Descomposición en valores singulares: Introducción y aplicaciones

Contextualización y objetivos para la ciencia de datos

PID_00262385

Francesc Pozo Montero
Jordi Ripoll Missé

Universitat Oberta de Catalunya

Francesc Pozo Montero

Licenciado en Matemáticas por la Universidad de Barcelona (2000) y doctor en Matemática Aplicada por la Universidad Politécnica de Cataluña (2005). Ha sido profesor asociado de la Universidad Autónoma de Barcelona y profesor asociado, colaborador y actualmente profesor agregado en la Universidad Politécnica de Cataluña. Además, es cofundador del Grupo de Innovación Matemática E-learning (GIMEL), responsable de varios proyectos de innovación docente y autor de varias publicaciones. Como miembro del grupo de investigación consolidado CoDAlab, centra su investigación en la teoría de control y las aplicaciones en ingeniería mecánica y civil, así como en el uso de la ciencia de datos para la monitorización de la integridad estructural y para la monitorización de la condición, sobre todo en turbinas eólicas.

Jordi Ripoll Missé

Licenciado en Matemáticas y doctor en Ciencias Matemáticas por la Universidad de Barcelona (2005). Profesor colaborador de la Universitat Oberta de Catalunya desde 2011 y profesor del Departamento de Informática, Matemática Aplicada y Estadística de la Universidad de Girona (UdG) desde 1996, donde actualmente es profesor agregado y desarrolla tareas de investigación en el ámbito de la biología matemática (modelos con ecuaciones en derivadas parciales y dinámica evolutiva). También ha sido profesor y tutor de la UNED en dos etapas, primero en el centro asociado de Terrassa y actualmente en el de Girona. Ha participado en numerosos proyectos de innovación docente, especialmente en cuanto al aprendizaje de las matemáticas en línea.

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por la profesora: Cristina Cano Bastidas (2019)

Primera edición: febrero 2019

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

Índice general

Introducción

En este módulo se presentan dos técnicas que están fuertemente relacionadas a pesar de que tienen aplicaciones diferentes. Por un lado, el análisis de componentes principales (*principal component analysis*, PCA) y, por el otro, la descomposición en valores singulares (*singular value decomposition*, SVD). Veremos que, en cierto modo, el PCA será un caso particular de SVD.

Desde un punto de vista matemático, ambas técnicas están fundamentadas en el cálculo de valores y vectores propios. Este hecho en particular demuestra la importancia de los conceptos de valor y vector propio presentados en el módulo «Aplicaciones lineales, diagonalización y vectores propios».

En el caso concreto del análisis de componentes principales, suponed que hemos medido m variables en un total de n muestras. Si estas variables son numéricas, la información resultante se puede almacenar en forma de matriz de la manera siguiente:

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

donde x_{ij} es el valor de la variable j-ésima para el i-ésimo elemento de la muestra.

Observad que el número de filas de la matriz anterior representa el tamaño total de la muestra (por ejemplo, personas), mientras que el número de columnas de la matriz representa el número de variables que medimos en cada una de las muestras (por ejemplo, altura, peso, edad, coeficiente intelectual o poder adquisitivo). Cuando el número de variables es pequeño, es posible que los datos se puedan tratar de forma sencilla, incluso, visualmente. Ahora bien, cuando el número de variables es muy grande, una representación gráfica de los datos es casi imposible o, cuando menos, una interpretación rápida de esta información. Además, un problema añadido puede venir dado por variables que tienen una alta correlación, como por ejemplo el peso y la altura. Si este

fuera el caso, es decir, si la primera columna representara el peso y la segunda columna representara la altura, la información de ambas columnas sería, en cierto modo, redundante. Así pues, el objetivo del análisis de componentes principales es doble:

i) por un lado, queremos definir unas *variables* nuevas de forma que cuando expresemos los datos originales en términos de las nuevas variables, estas no sean redundantes, es decir, no estén correlacionadas.

ii) por otro lado, queremos reducir la dimensión de los datos originales. Es decir, es posible que con un número $\ell < m$ de nuevas variables, no haya prácticamente ninguna pérdida de información respecto de los datos originales.

Para el caso de la descomposición en valores singulares, considerad la siguiente matriz:

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

Esta matriz tiene seis filas y seis columnas y, por lo tanto, un total de 36 elementos. Ahora bien, observad que esta matriz se puede expresar como el producto del siguiente vector columna por el siguiente vector fila:

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

Esto significa que la matriz de elementos 36 (que tiene rango 1) es igual al producto de un vector columna por un vector fila. En este caso, los vectores quedan definidos por 12 elementos. ¡Observad que 36 es el triple de 12!

Si la matriz inicial tuviera dimensión $300 \cdot 300$, la matriz estaría definida por un total de 90.000 elementos. Si pudiéramos expresar también la matriz como

producto de un vector columna por un vector fila, el número de elementos necesarios sería de $300 \cdot 2 = 600$. En este caso, 90.000 no es el triple de 600, ¡es 150 veces más! Si la matriz de dimensión $300 \cdot 300$ representara una imagen en escala de grises, donde cada elemento de la matriz representara la intensidad de gris, estaríamos reduciendo el *peso* de la imagen. A pesar de que este pueda parecer un ejemplo más bien del ámbito de la informática gráfica, muchas veces los datos se pueden representar en forma de imagen y la reducción de su peso es, pues, un problema de la ciencia de datos. De hecho, el Hospital Clínic y la Universitat Politècnica de Catalunya tienen patentado un método para el reconocimiento y la clasificación de células sanguíneas anormales que se basa en las imágenes microscópicas de estas células. La reducción del tamaño de estas imágenes es fundamental para poder tratar el volumen de datos involucrado.

El objetivo de la descomposición en valores singulares es reducir la dimensión de los datos originales y encontrar características mejores para clasificar la información. Otras aplicaciones de la descomposición en valores singulares, más alejadas de la ciencia de datos, son el cálculo del rango y el núcleo de una matriz; el cálculo de la pseudoinversa de una matriz; o la resolución de sistemas sobredeterminados de rango máximo (mínimos cuadrados) y no de rango máximo.

Objetivos

El objetivo general de este módulo es presentar dos técnicas de tratamiento de datos: el análisis de componentes principales y la descomposición en valores singulares.

En particular, los objetivos docentes que se pretenden conseguir con este módulo son los siguientes:

1) Comprender el problema de la maldición de la dimensionalidad en la ciencia de datos.

2) Comprender la utilidad de los conceptos de álgebra lineal que se han trabajado en los módulos anteriores aplicados a la ciencia de datos.

3) Ser capaz de resolver un problema utilizando el análisis de componentes principales utilizando datos reales o realistas.

4) Ser capaz de reducir el peso de una imagen utilizando la descomposición en valores singulares.

5) Entender la utilidad de utilizar un lenguaje de programación para el tratamiento de grandes volúmenes de datos.

6) Practicar el uso del lenguaje R para la resolución de problemas con un gran volumen de datos.