Lecture 9: Singular Value Decomposition.

Recall: If $A \in \mathbb{R}^{n \times n}$ is symmetric, then

$$A = \begin{bmatrix} & & \\ & U & \\ & & \end{bmatrix} \begin{bmatrix} \ddots & & \\ & D & \\ & & \ddots \end{bmatrix} \begin{bmatrix} & & \\ & U^T & \\ & & \end{bmatrix}$$

and if $u_1, \dots, u_n$ are columns of A, then

1). diagonal entries of D are eigenvalues of A,

2). u_i are eigenvectors of A, and are orthonormal.

$$\begin{bmatrix} u_1 & \cdots & u_n \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \begin{bmatrix} u_1^T \\ \vdots \\ u_n^T \end{bmatrix}$$

$$= \underbrace{\lambda_1 u_1 u_1^T}_{n \times n \text{ matrix whose } i,j^{th} \text{ entry is } u_{1i} u_{1j}.} + \lambda_2 u_2 u_2^T + \cdots + \lambda_n u_n u_n^T$$

In this lecture: the generalization of this to general matrices.
$A \in \mathbb{R}^{m \times n}$.

SVD for short

Thm: (Singular Value Decomposition)
Any $A \in \mathbb{R}^{m \times n}$ can be written as

$$\begin{array}{c} \\ m \end{array} \overset{n}{\begin{bmatrix} & & \\ & A & \\ & & \end{bmatrix}} = {}_m \underbrace{\overset{m}{\begin{bmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_n \\ | & | & & | \end{bmatrix}}}_{U} {}_m \underbrace{\overset{n}{\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix}}}_{S} \underbrace{\overset{n}{\begin{bmatrix} v_1^T \\ \vdots \\ v_n^T \end{bmatrix}}}_{V} n$$

where:

1. $U \in \mathbb{R}^{m \times m}$ is orthogonal ($u_1, \dots, u_n$ form orthonormal basis)
2. S is "diagonal", $\sigma_i \geq 0 \ \forall i$.
3. $V \in \mathbb{R}^{n \times n}$ is also orthogonal

$\sigma_1 \geq \cdots \geq \sigma_n \geq 0$ are called the singular values of A.

$u_1, \cdots, u_m$ are left singular vectors of A
$v_1, \cdots, v_n$ are right singular vectors of A.

Equivalently:

$$A = \sum_{i=1}^{\min(m,n)} \sigma_i u_i v_i^T$$

What is the action of A?

$A v_i = \sigma_i u_i, \quad \forall i.$ so A maps $\{v_1, \cdots, v_n\}$ to $\{u_1, \cdots, u_n\}$, and scales the i^{th} coordinate by σ_i.

Similarly, $u_i^T A = \sigma_i v_i$ so A^T does the "opposite" mapping, with the same scaling.

Connection to eigenvectors / eigenvalues, PCA

SVD is for general matrices, but spectral decomposition is only for symmetric.

However, for symmetric, SVD follows directly from spectral:

If $A \in \mathbb{R}^{n \times n}$ is symmetric, then

$$A = U D U^T$$

If $\underbrace{\text{all entries of } D \text{ are } \geq 0}_{A \text{ is positive semi-definite (PSD)}}$, then this is an SVD

Otherwise, take

$$A = \begin{bmatrix} u_1 & \cdots & u_n \end{bmatrix} \begin{bmatrix} |\lambda_1| & & \\ & \ddots & \\ & & |\lambda_n| \end{bmatrix} \begin{bmatrix} \operatorname{sign}(\lambda_1) u_1^T \\ \vdots \\ \operatorname{sign}(\lambda_n) u_n^T \end{bmatrix}$$

and this is an SVD.

On the other hand, if $A \in \mathbb{R}^{m \times n}$ is arbitrary, we can relate its

SVD to the spectra of some related matrices:

$$\begin{bmatrix} A \end{bmatrix} = \begin{bmatrix} u_1 & \cdots & u_m \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ & & \end{bmatrix} \begin{bmatrix} v_1^T \\ \vdots \\ v_n^T \end{bmatrix}$$

$$A^TA = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_n & \end{bmatrix} \underbrace{\begin{bmatrix} u_1^T \\ \vdots \\ u_m^T \end{bmatrix} \begin{bmatrix} u_1 & \cdots & u_m \end{bmatrix}}_{= I \text{ since } \{u_i\} \text{ are orthonormal}} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ & & \end{bmatrix} \begin{bmatrix} v_1^T \\ \vdots \\ v_n^T \end{bmatrix}$$

$$= \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} \underbrace{\begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_n & \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ & & \end{bmatrix}}_{= \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{bmatrix}} \begin{bmatrix} v_1^T \\ \vdots \\ v_n^T \end{bmatrix}$$

$$= \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{bmatrix} \begin{bmatrix} v_1^T \\ \vdots \\ v_n \end{bmatrix}$$

This is spectral decomposition of A^TA!

So right singular vectors of $A \iff$ eigenvectors of A^TA

(nonzero) singular values of $A \iff \sqrt{\text{eigenvalues}}$ of A^TA.

Similarly, left singular vectors of $A \iff$ eigenvectors of AA^T

" $\iff$ " " of AA^T.

Relationship w/ PCA:

Note that in PCA, given data matrix X, the PC of X are top k eigenvectors of X^TX.

So PCA is just pretty much SVD!

Strictly speaking, PCA only cares about the right singular vectors of X, so SVD is strictly more general than PCA. But often are used interchangeably.

Algos for SVD: Given $A \in \mathbb{R}^{m \times n}$, how to find SVD?

Idea: just use power method on A^TA and/or AA^T.

Once we've found right singular vectors $\{v_1, \dots, v_n\}$, left singular vectors are given by

$$u_1 = Av_1$$
$$\vdots$$
$$u_n = Av_n$$

$$\left.\begin{matrix} u_{n+1} \\ \vdots \\ u_m \end{matrix}\right\} \longrightarrow \text{any basis of space perpindicular to } u_1, \dots, u_n.$$

Note: to apply power method to A^TA, we don't need to compute A^TA, which can be expensive. Instead:

$A^T A u = A^T(Au)$ so only need to do 2 matrix-vector multiples, which is usually faster.

best algo in practice:

np.linalg.svd(A) (or equivalent pkg).

Application: Low-rank approximation.

Q: How can we fill in the following matrix?

$$A = \begin{bmatrix} 7 & ? & ? \\ ? & 8 & ? \\ ? & 12 & 6 \\ ? & ? & 2 \\ 21 & 6 & ? \end{bmatrix} \Rightarrow \begin{bmatrix} 7 & 2 & 1 \\ 56 & 8 & 4 \\ 42 & 12 & 6 \\ 28 & 4 & 2 \\ 21 & 6 & 3 \end{bmatrix}$$

Obviously impossible in worst case!

But what if A has structure? i.e. what if all rows are multiples of each other?

An example of a low-rank matrix.

Def: $A \in \mathbb{R}^{m \times n}$ has rank 0 if $A =$ all zeros.

Def: $A \in \mathbb{R}^{m \times n}$ has rank 1 if $A = uv^T$ $\quad \begin{bmatrix} u \end{bmatrix}\begin{bmatrix} & v^T & \end{bmatrix}$

__Def:__ A has rank k if A can be written as a sum of k rank 1 matrices, and cannot be written as a sum of k-1.

$$A = \sum_{i=1}^{k} u_i v_i^T.$$ ← not necessarily unit vectors or orthogonal.

$$A = \begin{bmatrix} \\ u_1 \cdots u_a \\ \\ \end{bmatrix} \begin{bmatrix} v_1^T \\ \vdots \\ v_k^T \end{bmatrix}$$

"tall, skinny" "short, long"

Many equivalent definitions of rank:

1. The largest set of linearly independent columns of A has size k.
2. The largest set of linearly independent rows has size k.
3. A has k non-zero singular values.

Low-rank approximation.

Real-world data is unlikely to be exactly low-rank.

We can still ask for best rank-k approximation.

Q: Given A, find B which has rank-k that is "closest" to A.

A natural candidate: let

$$A = \sum_{i=1}^{\min(m,n)} \sigma_i u_i v_i^T$$ be the SVD of A

(recall $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$)

and take $$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$ as your rank-k approx.

It turns out that in many natural ways, this is __optimal__:

__Thm:__ For any $A \in \mathbb{R}^{m \times n}$, any B rank k:

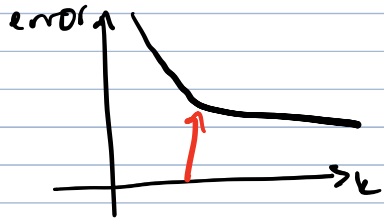
$$\|A - A_k\|_F \leq \|A - B\|_F.$$

($\|M\|_F = \sqrt{\sum_{ij} M_{ij}^2}$ is Frobenius norm of A).

Closely related to optimality of PCA!

How to choose k? k trades off size of representation vs quality.

often "elbow" phenomena



Another rule of thumb:

choose k s.t.

$$\sum_{i=1}^{k} \sigma_i \geq C \cdot \sum_{i=k+1}^{n} \sigma_i$$

3? 10?

Applications:

1. Compression
2. Denoising
3. Data completion.