Descomposició en valors singulars: Introducció i aplicacions

Contextualització i objectius per a la ciència de dades

PID_00262384

Francesc Pozo Montero
Jordi Ripoll Missé

Universitat Oberta de Catalunya

Francesc Pozo Montero

Llicenciat en Matemàtiques per la Universitat de Barcelona (2000) i doctor en Matemàtica Aplicada per la Universitat Politècnica de Catalunya (2005). Ha estat professor associat a la Universitat Autònoma de Barcelona i professor associat, col·laborador i actualment professor agregat a la Universitat Politècnica de Catalunya. A més, és cofundador del Grup d'Innovació Matemàtica E-learning (GIMEL), responsable de diversos projectes d'innovació docent i autor de diverses publicacions. Com a membre del grup de recerca consolidat CoDAlab, centra la recerca en la teoria de control i les aplicacions en enginyeria mecànica i civil, com també en l'ús de la ciència de dades per al monitoratge de la integritat estructural i per al monitoratge de la condició, sobretot en turbines eòliques.

Jordi Ripoll Missé

Llicenciat en Matemàtiques i doctor en Ciències Matemàtiques per la Universitat de Barcelona (2005). Professor col·laborador de la Universitat Oberta de Catalunya des del 2011 i professor del Departament d'Informàtica, Matemàtica Aplicada i Estadística de la Universitat de Girona (UdG) des del 1996, on actualment és professor agregat i desenvolupa tasques de recerca en l'àmbit de la biologia matemàtica (models amb equacions en derivades parcials i dinàmica evolutiva). També ha estat professor i tutor de la UNED en dues etapes, primer al centre associat de Terrassa i actualment al de Girona. Ha participat en nombrosos projectes d'innovació docent, especialment pel que fa a l'aprenentatge de les matemàtiques en línia.

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats per la professora: Cristina Cano Bastidas (2019)

Primera edició: febrer 2019

Av. Tibidabo, 39-43, 08035 Barcelona
Disseny: Manel Andreu
Realització editorial: Oberta UOC Publishing, SL

Índex

Introducció

En aquest mòdul es presenten dues tècniques que estan fortament relacionades tot i que tenen aplicacions diferents. D'una banda, l'anàlisi de components principals (*principal component analysis*, PCA) i, de l'altra, la descomposició en valors singulars (*singular value decomposition*, SVD). Veurem que, en certa manera, PCA serà un cas particular de SVD.

Des d'un punt de vista matemàtic, ambdues tècniques estan fonamentades en el càlcul de valors i vectors propis. Aquest fet demostra, en particular, la importància dels conceptes de valor i vector propi presentats en el mòdul «Aplicacions lineals, diagonalització i vectors propis».

En el cas concret de l'anàlisi de components principals, suposeu que hem mesurat m variables a un total de n mostres. Si aquestes variables són numèriques, la informació resultant es pot emmagatzemar en forma de matriu de la manera següent:

$$\begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nm} \end{bmatrix}$$

on x_{ij} és la mesura de la variable j-èssima a l'i-èssim element de la mostra.

Noteu que el nombre de files de la matriu anterior representa la mida total de la mostra (per exemple, persones), mentre que el nombre de columnes de la matriu representa el nombre de variables que mesurem a cada una de les mostres (per exemple, alçada, pes, edat, coeficient intel·lectual o poder adquisitiu). Quan el nombre de variables és petit, és possible que les dades es puguin tractar de forma senzilla, fins i tot, visualment. Ara bé, quan el nombre de variables és molt gran, una representació gràfica de les dades esdevé quasi impossible o, si més no, una interpretació ràpida d'aquesta informació. A més, un problema afegit pot venir donat per variables que tenen una alta correlació, com ara el pes i l'alçada. Si aquest és el cas, és a dir, si la primera columna representa el pes i la segona columna representa l'alçada, la informació d'ambdues columnes seria, en certa manera, redundant. Així doncs, l'objectiu de l'anàlisi de components principals és doble:

i) d'una banda, volem definir unes *variables* noves de manera que quan expressem les dades originals en termes de les noves variables, aquestes no siguin redundants, és a dir, no estiguin correlacionades.

ii) d'altra banda, volem reduir la dimensió de les dades originals. És a dir, és possible que amb un nombre $\ell < m$ de les noves variables, no hi hagi pràcticament cap pèrdua d'informació respecte de les dades originals.

Per al cas de la descomposició en valors singulars, considereu la següent matriu:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Aquesta matriu té sis files i sis columnes i, per tant, un total de 36 elements. Ara bé, noteu que aquesta matriu es pot expressar com el producte del següent vector columna pel següent vector fila:

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Això significa que la matriu de 36 elements (que té rang 1) és igual al producte d'un vector columna per un vector fila. En aquest cas, els vectors queden definits per 12 elements. Noteu que 36 és el triple que 12!

Si la matriu inicial tingués dimensió $300 \cdot 300$, la matriu estaria definida per un total de 90.000 elements. Si poguéssim també expressar la matriu com a producte d'un vector columna per un vector fila, el nombre d'elements necessaris seria de $300 \cdot 2 = 600$. En aquest cas, 90.000 no és el triple de 600, és 150 vegades més! Si la matriu de dimensió $300 \cdot 300$ representés una imatge en escala de grisos, on cada element de la matriu representés la intensitat de gris, el que estaríem fent és reduir el *pes* de la imatge. Tot i que aquest pugui semblar un exemple més aviat de l'àmbit de la informàtica gràfica, moltes ve-

gades les dades es poden representar en forma d'imatge i la reducció del seu pes és, doncs, un problema de la ciència de dades. De fet, l'Hospital Clínic i la Universitat Politècnica de Catalunya tenen patentat un mètode pel reconeixement i la classificació de cèl·lules sanguínies anormals que es basa en les imatges microscòpiques d'aquestes cèl·lules. La reducció en la mida d'aquestes imatges és fonamental per poder tractar el volum de dades involucrat.

L'objectiu, doncs, de la descomposició en valors singulars és reduir la dimensionalitat de les dades originals alhora que trobar característiques millors per classificar la informació. Altres aplicacions de la descomposició en valors singulars, més allunyades de la ciència de dades, són el càlcul del rang i el nucli d'una matriu; el càlcul de la pseudoinversa d'una matriu; o la resolució de sistemes sobredeterminats de rang màxim (mínims quadrats) i no de rang màxim.

Objectius

L'objectiu general d'aquest mòdul és presentar dues tècniques de tractament de dades: l'anàlisi de components principals i la descomposició en valors singulars.

En particular, els objectius docents que es pretenen aconseguir amb aquest mòdul són els següents:

1) Comprendre el problema de la maledicció de la dimensionalitat en la ciència de dades.

2) Comprendre la utilitat dels conceptes d'àlgebra lineal que s'han treballat en els mòduls anteriors en l'aplicació en l'àmbit de la ciència de dades.

3) Ser capaç de resoldre un problema utilitzant l'anàlisi de components principals en un cas d'ús utilitzant dades reals o realistes.

4) Ser capaç de reduir el pes d'una imatge utilitzant la descomposició en valors singulars.

5) Entendre la utilitat d'utilizar un llenguatge de programació pel tractament de grans volums de dades.

6) Agafar destresa en la utilització del llenguatge R per a la resolució de problemes amb un gran volum de dades.