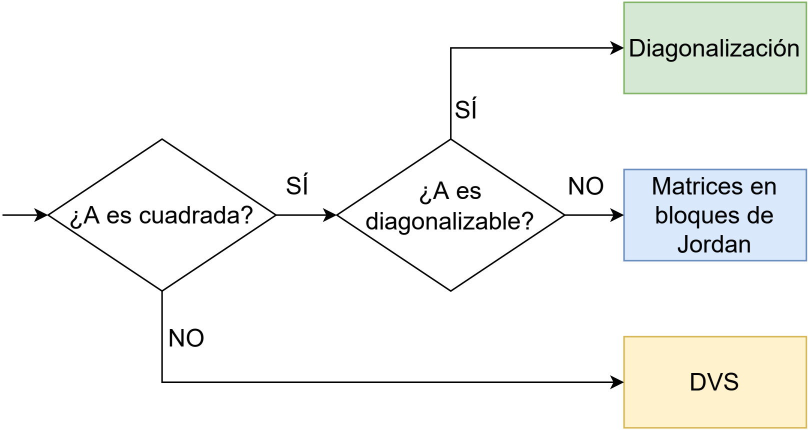
Análisis Matemático para Inteligencia Artificial

Martín Errázquin (merrazquin@fi.uba.ar)

Especialización en Inteligencia Artificial

Clase 4

¿Qué puede pasar al intentar diagonalizar?

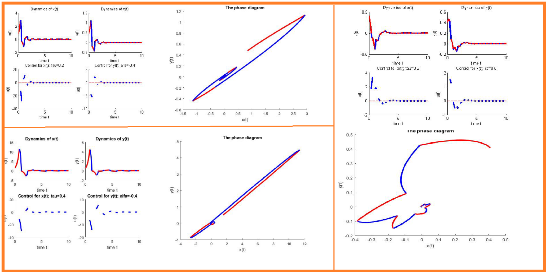


Sea $L : \mathbb{V} \to \mathbb{V}$ endomorfismo sobre $\mathbb{K} - e.v.$ con $dim(\mathbb{V}) = n$. Si $p(\lambda)$ se factoriza en $\mathbb{K}$ entonces existe una base que viene dada por la matriz en bloques $(m < n)$.

$$J = \begin{bmatrix} A_1 & . & . & . & 0 \\ 0 & . & & & . \\ . & & . & & . \\ . & & & . & . \\ 0 & . & . & . & A_m \end{bmatrix} \in \mathbb{K}^{n \times n}, \; A_k = \begin{bmatrix} \lambda_k & 1 & 0 & \dots & 0 \\ 0 & . & . & & . \\ . & & . & . & . \\ . & & & . & 1 \\ 0 & . & . & . & \lambda_k \end{bmatrix}, \; \sum_{k=1}^{m} \lambda_k = n.$$

Obs.: Cuando es diagonalizable $m = n$.

Aplicación en Control



¿Qué pasa cuando A no es cuadrada?

$S^T = (A^T A)^T = A^T \cdot (A^T)^T = A^T \cdot A = S$

$(AB)^T = B^T \cdot A^T$

$x^T S x = x^T A^T A x = (Ax)^T (\underbrace{Ax}_{y}) = y^T y = \|y\|^2 \geq 0$

$\forall x \in \mathbb{R}^n \setminus \{0_v\}$

Teorema: Dada $A \in \mathbb{R}^{m \times n}$, siempre se puede obtener una matriz simétrica, semidefinida positiva $S \in \mathbb{R}^{n \times n}$ como:

si $X \in \mathbb{R}^{n \times n}$

n m n $m < p$ $n < p$

$$S \doteq \underset{n \times m}{A^T} \underset{m \times n}{A}$$

$X^2 = X \cdot X$

Si X es diagonizable $= P \cdot D^2 \cdot P^{-1}$

Obs.: $\tilde{S} \in \mathbb{R}^{m \times m}$ definida como $\tilde{S} \doteq \underset{m \times n}{A}\underset{n \times m}{A^T}$ también resulta simétrica y semidefinida positiva.

A tiene rango r $\quad 0 \leq r \leq p$

Obs. (II): Como $S, \tilde{S}$ son semidef. positivas, todos sus autovalores son reales **no negativos** (¡incluso si A tomara valores complejos!).

$\lambda_i \geq 0$

↳ les puedo hacer $\sqrt{\lambda_i} = \sigma_i$

$p = \min(m,n)$

$\vec{\lambda} = (\lambda_1, \dots, \lambda_p)$ (ordenados de mayor a menor)

$\vec{\lambda}' = (\lambda_1, \dots, \lambda_p, 0, \dots, 0)$

DVS: Definición

Teorema: Sea $A \in \mathbb{R}^{m\times n}$ una matriz de rango $r \in \{0, \dots, \text{mín}\{m, n\}\}$. La descomposición en valores singulares (DVS, SVD en inglés) de A se define como:

$$X = P \cdot D \cdot P^{-1}$$

B→E diag E→B

$$\underset{m\times n}{A} = \underset{m\times m}{U} \cdot \underset{m\times n}{\Sigma} \cdot \underset{n\times n}{V^T}$$

sea $A \in \mathbb{R}^{3\times 5} \Rightarrow p = 3$

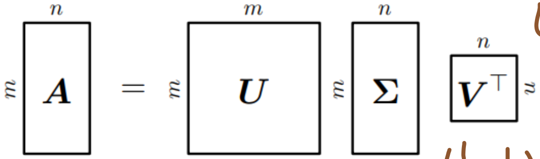
$$\Rightarrow \Sigma = \begin{pmatrix} \sigma_1 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_3 & 0 & 0 \end{pmatrix}$$

donde:

- $U \in \mathbb{R}^{m\times m}$ es la matriz ortogonal formada por los aves de AA^T. $= \tilde{S}$
- $\Sigma \in \mathbb{R}^{m\times n}$ es la matriz "diagonal" de valores singulares $\sigma_i = \sqrt{\lambda_i}$, con λ_i los primeros $p = \text{mín}\{m, n\}$ avas de AA^T y $A^T A$.
- $V \in \mathbb{R}^{n\times n}$ es la matriz ortogonal formada por los aves de $A^T A$. $= S$

si Q es ortogonal

$Q^T = Q^{-1}$



$U = \begin{pmatrix} u_1 & \cdots & u_m \end{pmatrix}$

$\hookrightarrow P_{\tilde{S}}$

$V = \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix} = P_S$

Sea $A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \in \mathbb{R}^{2\times 3}$

$p=2$

① $\tilde{S} = AA^t = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \in \mathbb{R}^{2\times 2}$ $\quad \lambda_1 = 2,\ \lambda_2 = 1$

$S = A^tA = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \in \mathbb{R}^{3\times 3}$ $\quad \lambda_1 = 2,\ \lambda_2 = 1,\ \lambda_3 = 0$

los p primeros ava son compartidos

el resto vale 0

② aves de $S \Rightarrow$ resolver $(S - \lambda I)v = \vec{0}$

$v_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix},\ v_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},\ v_3 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \Rightarrow V = \begin{pmatrix} | & | & | \\ \hat{v}_1 & \hat{v}_2 & \hat{v}_3 \\ | & | & | \end{pmatrix}$

$= \begin{pmatrix} 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 0 & 1 & 0 \end{pmatrix}$

$\Downarrow$

$\Sigma = \begin{pmatrix} \sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \in \mathbb{R}^{2\times 3}$

③ aves de $\tilde{S}$

$\hookrightarrow$ opc 1) $(\tilde{S} - \lambda I)u = \vec{0}$

$\hookrightarrow$ opc 2) $u_i = \dfrac{A \cdot v_i}{\sigma_i}$

$u_1 = \dfrac{\begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{pmatrix}}{\sqrt{2}} = \dfrac{\begin{pmatrix} 2/\sqrt{2} \\ 0 \end{pmatrix}}{\sqrt{2}} = \begin{pmatrix} 2/2 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

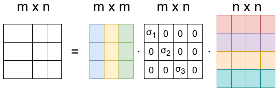
$u_2 = \dfrac{\begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}}{1} = \dfrac{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}{1} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

$\Rightarrow U = \begin{pmatrix} | & | \\ u_1 & u_2 \\ | & | \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

 $m=4, n=6 \Rightarrow p=4$

Sea $A \in \mathbb{R}^{m\times n}$ con DVS $A = U\Sigma V^T$:

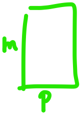
$\Sigma = \begin{pmatrix} 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{4\times 6}$



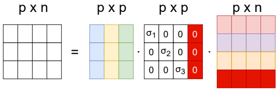
$\Sigma_p \in \mathbb{R}^{4\times 4}$

- Si $p = min\{m,n\}$, la DVS compacta es $A = U_p\Sigma_p V_p^T$

$r=3 \Rightarrow \Sigma_r \in \mathbb{R}^{3\times 3}$



como $m<n \Rightarrow p=m$



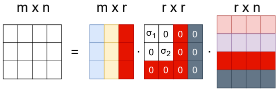
$k=1 \Rightarrow \Sigma_k = 3$

- Si $\sigma_1,\ldots,\sigma_r > 0, \sigma_{r+1},\ldots,\sigma_p = 0$, la DVS reducida es $A = U_r\Sigma_r V_r^T$

si $r<p$

$\sigma_1 = 3$ $\quad \sigma_3 = 10^{-1}$

$\sigma_2 = 1$ $\quad \cancel{\sigma_4 = 0}$ $\quad \cancel{\sigma_5 = 0}$

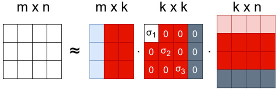


Hasta la reducida se mantiene el valor de A.
En la truncada hay perdida de información

- Sea $k \in \{1,\ldots,r-1\}$, la DVS truncada es $A \approx U_k\Sigma_k V_k^T$

$\sigma_1 = 10^5$ $\quad \sigma_3 = 10^{-2} \approx 0$

$\sigma_2 = 10^4$ $\quad \sigma_4 = 10^{-5} \approx 0$



Pseudoinversa de Moore-Penrose

$m \boxed{}$
n

Sea $A \in \mathbb{R}^{m \times n}$ y el sistema lineal $Ax = y$. La inversa de A no está definida, pero si $m > n$ una posible solución (cuadrados mínimos) era

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

donde Σ_r no tiene 0's

¿Pero qué ocurre si $A^T A$ no es invertible? Se define la pseudoinversa de Moore-Penrose como

$$A^\dagger = V_r \Sigma_r^{-1} U_r^T$$

$$\text{“}(U_r \cdot \Sigma_r \cdot V_r^T)^{-1}\text{”} = \text{“}A^{-1}\text{”}$$

y $\hat{x} = A^\dagger y$ es (si $m < n$) la solución de mínima norma euclídea o (si $m > n$) la aproximación de mínimo error cuadrático/distancia euclídea.

Obs.: Todo lo anteriormente visto también vale para $A \in \mathbb{C}^{m \times n}$, utilizando $\cdot^H$ en vez de $\cdot^T$. Se utilizaron matrices reales para facilitar la lectura.

Colab

Una aplicación obvia para la DVS es, como vimos recién, la "solución"de sistemas de ecuaciones para los cuales el método de cuadrados mínimos no es factible.

¿Pero es ese el único uso que le podemos dar? Las descomposiciones siempre nos brindan un criterio bajo el cual obtener la *mejor* aproximación de *orden* k para el objeto a descomponer, y eso podría llegar a ser muy útil.

Otras factorizaciones

$Ax=b$ $\begin{cases} 3x+4y+5z=6 \\ 2y+3z=4 \\ -2z=-5 \end{cases}$

Si bien la diagonalización y la DVS son las factorizaciones más importantes, existen otras útiles. Algunas de ellas (con aplicaciones en ML) son:

- **LU:** $PA = LU$ con P matriz de permutación de filas, L triangular inferior con diagonal de 1s, U triangular superior. Muy eficiente para resolver problemas del tipo $Ax = b$ cuando b va variando.
- **QR:** $A = QR$ con Q matriz ortogonal y R triangular superior. Algoritmos más simples pero un poco menos eficientes que LU para resolver $Ax = b$, también provee ortogonalización.
- **Cholesky:** $A = LL^T$ para A simétrica (semi)definida positiva, con L triangular inferior. Permite eficientizar muchos cálculos como A^{-1}, updates a A, etc. Muy usado en modelos probabilísticos con matrices de covarianza.

¿Por qué hacer una factorización?

Tomemos por ejemplo la factorización de Cholesky con $A = LL^T$. Si queremos resolver la ecuación $Ax = b$, debemos:

1. Invertir A, para obtener A^{-1}: orden $2n^3 + \mathcal{O}(n^2)$
2. Computar $x = A^{-1}b$: orden $2n^2$

En total, está dominado por $2n^3$.

Si en cambio usamos la factorización de Cholesky tenemos:

1. Computar L tal que $LL^T = A$: orden $\frac{1}{3}n^3$
2. Resolver sistema triangular $L^T y = b$: orden n^2
3. Resolver sistema triangular $Lx = y$: orden n^2

$2n^2$

Si bien el orden de magnitud es el mismo, ahora está dominado por $\frac{1}{3}n^3$.

Nota: hay opciones de pivoteo directo (o vía LU), que son del orden $\frac{2}{3}n^3$.

Presentación TP → OPCIONAL!

Recordatorio: consigna en Notebook en el Campus.

Vamos a tomar dos modelos clásicos de clasificación multiclase, LDA/QDA y vamos a:

- Implementarlos
- Tensorizarlos
- Explotar su estructura
- Medir las ganancias en tiempo y memoria

Los objetivos del TP son:

- Familiarizar al alumno con el ida y vuelta *cuenta-código*
- Familiarizar y "amigar" al alumno con la paralelización de operaciones matriciales usando tensores de rango 3+
- Concientizar al alumno sobre las ventajas del uso inteligente y eficiente de recursos

Mini-intro teórica

Ejemplo: tickets de compra de amigos Ana, Beto y Carlos.

Recordatorio: Encuesta de Clase

Recordar que está disponible la encuesta de clase! Completarla es cortito y sirve para ir monitoreando el estado del curso.

¿Dónde encontrarla? En la hoja de notas (e.g. "Notas CEIA 10Co2024"), abajo de todo, junto al link de las grabaciones de las clases.