CSCI B609: “Foundations of Data Science”

Lecture 8/9: 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 $x$ and compute $B^k x$
- $x = \sum_{i=1}^{d} c_i v_i$ (augment $v_i$’s to o.n.b. if $r < d$)
- $B^k x \approx \left( \sigma_1^{2k} v_1 v_1^T \right) \left( \sum_{i=1}^{d} c_i v_i \right) = \sigma_1^{2k} c_1 v_1$
  $B^k x = (A^T A)(A^T A) \ldots (A^T A)x$

**Theorem:** If $x$ is unit $\mathbb{R}^d$-vector, $|x^T v_1| \geq \delta$:
- $V = \text{subspace spanned by } v_i$’s for $\sigma_j \geq (1 - \epsilon) \sigma_1$
- $w = \text{unit vector after } k = \frac{1}{2\epsilon} \ln \left( \frac{1}{\epsilon \delta} \right) \text{ iterations of PM}$
  $\Rightarrow w$ has a component at most $\epsilon$ orthogonal to $V$
Faster Power Method: Analysis

- \( A = \sum_{i=1}^{r} \sigma_i u_i v_i^T \) and \( x = \sum_{i=1}^{d} c_i v_i \)
- \( B^k x = \sum_{i=1}^{d} \sigma_i^{2k} v_i v_i^T \sum_{j=1}^{d} c_j v_j = \sum_{i=1}^{d} \sigma_i^{2k} c_i v_i \)

\[
\|B^k x\|_2^2 = \left\| \sum_{i=1}^{d} \sigma_i^{2k} c_i v_i \right\|_2^2 = \sum_{i=1}^{d} \sigma_i^{4k} c_i^2 \geq \sigma_1^{4k} c_1^2 \geq \sigma_i^{4k} \delta^2
\]

- (Squared) component orthogonal to \( V \) is

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

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

- $y$ random spherical Gaussian with unit variance
- $x = \frac{y}{\|y\|_2}$
  
  $$
  \Pr \left[ |x^T v| \leq \frac{1}{20\sqrt{d}} \right] \leq \frac{1}{10} + 3e^{-d/64}
  $$

- $\Pr \left[ \|y\|_2 \geq 2\sqrt{d} \right] \leq 3e^{-d/64}$ (Gaussian Annulus)
- $y^T v \sim N(0,1) \Rightarrow \Pr \left[ \|y^T v\|_2 \leq \frac{1}{10} \right] \leq \frac{1}{10}$
- Can set $\delta = \frac{1}{20\sqrt{d}}$ in the “faster power method”
Singular Vectors and Eigenvectors

- Right singular vectors are eigenvectors of $A^T A$
- $\sigma_i^2$ are eigenvalues of $A^T A$
- Left singular vectors are eigenvectors of $AA^T$
- $A^T A$ satisfies $\forall x: x^T B x \geq 0$
  - $B = \sum_i \sigma_i^2 v_i v_i^T$
  - $\forall x: x^T v_i v_i^T x = (x^T v_i)^2 \geq 0$
  - 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 $A_i$ 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:

$$\frac{1}{n} \Sigma A_i$$

• 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: \( \ell = \mathbf{a} + \lambda \mathbf{v} \); distance \( \text{dist}(A_i, \ell) \)
- \( \|A_i - \mathbf{a}\|^2 = \text{dist}(A_i, \ell)^2 + \langle \mathbf{v}, A_i \rangle^2 \)
- Center so that \( \sum_{i=1}^n A_i = \mathbf{0} \) by subtracting the centroid
- \( \sum_{i=1}^n \text{dist}(A_i, \ell)^2 = \sum_{i=1}^n (\|A_i - \mathbf{a}\|^2 - \langle \mathbf{v}, A_i \rangle^2) \)
  \[ = \sum_{i=1}^n (\|A_i\|^2 + \|\mathbf{a}\|^2 - 2\langle \mathbf{a}, A_i \rangle - \langle \mathbf{v}, A_i \rangle^2) \]
  \[ = \sum_{i=1}^n \|A_i\|^2 + n\|\mathbf{a}\|^2 - 2\sum_{i=1}^n \langle \mathbf{a}, A_i \rangle - \sum_{i=1}^n \langle \mathbf{v}, A_i \rangle^2 \]
  \[ = \sum_{i=1}^n \|A_i\|^2 + n\|\mathbf{a}\|^2 - \sum_{i=1}^n \langle \mathbf{v}, A_i \rangle^2 \]
- Minimized when \( \mathbf{a} = \mathbf{0} \)
Principal Component Analysis

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

• Survey of 3-5 research papers
  – Closely related to the topics of the class
    • Algorithms for high-dimensional data
    • Fast algorithms for numerical linear algebra
    • Algorithms for machine learning and/or clustering
    • Algorithms for streaming and massive data
  – Office hours if you need suggestions
  – Individual (not a group) project
  – 1-page Proposal Due: October 31, 2016 at 23:59 EST
  – Final Deadline: December 09, 2016 at 23:59 EST

• Submission by e-mail to Lisul Islam (IU id: islammdl)
  – Submission Email Title: Project + Space + “Your Name”
  – Submission format: PDF from LaTeX
Separating mixture of $k$ Gaussians

- **Sample origin problem:**
  - Given samples from $k$ well-separated spherical Gaussians
  - Q: Did they come from the same Gaussian?
- $\delta =$ distance between centers
- For two Gaussians naïve separation requires
  \[ \delta > \omega \left( d^{1/4} \right) \]
- Thm. $\delta = \Omega \left( k^{-1/4} \right)$ suffices
- **Idea:**
  - Project on a $k$-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, the best fit line \( \{cv, c \in \mathbb{R}\} \) is:

\[
v = \arg \max_{|v|=1} \mathbb{E}_{x \sim p} \left[ (v^T x)^2 \right]
\]

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

- **Thm:** Best fit line for a Gaussian centered at \( \mu \) passes through \( \mu \) and the origin

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

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

- Best fit line maximizes \( (v^T \mu)^2 \)
Best fit subspace for one Gaussian

• Best fit $k$-dimensional subspace $V_k$:

$$V_k = \arg\max_{V: \text{dim}(V) = k} \mathbb{E}_{x \sim p} \left[ ||\text{proj}(x, V)||^2_2 \right]$$

• For a spherical Gaussian $V$ is a best-fit $k$-dimensional subspace iff it contains $\mu$

• If $\mu = 0$ then any $k$-dim. subspace is best fit

• If $\mu \neq 0$ then best fit line $v$ goes through $\mu$
  – Same greedy process as SVD projects on $v$
  – After projection we have Gaussian with $\mu = 0$
  – Any $(k - 1)$-dimensional subspace would do
Best fit subspace for $k$ Gaussians

- **Thm.** $p$ is a mixture of $k$ spherical Gaussians $\Rightarrow$ best fit $k$-dim. subspace contains their centers

- $p = w_1 p_1 + w_2 p_2 + \cdots + w_k p_k$

- Let $V$ be a subspace of dimension $\leq k$

$$\mathbb{E}_{x \sim p} \left[ \left\| \text{proj}(x, V) \right\|_2^2 \right] = \sum_{i=1}^{k} w_i \mathbb{E}_{x \sim p_i} \left[ \left\| \text{proj}(x, V) \right\|_2^2 \right]$$

- Each term is maximized if $V$ contains all $\mu_i$'s

- If we only have a finite number of samples then accuracy has to be analyzed carefully
HITS Algorithm for Hubs and Authorities

- Document ranking: project on 1st singular vector
- WWW: directed graph with links = edges
- \( n \) Authorities: pages containing original info
- \( d \) Hubs: collections of links to authorities
  - Authority depends on importance of pointing hubs
  - Hub quality depends on how authoritative links are

  Authority vector: \( v_j, j = 1, \ldots, n \): \( v_j \sim \sum_{i=1}^{d} u_i A_{ij} \)

  Hub vector: \( u_i, i = 1, \ldots, d \): \( u_i \sim \sum_{j=1}^{n} v_j A_{ij} \)

  Use power method: \( u = Av, v = A^T u \)

  Converges to first left/right singular vectors
Exercises

• Ex. 1: $A$ is $n \times n$ matrix with orthonormal rows
  – Show that it has orthonormal columns
• Ex. 2: Interpret the left and right singular vectors of the document x term matrix
• Ex. 3. Use power method to compute singular values of the matrix:
  \[
  \begin{pmatrix}
  1 & 2 \\
  3 & 4 \\
  \end{pmatrix}
  \]