Álgebra lineal para la ciencia de datos

PID_00262499

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

Introducción

Este material didáctico está diseñado y pensado para el grado de Ciencia de Datos Aplicada. Por un lado, contiene los aspectos fundamentales del álgebra lineal; por el otro, tiene un enfoque centrado en las aplicaciones del álgebra lineal en el ámbito de la ciencia de datos.

Según el famoso matemático alemán David Hilbert (1862-1943):

> "La matemática es el instrumento que vincula la teoría y la práctica, pensando y observando; establece el puente de conexión y las construye cada vez más fuerte. Por ello, nuestra cultura actual, siempre que pretende entender y aprovechar la naturaleza, coge como base la matemática."

Esta cita de Hilbert establece la importancia de la matemática como herramienta para entender el mundo. Es indiscutible la fuerza que tiene el álgebra lineal dentro de las matemáticas, por la estructura que proporciona a los problemas y porque es la base de muchas aplicaciones –análisis de riesgos, optimización de la producción, predicción de beneficios o simulación de sistemas, por citar algunos ejemplos–, especialmente de las técnicas y estrategias vinculadas a la ciencia de datos aplicada.

Los contenidos de este material didáctico se dividen en cinco retos:

1) ¿Por qué el álgebra lineal es importante en la ciencia de datos? ¿Qué elementos básicos tiene?

2) ¿Cómo podemos resolver problemas típicos de la ciencia de datos mediante sistemas de ecuaciones lineales?

3) ¿Qué son los valores y vectores propios (de matrices) y para qué los utiliza Netflix?

4) ¿Cómo podemos afrontar la maldición de la dimensionalidad en la ciencia de datos con el análisis de componentes principales y la descomposición en valores singulares?

5) ¿Cómo podemos modelar sistemas dinámicos con cadenas de Markov tal y como hace el algoritmo PageRank de Google?

Los primeros dos retos establecen los elementos básicos del álgebra lineal, como son la estructura de matriz, el concepto de determinante o los espacios vectoriales y sus operaciones. También se presenta la potencia de los sistemas de ecuaciones lineales y su representación en forma de matriz. A pesar de que el concepto de matriz sea sencillo, cabe destacar que es clave para el álgebra lineal y para la ciencia de datos aplicada.

El tercer reto introduce conceptos un poco más complejos, como las aplicaciones lineales, su representación matricial y los valores y vectores propios. En todos los retos anteriores se presenta una contextualización que enmarca cuáles son las aplicaciones de estos conceptos y procedimientos en el ámbito de la ciencia de datos aplicada.

Los dos últimos retos recogen estrategias que, a pesar de que no sean exclusivas del ámbito mencionado, tienen una clara aplicación. Por un lado, en el reto "¿Cómo podemos afrontar la maldición de la dimensionalidad en la ciencia de datos con el análisis de componentes principales y la descomposición en valores singulares?" se detallan dos técnicas –el análisis de componentes principales (PCA) y la descomposición en valores singulares (SVD)– que, aunque estén íntimamente relacionadas, pueden tener aplicaciones diferentes. En el caso de PCA, permite reducir la dimensionalidad de los datos y descubrir patrones o estructuras ocultas. En cuanto a SVD, una de las aplicaciones más destacadas es la compresión de imágenes.

Finalmente, en el reto "¿Cómo podemos modelar sistemas dinámicos con cadenas de Markov tal y como hace el algoritmo PageRank de Google?" se introducen las matrices estocásticas que permiten representar sistemas dinámicos discretos. Por medio de las técnicas recogidas en este reto, los creadores de Google generaron su algoritmo para evaluar, por ejemplo, la importancia de los documentos con enlaces mutuos, tales como las páginas web y los diferentes enlaces que contienen.

Los retos han sido pensados para que los estudiantes se centren en aprender los conceptos matemáticos y en la manera de aplicarlos a la resolución de problemas de la vida cotidiana. Los problemas, de carácter general, han sido contextualizados en el ámbito de la ciencia de datos aplicada. Los dos últimos retos incorporan casos de estudio y guías de resolución en el lenguaje de programación R. Así se pretende dar valor al aprendizaje de *software* matemático y estadístico sin perder de vista la importancia de comprender los conceptos explicados.

Todos los retos tienen una estructura similar, aunque puede que algunos de estos apartados no estén presentes:

- Conocimientos previos aconsejables para un buen aprovechamiento del aprendizaje del reto y ejercicios para que los estudiantes puedan comprobar su grado de adquisición.
- Ejemplo introductorio al tema del módulo. Con este elemento se pretende insistir en el enfoque aplicado de estos recursos docentes.
- Exposición de los conceptos y de las aplicaciones correspondientes, así como el uso de *software* matemático como ayuda al aprendizaje y numerosos ejemplos que los ilustran. Algunos módulos incluyen actividades sugeridas con la solución al final.

- Resumen de los conceptos más significativos del reto.
- Ejercicios de autoevaluación del aprendizaje de los conceptos fundamentales.
- Solucionario de los ejercicios de autoevaluación.
- Glosario de términos.
- Bibliografía recomendada.

Objetivos

Los cinco retos que hemos presentado en esta introducción forman parte de los recursos docentes de la asignatura de Álgebra para el grado de Ciencia de Datos Aplicada. Sus objetivos son:

1. Conocer y ser capaz de manipular elementos básicos del álgebra lineal (espacios vectoriales, independencia lineal, dimensión, matrices y determinantes) y de la geometría métrica (productos escalares, ortonormalidad, ángulos y distancias).
2. Comprender la importancia de los sistemas de ecuaciones lineales para resolver problemas típicos de la ciencia de datos.
3. Entender el concepto de vectores y valores propios, así como la manera de calcularlos y de interpretarlos geométricamente.
4. Conocer el análisis de componentes principales y ser capaz de aplicar esta estrategia a un caso de uso utilizando datos reales o realistas.
5. Saber resolver un problema mediante la descomposición de valores singulares en un caso de uso utilizando datos reales o realistas.
6. Ser capaz de resolver un problema con modelos matriciales en un caso de uso utilizando datos reales o realistas.
7. Coger destreza en la utilización del lenguaje R para la resolución de problemas con un gran volumen de datos.