Singular Value Decomposition

Based on
(Chapter 1 of) **Spectral Algorithms** by *Kannan and Vempala*
(Chapter 3 of) **Foundations of Data Science** by *Blum, Hopcroft and Kannan*

- Input to many computational problems can be represented as matrices.
- Spectrum of a matrix: eigenvalues, eigenvectors, singular values, singular vectors

Eigenvalues & eigenvectors

- For an $n \times n$ square matrix A, x is an eigenvector with eigenvalue λ if $Ax = \lambda x$.

$$Ax = \lambda x \text{ iff } (A - \lambda I)x = 0$$

- $\det(A - \lambda I) = 0$.
- Degree n polynomial has n complex roots (not necessarily distinct).
- A must be a square matrix for it to have eigenvalues & eigenvectors.

Eigenvalues & eigenvectors of symmetric matrices

If A is a symmetric matrix, then

- all its eigenvalues are real.
- Let x_1 and x_2 are eigenvectors with eigenvalue λ_1 and λ_2 respectively. If $\lambda_1 \neq \lambda_2$, then $\langle x_1, x_2 \rangle = 0$.
$$\lambda_1 \langle x_1, x_2 \rangle = \langle A x_1, x_2 \rangle = x_1^T A^T x_2 = x_1^T (\lambda_2 x_2) = \lambda_2 \langle x_1, x_2 \rangle$$
- If $\lambda_1 = \lambda_2$, then $c_1 x_1 + c_2 x_2$ is an eigenvector $\forall\ c_1, c_2 \in \mathbb{R}$
$$A(c_1 x_1 + c_2 x_2) = c_1 A x_1 + c_2 A x_2 = \lambda_1 (c_1 x_1 + c_2 x_2)$$

- Therefore, $A = V \Lambda V^T$, where columns of V are the eigenvectors and Λ is a diagonal matrix with Λ_{ii} being the eigenvalue of v_i (ith column of V)
$$A v_i = V \Lambda V^T v_i = V \Lambda e_i = V \Lambda_{ii} e_i = \Lambda_{ii} v_i$$

Singular values and vectors

For a matrix $A \in \mathbb{R}^{m\times n}$, σ is a singular value with corresponding singular vectors $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$ if they satisfy following two equations

- $Av = \sigma u$ and $u^T A = \sigma v^T$
- u is called a "left singular vector" of A, and v is called a "right singular vectors" of A.

- Without loss of generality, assume $\|u\| = \|v\| = 1$ since
$$\sigma\|u\|^2 = u^T(\sigma u) = u^T A v = (\sigma v^T)v = \sigma\|v\|^2$$

Singular values vs Eigenvalues

- Singular vectors of $A \equiv$ Eigenvectors of $A^T A$
- $(A^T A)v = A^T(\sigma u) = \sigma(u^T A)^T = \sigma(\sigma v^T)^T = \sigma^2 v$

- Let v be an eigenvector of $A^T A$ with eigenvalue λ. Then, $A^T A v = \lambda v$.
$$\lambda\|v\|^2 = v^T(\lambda v) = v^T(A^T A v) = (Av)^T(Av) = \|Av\|^2$$
- Therefore, $\lambda > 0$.
- Set $\sigma = \sqrt{\lambda}$ and $u = Av/\sigma$ and, we get $Av = \sigma u$ and
$$u^T A = \left(\frac{Av}{\sigma}\right)^T A = \frac{v^T A^T A}{\sigma} = \frac{(A^T A v)^T}{\sigma} = \frac{(\lambda v)^T}{\sigma} = \sigma v^T$$
- σ is a singular value of A iff σ^2 is an eigenvalue of $A^T A$.

Top singular value

- Theorem: $v_1' \stackrel{\text{def}}{=} \operatorname{argmax}_{x \in \mathbb{R}^n} \|Ax\|/\|x\|$ is a singular vector of A, and $\|Av_1'\|/\|v_1'\|$is the largest singular value of A.

- Let $v_1, \ldots, v_n$ be the (orthonormal) eigenvectors of $A^T A$ with eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_n^2$.
- For any $x \in \mathbb{R}^n$, let $x = \sum_{i \in [n]} c_i v_i$

$$\|Ax\|^2 = (Ax)^T(Ax) = x^T(A^T A)x = \left(\sum_j c_j v_j\right)^T (A^T A)\left(\sum_i c_i v_i\right)$$

$$= \left(\sum_j c_j v_j\right)^T \left(\sum_i c_i \sigma_i^2 v_i\right) = \sum_j \sum_i c_i c_j \sigma_i^2 v_j^T v_i = \sum_i c_i^2 \sigma_i^2$$

Top Singular value

- Therefore,

$$\frac{\|Ax\|}{\|x\|} = \sqrt{\frac{\sum_i c_i^2 \sigma_i^2}{\sum_i c_i^2}} \leq \sigma_1 \quad \forall\, x \in \mathbb{R}^n$$

- Choosing $x = v_1$ gives $\frac{\|Av_1\|}{\|v_1\|} = \sigma_1$.

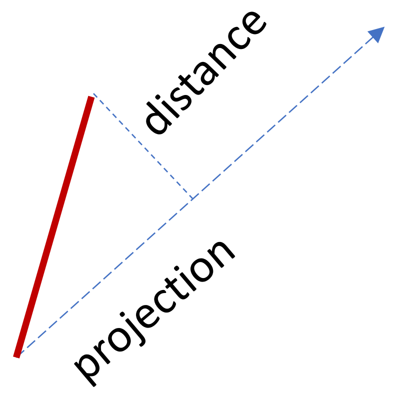
- v_1 is an eigenvector of $A^T A$, therefore, it is also a singular vector of A.
- $\|Av_1\|^2$ is the largest eigenvalue of $A^T A$. Therefore, $\|Av_1\|$ is the largest singular value of A.

Best fit line

- Given a set of points $a_1, \ldots, a_m$, find the "best fit" line.
- Find the direction v such the squared length of projections of the points on v is maximized

$$\operatorname{argmax}_{v:\|v\|=1} \sum_{i\in[m]} \langle a_i, v\rangle^2$$

- Find the direction v such the squared length of distances of points to v is minimized.



- Maximize projection on line $\equiv$ minimize distance to line
- Pythagoras theorem: Projection^2 + distance^2 = length^2
- $\sum_{i\in[m]} \text{Projection}(i)^2 + \sum_{i\in[m]} \text{distance}(i)^2 = \sum_{i\in[m]} \|a_i\|^2$

Best fit line

- Let A be the matrix with rows as $a_1, \ldots, a_m$. Then
$$\sum_{i\in[m]} \langle a_i, v\rangle^2 = \|Av\|^2$$
- Therefore, $\mathrm{argmax}_{v:\|v\|=1} \sum_{i\in[m]} \langle a_i, v\rangle^2 = \mathrm{argmax}_{v:\|v\|=1} \|Av\|^2 = v_1$

- Top singular vector gives the best fit line for a set of points.

Singular values

- Define $v_1 = \mathrm{argmax}_{x \in \mathbb{R}^n} \|Ax\|/\|x\|$ and $v_i' \stackrel{\text{def}}{=} \underset{x \perp v_1' \ldots v_{i-1}'}{\mathrm{argmax}} \|Ax\|/\|x\|$

- Theorem: v_k' is the kth singular vector of A.
- Proof by induction on k. Suppose, claim holds for all $i \leq k-1$, i.e. $v_i' = v_i \ \forall i \leq k-1$.
- Fix $x \perp v_1, \ldots, v_{k-1}$. Then $x = \sum_i c_i v_i$ where $c_1 = \cdots c_{k-1} = 0$.

$$\|Ax\|^2 = \cdots = \sum_i c_i^2 \sigma_i^2 = \sum_{i \geq k} c_i^2 \, \sigma_i^2$$

$$\underset{x \perp v_1 \ldots v_{k-1}}{\mathrm{argmax}} \|Ax\|/\|x\| = \sqrt{\frac{\sum_{i \geq k} c_i^2 \sigma_i^2}{\sum_{i \geq k} c_i^2}} \leq \sigma_k$$

- Therefore, v_k is the kth singular vector and $\|Av_k\|$ is the kth largest singular value of A.

Best fit subspace

- Given a set of points A and a number k, compute a k-dimensional subspace V_k' such that the sum of the squared lengths of the projections of the points on V_k' is maximized.
- Let $w_1, \dots, w_k$ be an orthonormal basis for V_k'. Sum of squared lengths of projections =

$$\sum_{i \in [m]} \left(\sum_{j \in [k]} \langle a_i, w_j \rangle^2 \right) = \sum_{j \in [k]} \left(\sum_{i \in [m]} \langle a_i, w_j \rangle^2 \right) = \sum_{j \in [k]} \|A w_j\|^2$$

- Theorem: Given a set of points A, the best fit k-dimensional subspace is given by span of the top k singular vectors (V_k).

- Proof by induction on k. Suppose claim is true for V_{k-1}.
- Suppose V_k' is the optimal rank k subspace. Let $w_1, \dots, w_k$ be an orthonormal basis for V_k' such that $w_k \perp V_{k-1}$.

Best fit subspace

- By optimality of V_{k-1} , we have
$$\|Aw_1\|^2 + \cdots + \|Aw_{k-1}\|^2 \leq \|Av_1\|^2 + \cdots + \|Av_{k-1}\|^2$$
- Since,
$$v_k = \operatorname*{argmax}_{x \perp V_{k-1}} \frac{\|Ax\|^2}{\|x\|^2}$$
- we have $\|Aw_k\|^2 \leq \|Av_k\|^2$. Therefore,
$$\|Aw_1\|^2 + \cdots + \|Aw_{k-1}\|^2 + \|Aw_k\|^2 \leq \|Av_1\|^2 + \cdots + \|Av_{k-1}\|^2 + \|Av_k\|^2$$
- Therefore, we V_k is as good as V_k'.

- *Singular value decomposition:* $A = \sum_{i \in [r]} \sigma_i u_i v_i^T = U \Sigma V^T$ (why?)
- Hint: For $X, Y \in \mathbb{R}^{m \times n}$, $X = Y$ iff $Xv = Yv \ \ \forall v \in \mathbb{R}^n$

Norms

- For a vector $x \in \mathbb{R}^n$, $\|x\| \stackrel{\text{def}}{=} \left(\sum_i x_i^2\right)^{1/2}$
- For an $m \times n$ matrix A, its Frobenius norm

$$\|A\|_F \stackrel{\text{def}}{=} \left(\sum_{i \in [m]} \sum_{j \in [n]} A_{ij}^2\right)^{1/2}$$

- Spectral norm

$$\|A\| \stackrel{\text{def}}{=} \max_{x \in \mathbb{R}^n} \frac{\|Ax\|}{\|x\|}$$

- Theorem: $\|A\| = \sigma_1(A)$

Norms

- $\|A\|_F^2 = \sum_j \sigma_j^2$

$$\|A\|_F^2 = \sum_{i\in[m]} \|a_i\|^2$$

- Since v_js form an orthonormal basis for the row-space of A, $\|a_i\|^2 = \sum_j \langle a_i, v_j\rangle^2$
- Therefore,

$$\|A\|_F^2 = \sum_{i\in[m]} \|a_i\|^2 = \sum_{i\in[m]} \sum_j \langle a_i, v_j\rangle^2 = \sum_j \sum_{i\in[m]} \langle a_i, v_j\rangle^2 = \sum_j \|Av_j\|^2 = \sum_j \sigma_j^2$$

Low rank matrix approximation

- Given a matrix A, compute a rank k matrix D that minimizes $\|A-D\|_F^2$.
- Theorem: Best rank k approximation is $A_k \stackrel{\text{def}}{=} \sum_{i\in[k]} \sigma_i u_i v_i^T$. Moreover,
$$\left\|A - \sum_{i\in[k]} \sigma_i u_i v_i^T\right\|_F^2 = \sum_{i>k} \sigma_i^2$$
- Let D be the optimal rank k matrix. $\|A-D\|_F^2 = \sum_{i\in[m]}\|A_i - D_i\|^2$, where A_i and D_i are the ith rows of A and D respectively.
- We may assume that D_i is the projection of A_i on a rank k subspace (why?)
- Therefore, $\|A_i - D_i\|^2 = \|A_i\|^2 - \|D_i\|^2$ (why?)

Low rank matrix approximation

$$\sum_{i\in[m]} \|A_i - D_i\|^2 = \sum_{i\in[m]} (\|A_i\|^2 - \|D_i\|^2) = \|A\|_F^2 - \sum_{i\in[m]} \|D_i\|^2$$

- Therefore, goal is to maximize $\sum_{i\in[m]} \|D_i\|^2$. Optimal subspace is given by top k singular vectors, i.e. $D_i = \sum_{j\in[k]} (A_i v_j) \cdot v_j^T = A_i \sum_{j\in[k]} v_j v_j^T$

$$D = A \sum_{j\in[k]} v_j v_j^T = \sum_i \sigma_i u_i v_i^T \sum_{j\in[k]} v_j v_j^T = \sum_i \sum_{j\in[k]} \sigma_i u_i v_i^T v_j v_j^T = \sum_{i\in[k]} \sigma_i u_i v_i^T$$

Spectral norm approximation

- Given a matrix A, compute a rank k matrix D that minimizes $\|A - D\|_2^2$
- Theorem: A_k gives the best rank k approximation, and $\|A - A_k\|_2^2 = \sigma_{k+1}^2$.
- Proof: $\|A - A_k\|_2^2 = \sigma_{k+1}^2$ (why?)
- Let D be optimal matrix. Let z be a unit vector in $\text{span}\{v_1, \ldots, v_{k+1}\}$ such that $z \perp \text{span}\, D$. Write $z = \sum_{i \in [k+1]} c_i v_i$

$$\|A - D\|_2^2 \geq \frac{\|(A - D)z\|^2}{\|z\|^2} = \frac{\|Az\|^2}{\|z\|^2} = \frac{z^T(A^T A)z}{\|z\|^2} = \frac{\sum_{i \in [k+1]} c_i^2 \sigma_i^2}{\sum_{i \in [k+1]} c_i^2} \geq \sigma_{k+1}^2$$

Power iteration

- Let A be a symmetric matrix with eigenvalues $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$ and $v_1, \dots, v_n$ the corresponding eigenvectors.
- Eigenvalues and eigenvectors can be irrational; therefore can not be computed exactly in general.
- Goal: given A and an error parameter ϵ, compute an “approximate” top-eigenvector.

- Let x_0 be a “random” unit vector. Write $x_0 = c_1 v_1 + \cdots + c_n v_n$. Then

$$\frac{Ax_0}{\|Ax_0\|} = \frac{c_1 A v_1 + \cdots + c_n A v_n}{\|Ax\|} = \frac{c_1 \sigma_1 v_1 + \cdots + c_n \sigma_n v_n}{\sqrt{c_1^2 \sigma_1^2 + \cdots + c_n^2 \sigma_n^2}}$$

Power iteration

$$\frac{A^k x}{\|A^k x\|} = \frac{c_1 \sigma_1^k v_1 + c_2 \sigma_2^k v_2 + \cdots + c_n \sigma_n^k v_n}{\sqrt{c_1^2 \sigma_1^{2k} + \cdots + c_n^2 \sigma_n^{2k}}}$$

- Define $x_k \stackrel{\text{def}}{=} A^k x_0 / \|A^k x_0\|$. As $k \to \infty$, then $x_k \to v_1$.
- If $\sigma_2 \ll \sigma_1$, then "fast" convergence (how many iterations?)

- Let p be the index such that $\sigma_p \geq (1-\epsilon)\sigma_1 > \sigma_{p+1}$. Let V_p be the subspace spanned by $v_1, \dots, v_p$. Compute a unit vector x whose projection on V_p is at least $1-\epsilon$.

Power iteration

- $\left\|A^k x_0\right\|^2 = \sum_i \sigma_i^{2k}\, c_i^2 \geq \sigma_1^{2k}\, c_1^2$
- $\sum_{i \geq p+1} \sigma_i^{2k} c_i^2 \leq (1-\epsilon)^{2k} \sigma_1^{2k} \sum_{i \geq p+1} c_i^2 \leq (1-\epsilon)^{2k} \sigma_1^{2k}$

- Therefore, the component of x_k orthogonal to V_p has squared length

$$\frac{\sum_{i \geq p+1} \sigma_i^{2k} c_i^2}{\|A^k x_0\|^2} \leq \frac{(1-\epsilon)^{2k} \sigma_1^{2k}}{\sigma_1^{2k}\, c_1^2} \leq \frac{e^{-2k\epsilon}}{c_1^2}$$

- Taking $k \geq \frac{1}{2\epsilon}\left(\ln \frac{1}{c_1^2 \epsilon}\right)$ suffices to ensure that $\frac{e^{-2k\epsilon}}{c_1^2} \leq \epsilon$

c_1?

- Taking a "random" unit vector will ensure that w.h.p. $c_1 = \Omega\left(\frac{1}{\sqrt{n}}\right)$.
- Proof in [Blum, Hopcroft, Kannan].

- Therefore, taking $k = \Theta\left(\frac{1}{\epsilon}\left(\log\frac{n}{\epsilon}\right)\right)$ will suffice.

- (H.W.)What is the value of $||Ax_k||$?

- Many other methods known for computing eigenvalues and eigenvectors.