Topics in ML:

(1) what is big data?

let ϕ be an algorithm

$S_\phi(n)$ = storage required by ϕ to act on input of size n.

$R_\phi(n)$ = runtime —"—

$S_{max}(t)$ = max storage available at time t.

Δt = doubling time $\quad S_{max}(t+\Delta t) = 2S_{max}(t)$

$n_{max} = \max\{n' : S_\phi(n') \leq S_{max}\}$

clearly $n_{max}(t+\Delta t) = 2n_{max}(t)$

so $R_\phi(n_{max}(t+\Delta t)) = R_\phi(2n_{max}(t))$

Suppose $R_\phi(n) = O(n^p)$ $\quad$ polynomial time!

then $R_\phi(n_{max}(t+\Delta t)) = R_\phi(2n_{max}(t))$

$= 2^p R_\phi(n_{max}(t))$

Now, since computer speed has doubled, computational time increases by $2^p/2 = 2^{p-1}$.

Def^n An algorithm is <u>scalable</u> if $p = 1$

(or $R_\phi = O(n \log n)$).

why?

<u>Answer</u>: big data is about <u>scalable algorithms</u>

②

(2) matrix factorizⁿ

$$X \simeq A \cdot B$$

$$(D \times N) \quad (D \times R) \cdot (R \times N)$$

→ storage cost: X : DN

A, B : $R(N+D)$

→ computational cost: $X \cdot \underline{v}$ $\quad D(2N-1)$

$AB\underline{v}$ $\quad R(2N-1) + D(2R-1)$

applications

(1) dimension reduction
(2) sparse coding
(3) collaborative filtering / recommender systems
(4) inpainting
(5) compression
(6) NNs.

D movies by N users $\quad X$

$A \; B \quad D$ movies $\times R$ latent features

R latent features $\times N$ users.

eg. recommender systems

$X =$

	alice	bob	edna	frank
star wars	3			
wall E		5		2
trainspotting		4	0	
metropolis				1
citizen kane				

Principal Component Analysis

mean: $\bar{\underline{x}} = \frac{1}{N}\sum_{n=1}^{N} \underline{x}_n$

covariance: $\Sigma = \frac{1}{N}\sum_{n=1}^{N} (\underline{x}_n - \bar{\underline{x}})(\underline{x}_n - \bar{\underline{x}})^T$

projected data:

mean: $\underline{u}_1^T \bar{\underline{x}}$

covariance: $\frac{1}{N}\sum_{n=1}^{N} (\underline{u}_1^T \underline{x}_n - \underline{u}_1^T \bar{\underline{x}})^2 = \frac{1}{N}\sum_{n=1}^{N} \left[\underline{u}_1^T (\underline{x}_n - \bar{\underline{x}}) \right]^2$

$= \frac{1}{N}\sum_{n=1}^{N} \underline{u}_1^T (\underline{x}_n - \bar{\underline{x}})(\underline{x}_n - \bar{\underline{x}})^T \underline{u}_1$

$= \underline{u}_1^T \Sigma \underline{u}_1$

Goal: maximise variance of projected data

ie "find most interesting direction in data"

$$\max_{\underline{u}_1} \underline{u}_1^T \Sigma \underline{u}_1 \quad \text{s.t.} \quad \|\underline{u}_1\|_2^2 = 1$$

Lagrangian: $\mathcal{L} = \underline{u}_1^T \Sigma \underline{u}_1 + \lambda_1 (1 - \underline{u}_1^T \underline{u}_1)$

set $\partial / \partial \underline{u}_1 = 0$, get $\Sigma \underline{u}_1 = \lambda_1 \underline{u}_1$

First principal direction: $\Sigma \underline{u}_1 = \lambda_1 \underline{u}_1$

$\underline{u}_1$ is evector of Σ w/ eval λ_1

λ_1 is variance of projected data $\boxed{\text{why?}}$

Minimum Error formulation

Define orthonormal basis $\{\underline{u}_n\}$

proj^n of $\underline{x}_n$ onto $\underline{u}_d$ is $z_{n,d} = \underline{x}_n^T \underline{u}_d$

$$\underline{x}_n = \sum_{d=1}^{D} z_{n,d}\, \underline{u}_d = \sum_{d=1}^{D} (\underline{x}_n^T \underline{u}_d)\underline{u}_d$$

why?

$$\tilde{\underline{x}}_n = \sum_{d=1}^{R} a_{n,d}\, \underline{u}_d + \underbrace{\sum_{d=R+1}^{D} b_d \cdot \underline{u}_d}_{\text{doesn't depend on } n.}$$

Approx^n error:

$$J(\{a_{n,d}\}, \{b_d\}) = \frac{1}{N}\sum_{n=1}^{N} \|\underline{x}_n - \tilde{\underline{x}}_n\|_2^2$$

minimize J wrt $a_{n,d}$ & b_d:

$$a_{n,d} = \underline{x}_n^T \underline{u}_d$$

$$b_d = \bar{\underline{x}}^T \underline{u}_d$$

why?

~~min wrt $\underline{u}_d$: $J =$~~

$$\underbrace{\underline{x}_n - \tilde{\underline{x}}_n}_{\text{displacement vector}} = \sum_{d=R+1}^{D} \left((\underline{x}_n - \bar{\underline{x}})^T \underline{u}_d\right) \underline{u}_d$$

is orthogonal to principal subspace.

Minimis^n wrt $\underline{u}_d$

$$J = \frac{1}{N}\sum_{n=1}^{N} \left(\underline{x}_n^T \underline{u}_d - \bar{\underline{x}}^T \underline{u}_d\right)^2$$

$$= \sum_{d=R+1}^{D} \underline{u}_d^T \Sigma\, \underline{u}_d$$

minimise by picking evectors $\underline{u}$ w/ smallest evals.

eg $D=2$, $K=1$

choose $\underline{u}_2$ to minimize $\underline{u}_2^T \Sigma \underline{u}_2$

subject to $\underline{u}_2^T \underline{u}_2 = 1$

$$\mathcal{L} = \underline{u}_2^T \Sigma \underline{u}_2 + \lambda_2 (1 - \underline{u}_2^T \underline{u}_2)$$

$\hookrightarrow \Sigma \underline{u}_2 = \lambda_2 \underline{u}_2$

$\rightarrow J = \lambda_2$.

Error is minimized by choosing smaller of the two evals.

<u>Matrix viewpoint:</u>

$X = [\underline{x}_1, \text{---}, \underline{x}_N]$ $\qquad$ $\underline{x}_n = D$-dim vec.

mean-centred data $\bar{X} = X - M$ $\qquad$ $M = [\bar{x}, -, \bar{x}]$

<u>Matrix factorizn of $\bar{X}$:</u>

$$\underset{(R \times N)}{\bar{Z}_R} = \underset{(R \times D)}{U_R^T} \cdot \underset{(D \times N)}{\bar{X}}$$

to approximate $\bar{X}$, return to original basis

$$\tilde{\bar{X}} = U_R \cdot \bar{Z}_R$$

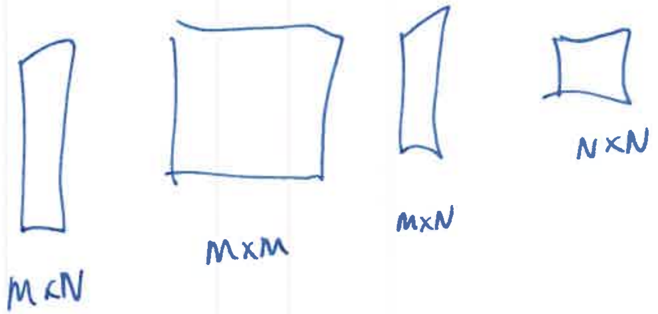
for $R = D$, get $\tilde{\bar{X}} = X$!

Singular Value Decomposition

<u>Thm</u>: let $A \in \mathbb{R}^{M \times N}$

then A can be decomposed (not nec. uniquely) as

$$A = U \cdot D \cdot V^T$$



$U^T U = I_M = U U^T$

D diagonal

$V^T V = I_N = V V^T$

Elements along diagonal of D are called <u>singular values</u>

Assume WLOG $M > N$.

$$\begin{cases} d_{i,i} > 0 & 1 \le i \le r \\ d_{i,i} = 0 & (r+1) \le i \le N \\ d_{i,j} = 0 & i \ne j \end{cases}$$

by convention order singular values from high-to-low.

$$d_1 \ge d_2 \ge \cdots \ge d_r > d_{r+1} = \cdots = d_N = 0.$$

Second principal direction:

$$\max_{\underline{u}_2} \ \underline{u}_2^T \Sigma \underline{u}_2 \quad \text{s.t.} \quad \|u_2\|_2^2 = 1 \ \& \ \underline{u}_2^T \underline{u}_1 = 0$$

$$\mathcal{L} = \underline{u}_2^T \Sigma \underline{u}_2 + \lambda_2 (1 - \underline{u}_2^T \underline{u}_2) + \eta (\underline{u}_2^T \underline{u}_1)$$

$$\frac{\partial \mathcal{L}}{\partial u_2} = 0 \implies 2 \Sigma \underline{u}_2 - 2\lambda_2 \underline{u}_2 + \eta \underline{u}_1 = 0$$

orthogonality $\implies \eta = 0$ Why?

choose $\underline{u}_2$ = evector w/ 2^{nd} largest eval.

More generally, decomposition $\Sigma = U \Lambda U^T$ contains all relevant info.

Why?

Spectral Theorem: If A is Hermitian then $\exists$ orthonormal basis of V consisting of evecs of A. Evals are real.

proof: FTA applied to char polyn $\implies$

$\exists$ eval λ_1 & evec $\underline{e}_1$. Why?

then $\lambda_1 \langle \underline{e}_1, \underline{e}_1 \rangle = \langle A\underline{e}_1, \underline{e}_1 \rangle$

$= \langle \underline{e}_1, A\underline{e}_1 \rangle = \bar{\lambda} \langle \underline{e}_1, \underline{e}_1 \rangle$

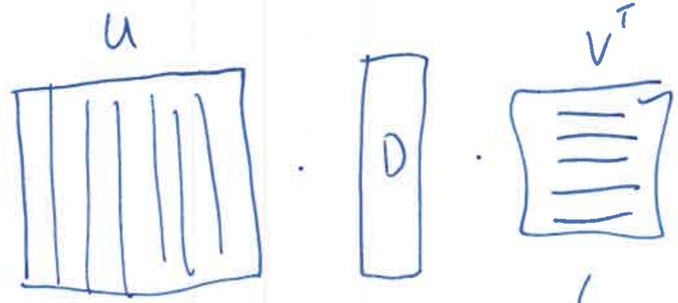
so $\lambda_1 \in \mathbb{R}$.

Now consider $K = \text{span}(\underline{e}_1)^\perp$

$AK \subset K$ Why?

induction. $\blacksquare$

Left & right singular vectors



left-singular vectors
orthogonal basis for
space spanned by columns of A

right singular vectors
orthogonal basis for
row space of A.

Singular values & vectors

$$\left.\begin{aligned} A\underline{v}_i &= d_i\, \underline{u}_i \\ A^T \underline{u}_i &= d_i\, \underline{v}_i \end{aligned}\right\} \quad i = 1 \ldots \min\{M,N\}.$$

Range, Null-space, Rank

* right singular vectors corresponding to vanishing singular values span null-space of A.
* left singular ~~values~~ vectors corresponding to non-zero singular values of A span range of A.

$$\Longrightarrow \quad rk(A) = \#\ \text{non-zero singular values}$$
$$= \#\{d_i > 0\}$$

Existence of SVD

let $\sigma(\underline{u},\underline{v}) = \underline{u}^T M \underline{v}$ where $\|\underline{u}\|_2 = 1 = \|\underline{v}\|_2$

Since σ is continuous & S^{M-1}, S^{N-1} are cpt, it follows that a maximum is attained & is ≥ 0. why?

Denote max by $\sigma_1 \geq 0$.

Thm: $\underline{u}_1, \underline{v}_1$ are left & right-singular vectors corresponding to σ_1.

proof: $\nabla_{\underline{u},\underline{v}} \sigma = \nabla_{\underline{u},\underline{v}} \underline{u}^T M \underline{v} - \lambda_1 \nabla_{\underline{u},\underline{v}} \underline{u}^T \underline{u} - \lambda_2 \nabla_{\underline{u},\underline{v}} \underline{v}^T \underline{v} = 0.$

$\hookrightarrow \nabla_{\underline{u}} \underline{u}^T M \underline{v} = \lambda_1 \nabla_{\underline{u}} \underline{u}^T \underline{u} \longrightarrow M\underline{v} = 2\lambda_1 \underline{u}_1$

& $\nabla_{\underline{v}} \underline{u}^T M \underline{v} = \lambda_2 \nabla_{\underline{v}} \underline{v}^T \underline{v} \longrightarrow M^T \underline{u} = 2\lambda_2 \underline{v}$

$M\underline{v}_1 = \sigma_1 \underline{u}_1$ & $M^T \underline{u}_1 = \sigma_1 \underline{v}_1$ $\longleftarrow$ $\sigma_1 = 2\lambda_1 = 2\lambda_2.$

Eigenvectors of AA^T & A^TA

$A = UDV^T$

$\therefore \quad AA^T = UD^2U^T$ columns of U are evectors of AA^T evals are σ_i^2 in both cases.

$A^TA = VD^2V^T$ cols of V ~~rows~~ A^TA

Why SVD & not PCA?

suppose D large.

cov. has dim D^2.

Pseudo Inverse

let D be diagonal

$$D^+_{ii} = \begin{cases} 1/D_{ii} & \text{if } D_{ii} \neq 0 \\ 0 & \text{else.} \end{cases}$$

Suppose $A = UDV^T$

then $A^+ = VD^+U^T$ pseudo-inverse.

(1) if A square & invertible then $A^+ = A^{-1}$

(2) if A is overdetermined, $A^+\underline{b}$ gives least-squares sol^n to $A\underline{x} \simeq \underline{b}$

(3) if A underdetermined, $A^+\underline{b}$ gives least squares sol^n to $A\underline{x} \simeq \underline{b}$ w/ minimal norm.