بسم الله الرحمن الرحیم

جلسه نهم

خلاصه‌سازی برای مه‌داده

مرور

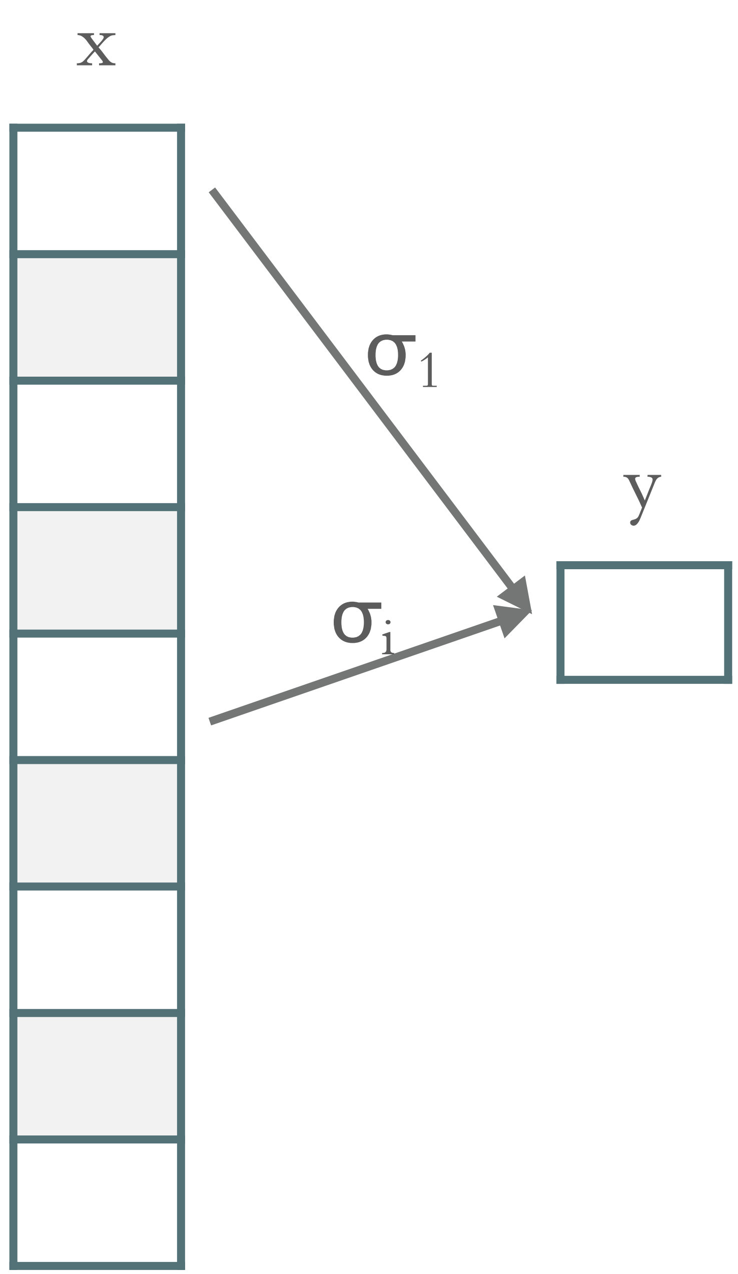
- خلاصه‌سازی خطی ==
  - به جای x : Πx
  - به‌روزرسانی ساده

تقریب نرم P:

- ساختمان داده: آرایه x با عملیات افزایش/کاهش یک خانه
- مسئله Fp: $\|x\|_p^p = \sum_{i=1}^{p} |x_i|^p$
  - F0: تعداد متفاوت‌ها (حالت فقط افزایش)
- هدف: تخمین‌گر a که با احتمال $1-\delta$ داشته باشیم $|a - F_p| \leq \epsilon F_p$
  - زیر هدف: تخمین‌گر a که با احتمال ۲/۳ داشته باشیم $|a - F_p| \leq \epsilon F_p$
- روش کلی: خلاصه‌سازی خطی
  - نگه‌داری $y = \Pi x$ (فقط m عدد به جای n عدد)
  - به‌روزرسانی ساده

ایده AMS:

$$EY = 0$$



$$EY^2 = E\left[\sum_{j,j'} \sigma_j \sigma_{j'} x_j x_{j'}\right]$$

$$= E\left[\sum_{j} \sigma_j^2 x_j^2 + \sum_{j \neq j'} \sigma_j \sigma_{j'} x_j x_{j'}\right]$$

$$= \sum_{j} E[x_j^2] + \sum_{j \neq j'} E[\sigma_j \sigma_{j'} x_j x_{j'}]$$

$$= ||x||_2^2$$

$E[\sigma_j]E[\sigma_{j'}]x_j x_{j'}$

و $E[\sigma_j] = 0$

$\sigma_i = +1$ یا -1 (تصادفی یکنواخت)

روش AMS

- شروع:
  - $\sigma \in \{-1,1\}^{m\times n}$ که استقلال ۴_طرفه
  - $\Pi_{i,j} = \sigma_{i,j}/\sqrt{m}$

حافظه: O(log (mn))

- به‌روزرسانی:
  - همیشه $y = \Pi x$ را نگه‌دار (برای x_i+=1، ستون i_ام Π را به y اضافه کن)
  - داریم: $y_i = \sum_{j=1}^{n} \sigma_{i,j} x_j/\sqrt{m}$
- تخمین‌گر:
  - $\|\Pi x\|_2^2 = \|y\|_2^2$

تحلیل AMS:

$$\begin{aligned}
\mathbb{E}\, y_r^2 &= \frac{1}{m}\,\mathbb{E}\left(\sum_{j=1}^{n}\sigma_{r,j}x_j\right)^2 \\
&= \frac{1}{m}\left[\|x\|_2^2 + \mathbb{E}\sum_{j\neq j'}\sigma_{r,j}\sigma_{r,j'}x_jx_{j'}\right] \\
&= \frac{1}{m}\left[\|x\|_2^2 + \sum_{j\neq j'}(\mathbb{E}\,\sigma_{r,j}\sigma_{r,j'})x_jx_{j'}\right] \\
&= \frac{1}{m}\left[\|x\|_2^2 + \sum_{j\neq j'}(\mathbb{E}\,\sigma_{r,j})(\mathbb{E}\,\sigma_{r,j'})x_jx_{j'}\right] \\
&= \frac{1}{m}\|x\|_2^2,
\end{aligned}$$

پس

$\|x\|_2^2 = \mathrm{E}[\,\|y\|_2^2\,]$

زیرا $\|y\|_2^2 = \sum_{r=1}^{m} y_r^2$

تحلیل AMS: واریانس تخمین‌گر

اندیس‌های $E[||y||_2^2]$
j=j' را حذف می‌کند

$$\mathbb{E}(\|y\|_2^2 - \mathbb{E}\,\|y\|_2^2)^2 = \frac{1}{m^2}\,\mathbb{E}(\sum_{r=1}^{m}\sum_{j\neq j'}\sigma_{r,j}\sigma_{r,j'}x_j x_{j'})^2$$

تحلیل AMS: واریانس تخمین‌گر

$E[||y||_2^2]$ اندیس‌های
j=j’ را حذف می‌کند

$$\mathbb{E}(\|y\|_2^2 - \mathbb{E}\,\|y\|_2^2)^2 = \frac{1}{m^2}\,\mathbb{E}(\sum_{r=1}^{m}\sum_{j\neq j'}\sigma_{r,j}\sigma_{r,j'}x_j x_{j'})^2$$

$$= \frac{1}{m^2}\sum_{r_1,r_2}\sum_{\substack{j_1\neq j_2\\ j_3\neq j_4}}(\mathbb{E}\,\sigma_{r_1,j_1}\sigma_{r_1,j_2}\sigma_{r_2,j_3}\sigma_{r_2,j_4})x_{r,j_1}x_{r,j_2}x_{r,j_3}x_{r,j_4}$$

تحلیل AMS: واریانس تخمین‌گر

$E[||y||_2^2]$ اندیس‌های j=j’ را حذف می‌کند

$$\mathbb{E}(\|y\|_2^2 - \mathbb{E}\,\|y\|_2^2)^2 = \frac{1}{m^2}\,\mathbb{E}(\sum_{r=1}^{m}\sum_{j\neq j'}\sigma_{r,j}\sigma_{r,j'}x_j x_{j'})^2$$

$$= \frac{1}{m^2}\sum_{r_1,r_2}\sum_{\substack{j_1\neq j_2\\ j_3\neq j_4}}(\mathbb{E}\,\sigma_{r_1,j_1}\sigma_{r_1,j_2}\sigma_{r_2,j_3}\sigma_{r_2,j_4})x_{r,j_1}x_{r,j_2}x_{r,j_3}x_{r,j_4}$$

اگر توان یکی فرد باشد،
عبارت = ۰

پس باید
$j_1=j_3$ و $j_2=j_4$
یا
$j_1=j_4$ و $j_2=j_3$

$$= \frac{2}{m}\sum_{j_1\neq j_2} x_{j_1}^2 x_{j_2}^2$$

تحلیل AMS: واریانس تخمین‌گر

اندیس‌های $E[||y||_2^2]$
j=j’ را حذف می‌کند

$$\mathbb{E}(\|y\|_2^2 - \mathbb{E}\,\|y\|_2^2)^2 = \frac{1}{m^2}\,\mathbb{E}(\sum_{r=1}^{m}\sum_{j\neq j'} \sigma_{r,j}\sigma_{r,j'}x_jx_{j'})^2$$

$$= \frac{1}{m^2}\sum_{r_1,r_2}\sum_{\substack{j_1\neq j_2\\ j_3\neq j_4}} (\mathbb{E}\,\sigma_{r_1,j_1}\sigma_{r_1,j_2}\sigma_{r_2,j_3}\sigma_{r_2,j_4})x_{r,j_1}x_{r,j_2}x_{r,j_3}x_{r,j_4}$$

اگر توان یکی فرد باشد،
عبارت = ۰

پس باید
$j_1=j_3$ و $j_2=j_4$
یا
$j_1=j_4$ و $j_2=j_3$

$$= \frac{2}{m}\sum_{j_1\neq j_2} x_{j_1}^2 x_{j_2}^2$$

$$\leq \frac{2}{m}\|x\|_2^4,$$

تحلیل نهایی AMS:

چبیشف

$$\forall\lambda > 0, \mathbb{P}(|X - \mathbb{E}\,X| > \lambda) < \frac{\mathbb{E}(X - \mathbb{E}\,X)^2}{\lambda^2}$$

$$P\left[\left|\ ||y||_2^2 - ||x||_2^2\ \right| > \epsilon||x||_2^2\right] < Var(||y||_2^2)/(\epsilon||x||_2^2)^2$$

تحلیل نهایی AMS:

چبیشف

$$\forall\lambda > 0, \mathbb{P}(|X - \mathbb{E}\,X| > \lambda) < \frac{\mathbb{E}(X - \mathbb{E}\,X)^2}{\lambda^2}$$

$$P\left[\left|\ ||y||_2^2 - ||x||_2^2\ \right| > \epsilon||x||_2^2\right] < Var(||y||_2^2)/(\epsilon||x||_2^2)^2$$

$$< \frac{2}{m}||x||_2^4 \frac{1}{\epsilon^2||x||_2^4}$$

$$\text{Var} \leq \frac{2}{m}\|x\|_2^4,$$

تحلیل نهایی AMS:

چبیشف

$$\forall \lambda > 0, \mathbb{P}(|X - \mathbb{E}\,X| > \lambda) < \frac{\mathbb{E}(X - \mathbb{E}\,X)^2}{\lambda^2}$$

$$P\left[\left|\ ||y||_2^2 - ||x||_2^2\ \right| > \epsilon ||x||_2^2\right] < Var(||y||_2^2)/(\epsilon ||x||_2^2)^2$$

$$< \frac{2}{m} ||x||_2^4 \frac{1}{\epsilon^2 ||x||_2^4}$$

$$= \frac{2}{\epsilon^2 m}$$

$\mathrm{Var} \leq \frac{2}{m}\|x\|_2^4,$

تحلیل نهایی AMS:

$$\forall \lambda > 0, \mathbb{P}(|X - \mathbb{E}\,X| > \lambda) < \frac{\mathbb{E}(X - \mathbb{E}\,X)^2}{\lambda^2}$$

چبیشف

$$P\left[\left|\ ||y||_2^2 - ||x||_2^2\ \right| > \epsilon ||x||_2^2\right] < Var(||y||_2^2)/(\epsilon ||x||_2^2)^2$$

$$< \frac{2}{m} ||x||_2^4 \frac{1}{\epsilon^2 ||x||_2^4}$$

Var $\leq \frac{2}{m}\|x\|_2^4,$

$$= \frac{2}{\epsilon^2 m}$$

$$\leq ⅓$$

m بزرگ

$(m = 6/\epsilon^2)$

تحلیل نهایی AMS:

چبیشف

$$\forall \lambda > 0, \mathbb{P}(|X - \mathbb{E}\,X| > \lambda) < \frac{\mathbb{E}(X - \mathbb{E}\,X)^2}{\lambda^2}$$

$$P\left[\Big|\ ||y||_2^2 - ||x||_2^2\ \Big| > \epsilon ||x||_2^2\right] < Var(||y||_2^2)/(\epsilon ||x||_2^2)^2$$

$$\mathrm{Var} \le \frac{2}{m}\|x\|_2^4,$$

$$< \frac{2}{m}||x||_2^4 \frac{1}{\epsilon^2 ||x||_2^4}$$

$$= \frac{2}{\epsilon^2 m}$$

$$\leq 1/3$$

m بزرگ

$(m = 6/\epsilon^2)$

پس

تقریب نرم ۲ (روش AMS)

حافظه: $\epsilon^{-2} \log 1/\delta$ تا عدد

ارتقاء AMS:

- زمان اجرای به‌روزرسانی: $y = \Pi x$
- $h : [n] \to [m]$ استقلال ۲ـ طرفه
- $\sigma \in \{-1, 1\}^n$ استقلال ۴ـ طرفه
- هر ستون j:
  - $\Pi_{h(j),j} = \sigma_j$
  - بقیه ستون صفر
- داریم
  - $\mathbb{E}\,\|\Pi z\|_2^2 = \|z\|_2^2$ $\qquad Var[\|\Pi z\|_2^2] = O(1/m)\|z\|_2^4$
  - پس: با احتمال حداقل ۲/۳ داریم $\|\Pi z\|_2^2$ یک $1+\epsilon$ تقریب است.