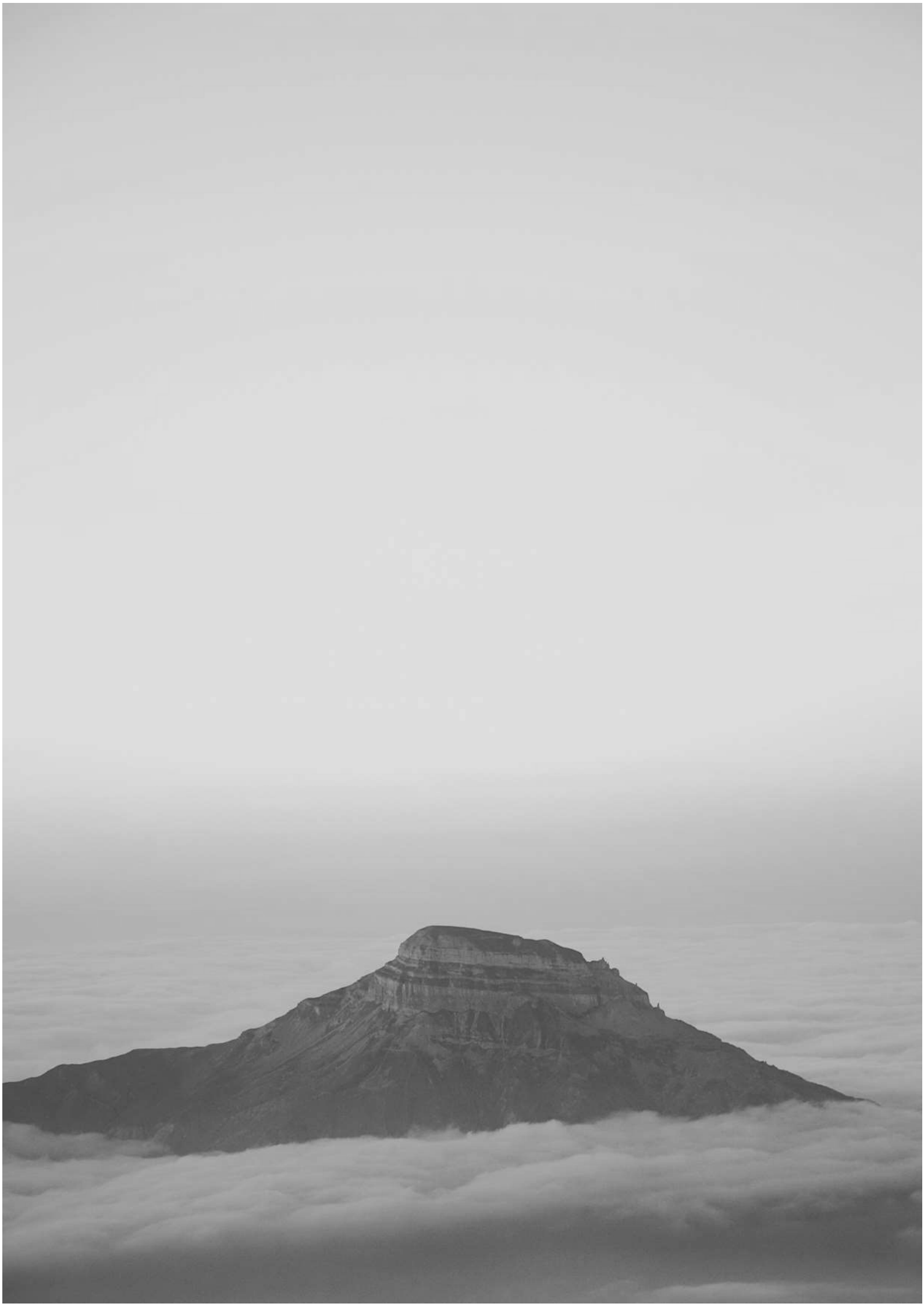


§1

Dimension Reduction

日期: Lecture 1: Singular Value Decomposition (SVD).

Consider data as vectors $v=\begin{pmatrix} v_1 \\ \vdots \\ v_d \end{pmatrix} \in \mathbb{R}^d$.

then, $A \in \mathbb{R}^{n\times d}$, 有两种看法: ① Matrix.

$=\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$.

② n points in d-dimension space.

Fact: SVD finds the "best-fitting" k-dim subspace of A.

Aim: "best" $\Rightarrow$ ① $\min_{S:k\text{-dim space}} \sum_{i=1}^{n} dist^2(a_i, S)$

$\Leftrightarrow$ ② $\max_{S:k\text{-dim space}} \sum_{i=1}^{n} length^2(proj(a_i, S))$.

范数:

$\|x\|_p = (\sum_{i=1}^{n} x_i^p)^{\frac{1}{p}}$.

$\|x\|_\infty = \max_{i=1}^{n} |x_i|$

$\|x\| := \|x\|_2$.

Algorithm:

Consider $k=1$: $\max_{\|v\|=1} \sum_{i=1}^{n} length^2(proj(a_i, span(v)))$.

$= \max_{\|v\|=1} \sum_{i=1}^{n} |\langle a_i, v\rangle|^2$.

$= \max_{\|v\|=1} \|Av\|^2$.

Define $\sigma_1(A) = \max_{\|v\|=1} \|Av\|$

v_1 is "the first" singular vector if $\|Av_1\| = \sigma_1(A)$. (注:不一定唯一).

贪心算法 ↓

Consider $k=2$:

在 $k=1$ 基础上, 剔除 v_1 影响后继续 $\max_{\substack{\|v\|=1,\\ v\perp v_1}} \|Av\|$.

Define: the second singular value: $\sigma_2(A) = \max_{\substack{\|v\|=1\\ v\perp v_1}} \|Av\|$.

"the second" singular vector: $v_2 = \arg\max_{\substack{\|v\|=1\\ v\perp v_1}} \|Av\|$.

日期: /

... (这里指的求"best-fitting"的空间维数).

Until: $k=r+1$. (或 $k=k+1$)

find $\sigma_{r+1}(A)=0$. $\Rightarrow$ 算法停止.

Problem Remained: Definition.

1. How to compute $\arg\max_{\|v\|=1} \|Av\|$.

2. Greedy's correctness. $v\perp v_1,\cdots,v\perp v_k$

Thm 1: $A\in\mathbb{R}^{n\times d}$. singular vectors $v_1,\cdots,v_r$.

$\forall 1\le k\le r$, let $V_k=\mathrm{span}(v_1,\cdots,v_k)$. $\Rightarrow$ Correctness

Then, $\forall k$. V_k is the "best-fitting" k-dim subspace of A.

Pf: $k=1$ 时, 由前已证.

by induction: Suppose $k=k-1$ 成立.

$k=k$ 时, Suppose W is the "best-fitting" k-dim space.

总可选出一组 orthonormal basis $w_1,\cdots,w_k$. st. $v_1\perp w_k,\cdots,v_{k-1}\perp w_k$.

由归纳, $\mathrm{span}(v_1,\cdots,v_{k-1})$ is "best-fitting".

$$\|Aw_1\|^2+\cdots+\|Aw_{k-1}\|^2\le\|Av_1\|^2+\cdots+\|Av_{k-1}\|^2.$$

又由 v_k 选择, $\|Aw_k\|^2\le\|Av_k\|^2$.

$\therefore \|Aw_1\|^2+\cdots+\|Aw_k\|^2\le\|Av_1\|^2+\cdots+\|Av_k\|^2$. □

日期: /

(Rmk: 将A作用在 $span(v_1,\cdots v_r)$ 上，在旋转之外，相当于将 $v_1,\cdots v_r$ 伸长 $\sigma_i(A)$ 倍)

(且 $A\perp(span(v_1,\cdots v_r))^C \Rightarrow A = span(v_1,\cdots v_r)$

$\hookrightarrow \dim A = r$.

Def: (Frobenius Norm / F-Norm).

$$\|A\|_F := \sqrt{\sum a_{ij}^2} \quad (= \sqrt{\sum_{i=1}^n \|a_i\|^2}).$$

Rmk: $\|UAV^T\|_F = \|A\|_F$ (即旋转不影响 F-Norm).

Lem: $\sum \sigma_i^2(A) = \|A\|_F^2$.

(Geometric Exp: $\|A\|_F^2 = \sum_{i=1}^n \|a_i\|^2$. $a_i \in R^d$. (标准基下的模长计算).

换 basis 不改变模长.

而 $\sigma_i^2(A) = \|Av_i\|^2$ 即 $a_1,\cdots a_n$ 在 v_i 方向上投影的平方和.

而 $span(v_1,\cdots v_r) = span(a_1,\cdots a_n)$.

∴ LHS 即在 $\{v_1,\cdots v_r\}$ 基下计算 $a_1,\cdots a_n$ 模长平方和.

∴ LHS = RHS. □.)

Def: $A \in R^{n\times d}$. define the left singular vector as:

$$\star\ u_i = \frac{1}{\sigma_i(A)} Av_i$$

{A's singular value}

Lem: $\{u_i\}_{i=1,\cdots,r}$ 两两正交.

Pf: 反证. let i be the smallest index. s.t.

$\langle u_i, u_j\rangle \neq 0$ for some $j>i$.

日期: /

WLOG suppose $\langle u_i, u_j \rangle = \delta > 0$.
(不失一般性).

for some $\varepsilon > 0$, define $v_i' = \frac{v_i + \varepsilon v_j}{\|v_i + \varepsilon v_j\|}$. 则 $\|v_i'\| = 1$.

① $\|Av_i'\| \geq u_i^T \cdot Av_i'$.

$$> \frac{u_i^T \sigma_i u_i + u_i^T \varepsilon \sigma_j u_j}{\sqrt{1+\varepsilon^2}}$$

$$= \frac{\sigma_i + \sigma_j \varepsilon \delta}{\sqrt{1+\varepsilon^2}}$$

$\frac{1}{\sqrt{1+\varepsilon^2}} \geq 1 - \frac{\varepsilon^2}{2}$, $\varepsilon \to 0$.

$$\geq (\sigma_i + \sigma_j \varepsilon \delta)(1 - \tfrac{\varepsilon^2}{2}) > \sigma_i.$$

② $i = 1$ 时, $\|Av_i'\| \overset{①}{>} \sigma_1 = \max_{\|v\|=1} \|Av\|$. 矛盾.

$i > 1$ 时,

by suppose
$\Rightarrow$ 由 $v_i' = \frac{v_i + \varepsilon v_j}{\|v_i + \varepsilon v_j\|}$ is orthogonal to $v_1, \cdots, v_{i-1}$.

$\therefore \|Av_i'\| \leq \sigma_i$. 矛盾. □.

Thm: $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$ (SVD)

Pf: for $\forall i$, $Av_i = \sigma_i u_i$.

(Rmk: for $\forall v \in R^d$, $Av = Bv$, then $A = B$.).

$\therefore$ for $\forall v \in R^d$, suppose $v = a_1 v_1 + \cdots + a_r v_r + \cdots + a_d v_d$

$$\therefore Av = \sum_{i=1}^{r} a_i Av_i = \sum_{i=1}^{r} a_i \sigma_i u_i$$

而 $(\sum_{i=1}^{r} \sigma_i u_i v_i^T) v = \sum_{i=1}^{r} \sigma_i u_i v_i^T v_i a_i = \sum_{i=1}^{r} a_i \sigma_i u_i$

$\therefore A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$. □.

日期: /

Cor1: $U=\underset{\in R^{n\times r}}{(u_1,\cdots,u_r)}$. $V=\underset{\in R^{d\times r}}{(v_1,\cdots,v_r)}$. $D=\underset{\in R^{r\times r}}{\begin{pmatrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_r\end{pmatrix}}$.

$\Rightarrow A=UDV^T$.

Cor2: $U=\underset{\in R^{n\times n}}{(u_1,\cdots,u_r,\overbrace{\cdots}^{用U补满正交基补全},u_n)}$. $V=\underset{\in R^{d\times d}}{(v_1,\cdots,v_r,\overbrace{\cdots}^{\cdots},v_d)}$ $D=\underset{\in R^{n\times d}}{\begin{pmatrix}\sigma_1 & & & \\ & \ddots & & \\ & & \sigma_r & \\ & & & 0\end{pmatrix}}$

$\Rightarrow A=UDV^T$.

I. Singular vectors & eigenvectors.

Suppose $A\in R^{n\times d}$.

① If A has SVD: $A=\sum_{i=1}^{r}\sigma_i u_i v_i^T$

then $Av_i=\sigma_i u_i$, $A^Tu_i=\sigma_i v_i$

Pf: 由 $A^T=\sum_{i=1}^{r}\sigma_i v_i u_i^T$ $\therefore A^Tu_i=\sigma_i v_i$ □.

(Rmk: 类似于 eigenvectors 之与对应 eigenvalue 之的性质

$A\alpha=\lambda\alpha$ (for square A).).

② Let $B=A^TA$, then v_i is eigenvectors of B,

对应 eigenvalue is σ_i^2.

同理, AA^T, then u_i is eigenvectors of B,

对应 eigenvalue is σ_i^2.

Pf: $B^Tu_i=(\Sigma\sigma_j u_j v_j^T)(\Sigma\sigma_k v_k u_k^T)u_i=(\Sigma\sigma_j^2 u_j u_j^T)u_i=\sigma_i^2 u_i$ □.

日期: /

II. low rank approximation.

$A_{n\times d} \leftarrow A'_{n\times d}$ (low rank).

eg. customer-movie.

hypothesis: 有k个 underlying factors for a_{ij}.

$\Rightarrow$ low rank.

$$A \approx (x_{ij})_{n\times k} \; (y_{ij})_{k\times d}.$$

customer-factors (偏好矩阵) factors-movie (属性矩阵).

Consider SVD: $A = \underbrace{U}_{X} P \underbrace{V^T}_{Y^T}$ (rmk: 可能出现负元(但可忍受))

Goal: find B with rank k, s.t. B is close to A

(体现为优化问题: eg. s.t. $\|A-B\|_F$ or $\|A-B\|_2^2$ is minimal.).

Def: def A_k as: rows of A_k are projections ($r(A_k)=k$)

of rows of A onto the subspace $V_k = \text{span}\{v_1, \cdots, v_k\}$

即. $A_k = (a_i \sum_{j=1}^{k} v_j v_j^T) \rightarrow \sum_{j=1}^{k} \langle a_i, v_j \rangle v_j^T$ (单行的proj.).

$$= A \sum_{j=1}^{k} v_j v_j^T$$

$$= \left(\sum_{i=1}^{n} \sigma_i u_i v_i^T\right)\left(\sum_{i=1}^{k} v_i v_i^T\right)$$

$$= \sum_{i=1}^{k} \sigma_i u_i v_i^T.$$

日期: /

By above, $\|A-A_k\|_F^2 = \sum_{i=1}^{n} dist^2(a_i, V_k) = \min_{\substack{S:\text{space.} \\ dim(S)=k.}} \sum_{i=1}^{n} dist^2(a_i, S)$.

Thm1: A_k is the best rank approximation w.r.t. F-Norm. (w.r.t. = with respect to.)

即. $\forall B$ with $rank(B)\le k$, $\|A-A_k\|_F \le \|A-B\|_F$.

Pf: Suppose $B \in \mathbb{R}^{n\times d}$ s.t. $\|A-B\|_F^2$ min. → Geometric.

设 V 为 span(rows of B). (k-dim.)

then $\|A-B\|_F^2 = \sum_{i=1}^{n} dist(a_i, V) \ge \|A-A_k\|_F^2$. □

Def: (2-Norm): $\|A\|_2 = \max_{\|x\|\le 1} \|Ax\|$

(Note that 1° $\|A\|_2 = \sigma_1(A)$ 2° $\|A\|_2$ 与向量 v 的 $\|v\|_2$ 定义相容.

Thm2: A_k is the best rank approximation w.r.t. 2-Norm.

即. $\forall B$ with rank k, $\|A-A_k\|_2 \le \|A-B\|_2$

Lem: $\|A-A_k\|_2^2 = \sigma_{k+1}^2$.

Note that, $A-A_k = \sum_{i=k+1}^{r} \sigma_i u_i v_i^T$.

For $\forall v \in \mathbb{R}^d$. 不妨 $v = \sum_{j=1}^{r} \alpha_j v_j$. $(\alpha_j \in \mathbb{R})$

then. $(A-A_k)v = \sum_{i=k+1}^{r} \alpha_i \sigma_i u_i$

$\therefore \|(A-A_k)v\| = \|\sum_{i=k+1}^{r} \alpha_i \sigma_i u_i\| = \sqrt{\sum_{i=k+1}^{r} \alpha_i^2 \sigma_i^2}$.

日期: /

let v^* is the unit vector maximizing above.

$\therefore \sum_{i=1}^{r} a_i^2 = 1$.

又: $\sigma_{k+1} \geq \cdots \geq \sigma_r$.

$\therefore v^*$ satisfy $a_{k+1} = 1$, $a_i = 0$ $(i \neq k+1)$

$\therefore \|A - A_k\|_2 = \|(A - A_k)v^*\| = \sigma_{k+1}$. □.

Thm-Pf: If $\text{rank}(A) \leq k$ then $\|A - A_k\|_2 = 0$.

Assume $\text{rank}(A) > k$. 則 $\|A - A_k\|_2^2 = \sigma_{k+1}^2$.

Consider the null space of B, i.e. $\text{Null}(B) = \{v : Bv = 0\}$.

Note that $\dim(\text{Null}(B)) \geq d - k$. ($\because \text{rank}(B) \leq k$).

then $\exists$ vector $z \neq 0$ such that $z \in \text{Null}(B) \cap \text{Span}(v_1, \cdots, v_{k+1})$

且 $\|z\| = 1$.

Note that $\|A - B\|_2^2 \geq \|(A - B)z\|_2^2$ ← $z \in \text{Null}(B)$. ← $Bz = 0$

$= \|Az\|_2^2$

$z \in \text{Span}(v_1, \cdots, v_k)$ ↓

$= \|\sum_{i=1}^{k+1} \sigma_i (v_i^T z) u_i\|_2^2$ (number)

$= \sum_{i=1}^{k+1} \sigma_i^2 (v_i^T z)^2$

$\geq \sigma_{k+1}^2 \underbrace{\sum_{i=1}^{k+1} (v_i^T z)^2}_{= \|z\|}$

$= \sigma_{k+1}^2$. □.

(Rmk: 即 A_k is best approximation over 2-form & F-form.)

日期: /

Ⅲ. Compute SVD:

1. Exact Computation: $O(nd^2)$. → 太大，考虑近似的.

2. Approximate SVD:

1. Start with v_1: (Power Method). → 通过power扩大σ_i的impact
   ↳幂.

① $B = A^TA = (\sum \sigma_i u_i v_i^T)(\sum \sigma_j u_j v_j^T) = \sum \sigma_i^2 v_i v_i^T$.

$B^2 = (\sum \sigma_i^2 v_i v_i^T)(\sum \sigma_i^2 v_i v_i^T) = \sum \sigma_i^4 v_i v_i^T$.

...

$B^k = \sum \sigma_i^{2k} v_i v_i^T$.

If $\sigma_1 > \sigma_2$ & k large enough, then, $B^k \to \sigma_1^{2k} v_1 v_1^T$.

$\Rightarrow$ B^k's 1st 列 $b \approx c \cdot v_1$. (c is Const.).

$\Rightarrow$ Output $\frac{b}{\|b\|}$ as v_1.

② Compute B^k too expansive!

Idea: matrix · matrix $O(d^3)$ → matrix · vector $O(d^2)$.

Choose a random vector $x \in \mathbb{R}^d$.

then compute $B^k x = B^{k-1} \cdot (Bx)$.

Then, $B^k x \approx c \cdot v_1$

[直观-Pf: Assume $x = \sum_i^d a_i v_i$

$\Rightarrow B^k x = (\sum \sigma_i^{2k} v_i v_i^T)(\sum a_i v_i) = \sum \sigma_i^{2k} a_i v_i$

$b \to \sigma_1^{2k} a_1 v_1 = c' v_1$.

($\sigma_1 > \sigma_2$, k足够大.)]

日期: /

严格地有 Thm:

Let $A \in \mathbb{R}^{n\times d}$. x is a vector s.t. $\|x\|=1$ & $|x^T v_1| \geq \delta$ (→即 α_1.)

Let V be the space spanned by right singular vectors of A (→即 $V = \text{span}(v \mid \sigma(v) \to \sigma(v_1))$.)

corresponding to singular values σ_i s.t. $\sigma_i \geq (1-\varepsilon)\sigma_1$.

Let w be normalized vector after $k = \frac{\ln(1/\varepsilon\delta)}{2\varepsilon}$ iterations

of power method. i.e. $w = \frac{B^k x}{\|B^k x\|}$ (→归一化)

Then w has a component of at most ε perpendicular to V, (垂直)

i.e. $\|w - \text{proj}(w, V)\| \leq \varepsilon$.

(Rmk: 即. 若 $\varepsilon \to 0$, then $w \to V$, 而 $V \approx \text{span}(v \mid \sigma(v) \to \sigma(v_1))$.).

(Rmk: Pf method 在 spectual cluster 等中经常使用.)

Pf: Let $A = \sum^r \sigma_i u_i v_i^T = \sum^d \sigma_i u_i v_i^T$, 其中 $\sigma_{r+1} = \cdots = \sigma_d = 0$.

$$x = \sum^d \alpha_i v_i$$

then $B^k x = \sum^d \sigma_i^{2k} \alpha_i v_i$.

Suppose $\sigma_1, \cdots, \sigma_m \geq (1-\varepsilon)\sigma_1$, $\sigma_{m+1} < (1-\varepsilon)\sigma_1$.

Consider: $\|B^k x\|^2 = \sum^d \sigma^{4k} \alpha_i^2$

$$\geq \sigma_1^{4k} \cdot \alpha_1^2 = \sigma_1^{4k} \cdot \delta^2.$$

then, $\|w - \text{proj}(w, V)\|^2 = \frac{\|\sum_{i=m+1}^d \sigma_i^{2k} \alpha_i v_i\|^2}{\|B^k x\|^2} < \frac{(1-\varepsilon)^{4k} \sigma_1^{4k} \sum_{i=m+1}^d \alpha_i^2}{\sigma_1^{4k} \delta^2} \leq \frac{(1-\varepsilon)^{4k}}{\delta^2}$ (→$\leq \sum_{i=1}^d \alpha_i^2 = \|x\| = 1$)

日期: /

we want $\frac{(1-\varepsilon)^{4k}}{\delta^2} \le \varepsilon^2$.

$\Rightarrow k \ge \frac{\ln(1/\varepsilon\delta)}{2\varepsilon}$. □.

$1+\varepsilon \le e^{\varepsilon}$.

Rmk: 可通过 larger δ s.t. k 不过大 for some ε.

↳ How to find proper x:

1) Take k samples $\sim N(0,1)$: $x_1, \cdots, x_d$. 令 $x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$.

2) Normalize x. s.t. $\|x\|=1$. $\Rightarrow$ random projection.

$x^T v_i \sim N(0,1)$.

(以较大概率 s.t. $|x^T v_i| \ge \delta$.)

2. Approximate first k singular vectors:

Method 1:

1) select random vector r.

2) define $u_1, \cdots u_k, v_1, \cdots v_k$. as:

$u_1 \leftarrow r$, $u_2 \leftarrow Br$, $\cdots$, $u_k \leftarrow B^{k-1} r$.

$v_1, \cdots, v_k \leftarrow u_1, \cdots u_k$ 作施密特正交化.

3) Repeat.

update $u_1, \cdots u_k, v_1, \cdots v_k$. as:

$u_1 \leftarrow v_1$, $u_2 \leftarrow Bv_2$, $\cdots$, $u_k \leftarrow Bv_k$.

$v_1, \cdots, v_k \leftarrow u_1, \cdots u_k$ 作施密特正交化.

日期: /

Method 2:

1) select $d \times k$ matrix V s.t. $V^TV = I_k$. define $Z \leftarrow 0$.

2) Repeat:

update $Z \leftarrow BV$.

find Z 的 QR decomposition. (Q: 政阵. R: 上三角).

update $V \leftarrow Q$.

Method 3:

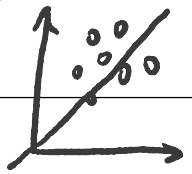
1). select w_1. as above.

2) Consider $A - \tilde{\sigma}_1 \tilde{u}_1 w_1$:

select w_2 as above.

Repeat.

IV. Principal Component Analysis (PCA)

之前考虑的是  $\max_{\|v\|=1} \|Av\|$.

best-fitting k-dim subspace.

Now Consider:

best-fitting k-dim affine subspace.

即 $\max_{\|v\|=1} \|[A - (\frac{1}{1})\cdot \vec{r}]v\|$.

Fact: best-fitting k-dim affine subspace. 必过它们的中心

日期: /

PCA Alg:

1) Input $A \in \mathbb{R}^{n \times d}$. (记 $A=(a_{ij})$.)

2) 令 $\bar{a}_j = \frac{1}{n}\sum_{i=1}^{n} a_{ij}$. $A' = A - \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} \cdot (\bar{a}_1, \cdots, \bar{a}_d)$. (中心化).

3) Compute A'. SVD.

↓ 中心化 ↓ 低秩近似(SVD) ↓

V. Clustering mixtures of Gaussian distributions.

↳ 基本分布的带权求和，本身也是一种分布.

$f = w_1 p_1 + \cdots + w_k p_k$. , $w_i > 0$, $\sum w_i = 1$.

Problem: find w_i, $\mu(p_i)_{(i=1,\cdots,k)}$, σ to fit n i.i.d. (f). samples.

(恢复类 ⇒ 每个类分别恢复Gaussian).

Assume $p_1, \cdots, p_k$ spherical. Gaussian (但未知 μ, σ^2)

TASK: ① Cluster the set of samples into k clusters $C_1, \cdots, C_k$.

② for $\forall$ clusters C_i 求 p_i.

Alg: ① for $\forall$ sample pair $\|x-y\|^2$.

② put x, y in same cluster if $\|x-y\|^2 \le T$.

2 Case:

i. if 2 samples x, y from same spherical. Gaussian with σ^2.

日期: /

then $\sup \|x-y\| \approx 2(\sqrt{d}\pm O(1))^2 \cdot \sigma^2$ (:=A)

II. if 2 samples x,y from diff. spherical Guassian with σ^2, $\Delta\mu$.

then $\inf \|x-y\| \approx 2(\sqrt{d}\pm O(1))^2 \cdot \sigma^2 + \Delta^2$ (:=B).

为了区分 x,y $\Rightarrow$ 要求 B>A. (否则无法存在分类).

$\Rightarrow$ $\Delta > C \cdot d^{1/4} \cdot \sigma$ (C is some Const).

↓

与d有关. 可能太强]. (考虑降维, s.t. 只要 $\Delta > C'\sigma$ 即可).

An SVD-based alg.: (Input: n samples w.p.d. f.).

① Compute A's SVD to find first k singular vectors, project n samples to span($v_1, \cdots, v_k$).

② Cluster them in span($v_1, \cdots, v_k$).

Thm: The above alg. can seperate k different Guassians with $\Delta > ck^{1/4}\sigma$, whp (不证).

一些直观: ① the set of samples can 近似 prob. density f.

② for the prob. density f, the best-fitting subspace contains the k centers. ?

(Rmk: $V_k = \underset{V: k\text{-dim}}{\operatorname{argmax}} E\|\text{proj}(x,V)\|$ $x\sim f$).

(Cor: μ_i 在投影下不变).

日期: /

③ the projection of a spherical Guassion remains the variance σ^2. ?

⇒ μ_i 与 G 都不变.

⇒ 从而只要 $\Delta > c \cdot k^{1/4} \cdot \sigma$.

Lecture 2. Random Projection & JL Lemma.

Thm 1: JL Lemma:

$\forall\ 0 < \varepsilon < 1$ (精度). $n > 0 \in \mathbb{Z}$. let $k = O(\frac{c \ln n}{\varepsilon^2})$ (即 k 不太小). for some Const. c.

$\forall$ set of n points $\{a_1, \dots, a_n\}$ in $\mathbb{R}^d$. Define random projection:

$f: \mathbb{R}^d \to \mathbb{R}^k$. $f(a) = \frac{1}{\sqrt{k}}(u_1^T a, \dots, u_k^T a)^T$. $\left(= \frac{1}{\sqrt{k}} F \cdot a,\ F = \begin{pmatrix} u_1^T \\ \vdots \\ u_k^T \end{pmatrix} \in \mathbb{R}^{k \times d}\right)$

where u_i is a d-dim vectors $\sim$ Guassian (0,1) at each coordinate.

Then w.p. $\geq 1 - \frac{c}{n}$. (for some fixed Const c) $\forall (a_i, a_j)$ (注意: 整体 Prob.).

$(1-\varepsilon)\|a_i - a_j\| \leq \|f(a_i) - f(a_j)\| \leq (1+\varepsilon)\|a_i - a_j\|$. whp. dist$(a_i, a_j)$ 变化不太大.

Thm 2: Random-Projection Thm

$\exists$ Const $c > 0$. s.t. for $\varepsilon \in (0,1)$. $\forall a$,

$P[\ |\|F(a)\| - \sqrt{k}\|a\|| \geq \varepsilon \sqrt{k}\|a\|\ ] \leq 3e^{-ck\varepsilon^2}$. whp. $\|a\|$ 变化不太大.

日期: /

Thm2 ⇒ Thm1: 取 $k=\frac{3\ln n}{c\varepsilon^2}$, $a=a_i-a_j$ 代入 Thm2:

s.t. $P\left[\left|\frac{1}{\sqrt{k}}\|F(a_i-a_j)\|-\|a_i-a_j\|\right|\geq\varepsilon\cdot\|a_i-a_j\|\right]\leq\frac{3}{n^3}$

by union bound,

w.p. $1-\frac{3}{2n}$ 有 Thm1 结论.

Thm3: Guassian Annulus (环) Thm.

For d-dim spherical Guassian with variance 1 各维,

for $\forall\ \beta\leq\sqrt{d}$. 至少 $1-3e^{-c\beta^2}$ of the prob. mass lie in the annulus

$\sqrt{d}-\beta\leq\|x\|\leq\sqrt{d}+\beta$. for fixed const. $c>0$. ($\approx\frac{1}{96}$?)

(Rmk: 1° for d-dim spherical Guassian, 大部分 mass 集中于中心. ($\sqrt{d}$ 两边).

2° for d-dim ball, $d\to\infty$, mass 集中于表面.)

归一化 ⇒ $\forall\ \beta\leq1$, w.p. $\geq 1-3e^{-c\beta^2 d}$.

s.t. $1-\beta\leq\|x\|\leq1+\beta$.

fix β. 当 $d\to\infty$. whp $\|x\|-1|\leq\beta$. 即 mass 集中于表面.

whp. 在中心. ↓ whp. 归一化后 $\|x\|\approx1$ 即集中于表面.

Thm3 ⇒ Thm2: WLOG assume $\|a\|=1$.

由 $a\cdot u_i=\sum_{j=1}^{d}a_j\cdot u_{ij}$. $u_{ij}\sim N(0,1)$.

$\therefore\ a\cdot u_i\sim N(0,\sum a_j^2 Var(u_{ij}))=N(0,\sum a_j^2)=N(0,1)$.

⇒ $F(a)\sim$ k-dim spherical Guassian with variance 1 各维.

by Thm3: 令 $\beta=\varepsilon\sqrt{k}$. then.

$(1-\varepsilon)\sqrt{k}\leq\|F(a)\|\leq(1+\varepsilon)\sqrt{k}$ w.p. $\leq 3e^{-ck\varepsilon^2}$. (注意: $\|a\|=1$).

日期: /

Thm3-Pf: (when $\beta \le \frac{\sqrt{d}}{5}$).

Assume $X=(x_1,\cdots,x_n)$. $x_i \sim N(0,1)$. let $r=\|X\|$.

$\sqrt{d}-\beta \le \|X\| \le \sqrt{d}+\beta \iff |r-\sqrt{d}| \le \beta$.

$\therefore$ Thm3 $\iff P[|r-\sqrt{d}| \ge \beta] \le 3e^{-c\cdot\beta^2}$.

$= P[|r^2-d| \ge \beta(r+\sqrt{d})]$

↓ 放缩.

只要证 $P[|r^2-d| \ge \beta\sqrt{d}] \le 3e^{-c\beta^2}$.

由 $r^2=\|X\|^2=\sum x_i^2 \sim \chi^2(d)$.

(Fact: $g_{\chi^2(d)}(t)=(1-2t)^{-\frac{d}{2}}$. $\boxed{g_{N(0,1)}=(1-2t)^{-\frac{1}{2}}}$)

$P[r^2 \ge d+\beta\sqrt{d}] = P[e^{\lambda r^2} \ge e^{\lambda(d+\beta\sqrt{d})}] \overset{\text{Markov}}{\le} \frac{E[e^{\lambda X}]}{e^{\lambda(d+\beta\sqrt{d})}}$

↘ positive. ↗ $(\lambda>0)$.

(矩母出数上用Markov).

$= \frac{\prod E[e^{\lambda x_i^2}]}{e^{\lambda(d+\beta\sqrt{d})}}$.

$= \frac{(1-2\lambda)^{-\frac{d}{2}}}{e^{\lambda(d+\beta\sqrt{d})}}$

$= e^{-\frac{d}{2}\ln(1-2\lambda)-\lambda(d+\beta\sqrt{d})}$

$\le e^{d(\lambda^2+2\lambda^3)-\lambda\beta\sqrt{d}}$. (对$\forall \lambda>0$).

($\ln(1-x) \ge x-\frac{x^2}{2}-\frac{x^3}{2}$, $\forall 0<x<\frac{1}{5}$)

取 $\lambda=\frac{\beta}{2\sqrt{d}}$, 即有 $P[x \ge d+\beta\sqrt{d}] \le e^{-\frac{\beta^2}{5}}$

$P[x \le d-\beta\sqrt{d}]$ 类似可证. 口.

(或者用 Master Tail Bound 可证).

日期: /

Rmk On JL Lemma:

① JL Lemma 针对 all pairs of a_i, a_j. (whp)

② 可降维至 $k=O(\frac{\ln n}{\varepsilon^2})$. (与d无关).

(data set 多会影响正确率).

③ Random Projection function f 与 data set (x_i) 无关.

④ Running Time: $O(n \cdot d \cdot k) \approx O(n \cdot d \cdot \frac{\ln n}{\varepsilon^2})$.

⑤ $f: \mathbb{R}^d \to \mathbb{R}^k$. bijection. w.h.p.

More Extension:

① [Ailon, Chazelle]. gives a new projection f. s.t.

each projection $f(a_i)$ takes $O(d \log d + \frac{(\ln n)^{3+O(1)}}{\varepsilon^2})$.

② [Achlioptas] Guassian → ±1 r.v.

即令 $f_{ij} = \begin{cases} 1 & \text{wp. } \frac{1}{2} \\ -1 & \text{wp. } \frac{1}{2} \end{cases}$ 或者 $f_{ij} = \begin{cases} 1 & \text{wp. } \frac{1}{6} \\ 0 & \text{wp. } \frac{2}{3} \\ -1 & \text{wp. } \frac{1}{6} \end{cases}$

(rmk: 归一化因子须对应修改).

③ [Larsen, Nelson].

$O(\frac{\ln n}{\varepsilon^2})$ 是 k 的 lower bound. (即不可再改进).

④ [Nelson] Strong JL Lemma:

for all $x \in A$, for all $y \in R^d$, (JL: for all $x, y \in A$).

$(1-\varepsilon)\|x-y\| \le \|f(x)-f(y)\| \le (1+\varepsilon)\|x-y\|$.

日期: /

Lecture 3. Nearest Neighbor Search (NNS).

Input: a. n points ($\in D^d$) database.
↳ (domain: $\mathbb{R}/\mathbb{Z}/[r]/\cdots$).

Query: x.

Problem: find nearest point a^* to x. i.e. $a^* = \underset{a \in \text{database}}{\operatorname{argmin}}\, d(x,a)$.

(Rmk: Hamming distance: (for 0-1 bits).
the number of positions with different values.).
(此时. $\Leftrightarrow \|x-y\|_1 = \Sigma|x_i-y_i|$.).

Solution:

① Preprocessing: process database and build a data structure.

② Query: upon query x, answer a^*.

Goal: ① query time: 尽量快 (ideally poly(logn)).
② used space for data structure: 尽量小.
③ 预处理时间可接受 (poly(n,d)).

Exact Solution:

curse of demensionality:
if beats $O(d \cdot n)$ query time requires $\Omega(2^d)$ space.

(Rmk: Exact solution isn't proper!).

日期: /

Approximate Solution:

Problem1: C-Approximate Nearest Neighbor Search. (C-ANN)

Goal: output a^* s.t. $d(x,a^*) \le c \cdot \min_i d(x,a_i)$. $(c>1)$

Problem 2: (C,r)-ANN.

Promise: $\exists a_{i_0}$ s.t. $d(x,a_{i_0}) \le r$. then.

Goal: output a_i st $d(x,a_i) \le c \cdot r$. (否则output no).

若得no, 则随机output a_i.
此时. $d(x,a_i) \ge \frac{1}{c-1} D_{max}$. $\forall i$. $\Rightarrow (c-1)d \ge D_{max}$.
有 $d(x,a_i) \le d(x,a_j) + d(a_i,a_j) \le c \cdot d$.
即$\forall a_i$均符合要求. $= d$.

等比搜索(c).

From (c,r)-ANN. to c-ANN:

① Preprocess: $D_{max} = \max(a_i,a_j)$. $D_{min} = \min(a_i,a_j)$. $R = \frac{D_{max}}{D_{min}} \cdot \frac{1}{c-1}$.

② Solve (c,r)-ANN for $r_0 = D_{min}$, $r_1 = c \cdot D_{min}$, $\cdots$, $r_{\log_c R} = \frac{1}{c-1} \cdot D_{max}$. (G.P. Search).

Method 1: (针对 Hamming distance: Ham(x,y)).

Thm: $\exists$ a data structure uses space $n^{O(1)}$ (p > in fact: $dn^{O(1)}$), query time $O(d \cdot \ln n)$,

solves (c,r)-ANN w.h.p.

Alg: (Use dimension reduction).

1) Parameters: $k = \frac{\log n}{(\frac{1}{8} - 2^{-(c+2)})^2}$. $p = \frac{1}{2} - (\frac{1}{2})^{1+\frac{1}{r}}$. $s = (\frac{3}{8} - 2^{-(c+2)}) \cdot k$.

2) Sample a $k \times d$ matrix $U \in \mathbb{F}_2^{k \times d}$:

日期: /　　思路：在低维上事先枚举query.(2^k).

$U_{ij} = \begin{cases} 1 & wp.\ p \\ 0 & wp.\ 1-p \end{cases}$ i.i.d.　　⇒ JL lemma's variant

3) for each $i \le n$, compute:

$z_i = U \cdot a_i \in \{0,1\}^k$.

4) store all s-balls: $B_s(u) = \{a_i \mid Ham(u, z_i) \le s\}$ for $\forall u \in \{0,1\}^k$.
(共2^k个).

5) Upon a query $x \in \{0,1\}^d$:

compute Ux:

if $B_s(Ux) \neq \emptyset$, return any $a_i \in B_s(Ux)$.
if $B_s(Ux) = \emptyset$, return no.

Analysis:

$P[U_i x \neq U_i y] = U_i(x-y) \neq 0$.

① Space: 2^k个 s-balls. 每个 $n \times d$: $O(nd \cdot 2^k) = n^{O(1)}$.

② Time: $O(k \cdot d) = O(d \cdot \ln n)$

③ Correctness:

Lemma:

$\forall x, y \in \{0,1\}^d$.
$Ham(x,y) \le r \Rightarrow P[Ham(Ux, Uy) \ge s] < e^{-\Omega(k)}$.
$Ham(x,y) \ge c \cdot r \Rightarrow P[Ham(Ux, Uy) < s] < e^{-\Omega(k)}$.

Pf: $Ham(Ux, Uy) = \sum_{i=1}^{k} x_i$. where $x_i = \begin{cases} 1, & (Ux)_i \neq (Uy)_i \\ 0, & (Ux)_i = (Uy)_i \end{cases}$.

Lem: $P[(Ux)_i \neq (Uy)_i] = \frac{1}{2}(1 - (1-2p)^{Ham(x,y)}) = \frac{1}{2}(1 - (\frac{1}{2})^{Ham(x,y)/r})$.
(由递推)

日期: /

Apply Chernoff Bound:

$$P[Ham(U_x U_y) \geq S] = P[\sum_{i=1}^{k} X_i \geq S] \quad ?$$

$$\leq P[|\sum_{i=1}^{k} X_i - \mu k| \geq s - ku]$$

$$\leq 3e$$

对比: Method 1: 降维有限 $(\log n)$. 精度 w.h.p. 一次到位(故枚举 2^k query) w.h.p. $dn^{O(1)}$ space

日期: Method 2: 降维任意. 精度常数. L次到位(故建立n点的hash表). w.p. ≥ 0.6 $n^{1+\frac{1}{c}}$ space

Me 思路: LSH + 查询至多L次, 4L个.

Method 2. (基于 LSH: Locality Sensitive Hashing). (针对 Hamming).

Goal: $\exists$ a randomized data structure solves (c,r)-ANN w.p. ≥ 0.6.

space: $O(n^{1+\frac{1}{c}} + nd)$ query time: $O(d \cdot n^{1/c})$.

Def: (LSH): random function: $h: \{0,1\}^d \to \{0,1\}^k$ is (p_1, p_2)-LSH. if:

1) if $\text{Ham}(x,y) \leq r$. then $P[h(x)=h(y)] \geq p_1$ (p_1 not-so-small).

2) if $\text{Ham}(x,y) \geq c \cdot r$. then $P[h(x)=h(y)] \leq p_2$. (p_2 small).

Def: (Hash family $\mathcal{H}$) ($g \in \mathcal{H}$ is called LSH)

$\mathcal{H} = \{g: \{0,1\}^d \to \{0,1\}^k\}$. for $p \in \{0,1\}^d$. $g(p) = \{p_{i_1}, \cdots, p_{i_k}\}$.

where i_j is independently u.a.r. chosen from $[d]$.

def h_j: $h_j(p) = p_{i_j}$. i_j u.a.r. chosen from $[d]$. then $g(p) = (h_1(p), \cdots, h_k(p))$.

def: parameter $\rho = \frac{\log(1/p_1)}{\log(1/p_2)}$.

Lemma: $\rho_g = \rho_h \leq \frac{1}{c}$. (Rmk: $P_{1,g} = P_{1,h}^k$, $P_{2,g} = P_{2,h}^k$).

Pf: $P_{1,g} = \prod_{i=1}^{k} P[h_i(p) = h_i(q)] = P_{1,h}^k$. if $\text{Ham}(p,q) \leq r$.

$P_{2,g} = \prod_{i=1}^{k} P[h_i(p) = h_i(q)] = P_{2,h}^k$. if $\text{Ham}(p,q) \geq c \cdot r$.

$\therefore \rho_g = \frac{\log(1/p_{1,h}^k)}{\log(1/p_{2,h}^k)} = \frac{\log(1/p_{1,h})}{\log(1/p_{2,h})} = \rho_h$.

又 $\forall p,q \in \{0,1\}^d$ $\therefore P[h(p)=h(q)] = \frac{d - \text{Ham}(p,q)}{d}$.

$\therefore \text{Ham}(p,q) \leq r \Rightarrow P[h(p)=h(q)] \geq 1 - \frac{r}{d} = P_{1,h}$.

日期: /

$Ham(p,q) \geq cr \Rightarrow P[h(p)=h(q)] \leq 1-\frac{cr}{d} = P_{2,h}$

$\therefore \rho_h \leq \frac{-\log(1-r/d)}{-\log(1-cr/d)} \leq \frac{1}{c}$ (as: $(1-\frac{r}{d})^c \geq 1-\frac{cr}{d}\ \forall c>1$).

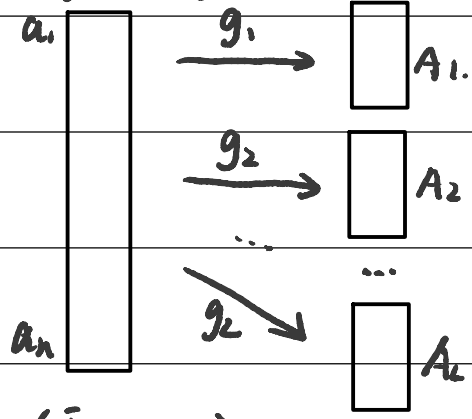
Alg.

Preprocessing:

1) set $k = \log n / \log(1/p_{2,h})$, $L = 2n^{\rho}$ where $\rho = \rho_g (\leq \frac{1}{c})$. (Lemma)

2) Allocate L hash tables $A_1, \cdots, A_L$, each a Hamming LSH g_i $(i \in [L])$.

3) Hash $\{a_1, \cdots, a_n\}$ into tables $A_1, \cdots, A_L$.



Query: (Input x)

顺序遍历 $A_1[g_1(x)], \cdots, A_L[g_L(x)]$, return first a s.t. $d(x,a) \leq cr$.

if no such a is found before $4L = 8n^{\rho}$ elements, output no.

Thm: The alg solves (c,r)-ANN w.p. ≥ 0.6.

Pf: ① if a^* is a point s.t. $Ham(x,a^*) \leq r$.

若有 $\leq r$ 的 a^*, w.p. $< \frac{1}{3}$ 存在其中.

则 $P[a^* \notin A_1[g_1(x)] \cup \cdots \cup A_L[g_L(x)]]$

$= \prod_{i=1}^{L} P[a^* \notin A_i[g_i(x)]]$

日期: /

$= \prod_{i=1}^{L} P[g_i(x) \neq g_i(a^*)]$

$\leq (1-p_{1,g})^L$.

ρ's origin.

Consider, $p_{1,g} = p_{1,h}^k = p_{1,h}^{\log n / \log(1/p_{2,h})} = n^{-\log(1/p_{1,h})/\log(1/p_{2,h})} = n^{-\rho}$.

从而 $P[a^* \in A_1[g_1(x)] \cup \cdots \cup A_L[g_L(x)]] \leq (1-n^{-\rho})^L = (1-\frac{1}{n^\rho})^{2n^\rho} \leq (\frac{1}{e})^2 < \frac{1}{7}$.

L's origin. (erase n's impact).

② let Y be the number of pairs (a, j). s.t.

$d(a,x) > cr$ 且 $g_j(a) = g_j(x)$. 即坏点数 w.p. $\leq \frac{1}{4} \cdot \geq 4L$.

(Rmk: 即放在 L tables 中的 distance far 的 points 的 num).

Pf: $E[Y] \leq L \cdot n \cdot P[g_j(a) = g_j(x) \text{ for some fixed } j]$.

union.

$= n \cdot L \cdot p_{2,g}$.

Parameters: $\Rightarrow p_{2,g} \Rightarrow \rho \Rightarrow L$.

$= L$. ← $p_{2,g}$'s value's origin. (erase n's impact).

$p_{2,g} = \frac{1}{n} \cdot \searrow$

$k = \frac{\log(1/n)}{\log(1/p_{2,h})}$.

By Markov, $P[Y \geq 4L] = P[Y \geq 4EY] \leq \frac{1}{4}$.

即 w.p. $\leq \frac{1}{4}$, $Y \geq 4L$.

③ Assume $\exists\, a^*$ s.t. $Ham(a^*, x) \leq r$.

by ①, ② w.p. $\geq 1 - \frac{1}{7} - \frac{1}{4} > 0.6$.

alg finds a s.t. $d(x,a) \leq c \cdot r$.

(Rmk: 无法保证 find a s.t. $d(x,a) \leq r$.)

Thm: space: $O(n^{1+\frac{1}{c}} + nd)$. (↑ $n^{1/c}$ 表, → n 点.) query time: $O(d \cdot n^{1/c})$. (→ compute Hamming) $L = n^\rho$.

§2

Streaming Algorithom

日期: /

Lecture 4: Counting.

(Rmk: the important point is space).

Problem: A sequence of events, at any given time, output the number of events.

Goal: find a data structure maintaing a value n and supporting:

① update: ② query:

Aim: Use as small space as possible.

Alg:

1. Exact Solutions:

maintain n. Space $\Theta(\log n)$.

2. Approximate Solutions:

Goal: Maintain $\tilde{n}$ s.t.

$P[|n-\tilde{n}|>\varepsilon \cdot n]<\delta$ $\quad \varepsilon, \delta \in (0,1)$. $\Rightarrow$ w.p. $1-\delta$, $(1-\varepsilon)n \le \tilde{n} \le (1+\varepsilon)n$.

① Morris Alg:

1) $X \leftarrow 0$.

2) For each update, increment x by 1 w.p. $\frac{1}{2^x}$.

3) For query, output $\tilde{n} = 2^x - 1$.

Space: $O(\log\log n)$.

↑ 期望上.

Jensen Ineq:

$2^{E[X_n]} \le E[2^{X_n}] = n+1$.

日期: /

$\Rightarrow EX_n = O(\log n)$.

$\xrightarrow{\text{Jesen}} E\log X_n = O(\log\log n)$.

Correctness Analysis:

let X_n be X after n update.

(Aim: $E[X_n]$. $Var[X_n]$.)

Lemma1: $E[2^{X_n}] = n+1$. $\Rightarrow E[\hat{n}] = n$

Pf: By Induction.

$$E[2^{X_{n+1}}] = \sum P[X_n=j]\cdot E[2^{X_{n+1}}|X_n=j].$$

$$= \sum P[X_n=j]\cdot\left[2^j\cdot(1-\frac{1}{2^j}) + \frac{1}{2^j}\cdot 2^{j+1}\right].$$

$$= \sum P[X_n=j]\,2^j + \sum P[X_n=j].$$

$$= E[2^{X_n}] + 1 = n+2. \quad \square.$$

(↑ induction.)

Lemma2: $Var[\hat{n}]$.

$Var[\hat{n}] = E[(2^{X_n}-1-n)^2]$. $\left(E[2^{2X_n}] = \frac{3}{2}n^2+\frac{3}{2}n+1.\right)$

$$= \frac{n^2}{2} - \frac{n}{2} < \frac{n^2}{2}$$

By Chebyshev's Ineq:

$$P[|\hat{n}-n| > \varepsilon n] \le \frac{Var[\hat{n}]}{\varepsilon^2 n^2} < \frac{1}{2\varepsilon^2}.$$

(太大. 不能 s.t. $\varepsilon, \delta \in (0,1)$ 且可任意小. (即 δ 最好与 ε 无关).)

$\Rightarrow$ Bound 不够. 改进:

② Morris + alg:

日期: /

① Independently run s Morris alg. $(s \geq \frac{1}{2\delta\varepsilon^2})$.

② Output $\hat{n} = \frac{1}{s}\sum_{i=1}^{s}\tilde{n}_i$. Space: $O(\frac{1}{2\delta\varepsilon^2}\log\log n)$.

Correctness Analysis:

$E[\hat{n}] = E[\frac{1}{s}\sum_{i=1}^{s}\tilde{n}_i] = n$.

w.p. $\frac{1}{(1+2)^x}\uparrow$ $\tilde{n} = \frac{(1+2)^x - 1}{2-1}$

$Var(\hat{n}) = \frac{1}{s^2}\sum_{i=1}^{s}Var(\tilde{n}_i) < \frac{1}{s^2}\cdot s\cdot\frac{n^2}{2} < \frac{n^2}{2s}$ $\quad \frac{2n^2}{2s}$

By Chebyshev's Ineq:

$P[|\hat{n}-n| > \varepsilon n] \leq \frac{Var[\hat{n}]}{\varepsilon^2 n^2} < \frac{1}{2\varepsilon^2 s}$. $\quad \frac{2}{2\varepsilon^2 s}$

let $s \geq \frac{1}{2\delta\varepsilon^2}$. then $P[|\hat{n}-n| > \varepsilon n] < \delta$. □.

③ Morris ++ alg: $\quad s = \frac{32}{2\varepsilon^2}$

① let $s = \frac{1}{2\cdot\frac{1}{3}\varepsilon^2}$. Independently run t Morris + alg. $(t \geq 18\ln(1/\delta))$.

② Output the median of $\tilde{n}_1, \cdots, \tilde{n}_t$.

Analysis:

def $Y_i = \begin{cases} 1, & (1-\varepsilon)n \leq \tilde{n}_i \leq (1+\varepsilon)n \\ 0, & \text{otherwise.} \end{cases}$ $\Rightarrow E[Y_i] = P[Y_i] \geq \frac{2}{3}$ 示性变量

def $Y = \sum_{i=1}^{t} Y_i$

$E[Y] = \sum_{i=1}^{t} E[Y_i] \geq \frac{2}{3}t$

$P[\hat{n} \text{ bad}] \leq P[Y \leq \frac{t}{2}] \leq P[Y - E[Y] < -\frac{t}{6}]$.

Chernoff $\leq e^{-2(\frac{1}{6})^2 t}$.

计算 $E[\sum Y_i]$.

取中位数.

转化为 Chernoff.

日期: /

let $t \geq 18\ln(1/\delta) \Rightarrow P[\hat{n} \text{ bad}] \leq \delta$. □.

Space: $\Theta(s \cdot t \cdot \text{space}(\tilde{x})) = \Theta\left(\frac{\ln(1/\delta)}{\varepsilon^2} \text{space}(\tilde{x})\right)$.

If $\tilde{x}$ reaches $\log(stn/\delta)$, then at any time, $\tilde{x}$ increase wp. $\frac{1}{2^{\tilde{x}}} \leq \frac{\delta}{stn}$

then $P[\tilde{x} \text{ increase in the next } n \text{ updates}] \leq \frac{\delta}{stn} \cdot n = \frac{\delta}{st}$

By Union Bound, wp. $\geq 1-\delta$,

$\tilde{x_i} \leq \log(stn/\delta)$, $i=1,\cdots,st$. $(stn/\delta = \frac{n}{\varepsilon^2\delta}\log(\frac{1}{\delta}) \leq \frac{n^2}{\varepsilon^2\delta^2})$

$\Rightarrow$ Space $\leq O(st \cdot \log\log(stn/\delta)) = O\left(\frac{1}{\varepsilon^2}\log(\frac{1}{\delta})\log\log(\frac{n}{\varepsilon\delta})\right)$. wp. $\geq 1-\delta$.

其他估计: $2st \log\log(2st^n/\delta)$. $(2+1)$

Flajolet 85: $O(\log(1/\varepsilon) + \log\log n + \log(1/\delta))$.

Nelson 2020: $\Theta(\log\frac{1}{\varepsilon} + \log\log n + \log\log(1/\delta))$ (需要对 Morris++ 稍作改进)

Lecture 5: Reservoir Sampling.

Input: A Sequence of numbers $a_1, a_2, \cdots, a_m$ (未知总数) revealed over time.

Problem: Output a uniform sample s from all the revealed numbers.

Algorithm: ① 维护一个数字 s. initially let $s=0$.

② on seeing the m-th number a_m, then set $s = a_m$ wp. $\frac{1}{m}$,

日期: /

else, maintain $s=s$.

Correctness: $P[s=a_i] = \frac{1}{i}\cdot(1-\frac{1}{i+1})\cdots(1-\frac{1}{m}) = \frac{1}{m}$. $\forall i$.

Space: (If $a_i \in n$, $\forall i$). then $O(\log n) + O(\log m)$ bits.
↓ store n_i. ↳ store m.

(Rmk: ① k samples ⇒ Independently run k 个 Alg.

② m可用 Morris Alg 优化.).

Lecture 6: Estimating the number of distinct elements.

Input: a stream of integers $i_1, \cdots, i_m \in [n] = \{1, \cdots, n\}$.

Output: the number of distinct elements seen in the stream.

(Rmk: denoted as 数据流的一范数问题.).

Exact Solutions:

① store an array $[n]$: $O(n)$

② store all the elements. $m \cdot O(n)$.

Approximate:

Maintain a $\tilde{t}$ st. $P[|\tilde{t}-t| > \varepsilon t] < \delta$. where t is the number of distinct elements in stream $\varepsilon, \delta \in (0,1)$.

Alg 1: Idealed FM alg:

日期: /

① Pick a random function : $h[n] \to [0,1]$.

② Maintain a counter $X = \min_{i \in \text{Stream}} h(i)$.

③ Output $\tilde{t} = 1/X - 1$.

Correctness Pf:

Claim1: $\underline{E[X] = \frac{1}{t+1}}$.

Lem1: $\underline{E[X]} = \int_0^\infty f(y) y \, dy$

$= \int_0^\infty \int_0^y f(y) \, d\lambda \, dy$

$= \int_0^\infty \int_\lambda^\infty f(y) \, dy \, d\lambda$

$= \underline{\int_0^\infty P[X \geq \lambda] \, d\lambda}$.

由Lem1, $E[X] = \int_0^\infty P[X \geq \lambda] \, d\lambda$

$= \int_0^\infty P[\forall a_i \in \text{stream}, h(a_i) \geq \lambda] \, d\lambda$

$= \int_0^\infty \prod P[h(a_i) \geq \lambda] \, d\lambda$.

$= \int_0^1 (1-\lambda)^t \, d\lambda$.

$= \frac{1}{t+1}$.

Claim2: $\underline{Var(X) = \frac{t}{(t+1)^2(t+2)} < \frac{1}{(t+1)^2}}$.

Pf: $E[X^2] = \int_0^\infty P[X^2 \geq \lambda] \, d\lambda$.

$= \int_0^1 (1-\sqrt{\lambda})^t \, d\lambda$.

日期: /

$$= 2\int_0^1 u^t(1-u)\,du = \frac{2}{(t+1)(t+2)}.$$

$$\therefore \mathrm{Var}[X^2] = E[X^2] - (E[X])^2 = \frac{t}{(t+1)^2(t+2)} < \frac{1}{(t+1)^2}.$$

Alg2: FM + Alg. (Mean Trick) → 降低 Var 以获得 Correctness.

① Independently run $s = \frac{25}{\varepsilon^2\delta}$ copies of FM Alg. and obtain counters $X_1, \cdots, X_s$.

② Output $\tilde{t} = \frac{1}{Z} - 1$, where $Z = \frac{1}{s}\sum_{i=1}^{s} X_i$.

Correctness Pf:

Lem1: $P\left[\left|Z - \frac{1}{t+1}\right| > \frac{\varepsilon}{5(t+1)}\right] < \delta$

Pf: $E[Z] = \frac{1}{t+1}$. $\mathrm{Var}[Z] = \frac{1}{s^2}\sum \mathrm{Var}[X_i] < \frac{1}{s(t+1)^2}$

By Chebyshev ineq.

$$P\left[\left|Z - \frac{1}{t+1}\right| > \frac{\varepsilon}{5(t+1)}\right] < \frac{\mathrm{Var}[Z]}{\left(\frac{\varepsilon}{5(t+1)}\right)^2} < \frac{25}{\varepsilon^2 s} = \delta.$$

Claim: $P[|\tilde{t} - t| > \varepsilon t] < \delta$

Pf: if $t = 0$ ✓.

if $t \geq 1$. by Lem 1. w.p. $\geq 1-\delta$,

$$\frac{1-\varepsilon/5}{(t+1)} < Z < \frac{1+\varepsilon/5}{(t+1)}$$

$$\Rightarrow \frac{t-\varepsilon/5}{1+\varepsilon/5} < \tilde{t} = \frac{1}{Z} - 1 < \frac{t+\varepsilon/5}{1-\varepsilon/5}$$

$$\frac{t-\varepsilon/5}{1+\varepsilon/5} > (t-\varepsilon/5)(1-\varepsilon/5) > t - 2\varepsilon t/5.$$

日期: /

另一边同理, $\Rightarrow P(|\hat{t}-t|>\varepsilon t)<\delta$. □.

Alg3. FMtt Alg. (Median Trick) → 降低 space complexity.

① Independently run $q=8\ln(1/\delta)$ copies FMt Alg. with $\delta'=\frac{1}{3}$. Obtain $\hat{t}_1,\cdots,\hat{t}_q$.

② Output median of $\hat{t}_1,\cdots,\hat{t}_q$ as $\hat{t}$.

Correctness Pf:

Let $Y=\sum_{i}^{q} Y_i$ where $Y_i=\begin{cases}1, & |\hat{t}-t|<\varepsilon t\\ 0, & \text{o.w.}\end{cases}$

Note $EY\geq\frac{2}{3}q$. <= Median Trick 对于连续变化的量(如t)均可行

$\therefore P[|\hat{t}-t|>\varepsilon t]\leq P[Y\leq\frac{q}{2}]$

Chernoff ↓

$\leq e^{-2\cdot\frac{1}{6}\cdot q}\leq\delta$. □.

Space: #counters $=O(q\cdot s)=O(\frac{1}{\varepsilon^2}\ln(1/\delta))$.

---

way to get hash functions.

I. k-wise independent hash families:

Def: a family $\mathcal{H}$ of functions mapping $[a]\to[b]$ is k-wise independent if:

$\forall j_1,\cdots,j_k\in[b]$, distinct $i_1,\cdots,i_k\in[a]$ 有

$$P_{h\in\mathcal{H}_k}[h[i_1]=j_1\wedge\cdots\wedge h[i_k]=j_k]=\frac{1}{b^k}.$$

(Rmk: $P[h[i]=j]=\frac{1}{b}$ $\forall i,j$., k-wise independent 随k↑增强).

eg1: $\mathcal{H}=\{\text{all functions: } [a]\to[b]\}$ is k-wise independent.

日期: /

存储 $h \in H_{poly}$ need space $O(a \log b)$.

eg2: let $a=b=q$. where $q=p^r$. p.prime.

$H_{poly} = \{$ polynomials $\in \mathbb{F}_q[x]$ | degree $\le k-1\}$.

H_{poly} is k-wise independent.

存储 $h \in H_{poly}$ need space $O(k \log q)$. (k是数字.)

(Rmk: 用小空间存储 hash 函数).

II. KMV alg (BJKST). (k-minimum values). space: $O(\frac{\log n}{\varepsilon^2})$

① 2-wise independent hash $h:[n] \to [M]$. $M=n^3$.

② Maintain k ($O(\frac{1}{\varepsilon^2})$) smallest hash values of numbers in stream.

③ if $< k$ distinct hash values, output # distinct hash values.

o.w. output $\tilde{t} = k \cdot M / X$. (X is k-th smallest value)

Analysis:

If $\frac{1}{\sqrt{n}} < \varepsilon < \frac{1}{2}$, then w.p. $\ge \frac{2}{3}$. $(1-\varepsilon)t < \tilde{t} < (1+\varepsilon)t$.

(Rmk: use median trick, run $O(\ln \frac{1}{\delta})$次, to boost prob. to $1-\delta$.)

Pf: ① $P[\tilde{t} > (1+\varepsilon)t]$

$= P[X < \frac{kM}{(1+\varepsilon)t}]$.

$= P[Y \ge k]$.

def 指示变量 $Y_i = \begin{cases} 1, & h(a_i) < \frac{kM}{(1+\varepsilon)t} \\ 0, & h(a_i) \ge \frac{kM}{(1+\varepsilon)t} \end{cases}$.

$Y = \sum_{i=1}^{t} Y_i$

Note: $EY_i < \frac{k}{(1+\varepsilon)t}$ $\Rightarrow$ $EY < \frac{k}{1+\varepsilon}$.

$\mathrm{Var}Y < \frac{k}{1+\varepsilon}$.

日期: /

$\leq P[Y-E[Y] \geq k - \frac{k}{1+\varepsilon}]$.

Chebyshev $< \frac{k}{1+\varepsilon} \cdot \frac{(1+\varepsilon)^2}{k^2\varepsilon^2} = \frac{1+\varepsilon}{\varepsilon^2 k} < \frac{1}{6}$.

② $P[\tilde{t} < (1-\varepsilon)t]$

$= P[X > \frac{km}{(1+\varepsilon)t}]$.

$= P[Z < k]$.

...

$< \frac{1}{6}$. □.

def 指标变量 $Z_i = \begin{cases} 1, & h(a_i) < \frac{km}{(1-\varepsilon)t} \\ 0, & h(a_i) \geq \frac{km}{(1-\varepsilon)t} \end{cases}$.

$Z = \sum_{i=1}^{t} Z_i$.

Note: 由 $P[Z_i=1] > \frac{k}{(1+\varepsilon)t} - \frac{1}{m}$.

$\Rightarrow E[Z_i] > \frac{k}{(1-\varepsilon)t} - \frac{1}{m}$

$\geq \frac{(1+\varepsilon)k}{t} - \frac{\varepsilon k}{4t}$ $(m = n^3 > \frac{4}{\varepsilon k})$.

Lecture 7: Frequent items

§1: Finding the majority:

Input: a stream. $a_1, \cdots, a_n \in [m]$.

Problem: Check if $\exists\, i \in [n]$ s.t. appears $> \frac{n}{2}$ times. if yes, output it.

Misra-Gries alg: Space: $O(\log m + \log n)$.

Maintain an ID and a counter C. (initial ID=1, C=0).

if $a = ID \Rightarrow C \leftarrow C+1$.

else if $C=0 \Rightarrow ID \leftarrow a, C \leftarrow 1$.

else $\Rightarrow C \leftarrow C-1$.

(Rmk: 若不$\exists$ majority, 会输出 any number.

日期: /

Way to 排除：再跑一遍stream (以验证.).

§2: Finding frequent items. $O(k(\log n+\log m))$.

Input: a stream. $a_1,\cdots,a_n\in[m]$., $k\geq 1$. (Addition Stream)

Problem: Check if $\exists\ i\in[n]$ s.t. appears $>\frac{m}{k+1}$ times. if yes, output all. (最多k个).

Misra-Gries alg: (General).

Maintain k ID and k couter C. (initial $ID_i=\perp$, $c_i=0$).

if $\exists\ i$ s.t. $a=ID_i$ $\Rightarrow$ $c_i\leftarrow c_i+1$

else if $\exists i$ st $c_i=0$. $\Rightarrow$ $ID_i\leftarrow a$. $c_i\leftarrow 1$

else $c_j\leftarrow c_j-1$. $\forall j$. (decrement-all (DA) step).

(Rmk: the output 不一定都满s.t. condition. 可通过 second pass).

Proof:

Denote f_i as the times i appears in the sream.

$$\tilde{f}_i \text{ as } \begin{cases} c_j, & \text{if } i=ID_j. \\ 0, & \text{o.w.} \end{cases}$$

Thm: for $\forall i\in[n]$, $f_i-\frac{m}{k+1}\leq\tilde{f}_i\leq f_i$. 即要求到 $\frac{m}{2(k+1)}$ 次DA.

Pf: the alg $\Leftrightarrow$ Maintain counter c_i for $\forall i\in[n]$.

日期: /

input i:

if $C_i \neq 0$, $\Rightarrow$ $C_i \leftarrow C_i + 1$

else. if $<k$ unzero counters. $\Rightarrow C_i \leftarrow 1$

else $\Rightarrow C_j \leftarrow C_j - 1$ for $\forall C_j \neq 0$.

$(\tilde{f}_i)$ 对于 C_i input i · input j		$\tilde{f}_i - f_i$
+1	0	0
+1	0	0
0	-1	-1

由上易见, $\tilde{f}_i \leq f_i$. 要证 $f_i - \frac{m}{k+1} \leq \tilde{f}_i$. 只要证最多发生 $\frac{m}{k+1}$ 次 DA steps.

Pf: 设发生了 t 次. 则 $(m-t) - kt \geq 0 \Rightarrow t \leq \frac{m}{k+1}$. □.

势 $\varphi(m) = \sum \tilde{f}_i^{in}$ 势.

Cor: $HH_{k+1}(S) \subseteq H$.

Extension:

denote $HH_{k+1}(\overset{\text{stream}}{\rightarrow} S) = \{j \in [n] : f_j > \frac{1}{k+1} m\}$ Alg output H.

for Misra-Gries alg, $HH_{k+1}(S) \subseteq H$.

Here's an alg, s.t. $HH_{k+1}(S) \subseteq H \subseteq HH_{2(k+1)}(S)$.

(aim: $f_i - \frac{m}{k+1} \leq \tilde{f}_i \leq f_i - \frac{m}{2(k+1)}$.)

§3: turnstile streaming model:

General Model:

Input $i_1, \cdots, i_m \in [n]$.

let $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$. denote each e_j in stream as (i_j, Δ_j). $\Delta_j \in R$.

update: $x_{i_j} \leftarrow x_{i_j} + \Delta_j$.

日期: /

Cash Register Model: $\Delta_j > 0$. (especially $\Delta_j = 1$).

Strict Turnstile Streaming Model: Δ_j arbitrary, but $x \geq 0$. at all time.

Problems: (for General Model).

I: (k, L_1)-point query problem:

query i

⇓ return $\tilde{x}_i$ s.t. $|x_i - \tilde{x}_i| \leq \frac{1}{k}\|x\|_1$. w.h.p.

II: (k, L_1)-heavy hitters:

query.

return a set L s.t. 1) $L = O(k)$. w.h.p.

2) if $|x_i| > \frac{1}{k}\|x\|_1$. then $i \in L$.

Lemma: If A is an alg. for $(3k, L_1)$-point query problem w.p. $\geq 1 - \frac{\delta}{n}$ & w.s. s bits. Then there's an alg. A' solve (k, L_1)-heavy hitters. w.p. $1-\delta$ & w.s. $s + O(k \log n)$ bits.

Proof: Define A' as:

for all $i \in [n]$. use A to answer query (i).

remember $3k$ indices with the $3k$ largest point query values.

Correctness: w.p. $\geq 1 - \frac{\delta}{n} \cdot n = 1 - \delta$.

日期: /

$\forall$ queries (i) return $\tilde{x_i}$, s.t. $|\tilde{x_i} - x_i| \le \frac{\|x\|_1}{3k}$.

then for $\forall i$ s.t. $x_i > \frac{\|x\|_1}{k}$, 有: $\tilde{x_i} > \frac{\|x\|_1}{k} - \frac{\|x\|_1}{3k} = \frac{2\|x\|_1}{3k}$.

$\forall i$ s.t. $x_i \le \frac{\|x\|_1}{3k}$, 有 $\tilde{x_i} \le \frac{\|x\|_1}{3k} + \frac{\|x\|_1}{3k} = \frac{2\|x\|_1}{3k}$.

Note that, $\#\{i: x_i > \frac{\|x\|_1}{3k}\} \le 3k$.

$\Rightarrow \#\{i: \tilde{x_i} > \frac{2\|x\|_1}{3k}\} \le \#\{i: x_i > \frac{\|x\|_1}{3k}\} \le 3k$.

而 $\{i: x_i > \frac{\|x\|_1}{k}\} \subseteq \{i: \tilde{x_i} > \frac{2\|x\|_1}{3k}\}$. □.

Count Min Sketch: (for: $(\varepsilon k, L_1)$-point query).

Let $w > 2k$, $d = \Omega(\log(1/\delta))$ be parameters.

① Choose 2-wise independent hash functions $h_1, \cdots, h_d: [n] \to [w]$.

let $C[l, s] = 0$. $1 \le l \le d$, $1 \le s \le w$.

② for each item $e_t = (i_t, \Delta_t)$ in stream.

$C[l, h_l(i_t)] \leftarrow C[l, h_l(i_t)] + \Delta_t \quad \forall l$

③ for $\forall i \in [n]$, set $\tilde{x_i} = \min_{l \le d} C[l, h_l(i)]$ MIN.

④ for a query i, output $\tilde{x_i}$.

Lemma (for strict turnstile model):

For $\forall$ fixed i, $x_i \le \tilde{x_i}$ and $P[\tilde{x_i} \ge x_i + \|x\|_1/k] \le \delta$.

Pf: ① $\tilde{x_i} = \min_{j \le d} \tilde{x_{ji}} = \min_{j \le d} (x_i + \underbrace{\sum_{i_s \in I_j} x_{i_s}}_{\ge 0}) \ge x_i$.

日期: /

② $\forall i$, $\underset{(l)}{E}[\underbrace{C[l, h_l(i)]}_{:=Z_l}] = x_i + \sum_{j \neq i} Pr[h_l(j)=h_l(i)] \cdot x_j \geq x_i$.

$$= x_i + \frac{1}{w}\sum_{j\neq i} x_j$$

$$< x_i + \frac{\|x\|_1}{2k}.$$

$$\Rightarrow E[Z_l - x_i] < \frac{\|x\|_1}{2k}$$

$$\underset{Z_l - x_i \geq 0}{\overset{\text{Markov}}{\Rightarrow}} Pr[Z_l - x_i > \frac{\|x\|_1}{k}] < \frac{\|x\|_1/2k}{\|x\|_1/k} = \frac{1}{2}.$$

$$\Rightarrow Pr[\tilde{x}_i - x_i > \frac{\|x\|_1}{k}] < (\frac{1}{2})^d < \delta.$$

d是决定误差的参数.

Rmk: let $d = \Theta(\ln n)$. $w = 2k$. then w.p. $\geq 1 - \frac{1}{n}$ s.t. $\tilde{x}_i \leq x_i + \frac{\|x\|_1}{k}$ for some x_i.

Space Complexity: dw counters $= O(k \log n \cdot \log m)$. ($\|x\|_1 \leq m$).

Count Sketch (for (k, L_2)-point query).

Let $w > 3k^2$, $d = 4(\log(1/\delta))$. be parameters.

① Choose 2-wise independent hash functions $h_1, \cdots, h_d : [n] \to [w]$

let $C[l,s] = 0$. $1 \leq l \leq d$, $1 \leq s \leq w$. hash functions $g_1, \cdots, g_d : [n] \to \{\pm 1\}$

② for each item $e_t = (i_t, \Delta_t)$ in sream.

$$C[l, h_l(i_t)] \leftarrow C[l, h_l(i_t)] + \Delta_t \cdot g(i_t).$$

③ for $\forall i \in [n]$, set $\tilde{x}_i = \underset{l \leq d}{\text{median}} C[l, h_l(i_t)] \cdot g(i_t)$. Median.

④ for a query i, output $\tilde{x}_i$.

日期: /

Lemma: for $\forall$ fixed i, ① $E[\tilde{x}_i] = x_i$.

② $Pr[|\tilde{x}_i - x_i| \le \frac{\|x\|_2}{k}] \ge 1-\delta$.

Pf: fix i. ①: let $Z_i = C[i, h_i(i)]$.

$$\Rightarrow EZ_i = x_i \cdot g(i) + \sum_{j \neq i} Pr[h_i(j) = h_i(i)] \cdot x_j \cdot g(j)$$

$$= x_i g(i) + \sum_{j \neq i} \frac{1}{w}\left(\frac{1}{2}x_j - \frac{1}{2}x_j\right)$$

$$= x_i g(i).$$

$$\Rightarrow E[\tilde{x}_i] = x_i.$$

② $EZ_i^2 = x_i^2 + \sum_{j \neq i} Pr[h_i(j) = h_i(i)] \cdot x_j^2 + 2\sum Pr[h_i(j) = h_i(i)] \cdot x_i g(i) \underbrace{E[x_j g(j)]}_{=0}$

$$< x_i^2 + \frac{\|x\|_2^2}{w}.$$

$$\Rightarrow Var Z_i = EZ_i^2 - (EZ_i)^2 < \frac{\|x\|_2^2}{w}.$$

$$Pr[|Z_i - x_i| \ge \frac{\|x\|_2}{k}] \le \frac{\|x\|_2^2/w}{\|x\|_2^2/k^2} = \frac{k^2}{w} = \frac{1}{3}$$

由 Median Trick, $Pr[|\tilde{x}_i - x_i| \ge \frac{\|x\|_2}{k}] \le \delta$.
(手性 + Chernoff).

(Rmk: 记 Count Sketch 为 g, 则 g=linear. 即 $g(x + \Delta_j \vec{I}_j) = g(x) + \Delta_j \cdot g(\vec{I}_j)$.)

Applications:

I. $\Rightarrow$ Point query. $\Rightarrow$ heavy hitters.

II. Range Queries:

日期: /

Count Min Sketch + 线段树

given $i,j\in[n]$, estimate $\sum_{i\le l\le j} x_l$.

Ideal $O(\log(nm))$.

(Rmk: 不能直接 $\tilde{x}_{ij}=\sum_{i\le l\le j}\tilde{x}_l$.)

Def: Dyadic interval:

$[i,j]$ is a dyadic interval/range if:

① $j-i+1=2^k$. ② 2^k divides $i-1$ for some $k\ge 0$.

Alg: (线段树).

① 对线段树每层用 Count Min Sketch. (with $w=2k$, $d=\log n\log 1/\delta$.)

② upon a query $[i,j]$, decomposite it to $[i,j_1],\cdots,[i_s,j]$, $s\le 2\log n$.

③ 对 s个 dyadic intervals, maintain $\tilde{x}_{ij_1},\cdots,\tilde{x}_{j_sj}$.

④ output $\tilde{x}_{ij}=\tilde{x}_{ij_1}+\cdots+\tilde{x}_{j_sj}$.

Space: $O(k\log^2 n\log m)$. update $O(1)$. (均摊). query: $O(\log n)$.

III. Sparse Recovery.

Given a vector $x\in\mathbb{R}^n$ (Streaming model) find z s.t. z has $\le k$ non-zeros and $\|x-z\|_p$ is minimized for some $p\ge 1$.

(Rmk: Offline Model: z picks the largest k entries of x.)

Aim: turnstile streams model. $p=2$. $\tilde{O}(k)=O(k\cdot\log^c k)$ space.

Define: $err_2^k(x)=\min_{z:\|z\|_0\le k}\|x-z\|_2^2$.

日期: /

Thm: The Algorithm use $O(\frac{k}{\varepsilon^2}\text{poly}(\log n))$ spaces and return Z, s.t.

$\|Z\|_0 \le k$ 且 w.p. $\ge 1-\frac{1}{n}$, $\|x-z\|_2 \le (1+\varepsilon)\,err_2^k(x)$.

(Rmk: in the algorithm, if x is k-sparse, then $x=z$)

Alg: Count Sketch + 选前 k 个.

↳ k len. queue 形式.

① let $w = 3k/\varepsilon^2$. $d = \Omega(\log n)$. use Count Sketch. (Maintain $\hat{x}_1, \cdots, \hat{x}_n$).

② Output the k coordinate of the k maximal estimate.

Rmk: space: $O(k/\varepsilon^2 \cdot \log n \log m)$. update: $O(n \cdot \log k)$.

Lemma1: Count Sketch with $w = 3k/\varepsilon^2$. $d = \Omega(\log n)$. ensures.

$\forall i$, $|\hat{x}_i - x_i| \le \frac{\varepsilon}{\sqrt{k}} err_2^k(x)$. w.p. $\ge 1-\frac{1}{n}$.

Lemma2:

let x, y be vectors s.t. $\|x-y\|_\infty \le \frac{\varepsilon}{\sqrt{k}} err_2^k(x)$.

Then let T be the set of k largest indices of y.

let $z_i = y_i$ if $i \in T$. then $\|z-x\|_2 \le (1+5\varepsilon)\,err_2^k(x)$.

Pf1: denote $T_{big} = \{i \mid x_i \text{ is } k \text{ largest in } x\}$. $T_{small} = [n] \setminus T_{big}$.

Fix $i \in [n]$. Consider l s.t. $1 \le l \le d$

① denote A_l as event: $h_l(i) = h_l(i')$ for some $i' \in T_{big}$. $(i' \ne i)$.

To prove $Pr[A_l] \le \frac{\varepsilon^2}{3}$.

日期: /

② denote $Z_i = g_{t(i)} C[t, h_t(i)]$.

then $\hat{x_i} = x_i + \underbrace{\sum_{j \in T_{big} \setminus i} g_{t(i)} g_{t(j)} Y_j x_j}_{:= Z_i''} + \underbrace{\sum_{j \in T_{small} \setminus i} g_{t(i)} g_{t(j)} Y_j x_j}_{:= Z_i'}$

To prove $Pr[|Z_i'| \geq \frac{\varepsilon}{\sqrt{k}} err_2^k(x)] \leq \frac{1}{3}$.

③ Assuming ①② holds, then.

$|Z_i - x_i| = |Z_i'' + Z_i'|$

$= |Z_i'|$ w.p. $\geq 1 - \frac{\varepsilon^2}{3}$ $\Rightarrow$ w.p. $\leq \frac{2}{3}$.

$< \frac{\varepsilon}{\sqrt{k}} err_2^k(x)$ w.p. $\geq \frac{2}{3}$

Then by median Tick, $|\hat{x_i} - x_i| < \frac{\varepsilon}{\sqrt{k}} err_2^k(x)$ w.p. $1 - \frac{1}{n}$. ($d = \Omega(\log n)$).

Prove of ①: 设 $Y_j = Pr[h_t(j) = h_t(i)] \Rightarrow Pr[Y_j = 1] = \frac{1}{w} \leq \frac{\varepsilon^2}{3k}$.
$= E[Y_j]$.

Markov $\Rightarrow$ $Pr[\sum Y_j \geq 1] \leq k \frac{\varepsilon^2}{3k} = \frac{\varepsilon^2}{3}$.

Prove of ②: $E[Z_i'] = 0$. $Var[Z_i'] \leq \frac{\varepsilon^2}{3k} \sum_{j \in T_{small}} x_j^2 = \frac{\varepsilon^2}{3k} err_2^k(x)$.

$\Rightarrow Pr[|Z_i'| \geq \frac{\varepsilon}{\sqrt{k}} err_2^k(x)] \leq \frac{1}{3}$.

Pf2: $\|x - z\|_2^2 = \underbrace{\sum_{i \in T} |x_i - z_i|^2}_{(i)} + \underbrace{\sum_{i \in S \setminus T} |x_i - z_i|^2}_{(ii)} + \underbrace{\sum_{i \in [n] \setminus S \cup T} |x_i - z_i|^2}_{(iii)}$ ($S \triangleq T_{big}$)

$\leq k \cdot (\frac{\varepsilon}{\sqrt{k}} err_2^k(x))^2$ $\leq k \cdot (\frac{\varepsilon}{\sqrt{k}} err_2^k(x))^2$ $\leq err_2^k(x)$.

IV. Matrix Sketch (The Frequent Directions Problem).

Input: n rows of $A \in R^{n \times d}$ one by one (流 in streaming model).

日期: /

(Streaming 下做SVD)

Output: $B \in R^{k \times d}$. where $k << n$. s.t. $A^TA \approx B^TB$.

Precisely, $\forall \|x\|=1$, $0 \le \|Ax\|^2 - \|Bx\|^2 \le 2\|A\|_F^2/k$.

i.e. $\|A^TA - B^TB\| \le 2\|A_F\|^2/k$. 且 $A^TA - B^TB \ge 0$.

Rmk: ① Frequent Items Problem's 推广.

② If x is a direction ($\|x\|=1$) that $\|Ax\|^2 \ge \varepsilon \cdot \|A\|^2$. ($\|A\| := \max_{\|v\|=1} \|Av\|$)

then the Alg can recover x by choosing $k \ge \frac{2}{\varepsilon} \cdot \frac{\|A\|_F^2}{\|A\|^2}$.

Alg. (需要整理)

① Initialize $B \in \mathbb{R}^{k \times d}$ to be all-zero matrix.

② For each row A_i ($i \in [n]$).

↓ 1° Insert A_i into a zero valued row of B

↓ 2° If B has no zero valued row, then,

1) Compute SVD of B: $B = UDV^T$

2) let $C = DV^T$, $\delta = \sigma_{k/2}^2$, $\tilde{D} = \sqrt{\max(D^2 - I_k \cdot \delta, 0)}$.

(至少一半以上变为0)

3) let $B = \tilde{D}V^T$

↳ 不要U. 因为保证下面$\frac{k}{2}$行为0.

③ Output B.

Correctness:

Claim1: for output B, $0 \le B^TB \le A^TA$.

日期: /

Pf: $0 \le B^TB$ 显然.

Denote $B^{(i)}, C^{(i)}$ as B and C after i loops.

Note that, $B^{(0)} = 0$. $B^{(n)} = B$.

then $\|Ax\|^2 - \|Bx\|^2 = \sum_{i=1}^{n} (\langle A_i, x\rangle^2 + \|B^{(i-1)}x\|^2 - \|B^{(i)}x\|^2)$.

$\overset{(*)}{=} \sum_{i=1}^{n} \|C^{(i)}x\|^2 - \|B^{(i)}x\|^2$

$\overset{(\star)}{\ge} 0$.

Why (*): $B^{(i-1)} + A_i \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} = U^T C^{(i)}$.

Why ($\star$): Case 1: $\|C^{(i)}x\|^2 - \|B^{(i)}x\|^2 = 0$.

Case 2: $x^T C^{(i)T} C^{(i)} x - x^T B^{(i)T} B^{(i)} x$

$= (V^Tx)^T D^{(i)T} D^{(i)} (V^Tx) - (V^Tx)^T \tilde{D}^{(i)T} \tilde{D}^{(i)} (V^Tx)$

$= (V^Tx)^T (D^{(i)} - \tilde{D}^{(i)})^T (D^{(i)} - \tilde{D}^{(i)}) (V^Tx)$

≥ 0 ($D^{(i)} - \tilde{D}^{(i)} \ge 0$).

Claim 2: $\|A^TA - B^TB\| \le 2\|A\|_F^2 / k$.

Pf: Let x be the unit eigenvector for the max eigenvalue of $A^TA - B^TB$.

Note that, $\|A^TA - B^TB\| = \max_{\|v\|\le 1} \|(A^TA - B^TB)x\| = \|(A^TA - B^TB)x\| = \lambda$.

$x^T(A^TA - B^TB)x = x^T \lambda x = \lambda$.

日期: /

$$\therefore \underline{\|A^TA - B^TB\| = \|Ax\|^2 - \|Bx\|^2}$$

$$\text{Claim} = \sum_{i=1}^{n} \|C^{(i)}x\|^2 - \|B^{(i)}x\|^2$$

$$\overset{(*)}{\leq} \sum_{i=1}^{n} \|C^{(i)T}C^{(i)} - B^{(i)T}B^{(i)}\|.$$

$$= \sum_{i=1}^{n} \|(\tilde{D}^{(i)})^2 - (D^{(i)})^2\|.$$

$$= \sum_{i=1}^{n} \left\| \begin{pmatrix} \delta^{(i)^2} & & & \\ & \ddots & & \\ & & \delta^{(i)^2} & \\ & & & \sigma_{k/2+1}^2 \\ & & & & \ddots \\ & & & & & \sigma_n^2+1 \end{pmatrix} \right\|$$

$$= \sum_{i=1}^{n} \delta^{(i)2}.$$

Why $(*)$: $\sum_{i=1}^{n} \|C^{(i)}x\|^2 - \|B^{(i)}x\|^2$

$$= \langle x, (C^{(i)T}C^{(i)} - B^{(i)T}B^{(i)})x \rangle$$

Cauchy Ineq $\leq \|x\| \cdot \|(C^{(i)T}C^{(i)} - B^{(i)T}B^{(i)})x\|.$

$$\leq \|(C^{(i)T}C^{(i)} - B^{(i)T}B^{(i)})\|.$$

Lem: $\|A\|_F^2 = Tr(A^TA)$.

$$\|B\|_F^2 = \sum_{i=1}^{n} \|B^{(i)}\|_F^2 - \|B^{(i-1)}\|_F^2$$

$$= \sum_{i=1}^{n} (\|C^{(i)}\|_F^2 - \|B^{(i-1)}\|_F^2) - (\|C^{(i)}\|_F^2 - \|B^{(i)}\|_F^2).$$

$$\overset{(\star)}{=} \sum_{i=1}^{n} \|A_i\|^2 - Tr(C^{(i)T}C^{(i)} - B^{(i)T}B^{(i)})$$

$$= \|A\|_F^2 - Tr((D^{(i)})^2 - (\tilde{D}^{(i)})^2)$$

$$\leq \|A\|_F^2 - k/2 \cdot \delta^{(i)2}.$$

日期: /

$\Rightarrow \delta^{(\omega)} \leq 2(\|A\|_F^2 - \|B\|_F^2)/k \leq 2\|A\|_F^2/k$

Space: $O(d \cdot k)$. ?

Time: $O(ndk)$

§ 3

Machine Learning

日期: /

Lecture 8. Classification PAC learning.

(Model, Optimization, Generalization).

Space $X \subseteq \mathbb{R}^d$

Goal: Find a "simple" rule performs well on training data and as long as training data are representative to future data, then it also performs well on future data.

I: The perception algorithm:

Problem: labels: $l_1, \cdots, l_n$
examples: $x_1, \cdots, x_n \in X$. (Assume hyperplane 过原点).

Goal: to find weight vector w, and a threshold t, s.t.

$w^T x_i > t$ if $l_i = 1$.

$w^T x_i < t$ if $l_i = -1$.

⇒ 要求线性可分性.

Solutions: (Assuming $\exists\ w, t$).

(1) linear programming algorithm:

$(\min\ c^T x)$.

s.t. $Ax \geq b$.

solvable but time-consuming.

(2) The Perception alg:

Define $\hat{x}_i = l_i \cdot \begin{pmatrix} x_i \\ 1 \end{pmatrix}$. $\hat{w} = \begin{pmatrix} w \\ -t \end{pmatrix}$

日期: /

then, $\begin{cases} w^T x_i > t & \text{if } l_i = 1 \\ w^T x_i < t & \text{if } l_i = -1 \end{cases} \Leftrightarrow \hat{w}^T \hat{x}_i > 0$.

Rmk: $\hat{l} = \text{sign}(\hat{w}^T \hat{x})$. 在训练集上总有 $\hat{l}_i = l_i$.

Alg:

(1) Intialize $\hat{w} \leftarrow 0$.

(2) if $\exists\, i$ st $\hat{w}^T \hat{x}_i \le 0$. update $\hat{w} \leftarrow \hat{w} + l_i \begin{pmatrix} x_i \\ 1 \end{pmatrix}$

Thm: If $\exists\, w^*$ st $w^{*T} x_i l_i \ge 1 \quad \forall i$ (→ 即相隔一定距离)

then the alg. finds a solution w st $w^T x_i l_i > 0 \quad \forall i$.

and at most $r^2 \|w^*\|^2$ updates where $r = \max_i \|x_i\|$.

(Rmk: ① Assume hyperplane 过原点

② $\text{margin} = \min_i \text{dist}(x_i, w^{*T} x = 0) = \min_i \frac{|w^{*T} x_i|}{\|w^*\|} \ge \min_i \frac{1}{\|w^*\|}$.

③ $r = \max_i \|x_i\|$ 即 radius of data ball.

$\|w^*\| = \frac{1}{\text{margin}}$.

即 radius越小, margin越大, alg越快.

④ the output w is $\{x_i\}$'s linear combination. i.e. $w = \sum c_i x_i$)

Pf: Each update, consider $w^T w^*$ 与 $\|w\|^2$:

① $(w + x_i l_i)^T w^* = w^T w^* + (x_i l_i)^T w^* \ge w^T w^* + 1$.

日期: /

② $\|w+x_i l_i\|^2 = \|w\|^2 + 2\underbrace{w^T x_i l_i}_{\leq 0} + \|x_i l_i\|^2$

$\leq \|w\|^2 + \|x_i l_i\|^2$

$\leq \|w\|^2 + r^2$.

Let m be numbers of updates. then,

$ww^* \geq m$. $\|w\|^2 \leq mr^2$.

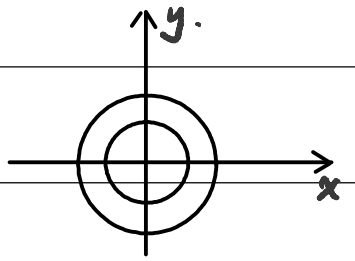
∴由 Cauchy Ineq. $m^2 \leq |ww^*|^2 \leq \|w\|^2\|w^*\|^2 \leq mr^2\|w^*\|^2$.

$\Rightarrow$ $m \leq r^2\|w^*\|^2$.

(注:可用反证法,得最终 $w^T x_i l_i > 0 \ \forall i$.)

Improvement:

Alg can't work when $\nexists\ w^*$. eg.



Solutions:

① use unlinear function to partition.

② 在高维 space 中 may have 划分.

eg. $(x,y) \xrightarrow{\varphi} (x, y, x^2+y^2)$. then $\exists w^*$.

缺点: φ 难以确定 (dim?, entries of new dim.?).

③ Kenel Method.

(Perception Alg. 只要有 $k(x_i, x_j) := \varphi(x_i)^T\varphi(x_j)$. 即可.)

日期: /

Reason:

dual perception Alg.

① Initialize $\alpha=0\in\mathbb{R}^d$. $w=\sum\alpha_i l_i x_i$. $k(x_i,x_j)$. ↑φ.

② Iterate: $\forall i$, let $\hat{y}_j=\operatorname{sgn}(w^Tx_j)=\operatorname{sgn}(\sum\alpha_i l_i \underbrace{x_i^Tx_j}_{\text{}})$

if $\exists\ i$ s.t $\hat{y}_i\neq l_i$. then $\alpha_i\leftarrow\alpha_i+1$.

else return α and w.

For Prediction of x. output $\operatorname{sgn}(\sum\alpha_i l_i \underbrace{x_i^Tx}_{\text{}})$. $k(x_i,x)$. ↑φ.

Kernel Matrix: (for finite data-set).

Define $\varphi:\mathbb{R}^d\to\mathbb{R}^{d'}$ $(d'>d)$. define $k(x,y)=\varphi(x)^T\varphi(y)$.

Kernel Matrix $K=(K_{ij})_{\substack{i\le n\\ j\le n}}$, $K_{ij}=k(x_i,x_j)$.

Lemma: A matrix K is a Kernel Matrix iff K is PSD. ↑半正定.

Pf: (FACT: PSD $K\Leftrightarrow K=B^TB$. for some B.).

Define $\varphi(x_i)$ is i-th column of B.

then $K_{ij}=(B^TB)_{ij}=\varphi(x_i)^T\varphi(x_j)$.

反之亦然.

日期: /

Thm: 记 $\mathcal{K}=\{\text{Kernel functions } k(x,y)\}$.

① $\mathcal{K}$ 是一个环. ? 几何.

② $\forall$ scalar function f, $\forall k\in\mathcal{K}$. (↑标量)

$k'(x,y)=f(x)f(y)k(x,y)$

常用 eg.: 1) Polynomial Kernel: $\boxed{k(x,y)=(1+x^Ty)^t}$. for $\forall t\geq 0$.

2) Guassian kernel: $\boxed{e^{-c\|x-y\|^2}}$ $(=f(x)f(y)e^{-2cx^Ty})$.

($f(x)=e^{-c\|x\|^2}$ scalar function. e^{-2cx^Ty} → Taylor展开.)

II. SVM (支持向量机). (optimum hyperplane)

由 Perception. Alg. Assume (x_i,l_i) is linear seperable (have margin).

we can find w,t. s.t.

$$(w^Tx_i+t)l_i\geq 1.\quad\forall i\quad(\Leftarrow (w^Tx_i+t)l_i>0.\quad\forall i.)$$

SVM with hard margin:

Assume linear seperable, find (w,t) s.t.

$$\begin{cases}\text{maximize margin } \frac{1}{\|w\|}.\ (\Leftrightarrow \text{minimize } \frac{1}{2}w^Tw.)\\ (w^Tx_i+t)l_i\geq 1.\quad\forall i\end{cases}$$

(Rmk: It's a convex programming (凸优化). 有 poly. time. solution.).

日期: /

Def: support vector: 恰在 $(w^Tx_i+t)l_i=1$ 线上的 vectors.

SVM with soft margin:

No linear seperable assumption.

Try: ① Kernel method.

② $\text{minimize}_{w,t}$ #mistakes. under $(w^Tx_j+t)l_j \geq 1 \cdot \forall j$.

③ $\text{minimize}_{w,t}$ $(\frac{1}{2}w^Tw + C\cdot \#\text{mistakes})$. ⇒ NP-hard Problem.

考虑到 $(w^Tx_j+t)l_j \geq 1 \cdot \forall j$.

$\text{loss}_j = \max(0, 1-(w^Tx_j+t)l_j)$.

⇒ $\text{minimize}_{w,t}$ $\frac{1}{2}w^Tw + C\cdot\sum_j \max(0,1-(w^Tx_j+t)l_j)$.

↓

$$\begin{cases} \text{minimize}_{w,t} & \frac{1}{2}w^Tw + C\cdot\sum \varsigma_j \\ (w^Tx_j+t)l_j \geq 1-\varsigma_j. \quad \forall j. & \varsigma_j \geq 0. \end{cases}$$

有 $L(w,t,\varsigma,\alpha,\beta) = \frac{1}{2}w^Tw + C\sum\varsigma_j - \sum\alpha_j[(w^Tx_j+t)l_j - 1+\varsigma_j] - \sum\beta_j\varsigma_j$.

↓ KKT条件

$$\begin{cases} \nabla_w L = w - \sum\alpha_j l_j x_j = 0. & (1^\circ) \\ \nabla_t L = -\sum\alpha_j l_j = 0. & (2^\circ) \\ \nabla_{\varsigma_j} L = C-\alpha_j-\beta_j = 0. \quad \forall j. & (3^\circ) \\ \alpha_j((w^Tx_j+t)l_j - 1+\varsigma_j) = 0. \quad \forall j. & (4^\circ) \\ \beta_j\varsigma_j = 0. \quad \forall j. & (5^\circ) \end{cases} \quad (*1).$$

又当 $\alpha_i \in (0,C)$. $\overset{(3^\circ)}{\Rightarrow} \beta_j > 0. \overset{(5^\circ)}{\Rightarrow} \varsigma_j = 0. \overset{(4^\circ)}{\Rightarrow} t = l_j - w^Tx_j$. (6°)

日期:

$$\sum \xi_j \overset{3^\circ}{=} \sum(\alpha_j+\beta_j)\xi_j \overset{5^\circ}{=} \sum \alpha_j \xi_j \overset{4^\circ}{=} \sum \alpha_j(1-l_j(w^Tx_j+t)) \overset{1^\circ}{=} \sum \alpha_j - w^Tw - \underbrace{\sum \alpha_i l_i}_{=0\ (2^\circ)} t = \sum \alpha_j - w^Tw \quad (7^\circ)$$

$$= \frac{1}{2}$$

$\Rightarrow$ Dual Problem: $\quad K(x_i, x_j)$

$$\begin{cases} \text{maximize}_\alpha\ L(w,t,\xi,\alpha,\beta) = \sum \alpha_i - \frac{1}{2}\sum \alpha_i \alpha_j l_i l_j \underbrace{x_i^T x_j}\ \ (7^\circ) \\ \sum \alpha_i l_i = 0 \quad (2^\circ) \end{cases} \quad (*2)$$

求解 $(*2)$ 后, 由 $(*1)$. $w = \sum \alpha_j l_j x_j$ (1°)

(凸优化在多项式时间内可解. 多用椭球法等方法).

$t = l_{j_0} - w^T x_{j_0}$ (for some j_0 st $\alpha_{j_0} \in (0,C)$) (6°)

$\Rightarrow$ classification function:

$$f(x) = \text{sgn}(w^Tx+t)$$

$$= \text{sgn}\Big(\underbrace{\sum \alpha_j l_j x_j^T x}_{K(x_j,x)} + l_{j_0} - \underbrace{\sum \alpha_i l_i x_i^T x_{j_0}}_{K(x_i,x_{j_0})}\Big)$$

(Rmk: SVM 同样可用 Kernel Method 优化).

III PAC learning (Probably Approximately Correct) (适用于离散假设集)

Want to generalize the rule learned from training data.

$\Rightarrow$ need data to be "representative".

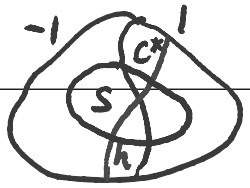
推广 $\Rightarrow$ 均匀抽样 $\Rightarrow$ 任意分布抽样

Assumption: $\exists$ a prob. distribution D over instance space X.

s.t. 1) training set S consists of n points drawn independently at random from D.

2) Goal is to predict well on new points drawn from D.

日期: /



Def: 1). C^*: target concept:

a subset of X corresponding to positive class in X.

2) h: hypothesis

3) true/generalization error of h:

$$err_D(h) = \Pr_{x\sim D}[x \in h \triangle C^*].$$

4) training/empirical error:

$$err_S(h) = \frac{\#S \cap (h \triangle C^*)}{\#S}.$$

5) Hypothesis class $\mathcal{H}$ over X. ($\forall h \in \mathcal{H}$ is a hypothesis.).

Goal: to produce h with low true error.

Intuition: If S is large enough, then w.h.p. $\forall h \in \mathcal{H}$ have:

$$err_S(h) \approx err_D(h).$$

Thm (PAC Learning). ★

Let $\mathcal{H}$ be a hypothesis class, $\varepsilon, \delta \in (0,1)$.

If training set S of size $n \geq \frac{1}{\varepsilon}(\ln|\mathcal{H}| + \ln(1/\delta))$ is drawn from distribution D,

then w.p. $\geq 1-\delta$. every $h \in \mathcal{H}$ with training error 0 has true error $\leq \varepsilon$.

(Rmk: $h \in \mathcal{H}$ with 0 training error is Probably Approximately Correct (PAC))

Pf: 逆否命题: w.p. $\geq 1-\delta$. $\forall h \in \mathcal{H}$ with $err_D(h) \geq \varepsilon$. has $err_S(h) > 0$.

日期: /

fix h with $err_D(h) \geq \varepsilon$. then.

$$\Pr_{S\sim D}[err_S(h)=0] \leq (1-err_D(h))^{|S|} \leq (1-\varepsilon)^n.$$

then $\Pr[\exists h\in H$ with $err_D(h)\geq\varepsilon$ has $err_S(h)=0]$

$$\leq |H| \cdot \Pr_{S\sim D}[err_S(h)=0]$$

$$= |H|\cdot(1-\varepsilon)^n$$

$$\leq |H|\cdot e^{-\varepsilon n}$$

$$\leq |H|\cdot e^{-(\ln|H|+\ln(1/\delta))} = \delta.$$ □.

Thm (PAC Learning 2): ☆

Let H be a hypothesis class, $\varepsilon, \delta \in (0,1)$.

If training set S of size $n \geq \frac{1}{2\varepsilon^2}(\ln|H|+\ln(2/\delta))$ is drawn from distribution D,

then w.p. $\geq 1-\delta$. $\forall\ h\in H$ 有 $|err_S(h)-err_D(h)|\leq\varepsilon$.

Pf: (思路同上: fix $h \Rightarrow$ Pr[opposite] $\leq *$ $\Rightarrow$ Union bound).

fix $h\in H$. for j-th element in S. define

$$X_j = \begin{cases} 1. & \text{if } h \text{ disagree with } c^* \text{ on } j\text{-th element} \\ 0. & \text{o.w.} \end{cases}$$

then. $P[X_j=1] = err_D(h)$. $err_S(h) = \frac{\sum_S X_j}{|S|}$.

$$E[err_S(h)] = \frac{1}{|S|}\sum_S E[X_j] = err_D(h).$$

By Chernoff-Hoeffding,

日期: /

$$\Pr[|err_S(h) - err_D(h)| > \varepsilon] \le 2e^{-2\varepsilon^2 n}.$$

Then, $\Pr[\exists h \in \mathcal{H}. s.t. |err_S(h) - err_D(h)| > \varepsilon]$

$\le 2|\mathcal{H}| \cdot e^{-2n\cdot\varepsilon^2}$

$\le \delta.$ □

IV. Occam's Razor:

Thought: Simple explanations are better than complicated ones.

Thm (Occam's Razor)

w.p. $\ge 1-\delta$, $\forall h$ with $err_S(h)=0$ that can be described using $< b$ bits.

has $err_D(h) \le \varepsilon$ if $|S| \ge \frac{1}{\varepsilon}(b\ln 2 + \ln(1/\delta))$.

(Rmk: 即 w.p. $\ge 1-\delta$, $\forall h$ with $err_S(h)=0$ that can be described using $< b$ bits. has $err_D(h) \le \frac{b\ln 2 + \ln(1/\delta)}{|S|}$

Pf: let $\mathcal{H} = \{h : \forall h$ can be described using $< b$ bits$\}$.

then $|\mathcal{H}| \le 2^b$. plugging it into Thm(PAC). □

V. VC dimension ☆ (适用于连续假设集)

Def: System Set $(X, \mathcal{H})$.

X: a set, instance space.

$\mathcal{H}$: a class of subsets of X. hypothesis class.

Def: $(X, \mathcal{H})$ shatters a set $A \subseteq X$:

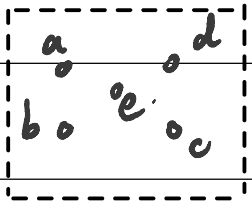
日期: /

$\forall$ subset of A can be expressed as $A \cap h$ for some $h \in \mathcal{H}$.

Def: VC dimension of $(X, \mathcal{H})$:

$$\max_{(X,\mathcal{H}) \text{ shatters } A} \#A.$$

eg1: $X = \mathbb{R}^2$. $\mathcal{H} = \{\text{all rectangles}\}$.



$VC(X, \mathcal{H}) = 4$.

当 $A = \{a, b, c, d, e\}$. 必有一个在内. 不妨为 e. 则 $\{a, b, c, d\} \neq A \cap h$ for some h

eg2: $X = \mathbb{R}^2$. $\mathcal{H} = \{[a, b] \mid a \leq b\}$.

$VC(X, \mathcal{H}) = 2$.

---

eg3: Consider $X = \mathbb{R}^d$

linear seperators: $\mathcal{H} = \{w^T x + b \geq 0, w \in \mathbb{R}^d, b \in \mathbb{R}\}$

☆ (Rmk: $\#\mathcal{H} = +\infty$, 不能直接用 PAC Learning)

Thm: $VC(X, \mathcal{H}) = d + 1$.

Pf: ① 取 $A = \{e_1, e_2, \cdots, e_d, 0_d\}$. $e_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$ $0_d = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$.

1° for $\forall$ subset $A' = \{e_{i_1}, \cdots, e_{i_j}, 0_d\}$.

取 $w' = 1_d - \sum_{k=1}^{j} e_{i_k}$ $b = 0$.

则 $A' = A \cap \{w'^T x + b \leq 0\}$.

日期: /

2° for $\forall$ subset $A'=\{e_{i_1},\cdots,e_{i_j}\}$.

取 $w'=1_d-2\sum_{k=1}^{j}e_{i_k}$. $b=1$.

则 $A'=A\cap\{w'^Tx+b\le 0\}$. $\therefore VC(X,\mathcal{H})\ge d+1$.

② 由 Radon Thm, $VC(X,\mathcal{H})\le d+1$. □.

Def: S is a set of s vertices, then

convex hull of S: $convex(S)=\{\sum_{i}^{s}a_ia_i \mid a_i\ge 0, \sum_{i}^{s}a_i=1, a_1,\cdots,a_s\in S\}$.

Thm(Radon): Any set $S\subseteq\mathbb{R}^d$ with $|S|\ge d+2$ can be partitioned into two disjoint subsets S_1, S_2 s.t. $convex(S_1)\cap convex(S_2)\neq\emptyset$.

Pf: Assume $S=\{a_1,\cdots,a_{d+2}\}$.

记 $A=(a_1,\cdots,a_{d+2})\in\mathbb{R}^{d\times(d+2)}$.

$B=\begin{pmatrix}A\\1_{d+2}^T\end{pmatrix}\in\mathbb{R}^{(d+1)\times(d+2)}$.

Consider $BX=0$. 由 $rank(B)\le d+1$. $X\in\mathbb{R}^{d+2}$.

$\Rightarrow$ has a non-zero solution $X=(x_1,\cdots,x_{d+2})$.

Reorder columns st $x_1,\cdots,x_s\ge 0$. $x_{s+1},\cdots,x_{d+2}<0$.

$\Rightarrow$ normalize X s.t. $\sum_{i=1}^{s}|x_i|=1$.

记 $b_i=\begin{pmatrix}a_i\\1\end{pmatrix}$. then $B=(b_1,\cdots,b_{d+2})$. then $\sum_{i=1}^{b+2}b_ix_i=0$

$\Rightarrow \sum_{i=1}^{s}b_i|x_i|=\sum_{i=s+1}^{d+2}b_i|x_i|$.

日期: /

$b_i=\binom{a_i}{1}$. $\Rightarrow$ $\begin{cases}\sum_{i=1}^{s} a_i|x_i| = \sum_{i=s+1}^{d+2} a_i|x_i| := a. \\ \sum_{i=1}^{s}|x_i| = \sum_{i=s+1}^{d+2}|x_i| \ (=1).\end{cases}$

则 $a \in \text{convex}(a_1,\cdots,a_s) \cap \text{convex}(a_{s+1},\cdots,a_{d+2})$. 口.

---

eg.4: spheres in d-dimesion:

$X=\mathbb{R}^d$. $\mathcal{H}=\{x \mid |x-x_0|\le r\}$. $VC(X,\mathcal{H})=d+1$ (不证).

eg.5. $X=\mathbb{R}$. $\mathcal{H}=\{h \mid h \text{ contains finite number of real number}\}$.

$VC(X,\mathcal{H})=\infty$.

Rmk: VC-dimension 可用于描述 $(X,\mathcal{H})$'s complexity.

Shatter function (extension of VC dimension).

Def: Given $(X,\mathcal{H})$. the shatter function

$$\pi_{\mathcal{H}}(n) := \max_{|A|=n} \#\{A' \subseteq A \mid A' = A \cap h \text{ for some } h \in \mathcal{H}\}.$$

(Rmk: if $VC(X,\mathcal{H})=n_0$. then $\pi_{\mathcal{H}}(n_0)=2^{n_0}$. $\pi_{\mathcal{H}}(n_0+1)<2^{n_0+1}$.

Lemma (Sauer) (用 VC dimension 估算 $\pi_{\mathcal{H}}(n)$).

$VC(X,\mathcal{H})=d$. then $\pi_{\mathcal{H}}(n) \le \binom{n}{\le d} \triangleq \binom{n}{0}+\binom{n}{1}+\cdots+\binom{n}{d} \le n^d+1$.

Thm (Key Thm): $(X,\mathcal{H})$. $x \sim D$. let n s.t $n \ge \frac{8}{\varepsilon}$ 且 $n \ge \frac{2}{\varepsilon}[\log_2(\underbrace{2\pi_{\mathcal{H}}(2n)}_{|\mathcal{H}| \to 2\pi_{\mathcal{H}}(2n)})+\log_2(1/\delta)]$.

let $S \sim D$ 且 $\#S=n$. w.p. $\ge 1-\delta$ $\forall h \in \mathcal{H}$ w.p. $\ge 1-\varepsilon$ (针对 D) intersects S.

日期:

Double Sampling.

Pf: $S_1 \sim D$. $S_2 \sim D$.

def A: $\exists h \in H$ w.p. $\geq \varepsilon$ disjoint from S_1

B: $\exists h \in H$ w.p. $\geq \varepsilon$ s.t.

$h \cap S_1 = \emptyset$ 且 $\# h \cap S_2 \geq \frac{\varepsilon n}{2}$.

Lemma 1: $P_r[B|A] \geq \frac{1}{2}$

Lemma 2: $P_r[B] \leq \frac{\delta}{2}$.

Assume the above correct, then

$$\frac{\delta}{2} \geq P_r[B] \geq P_r[AB] = P_r[B|A]\, P_r[A] \geq \frac{1}{2} P_r[A]. \quad \square$$

Lemma 2 Pf:

$S_1 \sim D$. $S_2 \sim D$

st. S_1, S_2 产生 connection.

fix S_1, S_2 random h. → fix h, random S_1, S_2.

依序抽样 → 混合抽样 + 划分

<=> double sampling:

转换随机对象.

1° $S_3 \sim D$ st. $\# S_3 = 2n$.

2° randomly partition S_3 into S_1, S_2. st. $\# S_1 = \# S_2$

于是, 有 $\# \{ S_3 \cap h \mid h \in H \} \leq \pi_H(2n)$.

Lemma 3: Given $h \in H$ st $h \cap S_3 \neq \emptyset$. 记 $h' = h \cap S_3$.

randomly partition S_3 into S_1, S_2.

then $\# S_1 \cap h' = 0$. $\# S_2 \cap h' \geq \frac{\varepsilon n}{2}$ w.p. $\leq \frac{\delta}{2\pi_H(2n)}$

日期: /

Assume correct. then $Pr[B] \le \frac{\delta}{2\tilde{\pi}_H(2n)} \cdot \tilde{\pi}_H(2n) = \frac{\delta}{2}$.

Pf of lemma 3:

case I: $\#h \cap S_3 < \frac{\varepsilon n}{2}$

$Pr[\#h' \cap S_1 = 0, \#S_2 \cap h' \ge \frac{\varepsilon n}{2}] = 0$.

case II: $\#h \cap S_3 \ge \frac{\varepsilon n}{2}$.

$Pr[\#h' \cap S_1 = 0, \#S_2 \cap h' \ge \frac{\varepsilon n}{2}]$

$= Pr[\#h' \cap S_1 = 0]$.

$\le (\frac{1}{2})^{\frac{\varepsilon n}{2}} \le \frac{\delta}{2\tilde{\pi}_H(2n)}$. By Union Bound. □.

Pf of lemma 1:

let h be a set w.p. $\ge \varepsilon$ $S_1 \cap h = \emptyset$.

let $X_i = \begin{cases} 1, & \text{the } i\text{-th point in } S_2 \text{ belong to } h. \\ 0, & \text{o.w.} \end{cases}$

let $X = \Sigma X_i$.

$\therefore \eta := E[X_i] \ge \varepsilon$. $Var[X_i] = \eta - \eta^2 < \eta$.

$\Rightarrow E[X] = n\eta$. $Var[X] < n\eta$.

$\Rightarrow Pr[|X - E[X]| \ge \frac{n\eta}{2}] \le \frac{n\eta}{(n\eta/2)^2} \le \frac{4}{n\eta} \le \frac{1}{2}$ ↓ $n \ge \frac{8}{\varepsilon} \ge \frac{8}{\eta}$

$\Rightarrow Pr[B|A] = Pr[X \ge \frac{\varepsilon n}{2}] \ge Pr[X \ge \frac{\eta n}{2}] \ge Pr[|X - \eta n| \le \frac{n\eta}{2}]$

$= Pr[|X - E[X]| \le \frac{n\eta}{2}] \ge \frac{1}{2}$.

日期: /

即预测错误的部分.

Prop: A target concept c^*. let $\mathcal{H}' = \{h \Delta c^* \mid h \in \mathcal{H}\}$

Fact: $VC(X, \mathcal{H}') = VC(X, \mathcal{H})$

and $(X, \mathcal{H}')$, $(X, \mathcal{H})$ have same shutter function.

Thm1: $(X, \mathcal{H})$. $x \sim D$. let n s.t $n \geq \frac{8}{\varepsilon}$ 且 $n \geq \frac{2}{\varepsilon}[\log_2(2\tau_{\mathcal{H}}(2n)) + \log_2(1/\delta)]$. ($|\mathcal{H}| \to 2\tau_{\mathcal{H}}(2n)$.)

$S \overset{i.i.d}{\sim} x$. Then w.p. $\geq 1-\delta$. $\forall h \in \mathcal{H}$ with $err_S(h) = 0$. has $err_D(h) \leq \varepsilon$.

Pf: $\forall h$ 反证: Assume $err_D(h) > \varepsilon$.

Apply Key Thm to $\mathcal{H}' = \{h \Delta c^* \mid h \in \mathcal{H}\}$.

Then, w.p. $\geq 1-\delta$, $err_S(h) = \Pr_{x \sim D}[x \in h \Delta c^*] > \varepsilon$. has

$|S \cap (h \Delta c^*)| > 0$. 这与 $err_S(h) = 0$ 矛盾. □.

Thm2: $(X, \mathcal{H})$. $x \sim D$. let n s.t $n \geq \frac{8}{\varepsilon}$ 且 $n \geq \frac{8}{\varepsilon^2}[\log_2(2\tau_{\mathcal{H}}(2n)) + \log_2(1/\delta)]$. ($|\mathcal{H}| \to 2\tau_{\mathcal{H}}(2n)$.)

$S \overset{i.i.d}{\sim} x$. Then w.p. $\geq 1-\delta$. $\forall h \in \mathcal{H}$, $|err_S(h) - err_D(h)| \leq \varepsilon$.

Rmk: 需要将 Key Thm 修改B为 $|S_1 \cap h| - |S_2 \cap h| \geq \frac{\varepsilon n}{2}$?

并用 Chernoff Bound 做改进.

Rmk: ① By Sauer Bound, denote $d = VC(X, X)$. Improvement:

Thm1: $n \geq \frac{c}{\varepsilon}(d \log n + \log(\frac{1}{\delta}))$.

$n \geq \frac{c'}{\varepsilon}(d \log \frac{1}{\varepsilon} + d \log \frac{1}{\delta})$ ↩

Thm2: $n \geq \frac{c'}{\varepsilon^2}(d \log \frac{1}{\varepsilon} + d \log \frac{1}{\delta})$.

② $|\mathcal{H}| < +\infty$ 时, $\tau_{\mathcal{H}}(2n)$ 版本 Thm2 表现很可能不如 $|\mathcal{H}|$ 版本

(除非 $\tau_{\mathcal{H}}(2n) < \log|\mathcal{H}|$).

§4

Clustering

日期: /

Input: n points $\in \mathbb{R}^d / \{0,1\}^d$.

Goal: k clusters.

- centerbased clustering. eg. k-means, k-median, k-centers.
- high-density clustering. eg. PBSCAN.

I. Centerbased clustering.

Form in Euclidean space:

$A \subseteq M = \mathbb{R}^d$. $D(x,y) = \|x-y\|_2 \ \forall x,y \in \mathbb{R}^d$

Find $C = \{c_1, \cdots, c_k\}$ s.t. $\sum_i D(a_i, \{c_1, \cdots, c_k\})^q$ for some q. is minimized.

then have $C_i = \{a \in A \mid \forall j, D(a, c_i)^q \le D(a, c_j)^q\}$

ties broken arbitrary. ($\Rightarrow \nexists D(a,c_i) = D(a,c_j) = \min$, randomly choose one)

k-centers: $q = +\infty \Rightarrow \min_{\{c_1, \cdots, c_k\}} \max_i D(a_i, \{c_1, \cdots, c_k\})$. (∞非常数)

k-median: $q = 1 \Rightarrow \min_{\{c_1, \cdots, c_k\}} \sum_i D(a_i, \{c_1, \cdots, c_k\})$.

k-means: $q = 2 \Rightarrow \min_{\{c_1, \cdots, c_k\}} \sum_i D(a_i, \{c_1, \cdots, c_k\})^2$

Form in general space:

Def: (Metric Space).

$M = \mathbb{R}^d / \{0,1\}^d /$ other set

$D: M \times M \to \mathbb{R}$. for $\forall x,y,z \in M$. st.

日期: /

1° $D(x,y)=0. \Rightarrow x=y.$

2° $D(x,y)=D(y,x).$

3° $D(x,y) \le D(x,z)+D(y,z).$

then (M,D) is a Metric Space.

eg. $D_{L_2}(x,y)=\|x-y\|_2$ → Euclidean space.

$D_{L_p}(x,y)=\|x-y\|_p.$ → L_p space.

(Rmk: form in general space 多需 replace (M,D) by a Metric Space).

The exact opt problem is NP-hard.

Lloyd's algorithm (for k-means clustering. in Euclidean space).

$A \subseteq M=\mathbb{R}^d.$ $k \ge 1.$

① choose k centers $c_1,\cdots,c_k \in \mathbb{R}^d.$ arbitrarily.

② Repeat: (Until convergence). 收敛. → eg. loss变化不大.

1° find clusters $C_1,\cdots,C_k$ → (assignment)

2° compute centroid $c(C_i)=\frac{1}{|C_i|}\sum_{a\in C_i} a$ } (update).

3° $c_i \leftarrow c(C_i).$

③ Return $c_1,\cdots,c_k$, $C_1,\cdots,C_k.$

日期: /

☆

Lemma1: $\forall x \in \mathbb{R}^d$.

$$\sum_{a \in A'} \|a-x\|^2 = \sum_{a \in A'} \|a-C(A')\|^2 + |A'| \cdot \|C(A')-x\|^2.$$

Pf: $LHS = \sum_{a \in A'} \|a-C(A')+C(A')-x\|^2$

$$= \sum_{a \in A'} (\|a-C(A')\|^2 + \|C(A')-x\|^2 + 2\cdot(C(A')-x)^T\cdot(a-C(A'))).$$

$$= \sum_{a \in A'} \|a-C(A')\|^2 + |A'|\cdot\|C(A')-x\|^2 + 2(C(A')-x)^T \underbrace{\sum_{a \in A'}(a-C(A'))}_{=0}.$$

$= RHS$. 口.

Lemma2: halt after finite steps. the running time is $O(nkdR^*)$.

R^* is the number of iterations until convergence.

(Rmk: 有人证明, $R^* \le 2^{\Omega(\sqrt{n})}$.).

Pf: Loss ↓ each iteration. 略.

(Rmk: 某些实例中 k-means 的结果与最优解相差较远, 但一般效果不错).

(Rmk: 注意到主要是初始点集较差时效果差, 尝试改变抽样方式.).

补充:

Thm: k-means Problem is NP-hard.

Thm: If P≠NP, then 无 poly. time alg. can approximate k-means Problem within 1.0013 multiplicative factor.

日期: /

D^2-sampling + Lloyd's.

II. k-means++ algorithm.

Input A, k.

Alg: ① Run D^2-sampling on A, k to obtain $C = \{c_1, \cdots, c_k\}$.

② Run k-means on A, k with C initially.

D^2-distribution: (针对 S).

for $\forall S \subseteq \mathbb{R}^d$, $|S| < \infty$. denote the distribution on A by:

$$P_S(a) = \frac{D(a,S)}{D(A,S)} = \frac{D(a,S)}{\sum_{b \in A} D(b,S)} = \frac{\min_{s \in S} D(a,s)}{\sum_{b \in A} \min_{s \in S} D(b,s)} \quad (\text{记 } D := D^{(2)}_{L_2}).$$

D^2-sampling:

① choose point $c_1 \in A$ u.a.r. set $C' \leftarrow c_1$.

② loop: (2~k):

draw point $c_i \in A$ according to D^2-distribution $P_{C^{i-1}}(\cdot)$.

set $C^i \leftarrow C^{i-1} \cup \{c_i\}$.

Thm: let C be output of D^2-sampling, C^* be the opt. solution for input A, k. Then,

$$E[D(A,C)] \leq 8(\ln k + 2) D(A, C^*). \quad \Leftrightarrow \quad E[\text{cost}] \leq 8(\ln k + 2) \cdot \text{OPT}.$$

(Rmk: This implies output C' of k-means also s.t. above ineq.)

Analysis:

日期: /

☆

Lem 1 : Let $S \subseteq \mathbb{R}^d$ be a point set, $c \in S$ a point chosen u.a.r. from S. Then,

$$E[D(S, \{c\})] = 2 \cdot D(S, c(S))$$

where $c(S)$ is centroid of S.

Pf: $E[D(S, \{c\})] = \sum_{b \in S} \frac{1}{|S|} D(S, \{b\})$.

$$= \frac{1}{|S|} \sum_{a \in S} \sum_{b \in S} \|a-b\|^2$$

$$= \frac{1}{|S|} \sum_{b \in S} (\sum_{a \in S} \|a - c(S)\|^2 + |S| \cdot \|c(S) - b\|^2)$$

$$= \sum_{b \in S} \|b - c(S)\|^2 + \sum_{a \in S} \|a - c(S)\|^2$$

$$= 2D(S, c(S)). \quad \square.$$

Cor: ① $\sum_{a \in S} \|a - c(S)\|^2 = \frac{1}{2|S|} \cdot \sum_{a \in S} \sum_{b \in S} \|a-b\|^2$.

② let S be a cluster from optimal clustering $\mathcal{C}^* = \{C_1^*, \dots, C_k^*\}$.

If $c \overset{\text{w.a.r}}{\sim} A$. then

$$E[D(S, \{c\}) \mid c \in S] = 2 \cdot D(S, c(S)).$$

Lem 2: $S \subseteq \mathbb{R}^d$ point set. $C \subseteq \mathbb{R}^d$ 随机 finite set.

x is chosen from S by $P_C(\cdot)$. Then,

$$E[D(S, C \cup \{x\})] \leq 8 D(S, c(S)).$$

看论文.

日期: /

Rmk: Time $\geq \Omega(d)$.

Ⅲ. Dimension Reduction of k-means:

Another Version: $A = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \in \mathbb{R}^{n\times d}$. centers set $\mathcal{C} \subseteq \mathbb{R}^d$. $\#\mathcal{C} = k$.

denote $S(\mathcal{C}) = \begin{bmatrix} \arg\min_{c\in\mathcal{C}} \|a_1 - c\|_2 \\ \vdots \\ \arg\min_{c\in\mathcal{C}} \|a_n - c\|_2 \end{bmatrix} \in \mathbb{R}^{n\times d}$.

$\arg\min_{B\in\mathbb{R}^k} \|A-B\|_F^2 = A_k$

then cost $D(A,\mathcal{C}) = \sum_{c\in\mathcal{C}} \min \|a_i - c\|^2$

$= \|A - S(C)\|_F^2$.

Partition-based k-means:

find a partition $A_1, \cdots, A_k$ of A

to minimize $D((A_i)_{i=1}^k) := \sum_{i=1}^k D(A_i - c(A_i))$

↳ centroid.

Rmk: Partition-based Problem's solutions $\supseteq$ Center-based Problem's solutions

(OPT-par. 不一定是 OPT-cen.).

且若 Center-based Problem find OPT.

then Partition-based Problem find OPT.

Apply JL-lemma:

$|A| = n$. let $\varepsilon \in (0,1)$. $d' \geq \Omega(\frac{\log n}{\varepsilon^2})$. f random projection.

then for any partition $A_1, \cdots, A_k$ of A,

$$|D((A_i)_{i=1}^k) - D(f(A_i)_{i=1}^k)| \leq \varepsilon \cdot D((A_i)_{i=1}^k)$$

(由 JL-lemma 易证)

日期: /

JL-DimRed-k-means(A, k, ε):

1° $\varepsilon' = \frac{\varepsilon}{62}$. $d' \geq \Omega(\frac{\log n}{\varepsilon^2})$. f from JL-lemma.

2° Use a 2-approximation alg. for k-means on $f(A)$. to obtain a partition $(A_{f_i})_{i=1}^k$.

3° Translate $(A_{f_i})_{i=1}^k$ to $(A_i)_{i=1}^k$. Return $(A_i)_{i=1}^k$.

≫ $(2+\varepsilon)$-approximation w.h.p.

IV. Other Problems.

Local Search.

LocalSearch k-means ($A \subseteq \mathbb{R}^d$, $k \subseteq \mathbb{N}$.).

① choose k centers arbitrarily $T = \{c_1, \cdots, c_k\}$.

② while $\exists\ c \in T, c' \in A$ s.t. $D(A, T+\{c'\}-\{c\}) < D(A, T)$.

$C \leftarrow C + \{c'\} - \{c\}$.

③ Return C.

Rmk: 50-approximation.

k-median:

$A \subseteq \mathbb{R}^d$. $d \geq 1$. find $\mathcal{C} \subseteq \mathbb{R}^d$. $|\mathcal{C}| = k$. $D := D_{L_2}$. minimize.

$$D(A, \mathcal{C}) = \sum_{a \in A} \min_{c \in \mathcal{C}} D(a, c).$$

日期: /

Rmk: ① NP-hard.

② Linear-programming ⇒ constant-app.

③ Local search ⇒ 6-app.

k-center:

$A \subseteq \mathbb{R}^d$. $d \geq 1$. find $C \subseteq A$. $|C| = k$. $D := D_{L_2}$. minimize.

$$\bar{D}(A, C) = \sum_{c \in C} \max_{a \in A} D(a, c).$$

Rmk: NP-hard.

Gonzalez's alg: (furthest-first-traversal)

① choose c_1 arbitrarily.

② For $i = 2, \cdots, d$. do:

choose c_i maximize. $\bar{D}(X, C^{i-1})$.

i.e. $c_i = \arg\max_{x \in A} D(x, C^{i-1})$.

$C^i \leftarrow C^{i-1} \cup \{c_i\}$.

Rmk: Time: $O(|A| \cdot k) \cdot O(d)$ → compute distance

Thm: alg above outputs C s.t. $\bar{D}(A, C) \leq 2\bar{D}(A, C^*)$.

V. Coreset by k-means clustering:

Goal of Coreset: n points ⇒ $O(\log n)$ / $O(\frac{k}{\varepsilon})$ points.

日期: /

(Rmk: 不同于 dimension reduction: $A \in \mathbb{R}^{n \times d} \to A' \in \mathbb{R}^{n \times d'}$ ($d' << d$))

Def: Let $A \subseteq \mathbb{R}^d$ n points. $\varepsilon \in (0,1)$. $k \geq 1$.

$S \subseteq \mathbb{R}^d$ with weight function $w: S \to \mathbb{R}^+$ is (k,ε)-Coreset

if for $\forall\ C \subseteq \mathbb{R}^d, |C| = k$. $|D(A,C) - D(S,w,C)| \leq \varepsilon D(A,C)$.

↳ 即 cost function 差别不大.

Construction Alg:

Tools:

① ε-ball-cover:

ε-ball-cover of unit sphere is a point set B

s.t. $\forall p \in$ unit sphere, $d(p,B) \leq \varepsilon$.

lemma: $U \subseteq \mathbb{R}^d$ unit sphere. Then, $\forall \varepsilon \in (0,1)$ $\exists$ ε-ball-cover B

of size $(1+\frac{2}{\varepsilon})^d$ s.t. $\forall p, \min_{b \in B} \|p-b\| \leq \varepsilon$.

Cor: ① $\exists$ alg. construct a cover of size $\varepsilon^{-O(d)}$.

② if U has radius r. then ... $\forall p, \min_{b \in B} \|p-b\| \leq \varepsilon \cdot r$.

② App-Alg: Thm: $\exists$ a 6.357-app. alg for k-means problems.

③ (Generalized Triangle Inequality)

$a, b, c \in \mathbb{R}^d$. for $\forall \varepsilon \in (0,1)$. has.

$$\left| \|a-c\|^2 - \|b-c\|^2 \right| \leq \frac{12}{\varepsilon} \|a-b\|^2 + 2\varepsilon \|a-c\|^2.$$

日期: /

i.e. $C \subseteq \mathbb{R}^d$. $|D(a,C)-D(b,C)| \le \frac{12}{\varepsilon} D(a,b) + 2\varepsilon D(a,C)$.

Alg: ① Input A, k.

② compute 10-app. k-means, maintain $C_1,\cdots,C_k$ $c_1,\cdots,c_k$.

③ for $j=1,\cdots,k$:

1° let $F = C_j$.

2° let B^i be ball with $r_i = \sqrt{\frac{2^i}{n}\sum_{x\in F}\|x-c_j\|^2}$, centered c_j.

and S^i be $\varepsilon/192$-ball-cover of B^i. $i=1,2,\cdots,\log(\Delta n)$.

let $S_j = \bigcup_{i=0}^{\log(\Delta n)} S^i$.

3° $\forall x \in C_j$ let $B(x)$ be the nearest point in S_j.

$\forall y \in S_j$, let $w(y)$ be $\#\{x : B(x)=y\}$

④ $S = \bigcup_{x\in A} B(x)$. Return (S, w).

(Rmk: Size of S: $k \cdot \log(\Delta n) \cdot \varepsilon^{-O(d)}$. ($k$: k-cluster; $\log(\Delta n)$: $\log(\Delta n)$ 个 S^i; $\varepsilon^{-O(d)}$: ε-ball-cover.)

Analysis:

Lemma: $F \subseteq \mathbb{R}^d$ n points. B^i is ball with $r_i = \sqrt{\frac{2^i}{n}\sum_{x\in F}\|x-c_j\|^2}$, centered c_j.

let S^i be $\varepsilon/3$-ball-cover of B^i. $S = \bigcup_{i=0}^{\log(\Delta n)} S^i$. Then,

$$\sum_{x\in F} \min_{s\in S} \|x-s\|^2 \le \varepsilon^2 \sum_{x\in F} \|x-c_j\|^2$$

日期: /

远近分别放缩.

Pf: denote $F_{close} := S^0 = \{y \in F \mid \|y\|^2 \le \frac{1}{n}\sum \|x-c_j\|^2\}$. $F_{far} = F \setminus F_{close}$.

① $\sum_{x\in F_{close}} \min_{s\in S^0} \|x-s\|^2 \le |F_{close}| \cdot \underbrace{\frac{1}{n}\sum_{x\in F}\|x-c_j\|^2}_{(\varepsilon r)^2} \cdot (\frac{\varepsilon}{3})^2 \le \frac{\varepsilon^2}{9}\sum_{x\in F}\|x-c_j\|^2$.

② $\min_{s\in S^i}\|x-s\|^2 \le \frac{\varepsilon^2}{9}\cdot r_i^2 \le \frac{4\varepsilon^2}{9} r_{i-1}^2 \le \frac{4\varepsilon^2}{9}\|x-c_j\|^2$. $(i\ge 1)$.

↓ 归纳(这里 S^0 过强).

$\Rightarrow \sum_{x\in F_{far}} \min_{s\in S}\|x-s\|^2 \le \frac{4\varepsilon^2}{9}\sum_{x\in F_{far}}\|x-c_j\|^2 \le \frac{4\varepsilon^2}{9}\sum_{x\in F}\|x-c_j\|^2$

$$\Rightarrow \sum_{x\in F}\min_{s\in S}\|x-s\|^2 \le \sum_{x\in F_{close}}\min_{s\in S}\|x-s\|^2 + \sum_{x\in F_{far}}\min_{s\in S}\|x-s\|^2$$
$$\le \frac{\varepsilon^2}{9}\sum_{x\in F}\|x-c_j\|^2 + \frac{4\varepsilon^2}{9}\sum_{x\in F}\|x-c_j\|^2$$
$$\le \varepsilon^2 \sum_{x\in F}\|x-c_j\|^2$$

Thm:

Pf: $\forall C \subseteq \mathbb{R}^d,\ |C|=k$.

$$|D(A,C) - D(S,w,C)| = \Big|\sum_{x\in A}\min_{c\in C}\|x-c\|^2 - \sum_{x\in A}\min_{c\in C}\|B(x)-c\|^2\Big|$$
$$\le \sum_{x\in A}\min_{c\in C}\Big|\|x-c\|^2 - \|B(x)-c\|^2\Big|$$
$$\le \sum_{x\in A}\min_{c\in C}\Big(\underbrace{\tfrac{12}{\varepsilon}\|x-B(x)\|^2}_{\sum \frac{12}{\varepsilon}\|x-B(x)\|^2 \le \frac{12}{\varepsilon}\varepsilon^2\sum\|x-c_j\|^2 \le 12\varepsilon\cdot 6.357\cdot\sum\min\|x-c\|^2.} + 2\varepsilon\|x-c\|^2\Big) \quad \text{(triangle ineq.)}$$
$$\le O(\varepsilon)\cdot D(A,C). \quad (\varepsilon\text{-ball-cover, lemma, } k\text{-means})$$

(Rmk: 除了 "ε-ball-cover", 还常用 "Importance Sampling" 构造 Coreset.

即. 为 $a\in A$ def importance (eg. by the prop. to be center).

sampling by the importance.).

日期: /

VI. Coreset in Streaming:

Lemma (Composability(可组合性)).

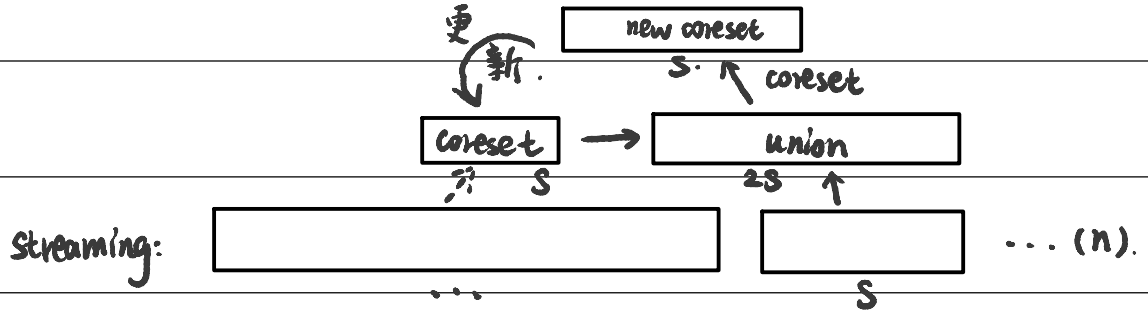
$A_1, A_2 \subseteq \mathbb{R}^d$ 不相交. 设 (S_1, w_1) is A_1's Coreset.

(S_2, w_2) is A_2's Coreset.

则 $(S_1 \cup S_2, w_1 \cup w_2)$ is $A_1 \cup A_2$'s coreset.

Merge & Reduce.

每流进s个 data, 求 coreset. 与之前的 coreset 合并(按树的方式)



(Rmk: need to store ≤2 coresets each level 同时.)

coreset size = $S = O(k \log n\, \varepsilon'^{-d}) = O(k(\log n)^{d+1} \varepsilon^{-d})$.

$\varepsilon' = \frac{\varepsilon}{\log n}$ (因为误差产生了累积). → 相当.

VII. Hierarchical Clustering:

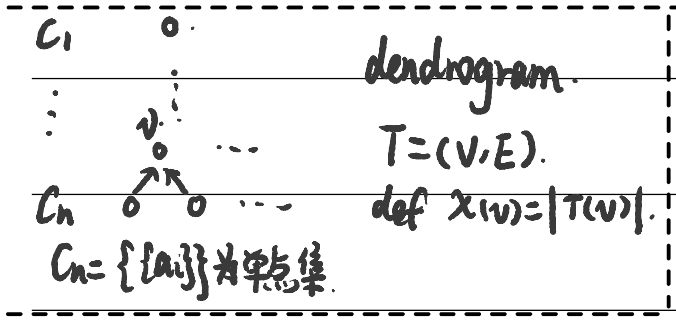
a sequence of clustering: $C_1, \cdots, C_n$. with $|C_k| = C_k$.

is called hierarchical clustering of A if $\forall H \in C_k$:

日期: /

1° $H \in C_{k+1}$

or 2° $\exists B, C \in C_{k+1}$, $H = B \cup C$, $C_k = (C_{k+1} \setminus \{B\} \cup \{C\}) \cup \{H\}$.



Construct 1: 自顶向下.

Main Idea: 每层对其中类做二分类, 直到划分为单点集.

Construct 2: 自底向上.

Main Idea:

① $C_n = \{C^1, \cdots, C^n\}$. $C^i = \{a_i\}$.

② each step, replace C^i, C^j (which are "closest") by $C^i \cup C^j$.

until get C_1.

③ return $C_1, \cdots, C_n$.

How to describe "closest"?

1. Agg CL: (基于 complete linkage).

用 $D_{CL}(C^1, C^2) = \max_{\substack{x \in C^1 \\ y \in C^2}} D(x, y)$ 衡量 "closest".

日期: /

Def: 直径: $diam^D(S) = \max_{x,y \in S} D(x,y)$

直径代价 $cost^D_{diam}(\mathbb{C}_k) = \max_{1 \le i \le k} diam^D(C_i)$.

Problem: $opt^{diam}_k(A) = \min_{|\mathbb{C}|=k} cost^D_{diam}(\mathbb{C})$.

=> CL 优化 直径代价

Thm: Agg CL computes a k-clustering $\mathbb{C}_k$ with:

$cost^D_{diam}(\mathbb{C}_k) \le O_d(\log k)\, opt^{diam}_k(A)$.

↳ 与 n 无关.

for $\forall\, k \le |A|$.

Rmk: ① $O_d(\cdot)$ 表示 hide 了 d 的依赖. 即 $O_d(\cdot) = g(d)\cdot(\cdot)$, 一般表示可将 d 视作常数.

② 对 $\forall k$ 可给出近似解.

2. Agg SL: (基于 single linkage).

用 $D_{SL}(C^1, C^2) = \min_{\substack{x \in C^1 \\ y \in C^2}} D(x,y)$ 衡量 "closest".

Rmk: SL 优化 Problem: maximize: $\min_{\substack{x \in C_i \\ y \in C_j \\ i \ne j}} D(x,y)$.

3. Agg ML: (基于 mean linkage).

用 $D_{SL}(C^1, C^2) = \text{mean}_{\substack{x \in C^1 \\ y \in C^2}} D(x,y)$ 衡量 "closest".

Construct 3: (Gonzalez's alg-based)

日期: /

full-furthest-first-traversal:

① choose c_1 arbitrarily.

② For $i=2,\cdots,|A|$ do:

choose c_i maximize $R_i = \max_{x \in A} D(x, C^{i-1})$.

$C^i \leftarrow C^{i-1} \cup \{c_i\}$.

③ maitain $c_1,\cdots,c_{|A|}$. $R_2,\cdots,R_{|A|}$. (rmk: $\{c_1,\cdots,c_{|A|}\} = A$).

---

④ define $L_0 = \{c_1\}$. $R = R_2$ (largest) $L_i = \{c_j \mid R_j \in (\frac{R}{2^i}, \frac{R}{2^{i-1}}]\}$

define $\text{parent}(c_i) = \arg\min_{y \in \bigcup_{j=0}^{i-1} L_j} D(c_i, y)$.

⑤ i from n to 1: merge c_i with parent(c_i). as parent(c_i).