\frac{c^2}{2}}
\end{aligned}
$$

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

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

高维单位球的特征

证明，存在一个边长趋近于 0 的正方体，它包含了高维球的大部分体积。

高维空间随机生成点的特征

从单位球中随机选取 $x_1, x_n, ..., x_n$ 个点，那么这些点有 $1 - O(\frac{1}{n})$ 的概率满足

- $$|x_i| \geq 1 - \frac{2\ln n}{d}$$
- $$|x_i \cdot x_j| \leq \frac{\sqrt{6\ln n}}{\sqrt{d-1}}$$

高维空间随机生成点的特征

- 将一个半径为 $\sqrt{d}$ 的 d 维球的体积投影到一个他的直径上，对于很大的 d，投影体积近似于什么分布？

Gaussian Annulus Theorem

对于 d 维标准高斯随机向量 $\mathbf{x}$，对于任意 $\beta \leq \sqrt{d}$，有：

$$P(||\mathbf{x}| - \sqrt{d}| \geq \beta) \leq 3e^{-c\beta^2}$$

随机投影

在 d 中选择 k 个高斯随机向量 $\mathbf{u_1}, \mathbf{u_2}, ..., \mathbf{u_k}$，则对于 $\mathbf{v} \in \mathbf{R}^d$，定义随机投影：

$$f(\mathbf{v}) = (\mathbf{u}_1 \cdot \mathbf{v}, \mathbf{u}_2 \cdot \mathbf{v}, ..., \mathbf{u}_k \cdot \mathbf{v})$$

对于上述定义的随机投影，我们可以借助 Gaussian Annulus Theorem 得到对于 $\exists c > 0, \forall \epsilon \in (0, 1)$:

$$P(||f(\mathbf{v}) - \sqrt{k}|\mathbf{v}|| \geq \epsilon\sqrt{k}|\mathbf{v}|) \leq 3e^{-ck\epsilon^2}$$

Johnson-Lindenstrauss Lemma

存在 c，对于任意的 $0 < \epsilon < 1$ 和正整数 n，令 $k > \frac{3}{c\epsilon^2} \ln n$，对于上述定义的随机投影，对于所有 $\mathbf{v}_i, \mathbf{v}_j \in \mathbf{R}^d$，有至少 $1 - \frac{3}{2n}$ 的概率下列不等式满足：

$$(1-\epsilon)\sqrt{k}|\mathbf{v}_i - \mathbf{v}_j| \leq |f(\mathbf{v}_i) - f(\mathbf{v}_j)| \leq (1+\epsilon)\sqrt{k}|\mathbf{v}_i - \mathbf{v}_j|$$

给出 d 维空间中的点，已知这些点分别属于两个不同的高维高斯随机分布，将其命名为红点和蓝点，现在我们希望区分哪些点是红点，哪些点是蓝点。

很显然，两个高斯随机分布的均值越接近，就越难分辨。我们假设两个均值相对方差来说分得比较开，更进一步的说，考虑两个由方差为 1 的高斯分布为分量的高维高斯分布，我们这一节的分析要求 $|\mu_1 - \mu_2| = \Omega(d^{\frac{1}{4}})$

基本的想法是高维空间中，属于同一个分布的 x 与 y，$|\mathbf{x} - \mathbf{y}|$ 相对稳定，且会明显有别于 x 和 y 不属于同一个分布的情况。