Àlgebra lineal per a la ciència de dades

PID_00262498

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

Aquest material didàctic està dissenyat i pensat per al grau de Ciència de Dades Aplicada. D'una banda, conté els aspectes fonamentals de l'àlgebra lineal; de l'altra, té un enfocament centrat en les aplicacions de l'àlgebra lineal a l'àmbit de la ciència de dades.

Segons el famós matemàtic alemany David Hilbert (1862-1943):

> "La matemàtica és l'instrument que vincula la teoria i la pràctica, tot pensant i observant; hi estableix una forta connexió. Per això, la nostra cultura actual, sempre que pretén entendre i aprofitar la natura, pren com a base la matemàtica."

Aquesta citació de Hilbert estableix la importància de la matemàtica com a eina per entendre el món. És indiscutible la força que té l'àlgebra lineal dins de les matemàtiques, per l'estructura que proporciona als problemes i perquè és la base de moltes aplicacions –anàlisi de riscos, optimització de la producció, predicció de beneficis o simulació de sistemes, per citar-ne alguns exemples–, especialment de les tècniques i estratègies vinculades a la ciència de dades aplicada.

Els continguts d'aquest material didàctic es divideixen en cinc reptes:

1) Per què l'àlgebra lineal és important en la ciència de dades? Quins elements bàsics té?

2) Com podem resoldre problemes típics de la ciència de dades mitjançant sistemes d'equacions lineals?

3) Què són els valors i vectors propis (de matrius) i per a què els utilitza Netflix?

4) Com podem afrontar la maledicció de la dimensionalitat en la ciència de dades amb l'anàlisi de components principals i la descomposició en valors singulars?

5) Com podem modelar sistemes dinàmics amb cadenes de Markov com fa l'algoritme PageRank de Google?

Els primers dos reptes estableixen els elements bàsics de l'àlgebra lineal, com ara l'estructura de matriu, el concepte de determinant o els espais vectorials i les seves operacions. També es presenta la potència dels sistemes d'equacions lineals i la seva representació en forma de matriu. Tot i que el concepte de

matriu és senzill, cal destacar que és clau per a l'àlgebra lineal i per a la ciència de dades aplicada.

El tercer repte introdueix conceptes una mica més complexos, com ara les aplicacions lineals, la seva representació matricial i els valors i vectors propis. En tots els reptes anteriors es presenta una contextualització que emmarca quines són les aplicacions d'aquests conceptes i procediments en l'àmbit de la ciència de dades aplicada.

Els dos últims reptes recullen estratègies que, tot i que no són exclusives de l'àmbit esmentat, hi tenen una clara aplicació. D'una banda, en el repte "Com podem afrontar la maledicció de la dimensionalitat en la ciència de dades amb l'anàlisi de components principals i la descomposició en valors singulars?" es detallen dues tècniques –l'anàlisi de components principals (PCA) i la descomposició en valors singulars (SVD)– que, malgrat que estiguin íntimament relacionades, poden tenir aplicacions diferents. En el cas de PCA, permet reduir la dimensionalitat de les dades i descobrir patrons o estructures ocultes. Pel que fa a SVD, una de les aplicacions més destacades és la compressió d'imatges.

Finalment, en el repte "Com podem modelar sistemes dinàmics amb cadenes de Markov com fa l'algoritme PageRank de Google?" s'introdueixen les matrius estocàstiques que permeten representar sistemes dinàmics discrets. Per mitjà de les tècniques recollides en aquest repte, els creadors de Google van generar el seu algoritme per avaluar, per exemple, la importància dels documents amb enllaços mutus, com ara les pàgines web i els diversos enllaços que contenen.

Els reptes han estat pensats perquè els estudiants se centrin a aprendre els conceptes matemàtics i la manera d'aplicar-los a la resolució de problemes de la vida quotidiana. Els problemes, de caràcter general, han estat contextualitzats en l'àmbit de la ciència de dades aplicada. Els dos últims reptes incorporen casos d'estudi i guies de resolució en el llenguatge de programació R. D'aquesta manera, es pretén donar valor a l'aprenentatge de programari matemàtic i estadístic sense perdre de vista la importància de comprendre els conceptes explicats.

Tots els reptes tenen una estructura similar, encara que alguns d'aquests apartats poden no ser-hi presents:

- Coneixements previs aconsellables per a un bon aprofitament de l'aprenentatge del repte i exercicis perquè els estudiants puguin comprovar el seu grau d'adquisició.

- Exemple introductori al tema del mòdul. Amb aquest element es pretén insistir en l'enfocament aplicat d'aquests recursos docents.

- Exposició dels conceptes i de les aplicacions corresponents, com també l'ús de programari matemàtic com a ajuda a l'aprenentatge i nombrosos exemples que els il·lustren. Alguns mòduls inclouen activitats suggerides amb la solució al final.

- Resum amb els conceptes més significatius del repte.

- Exercicis d'autoavaluació de l'aprenentatge dels conceptes fonamentals.

- Solucionari dels exercicis d'autoavaluació.

- Glossari de termes.

- Bibliografia recomanada.

Objectius

Els cinc reptes que hem presentat en aquesta introducció formen part dels recursos docents de l'assignatura d'Àlgebra per al grau de Ciència de Dades Aplicada. Els seus objectius són:

1. Conèixer i ser capaç de manipular elements bàsics de l'àlgebra lineal (espais vectorials, independència lineal, dimensió, matrius i determinants) i de la geometria mètrica (productes escalars, ortonormalitat, angles i distàncies).
2. Comprendre la importància dels sistemes d'equacions lineals per resoldre problemes típics de la ciència de dades.
3. Entendre el concepte de vectors i valors propis, com també la manera de calcular-los i interpretar-los geomètricament.
4. Conèixer l'anàlisi de components principals i ser capaç d'aplicar aquesta estratègia a un cas d'ús utilitzant dades reals o realistes.
5. Saber resoldre un problema mitjançant la descomposició de valors singulars en un cas d'ús utilitzant dades reals o realistes.
6. Ser capaç de resoldre un problema amb models matricials en un cas d'ús utilitzant dades reals o realistes.
7. Agafar destresa en la utilització del llenguatge R per a la resolució de problemes amb un gran volum de dades.