* Dimension Reduction :-

Recap :-

A $n \times n$ square matrix can be diagonalized iff A is normal

i.e. $AA^* = A^*A$

In real case $A^* = A^T$

& complex case $A^* = A^H = \overline{A^T}$

i.e. If a non singular matrix S consisting of eigen vectors of A such that

$$S^{-1} A S = \Lambda$$

where $\Lambda \Rightarrow$ diagonal mat with eigen values of A.

for other cases we need to study jordon-canonical form.

So let $A_{m \times n}$, $m \neq n$, surprisingly A can always be factorized. This is singular value decomposition

One can always find orthogonal matrices $U_{m \times m}$ & $V_{n \times n}$ such that

$$A_{m \times n} = U_{m \times m} \Sigma_{m \times n} V^T_{n \times n}$$

where,

Σ is $m \times n$ and has exactly "r" non zero entries. the singular values $\sigma_1, \sigma_2, \ldots, \sigma_r$ on diagonal & rest zero.

& r = rank of A

* The key observation is that both AA^T & A^TA are symm. & hence can be diagonalized

This can be used to obtain SVD

i.e. $A = U \Sigma V^T$

where U = eigen vectors of AA^T & is a orthogonal matrix

V = eigen vectors of A^TA & is an orthogonal matrix.

However the ranks of AA^T & A^TA are equal. & +ve eigen values of AA^T & A^TA are denoted by $\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2$.

Recall: $A \rightarrow$ positive definite $\Rightarrow$

$$x^TAx > 0 \quad \forall\ x \in R^n \quad x \neq 0$$

$$\boxed{x^TAx = \langle x, Ax \rangle}$$

if $B = AA^T$, $B^T = AA^T \Rightarrow B = B^T$

$\Rightarrow$ B is symmetric

$$\therefore \boxed{x^TBx = x^TAA^Tx.}$$

$$\therefore x^T B x = \underbrace{(A^T x)^T A^T x}_{y^T y}$$

$$\therefore \|x^T B x\| \geq 0$$ IMP

$\therefore$ B is a positive semi-definite.

$\therefore$ eigen values of B are either +ve or zero. Illy for A^TA.

Thus we can diagonalize both A^TA & AA^T by orthogonal matrices.

Now, we have,

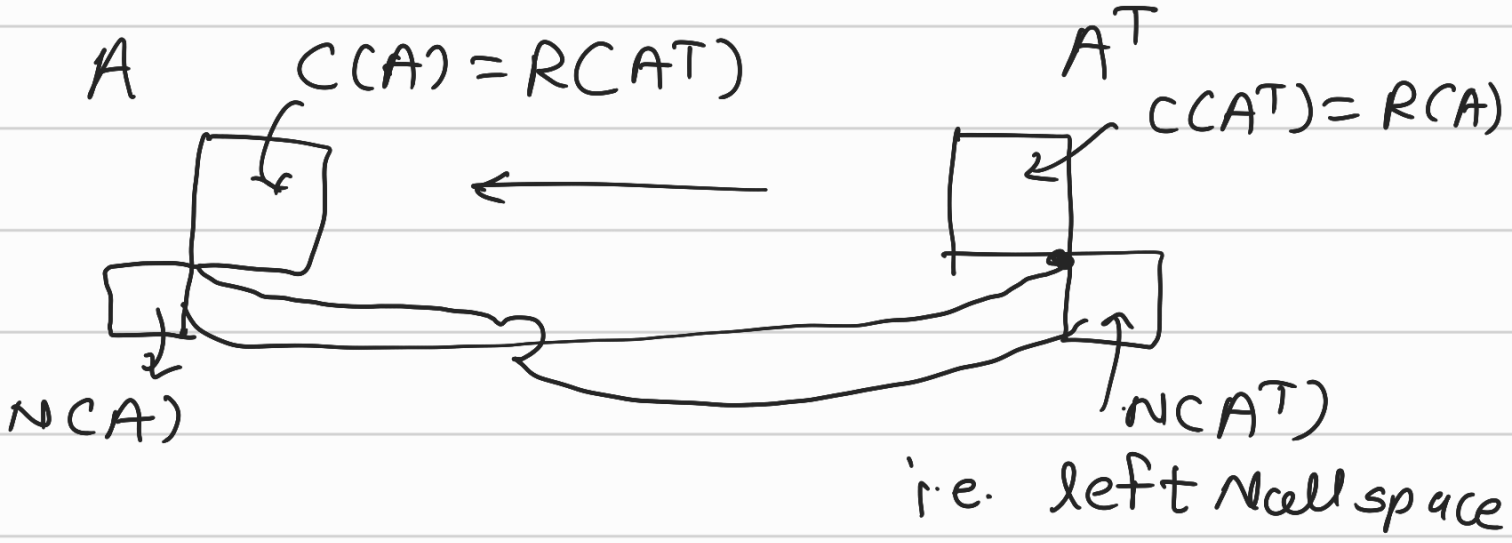
$$A = U \Sigma V^T$$

columns of U → eigen vectors of AA^T

Columns of V → eigen vectors of A^TA.

Imp Note:-

$A: R^n \longrightarrow R^m$



i.e. left Nullspace

So, SVD gives complete information about the four fundamental subspaces of A.

IMP

columnspace of A

① First r columns of $U \rightarrow C(A)$

② last $(m-r)$ cols of $U \rightarrow$ left Nullspace of $A = N(A^T)$

③ first r cols of $V \rightarrow R(A) = C(A^T)$

④ last $(n-r)$ cols of $V \rightarrow N(A)$

This is remarkable since in rare case only normal matrix can be diagonalized.

However every rectangular matrix $A_{m \times n}$ can be decomposed as

$$A = U \Sigma V^T$$

12/03/2021

Topics :- Pseudo-inverse or Moore-Penrose inverse or Generalized inverse.

Recall: If A is $m \times n$ & has rank r then

$$A = U \Sigma V^T$$

where, $U \Rightarrow m \times m$ & orthogonal

$V \Rightarrow n \times n$ & orthogonal

& $\Sigma \Rightarrow \text{diag}[\underbrace{\sigma_1, \sigma_2, \sigma_3, \ldots, \sigma_r}] \; \& \; m \times n$

eigen values of AA^T & A^TA.

(I) case of diagonal Σ i.e.

$0 < r \leq \min(m, n)$

then, by def^n moore-Penrose inverse or pseudo inverse Σ^+ is

$$\Sigma^+ = \begin{bmatrix} 1/\sigma_1 & & & \\ & 1/\sigma_2 & & \\ & & \ddots & \\ & & & 1/\sigma_r \\ & & & & \ddots \\ & & & & & 0 \end{bmatrix}_{n \times m}$$

So if Σ is $m \times n$ then Σ^+ is $n \times m$

& $\Sigma\Sigma^+$ is $m \times n$ & $\Sigma^+\Sigma$ is $n \times n$. Both will have I_r in the left corner.

i.e. $$\Sigma\,\Sigma^+ = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}_{m \times m}$$

$$\Sigma^+\Sigma = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}_{n \times n}$$

e.g. ① $\Sigma = \begin{bmatrix} 6_1 & 0 & 0 & 0 \\ 0 & 6_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}_{3\times 4}$

then $\Sigma^+ = \begin{bmatrix} 1/6_1 & 0 & 0 \\ 0 & 1/6_2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}_{4\times 3}$

$\therefore$ Σ is mapping : R^4 to R^3
& Σ^+ is mapping : R^3 to R^4

$\therefore \quad \Sigma \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$ & $b_3 = 0$
(Because last row of Σ is zero)

Thus if $b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$ & $b_3 \neq 0$ then

$\Sigma x = b$ cannot be solved.

Now what does Σ^+ do to b?

$$x^+ = \Sigma^+ \quad b = \begin{bmatrix} 1/\sigma_1 & 0 & 0 \\ 0 & 1/\sigma_2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

$$\therefore \quad x^+ = \begin{bmatrix} b_1/\sigma_1 \\ b_2/\sigma_2 \\ 0 \end{bmatrix} \Rightarrow \text{Pseudo soletion}$$

$$\& \; \Sigma\Sigma^+ = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \& \; \Sigma^+\Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

* Our main goal is dimension red^n using SVD.

Now, as $A = U\Sigma V^T$ then $\boxed{A^+ = V\Sigma^+ U^T}$

V VIMP

$$AA^+ = U\Sigma V^T . V\Sigma^+ U^T = U\Sigma I \Sigma^+ U^T$$

$$\therefore AA^+ = U\Sigma\Sigma^+ U^T$$

Note :- In case Σ is square & invertible we get A^+ as regular inverse i.e. A^{-1}.

IIIly.

$$A^+A = V\Sigma^+ U^T . U\Sigma V^T = V\Sigma^+\Sigma V^T$$

$$\therefore A^+A = V\Sigma^+\Sigma V^T$$

NOTE :-

$\therefore$ If A maps : $R^n \longrightarrow R^m$

then A^+ maps : $R^m \longrightarrow R^n$

Consider, $Ax = b$ given by

$$-x_1 + 2x_2 + 2x_3 = 18$$

$$\therefore \begin{bmatrix} -1 & 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 18 \end{bmatrix}$$

$$\therefore A = \underset{1\times 3}{\begin{bmatrix} -1 & 2 & 2 \end{bmatrix}} \longrightarrow \text{find SVD of } A$$

Ans $\Rightarrow$

$$A = \underset{1\times 1}{\begin{bmatrix} 1 \end{bmatrix}} \underset{1\times 3}{\begin{bmatrix} 3 & 0 & 0 \end{bmatrix}} \underset{3\times 3}{\begin{bmatrix} -1/3 & 2/3 & 2/3 \\ 2/3 & -1/3 & 2/3 \\ 2/3 & 2/3 & -1/3 \end{bmatrix}}$$

i.e. $U = [1]$

$$\Sigma = [3 \ 0 \ 0]$$

& $$V^T = \begin{bmatrix} -1/3 & 2/3 & 2/3 \\ 2/3 & -1/3 & 2/3 \\ 2/3 & 2/3 & -1/3 \end{bmatrix}$$

$\therefore$ By defn $A^+ = V \Sigma^+ U^T$

$$\therefore A^+ = \underset{3\times3}{\begin{bmatrix} -1/3 & 2/3 & 2/3 \\ 2/3 & -1/3 & 2/3 \\ 2/3 & 2/3 & -1/3 \end{bmatrix}} \underset{3\times1}{\begin{bmatrix} 1/3 \\ 0 \\ 0 \end{bmatrix}} \underset{1\times1}{\begin{bmatrix} 1 \end{bmatrix}}$$

$$\therefore A^+ = \begin{bmatrix} -1/9 \\ 2/9 \\ 2/9 \end{bmatrix}_{3\times1} \quad \text{whereas } A = \begin{bmatrix} -1 & 2 & 2 \end{bmatrix}$$

$\therefore A \Rightarrow$ Row vector But $A^+ \Rightarrow$ Column Vector.

To solve $Ax = [18]$ using A^+ we have

$$x^+ = A^+ [18] = \begin{bmatrix} -1/9 \\ 2/9 \\ 2/9 \end{bmatrix} [18]$$

$$\therefore x^+ = \begin{bmatrix} -2 \\ 4 \\ 4 \end{bmatrix} \rightarrow \text{Pseudo solution.}$$

$$\therefore Ax^{+} = \begin{bmatrix} -1 & 2 & 2 \end{bmatrix} \begin{bmatrix} -2 \\ 4 \\ 4 \end{bmatrix} = \begin{bmatrix} 18 \end{bmatrix}$$

Now we can easily check that,

$$A \overset{x_1}{\begin{bmatrix} -2 \\ 5 \\ 3 \end{bmatrix}} = \begin{bmatrix} -1 & 2 & 2 \end{bmatrix} \begin{bmatrix} -2 \\ 5 \\ 3 \end{bmatrix} = \begin{bmatrix} 18 \end{bmatrix}$$

Also, $A \overset{x_2}{\begin{bmatrix} -2 \\ 7 \\ 1 \end{bmatrix}} = [18]$ & $A \overset{x_3}{\begin{bmatrix} -6 \\ 3 \\ 3 \end{bmatrix}} = [18]$

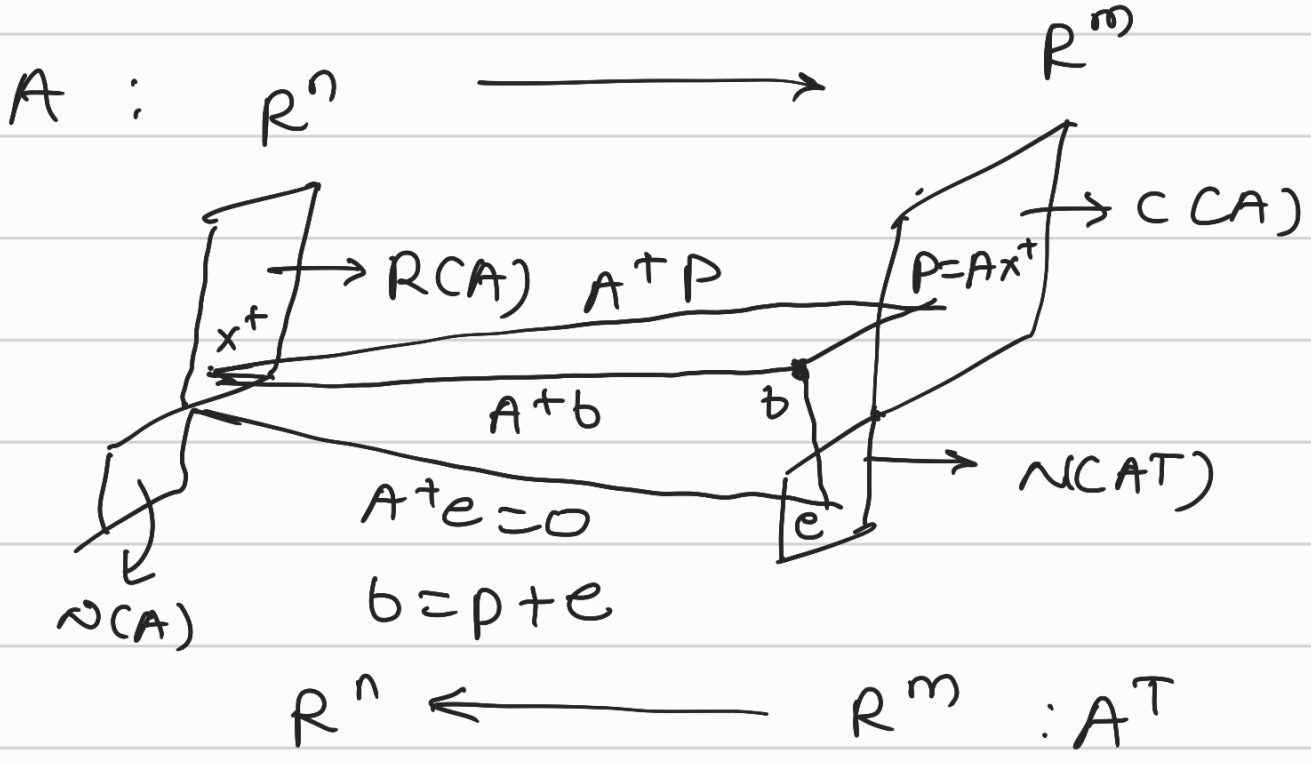
x_1, x_2 & x_3 are all solutions, so what's spl. about x^{+}.

However $\|x^{+}\| = (-2)^2 + 4^2 + 4^2 = 36$ which is least amongst all.

$\therefore$ The least square solution is coming from x^{+}.

& $\therefore$ $x^+ = A^+ b \Rightarrow$ least square soln.

*This is really significant, since even if $b \notin C(A)$ one can find best or nearest possible soln to $Ax = b$. using SVD of A & the pseudo-inverse. A^+.



So, essentially, A^+ inverts A where it can namely C(A) & in rest i.e. $N(A^T)$ A^+ sets all of them to 0.

So given 'b', decompose b, (orthogonally) into p & e

i.e. $b = p + e$ & $p \in C(A)$

$p \perp r$ to e & $e \in N(A^T)$

Then $x^+ = A^+ p$; since $A^T(e) = 0$

$\forall\ e \in N(A^T)$

16/03/2021

Topic : Least square solutions

Let $Ax = b$ & A is $m \times n$ & $b \notin C(A)$ then no solution is possible.

Still want to find nearest possible soln.

so,

Project b on $C(A)$, Therefore let $A\hat{x}$ = projection b on $C(A)$

then the error vector 'e' is

$e = b - A\hat{x}$ must be $\perp$ to $N(A^T)$.

$\therefore A^T e = 0$ or $A^T(b - A\hat{x}) = 0$

$\therefore A^T b = A^T A \hat{x}$

$\therefore \boxed{A^T A \hat{x} = A^T b} \Rightarrow$ Normal eqns

$A^T A$ is invertible only if the columns of A are independent.

In that case, we have

$$\boxed{\hat{x} = (A^T A)^{-1} A^T b}$$

which is the best solution & it's projection on C(A) is given by

$$\boxed{p = A(A^T A)^{-1} A^T b = P \cdot b}$$

where $\boxed{P = A(A^T A)^{-1} A^T} \Rightarrow$ Projection mat.

Recall: ① for projection matrix

(i) $p^2 = p$ (ii) $p = p^T$

② SVD: $A = U\Sigma V^T$ & $A^+ = V\Sigma^+ U^T$

& $x^+ = A^+ b \rightarrow$ pseudo-solution.

③ V & U are orthogonal matrices

∴ they preserve lengths.

∴ length of $(Ax - b)$ = length $(U\Sigma V^T x - b)$

∴ length of $(U^T(U\Sigma V^T x - b))$

$=$ length $(\Sigma V^T x - U^T b)$

Thus, $\boxed{\|Ax - b\| = \|\Sigma V^T x - U^T b\|}$

let, $V^T x = y$ ∴ $x = Vy$, ∵ $V^T = V^{-1}$

Minimizing $\|Ax-b\|$ is same as - minimizing $\|\Sigma y - U^T b\|$, But then Σ is diagonal. & we know Σ^+. Hence the best y is $y^+ = \Sigma^+ U^T b$

$\therefore$ The best x, denoted as x^+ is given by

$$x^+ = V y^+$$

$$\therefore x^+ = V(\Sigma^+ U^T b)$$

$$\therefore x^+ = (V\Sigma^+ U^T) b$$

$$\boxed{x^+ = A^+ b}$$

Indeed we get least square solution or best solution from A^+, via SVD.

In Data Science, we are in R^d 'd' very large. Also given 'n' pts $x_1, x_2, x_3, \ldots, x_n$ in R^d. 'n' is fairly large, so get nxd matrix (Rectangular).

To find SVD of this nxd matrix can be very costly.

So, fix some k ($k << n$) & some how find only k-vectors in R^d. & these k-linearly independent vectors in R^d. & these will span a k-dimensional subspace in R^d. This will be the best fitting subspace of R^d & project all data points on to this much smaller subspace & find solution.

Therefore, the best fitting subspace (via data points $x_1, x_2, \ldots, x_n$) are found by computing the singular vectors one at a time.

using the method of least squares.

Topic :- Best fit subspaces & SVD 17/3/2021

Consider each row of matrix $A_{n \times d}$ as point in d-dimensional subspace. The SVD finds the best-fitting k-dimensional subspace for $k = 1, 2, 3, \cdots$ for given set of 'n' data points.

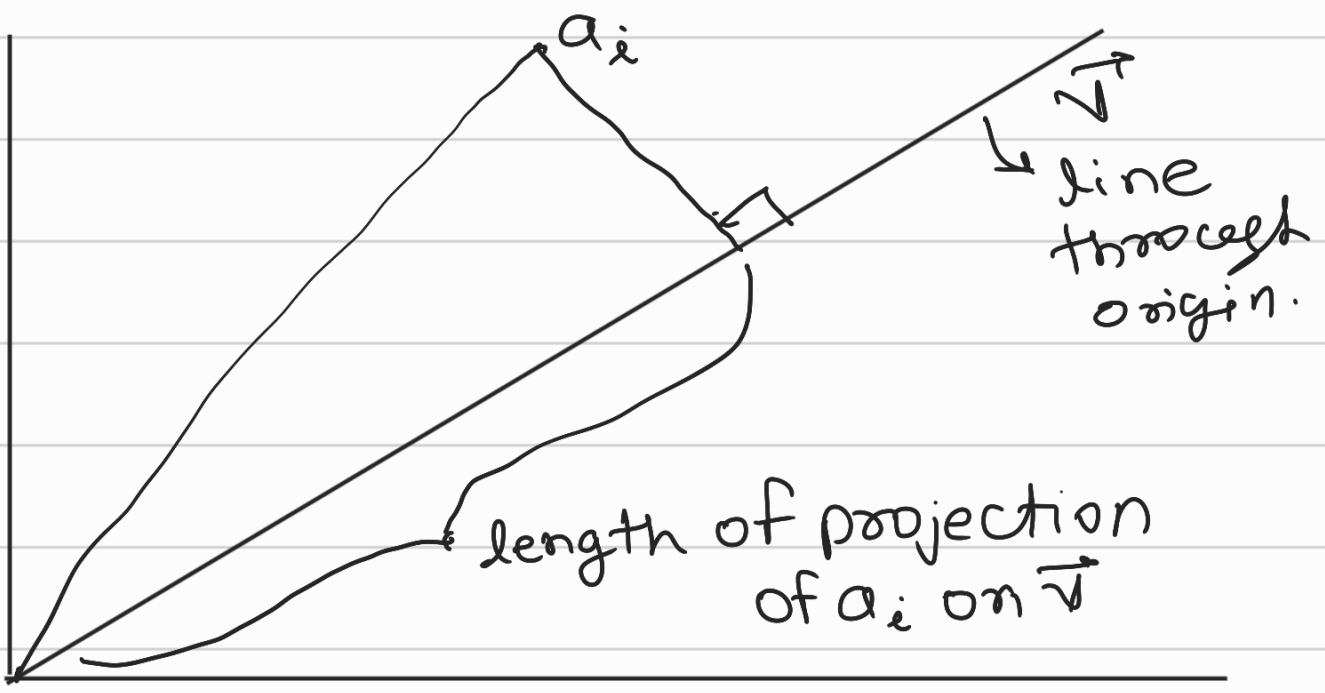
What does 'best' mean?

Consider projecting a point

$$a_i = (a_{i1}, a_{i2}, \cdots, a_{id}) \leftarrow R^d$$

on to a line through the origin then,

$$\|a_i\| = a_{i1}^2 + a_{i2}^2 + \cdots + a_{id}^2$$



By Pythagoras theorem,

$$a_{i1}^2 + a_{i2}^2 + a_{i3}^2 + \cdots + a_{in}^2$$

$$= (\text{1er dist. of pt. to line})^2 + (\text{length of projection})^2$$

$$\therefore d(\text{pt. to line})^2 = a_{i1}^2 + a_{i2}^2 + \ldots + a_{id}^2 - (\text{length of projection})^2$$

$\therefore$ minimizing d(Pt. to line) is equivalent maximizing (length of projection)

$\because \|a_i\|^2$ is fixed, we have 2-interpretations of best fit subspace.

(I) It minimizes the sum of squared distances of data points a_i to the subspace. This is the familiar least-square fit from calculas.

(II) Maximizing, the sum of "projections squared" of the data points on the subspace. This says that best fitting subspace contains maximum content of the data among all subspaces of the same dimension.

19/03/2021

$A \to n \times d$, think of rows of A as points in R^d.

If Rank(A) is r then rows of a will span a r-dimensional subspace in R^d.

Given 'k', s.t. $1 \leq k \leq r$, we want to find k-linearly independent vectors $v_1, v_2, \ldots, v_k \in R^d$ s.t. this subspace spanned by $v_1, v_2, \ldots, v_k$ is closest to A or 'Best fits' A

The subspace may not even contain the vectors or points $a_1, a_2, \ldots a_n$ etc. yet it is the best. because we look at sum of squared error. & it should be minimum.

Definition & Notations-

If 'f' is a real valued function defined on a S, $f: S \to R$, then we define

$$\text{Argmax } f = \{ x \in S : f(s) \leq f(x) \ \forall \ s \in S \}$$

i.e. set of all pts. at which f contains its maximum.

e.g. $f(x) = \cos x$ then,

$$\text{Argmax } f = \{ 2n\pi \} \quad n = 0, 1, 2, \ldots$$

Now, consider best-fit line through origin w.r.t. the pts. $a_1, a_2, a_3, \ldots, a_n$. in R^d. Let v be unit vector along this line. Length of the projection of a_i on to v is $|a_i \cdot v|$. From this we see that sum of squared length of predictions is exactly $\|Av\|^2$

$\therefore \|Av\|^2 = \sum |a_i \cdot v|^2$ ← maximize

$\therefore$ The best fit line is one which maximizes above f^n. & Hence minimizes sum of squared distances of the pts to the line with this mind, we define the first singular vector v_1 of A as

$$v_1 = \text{argmax } \|Av\| \text{ , } \|v\| = 1$$

Technically, there may be a tie so pick any one of them.

The value $\sigma_1(A) = \|A v_1\|$ is called the first singular value.

Next, the second singular vector v_2 is defined as best fit line ⊥ to v_1

$$\therefore v_2 = \text{argmax } \|Av\| \text{ , } v \perp v_1 \text{ \& } \|v\| = 1$$

& $\sigma_2(A) = \|Av_2\|$ & is second singular

value of A.

Ily, the third singular vector v_3

$$v_3 = \text{argmax } \|Av\| \;\&\; v \perp v_1, v_2 \;, \|v\| = 1$$

$\|Av_3\| = \sigma_3(A)$, the 3rd singular value of A.

Recall:- in SVD we arranged singular values $\sigma_1, \sigma_2, \ldots, \sigma_r$ in descending values.

i.e. $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \cdots \geq \sigma_r$

which are sq. roots of eigel vals. of AA^T (as well as A^TA)

Ily, we find all the singular vectors & singular values.

$$v_r = \text{argmax } \|Av\| \;, v \perp v_1, v_2, v_3, \ldots \;\&\; \|v\| = 1$$

$$\sigma_r(A) = \|Av_r\|$$

Theorem 3.1 :-

Greedy Algorithm

Let A be nxd matrix with singular vectors $v_1, v_2, \ldots, v_r$. Then for 'k', $1 \leq k \leq r$ let V_k be the subspace spanned by $v_1, v_2, v_3, \ldots, v_k$ for each k V_k is the best-fit k-dimensional subspace for A.

(Read proof from book)

A is nxd : $A: R^d \rightarrow R^n$

$Av_i \in R^n$ Think of Av_i as a list

Adding up the squares of the components of A along each of the

v_i gives the square of the whole content of A.

This is indeed true & is the matrix analogy of decomposing a vector into its conponent along orthogonal dir^n.

Consider one row, say a_j of A. Now $v_1, v_2, \ldots, v_r$ span all rows of A.

$\therefore a_j \cdot v = 0 \quad \forall\, v$ ⊥er to v_j's. Thus for each row a_j of A we let

$$\sum (a_j \cdot v_i)^2 = \|a_j\|^2$$

$\therefore$ summing over all j gives,

$$\sum_{j=1}^{n} \|a_j\|^2 = \sum_{j=1}^{n} \sum_{i=1}^{r} (a_j \cdot v_i)^2$$

$$\therefore \sum_{j=1}^{n} \|a_j\|^2 = \sum_{i=1}^{r} \underbrace{\left(\sum_{j=1}^{n} (a_j \cdot v_i)^2 \right)}_{\|Av_i\|^2}$$

$$\therefore \sum_{j=1}^{n} \|a_j\|^2 = \sum_{i=1}^{r} \sigma_i^2(A)$$

But then

$$= \sum_{j=1}^{n} \sum_{k=1}^{d} a_{jk}^2$$

This is denoted by $\|A\|_F^2$ called Frobennius norm of A.

where $$\|A\|_F = \sqrt{\sum_{j,k} a_{j,k}^2}$$

Lemma 3.2:-

for any matrix, the sum of square of singular values of A equals the Frobennius norm of A.

Vectors v_i's are technically right singular vectors of A.

Recap of Everything done till Now (Chapter: 3)

22/03/2021

A is $n \times d$ matrix & rows of A are thought of points in R^d (row-vectors) & We look at the best-fit line two these points. The first singular vector was

$$v_1 = \text{argMax} \; ||Av||, \; ||v||=1$$

& $||Av_1|| = \sigma_1(A) = 1^{st}$ singular vector of A.

Greedy algorithm says, repeated this procedure only with perpendicular vectors.

$$\text{i.e. } v_2 = \text{argmax}_{\perp v_1 \; \& \; ||v||=1} (||Av||)$$

& $\sigma_2(A) = ||Av_2|| \rightarrow 2^{nd}$ singular vector of A.

continue till . rank of matrix A i.e. r.

We saw that $v_1, v_2, \ldots, v_r$ span the space of all rows of A i.e. they give bound for row space of A.

for $\forall \quad v \perp (v_1, v_2, \ldots, v_r)$

$$a_j \cdot v = 0 \qquad (a_j = j^{th} \text{ row of } A)$$

Hence for each row a_j we have

thus summing over $\forall j$ we get

$$\sum_{j=1}^{n} \|a_j\|^2 = \sum_{j=1}^{n} \left(\sum_{i=1}^{r} (a_j \cdot v_i)^2 \right)$$

by interchanging orders of summation

$$\sum_{j=1}^{n} \|a_j\|^2 = \sum_{i=1}^{r} \sum_{j=1}^{n} (a_j \cdot v_i)^2$$

$$\sum_{j=1}^{n} \|a_j\|^2 = \sum_{i=1}^{r} \|Av_i\|^2 = \sum_{i=1}^{r} \sigma_i^2(A)$$

However,

$$\sum \|a_j\|^2 = \sum_{j=1}^{n} \sum_{k=1}^{d} a_{jk}^2$$

a_j = j^{th} row of A & a_j has d components.

$$a_{j_k} = (a_{j1}, a_{j2}, \ldots, a_{jd})$$

$$= \sum_{i=1}^{r} \sigma_i^2(A)$$

& we defined frobennius Norm of A as

$$\|A_F\|^2 = \sum_{j,k} a_{j,k}^2 \; ; \; \begin{matrix} j=(1,n) \\ k=(1,r) \end{matrix}$$

Since we have sum of all rows of A & A can be thought of nd components. i.e. A is like vector in R^{nd}.

Thus $\|A_F\|^2 = \sigma_i^2(A)$

The right singular vectors v_1 to v_r of A are such that

$$\|Av_i\| = \sigma_i(A)$$

∴ If we normalize Av_i we get unit vector & by def^n we get unit vectors of A, by def^n we have.

$$u_i = \frac{1}{\sigma_i(A)} \cdot Av_i$$

where u_i = non-sing. vector.

We can show that u_i non-sing. vector $\|u^T A\|$ are all vectors u s.t. $u \perp (u_1, u_2, u_3 \ldots, u_r)$

IMP fact :- These vectors are s.t. $\sigma_i v_i$ is a vector whose co-ordinate correspond to the projection of the rows of A onto v_i. Each $\sigma_i u_i v_i^T$

is a rank 'one' matrix whose rows are the v_i comps of rows of A.

The projection of the rows of A in the v_i direction. We will show that A can be decomposed as a sum of rank '1' matrices.

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

This is essentially SVD.

Geometrically, every point decomposed in A into its comp. along each of the 'r' orthorgonal dir^ (in orthogonal comp.) given by v_i.

Lemma:- Matrices A & B are identical if & only if $Av = Bv$ forall 'v'.

Pf: If $A = B$, clearly $Av = Bv$

by defⁿ Ae_i gives the ith column of A. By assumption $Av = Bv$ for $\forall\, v$

$\therefore Ae_i = Be_i$ for all i

i.e. all columns of A & B are the same. Hence $A = B$.

IMP theorem :-

If A is $m \times d$ with right singular vectors as, $v_1, v_2, v_3, \ldots, v_r$ & left singular vectors $u_1, u_2, \ldots, u_r$ & singular values $\sigma_1, \sigma_2, \sigma_3 \ldots, \sigma_r$ then,

$$A = \sum \sigma_i u_i v_i^T$$

Proof : Look at the effect of multiplying both sides by v_i

$$\sum \sigma_i \cdot u_i v_i^T v_j = \sigma_j u_j$$

But $\sigma_j u_j = A v_j$,

Thus $A v_j = \left(\sum_{i=1}^{r} \sigma_i u_i v_i^T \right) v_j$

& by the lemma we get

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

$$A_{n \times d} = U_{n \times r} \; D_{n \times r} \; V_{r \times d}.$$

23/03/2021

Ans: it is precisely the matrix A_k got by truncating the SVD of A to top k singular vectors.

i.e. if $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$

then $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$

It is clear that rank of A_k is k. Because the u_i's are orthonormal.

Lemma 3.5:-

The rows of A_K are the projection of the rows of A onto subspace spanned by the first K singular vectors of A.

Proof:-

Let 'a' be arbitrary row vector of A. The v_i's are orthonormal the projection of vector 'a' onto V_K is given by

$$\sum_{i=1}^{K} (a \cdot v_i) \cdot v_i^T$$

Thus the matrix whose rows are the projections of the rows of A onto V_K is given by $\sum A v_i v_i^T$

Recall that $Av_i = \sigma_i u_i$

$\therefore$ we get $\sum_{i=1}^{k} \sigma_i u_i v_i^T = A_K$

Theorem :-

For any matrix B of rank 'k',

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

Proof :-

Let B minimize $\|A - B\|_F^2$ among all rank k. Let V be the subspace spanned by the rows of B.

B is minimising $\|A - B\|_F^2$. Hence each row of B is projection of corresponding row of A onto V.

This is because we can replace the row of B with the projection of the corr. row of A onto V. This keeps the row space of B containing V & rank of B is atmost k.

But this reduces $\|A - B\|_F^2$, which is a contradiction.

Thus each row of B is projection of the corresponding row of A.

$\|A - B\|_F^2$ is the sum of squared distances

of the rows of A to V. Since A_k minimizes the sum of sq. distances of rows of A to any k-dimension. Subspace from th^m is function that

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

summary:-

Among all rank 'k' approximations to A, A_k got by k singular vectors of A is the best.

2- Norm or spectral Norm of A:-

It is denoted by, $\|A\|_2 = \max_{\|x\| \leq 1} \|Ax\|$

indeed this is the first singular value of A i.e. $\sigma_1(A)$

Recall Left singular vectors (u_i):-

$$u_i = \frac{1}{\sigma_i(A)} A v_i$$

Lemma –

$$\|A - A_k\|_2^2 = \sigma_{k+1}^2$$

24/03/2021

Spectral Norm :-

Also called 2-Norm, is defined as

$$\|A\|_2 = \max_{\|x\| \leq 1} \|Ax\| = \sigma_1(A)$$

i.e. $\|A\|_2$ = first singular value.

Lemma 3.8 :-

$$\boxed{\|A - A_k\|_2^2 = \sigma_{k+1}^2}$$

Proof :-

Let $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$, $r = \text{rank}(A)$

By defn for $k < r$ $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$ → truncated SVD

$$A - A_k = \sum_{i=k+1}^{r} \sigma_i u_i v_i^T$$

Denoted by v the top sing. vector of $A - A_k$. Express v as linear combination of the vectors $v_1, v_2, \ldots, v_r$, (possible since they span row space of $A \& A_k$).

i.e. let $v = \sum_{i=1}^{r} c_i v_i$

Then $\|(A - A_k)v\|$

$$= \left\| \left(\sum_{i=k+1}^{r} \sigma_i u_i v_i^T \right) \left(\sum_{j=1}^{r} c_j v_j \right) \right\|$$

As u_j, v_j are orthonormal.

$$\therefore \|(A - A_k)v\| = \left\| \sum \sigma_i u_i v_i^T v_i \right\|$$

$$= \left\| \sum_{i=k+1}^{r} c_i \sigma_i u_i \right\|$$

$$= \sqrt{\sum_{i=k+1}^{r} c_i^2 \sigma_i^2}$$

since u_i's are also orthonormal.

So the vector v maximizing the last quantity subject to the constraint that $\|v\| = \sum_{i=1}^{r} c_i^2 = 1$ occurs exactly when $c_{k+1} = 1$ & remaining c_i's are zero. (because σ_i's are in descending order) the summation begins at $i = k+1$; & all c_i's are b/w 0 & 1.

So this shows that $\|A - A_k\| = \sigma_{k+1}^2$

Hence proved

Theorem 3.9 :-

Let $A_{n \times d}$, for any matrix B of at most k,

$$\|A - A_k\| \leq \|A - B\|_2$$

(Note: the truncation got via SVD of A is the best rank k approximation even in 2-norm or spectral norm)

Proof :-

if A is rank k or less, nothing to prove $\because \|A - A_k\|_2 = 0$ for $s(A) \leq k$.

Assume $\text{rank}(A) > k$

As, $\|A - A_k\|_2^2 = \sigma_{k+1}^2$

$\because$ The null space of B, the set of

vectors 'v' such that $Bv = 0$ has dimension atleast $(d-k)$. Let $v_1, v_2, \ldots, v_{k+1}$ be the first $(k+1)$ singular vectors of A.

By dimension argument, there exists a vector '$z \neq 0$' in Null(B) & span $\{v_1, v_2, v_3, \ldots, v_{k+1}\}$

Normalize z to have unit length

so, $\|z\| = 1$

Then, $\|A - B\|_2^2 \geq \|(A-B)z\|^2$

/By def^n of 2-Norm.

So z is span of $\{v_1, v_2, \ldots, v_{k+1}\}$

$$\therefore \|Az\|^2 = \left\| \sum_{i=1}^{n} \sigma_i u_i v_i^T z \right\|^2$$

$$= \sum_{i=1}^{n} \sigma_i (v_i^T z)^2$$

$$\|Az\|^2 = \sum_{i=1}^{k+1} \sigma_i^2 (v_i^T z)^2 = \sum_{i=1}^{n} \sigma_i (v_i^T z)^2$$

$$\geq \sigma_{k+1}^2 \sum_{i=1}^{k+1} (v_i^T z)^2 = \sigma_{k+1}^2$$

$\because$ $z \rightarrow$ unit vector $\& \sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \cdots \geq \sigma_{k+1}$

From this we get that,

$$\boxed{\|A - B\|_2^2 \geq \sigma_{k+1}^2}$$

Hence the proof.