CSCI B609: “Foundations of Data Science”

Lecture 8: Faster Power Method and Applications of SVD

Slides at http://grigory.us/data-science-class.html

Grigory Yaroslavtsev

http://grigory.us

Faster Power Method

- PM drawback: $A^T A$ is dense even for sparse A
- Pick random Gaussian $\boldsymbol{x}$ and compute $B^k \boldsymbol{x}$
- $\boldsymbol{x} = \sum_{i=1}^{\boldsymbol{d}} c_i \boldsymbol{v}_i$(augment $\boldsymbol{v_i}$'s to o.n.b. if $r < \boldsymbol{d}$)
- $B^k \boldsymbol{x} \approx \left(\sigma_1^{2k} \boldsymbol{v}_1 \boldsymbol{v}_1^T\right)\left(\sum_{i=1}^{d} c_i \boldsymbol{v}_i\right) = \sigma_1^{2k} c_1 \boldsymbol{v}_1$

$$B^k \boldsymbol{x} = (A^T A)(A^T A) \ldots (A^T A)\boldsymbol{x}$$

- **Theorem:** If $\boldsymbol{x}$ is unit $\mathbb{R}^{\boldsymbol{d}}$-vector, $|\boldsymbol{x}^T \boldsymbol{v}_1| \geq \boldsymbol{\delta}$:
  - V = subspace spanned by $\boldsymbol{v}_i' s$ for $\sigma_j \geq (1 - \boldsymbol{\epsilon})\sigma_1$
  - $\boldsymbol{w}$ = unit vector after $k = \frac{1}{2\boldsymbol{\epsilon}} \ln\left(\frac{1}{\boldsymbol{\epsilon\delta}}\right)$ iterations of PM

  $\Rightarrow$ $\boldsymbol{w}$ has a component at most $\boldsymbol{\epsilon}$ orthogonal to V

Faster Power Method: Analysis

- $A = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$ and $\boldsymbol{x} = \sum_{i=1}^{\boldsymbol{d}} c_i \boldsymbol{v}_i$
- $B^k \boldsymbol{x} = \sum_{i=1}^{\boldsymbol{d}} \sigma_i^{2k} \boldsymbol{v}_i \boldsymbol{v}_i^T \sum_{j=1}^{\boldsymbol{d}} c_j \boldsymbol{v}_j = \sum_{i=1}^{\boldsymbol{d}} \sigma_i^{2k} c_i \boldsymbol{v}_i$

$$\left|\left|B^k \boldsymbol{x}\right|\right|_2^2 = \left\|\sum_{i=1}^{\boldsymbol{d}} \sigma_i^{2k} c_i \boldsymbol{v}_i\right\|_2^2 = \sum_{i=1}^{\boldsymbol{d}} \sigma_i^{4k} c_i^2 \geq \sigma_1^{4k} c_1^2 \geq \sigma_i^{4k} \delta^2$$

- (Squared) component orthogonal to V is

$$\sum_{i=m+1}^{\boldsymbol{d}} \sigma_i^{4k} c_i^2 \leq (1-\epsilon)^{4k} \sigma_1^{4k} \sum_{i=m+1}^{\boldsymbol{d}} c_i^2 \leq \ (1-\epsilon)^{4k} \sigma_1^{4k}$$

- Component of $\boldsymbol{w} \perp V \leq (1-\epsilon)^{2k} / \boldsymbol{\delta} \leq \epsilon$

Choice of $\boldsymbol{x}$

- $\boldsymbol{y}$ random spherical Gaussian with unit variance
- $\boldsymbol{x} = \frac{\boldsymbol{y}}{||\boldsymbol{y}||_2}$:
$$Pr\left[|\boldsymbol{x}^T\boldsymbol{v}| \le \frac{1}{20\sqrt{\boldsymbol{d}}}\right] \le \frac{1}{10} + 3e^{-\boldsymbol{d}/64}$$
- $Pr\left[||\boldsymbol{y}||_2 \ge \mathbf{2}\sqrt{\boldsymbol{d}}\right] \le 3e^{-\boldsymbol{d}/64}$ (Gaussian Annulus)
- $\boldsymbol{y}^T\boldsymbol{v} \sim N(0,1) \Rightarrow \Pr\left[\left||\boldsymbol{y}^T\boldsymbol{v}|\right|_2 \le \frac{1}{10}\right] \le \frac{1}{10}$
- Can set $\boldsymbol{\delta} = \frac{1}{20\sqrt{\boldsymbol{d}}}$ in the “faster power method”

Singular Vectors and Eigenvectors

- Right singular vectors are eigenvectors of $A^T A$
- σ_i^2 are eigenvalues of $A^T A$
- Left singular vectors are eigenvectors of AA^T
- $A^T A$ satisfies $\forall \boldsymbol{x}$: $\boldsymbol{x}^T B \boldsymbol{x} \geq 0$
  - $B = \sum_i \sigma_i^2 \boldsymbol{v}_i \boldsymbol{v}_i^T$
  - $\forall \boldsymbol{x}$: $\boldsymbol{x}^T \boldsymbol{v}_i \boldsymbol{v}_i^T \boldsymbol{x} = (\boldsymbol{x}^T \boldsymbol{v}_i)^2 \geq 0$
  - Such matrices are called positive semi-definite
- Any p.s.d matrix can be decomposed as $A^T A$

Application of SVD: Centering Data

- Minimize sum of squared distances from $\boldsymbol{A_i}$ to S_k
- **SVD**: best fitting S_k if data is centered
- What if not?
- **Thm.** S_k that minimizes squared distance goes through centroid of the point set:
$$\frac{1}{n}\sum \boldsymbol{A_i}$$
- Will only prove for $k = 1$, analogous proof for arbitrary k (see textbook)

Application of SVD: Centering Data

- **Thm.** Line that minimizes squared distance goes through the centroid
- Line: $\ell = \boldsymbol{a} + \lambda \boldsymbol{v}$; distance $dist(\boldsymbol{A_i}, \ell)$
- $\left||\boldsymbol{A_i} - \boldsymbol{a}|\right|_2^2 = dist(\boldsymbol{A_i}, \ell)^2 + \langle \boldsymbol{v}, \boldsymbol{A_i} \rangle^2$
- Center so that $\sum_{i=1}^{n} \boldsymbol{A_i} = \boldsymbol{0}$ by subtracting the centroid
- $\sum_i^n dist(\boldsymbol{A_i}, \ell)^2 = \sum_{i=1}^{n} (\left||\boldsymbol{A_i} - \boldsymbol{a}|\right|_2^2 - \langle \boldsymbol{v}, \boldsymbol{A_i} \rangle^2)$

$$= \sum_{i=1}^{n} (\left||\boldsymbol{A_i}|\right|_2^2 + \left||\boldsymbol{a}|\right|_2^2 - 2\langle \boldsymbol{A_i}, \boldsymbol{a} \rangle - \langle \boldsymbol{v}, \boldsymbol{A_i} \rangle^2)$$

$$= \sum_{i=1}^{n} \left||\boldsymbol{A_i}|\right|_2^2 + n\left||\boldsymbol{a}|\right|_2^2 - 2\langle \sum_{i=1}^{n} \boldsymbol{A_i}, \boldsymbol{a} \rangle - \sum_{i=1}^{n} \langle \boldsymbol{v}, \boldsymbol{A_i} \rangle^2$$

$$= \sum_{i=1}^{n} \left||\boldsymbol{A_i}|\right|_2^2 + n\left||\boldsymbol{a}|\right|_2^2 - \sum_{i=1}^{n} \langle \boldsymbol{v}, \boldsymbol{A_i} \rangle^2$$

- Minimized when $\boldsymbol{a} = \boldsymbol{0}$

Principal Component Analysis

- $\boldsymbol{n} \times \boldsymbol{d}$ matrix: customers×movies preference
- $\boldsymbol{n}$ = #customers, $\boldsymbol{d}$ = #movies
- A_{ij} = how much customer i likes movie j
- Assumption: A_{ij} can be described with k factors
  - Customers and movies: vectors in $\boldsymbol{u_i}$ and $\boldsymbol{v_i} \in \mathbb{R}^k$
  - $A_{ij} = \langle \boldsymbol{u_i}, \boldsymbol{v_j} \rangle$
- Solution: A_k

$$\text{customers}\begin{pmatrix} \\ \\ & A & \\ \\ \\ \end{pmatrix} = \overset{\text{factors}}{\begin{pmatrix} \\ \\ U \\ \\ \\ \end{pmatrix}} \overset{\text{movies}}{\begin{pmatrix} & V & \end{pmatrix}}$$

Separating mixture of k Gaussians

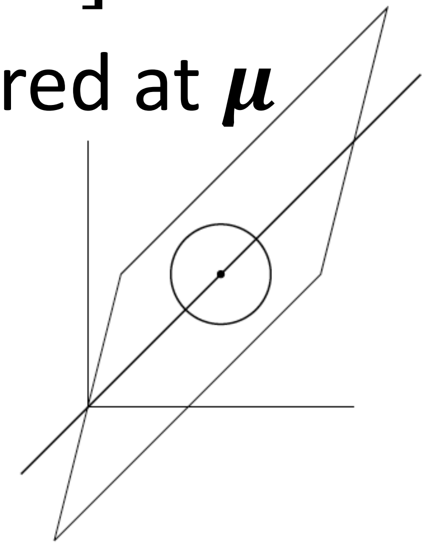
- **Sample origin problem**:
  - Given samples from $\boldsymbol{k}$ **well-separated** spherical Gaussians
  - **Q**: Did they come from the same Gaussian?
- δ = distance between centers
- For two Gaussians naïve separation requires
$$\delta > \omega\left(\boldsymbol{d}^{1/4}\right)$$
- **Thm.** $\delta = \Omega(\boldsymbol{k}^{\frac{1}{4}})$ suffices
- **Idea:**
  - Project on a $\boldsymbol{k}$-dimensional subspace through centers
  - **Key fact:** This subspace can be found via SVD
  - Apply naïve algorithm

Separating mixture of k Gaussians

- **Easy fact:** Projection preserves the property of being a unit-variance spherical Gaussian
- **Def.** If p is a probability distribution, **best fit line** $\{c\boldsymbol{v}, c \in \mathbb{R}\}$ is:

$$\boldsymbol{v} = argmax_{|\boldsymbol{v}|=1}\, \mathbb{E}_{\boldsymbol{x}\sim p}\left[\left(\boldsymbol{v}^{\boldsymbol{T}}\boldsymbol{x}\right)^{2}\right]$$

- **Thm:** Best fit line for a Gaussian centered at $\boldsymbol{\mu}$ passes through $\boldsymbol{\mu}$ and the origin



Best fit line for a Gaussian

- **Thm:** Best fit line for a Gaussian centered at $\boldsymbol{\mu}$ passes through $\boldsymbol{\mu}$ and the origin

$$\mathbb{E}_{x\sim p}\left[\left(\boldsymbol{v}^T\boldsymbol{x}\right)^2\right] = \mathbb{E}_{x\sim p}\left[\left(\boldsymbol{v}^T(\boldsymbol{x}-\boldsymbol{\mu}) + \boldsymbol{v}^T\boldsymbol{\mu}\right)^2\right]$$
$$= \mathbb{E}_{x\sim p}\left[\boldsymbol{v}^T(\boldsymbol{x}-\boldsymbol{\mu})^2 + 2(\boldsymbol{v}^T\boldsymbol{\mu})\boldsymbol{v}^T(\boldsymbol{x}-\boldsymbol{\mu}) + (\boldsymbol{v}^T\boldsymbol{\mu})^2\right]$$
$$= \mathbb{E}_{x\sim p}\left[\boldsymbol{v}^T(\boldsymbol{x}-\boldsymbol{\mu})^2\right] + 2(\boldsymbol{v}^T\boldsymbol{\mu})\mathbb{E}_{x\sim p}[\boldsymbol{v}^T(\boldsymbol{x}-\boldsymbol{\mu})] + (\boldsymbol{v}^T\boldsymbol{\mu})^2$$
$$= \mathbb{E}_{x\sim p}\left[\boldsymbol{v}^T(\boldsymbol{x}-\boldsymbol{\mu})^2\right] + (\boldsymbol{v}^T\boldsymbol{\mu})^2$$
$$= \sigma^2 + (\boldsymbol{v}^T\boldsymbol{\mu})^2$$

- Where we used:
  - $\mathbb{E}_{x\sim p}[\boldsymbol{v}^T(\boldsymbol{x}-\boldsymbol{\mu})] = \boldsymbol{0}$
  - $\mathbb{E}_{x\sim p}[\boldsymbol{v}^T(\boldsymbol{x}-\boldsymbol{\mu})^2] = \sigma^2$
- Best fit line maximizes $(\boldsymbol{v}^T\boldsymbol{\mu})^2$