Presentació

En aquesta pràctica resoldreu un cas d'ús proposat mitjançant l'anàlisi de components principals. Aquest cas d'ús us permetrà posar en pràctica els conceptes treballats en aquest repte, entendre i agafar destresa en la seva aplicació a un cas d'ús concret utilitzant dades reals o realistes. Veureu també la necessitat d'utilitzar un llenguatge de programació com, per exemple, `R` per a la seva resolució i agafareu destresa en la seva utilització.

Competències

En aquesta pràctica es treballen les següents competències del Grau en Ciència de Dades Aplicada:

- Que els estudiants hagin demostrat tenir i comprendre coneixements en un àrea d'estudi que parteix de la base de l'educació secundària general, i se sol trobar a un nivell que, si bé es recolza en llibres de text avançats, inclou també alguns aspectes que impliquen coneixements procedents de l'avantguarda del seu camp d'estudi.
- Utilitzar de forma combinada els fonaments matemàtics, estadístics i de programació per desenvolupar solucions a problemes en l'àmbit de la ciència de dades.
- Ús i aplicació de les TIC en l'àmbit acadèmic i professional.

Objectius

Els objectius concrets d'aquesta Pràctica són:

- Comprendre la utilitat dels conceptes d'àlgebra lineal que s'han treballat en els reptes 1-3 en l'aplicació en l'àmbit de la ciència de dades mitjançant l'anàlisi de components principals i la descomposició en valors singulars.
- Ser capaç de resoldre un problema utilitzant la descomposició en valors singulars o l'anàlisi de components principals en un cas d'ús utilitzant dades reals o realistes.
- Entendre la utilitat d'utilitzar un llenguatge de programació pel tractament de grans volums de dades.

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

Descripció de la Pràctica

Ser capaços de reduir la dimensionalitat de dades és molt important en l'àmbit de la ciència de dades on normalment treballem amb alts volums d'informació. En aquest repte veurem dues tècniques molt esteses que ens permetran reduir la dimensionalitat de les nostres dades: la descomposició en valors singulars i l'anàlisi de components principals, que estan molt relacionades. Ambdues tècniques, basades en els conceptes de l'àlgebra lineal analitzats en els reptes 1, 2 i 3, permeten considerar un conjunt de dades inicial i transformar-lo de manera que, o bé la dimensió resultant sigui inferior o bé la nova representació de les dades permeti desvetllar informació rellevant.

Per una banda, us demanem que respongueu un **qüestionari** (el podeu trobar a l'aula Moodle entrant a l'enllaç "Qüestionaris" a la part dreta de l'aula) en el què treballarem la part més instrumental d'aquest repte en una sèrie de preguntes genèriques.

Us demanem també que resolgueu la pràctica descrita en aquest document. Aquests exercicis us plantejaran escenaris propis de la ciència de dades i veureu com els conceptes treballats en aquest repte tenen rellevància en aquests contextos.

Recursos

Recursos Bàsics

- Document introductori a la descomposició en valors singulars per a la ciència de dades.
- Mòdul 4.
- Document de problemes sobre la descomposició en valors singulars enfocats a la ciència de dades.

Recursos Complementaris

- Cas d'ús i guia de resolució en R.

Criteris d'avaluació

- La pràctica s'ha de resoldre de manera individual.
- És necessari justificar tots els passos realitzats a la resolució de la Pràctica.

Tingueu en compte que les dues activitats que es plantegen en aquest repte (la resolució de la pràctica que es planteja en aquest document i el qüestionari) seran part de la nota de pràctiques (Pr=(Pr1+Pr2)/2). La nota d'aquestes activitats correspon a la Pr1 (amb un pes del 20% pel qüestionari i un 80% per a la pràctica). Per a més informació sobre el model d'avaluació de l'assignatura, consulteu el pla docent.

Format i data de lliurament

Cal lliurar un únic document en PDF que incorpori la resolució de la pràctica (memòria tècnica detallada), el codi `R` i les imatges o figures que se us demanen.

Com s'ha dit en l'apartar anterior, s'ha de lliurar tot en un únic fitxer PDF que tingui per nom Pr1Cognom1Cognom2Nom.zip, tot i que després el Registre d'Avaluació Continuada (RAC) canviarà el nom del fitxer per incloure-hi el dia i l'hora del lliurament. Aquest fitxer s'ha de lliurar en l'espai del registre de l'avaluació continuada (RAC) de l'aula abans de les 24:00 hores del dia 28/05/2021 (hora central europea d'estiu (CEST)). No s'acceptaran lliuraments fora de termini.

1 Mesures de qualitat de l'aire de Nova York, Estats Units d'Amèrica

El paquet `R Datasets` d'`R` és un conjunt de *datasets* molt útil i que es troba a totes les implementacions d'`R`. Entre d'altres, podem trobar el conjunt de dades `airquality` que conté mesures diàries de qualitat de l'aire a Nova York, de maig a setembre de 1973.

De forma més precisa, `airquality` és un *data frame* que conté 153 observacions (files) i 6 variables. Les 153 observacions corresponen a lectures diàries de la qualitat de l'aire a Nova York entre el dia 1 de maig de 1973 i el 30 de setembre del mateix any, distribuïts de la següent manera:

- maig de 1973, observacions de l'1 a la 31;
- juny de 1973, observacions de la 32 a la 61;
- juliol de 1973, observacions de la 62 a la 92;
- agost de 1973, observacions de la 93 a la 123;
- setembre de 1973, observacions de la 124 a la 153.

Les 6 variables són:

- `Ozone`: ozó mitjà en parts per mil milions de 13:00 a 15:00 hores a l'illa Roosevelt.
- `Solar.R`: radiació solar mesurat en Langleys a la banda de freqüències 4000–7700 Angstroms de 08:00 a 12:00 hores a Central Park.
- `Wind`: velocitat mitjana del vent en milles per hora a les 7:00 i a les 10:00 hores a l'aeroport de LaGuardia.
- `Temp` temperatura màxima diària en graus Fahrenheit a l'aeroport de LaGuardia.
- `Month`: mes de l'any.
- `Day`: dia del mes.

Les dades s'han obtingut del Departament de Conservació de l'Estat de Nova York (dades d'ozó) i del National Weather Service (dades meteorològiques).

Cal dir que la variable `Month` la farem servir com a etiqueta. Tindrem, per tant, cinc classes, que correspondran als mesos de maig a setembre. Descartarem la variable `Day` perquè no correspon a cap mesura física.

L'objectiu, doncs, d'aquesta pràctica és treballar els conceptes del Repte 4. Veureu com l'anàlisi de components principals ens permet, si és possible:

(1) reduir la dimensionalitat de les nostres dades; i

(2) millorar la representació de la informació. Serem capaços de *resumir* en un diagrama de dispersió de dues dimensions tota la informació que es tenia, originalment, en quatre magnituds físiques.

Anem a veure quins serien els passos que caldria fer per analitzar aquestes dades com un científic o científica de dades!

Per començar, no cal importar les dades. Només cal assignar el *data frame* `airquality` a la variable `nycair`:

```
> nycair <- airquality
```

1. [5%] Genereu un vector `label` que contingui la cinquena columna (`Month`) del *data frame* `airquality`. De la mateixa manera, genereu una matriu `W` de 153 files i 4 columnes del *data frame* `airquality`. Per a generar la matriu feu servir, per exemple, `as.matrix`. Assegureu-vos que tant `label` com `W` són matrius i no *data frames* fent servir la instrucció:

```
> class(label)
[1] "integer"
> class(W)
[1] "matrix" "array"
```

2. [5%] Una part important del tractament de les dades és el seu preprocessament. El primer pas és comprovar que no hi ha cap observació sense valor en alguna o algunes variables. Una manera de fer-ho és repassar visualment el conjunt de dades. Ara bé, donat que tenim $153 \times 6 = 612$ valors, val la pena fer-ho de forma automàtica. A `R`, la instrucció `is.na(W)` retornarà una matriu plena de `TRUE` (equivalent a un 1) o `FALSE` (equivalent a un zero) en funció de si falta alguna dada o no, respectivament. Finalment, si calculem la suma de tots els elements d'aquesta matriu, amb la instrucció `sum` obtindrem un 0 si no hi ha cap valor que falti o un altre nombre natural en funció de les dades que faltin. Què obteniu? Quantes dades falten?

3. [5%] Com haureu pogut veure en la pregunta anterior, falten una quantitat important de dades. El paquet `mice` proporciona una funció interessant, `md.pattern` que ens permet veure, amb un cop d'ull, el patró de dades que falten. Si carregueu aquesta llibreria i executeu les instruccions:

```
> library(mice)
> md.pattern(W)
```

obtindreu una matriu similar a aquesta (atenció! no obtindreu la mateixa matriu):

```
    Wind Temp Solar.R Ozone
98     1    1       1     1  0
45     1    1       1     0  1
6      1    1       0     1  1
4      1    1       0     0  2
       0    0       10   49 59
```

Aquesta matriu s'ha d'interpretar de la següent manera:

- hi ha 98 observacions on no hi falta cap dada;
- hi ha 45 observacions on hi falta la dada de la variable `Ozone`;
- hi ha 6 observacions on hi falta la dada de la variable `Solar.R`; i
- hi ha 4 observacions on hi falten les dades de les variables `Solar.R` i `Ozone`.

En total, faltarien 59 dades. Per a les dades d'aquesta pràctica, quina matriu obteniu?

4. [5%] Abans de poder continuar amb la nostra anàlisi, cal procedir a completar les dades no disponibles. Aquest procés rep el nom de *data imputation* (imputació de dades). Existeix una gran diversitat de maneres diferents d'imputar les dades que falten, com ara la mitjana aritmètica o la mediana, entre d'altres. En aquest cas, farem servir el mètode anomenat *predictive mean matching*[1]. Però no us preocupeu, ja que no haureu de programar aquest mètode! La funció `mice` s'encarregarà de tot, si escriviu:

```
> nycairi <- mice(data = W,m = 5, method= "pmm", maxit = 50, seed = 500)
> W <- complete(nycairi,1)
```

Noteu que hem redefinit la matriu `W`, que ara ja no conté dades que falten.

5. [5%] Representeu en un diagrama de dispersió, a mode d'exemple, la primera variable (`Ozone`) respecte de la segona (`Solar.R`). Representeu també la segona variable (`Solar.R`) respecte de la tercera (`Wind`). Feu servir la instrucció `plot` amb les opcions `type = "p"` (punts) i `col = label`. L'opció `col = label` us permetrà veure, de color diferent, les observacions dels diferents mesos. Què observeu? Les classes se superposen o es poden separar clarament?

[1] `https://stefvanbuuren.name/fimd/sec-pmm.html`

6. [5%] Escaleu les dades de la matriu `W` tal com s'explica a la Secció 2.1 del mòdul. Anomeneu a la matriu resultant `Ws`. Comproveu que la mitjana aritmètica de les quatre columnes de la matriu `Ws` és zero (de fet, obtindreu un valor molt petit, proper al zero de màquina) i que la desviació tipus de les quatre columnes és 1. Podeu fer servir les instruccions `mean` i `sd`. Quina d'aquestes dues instruccions us ajudarà a resoldre aquesta pregunta:

   – Aquesta? `apply(Ws,1,mean)`

   – O aquesta? `apply(Ws,2,mean)`

   Justifiqueu la resposta.

7. [10%] Calculeu la matriu de covariàncies de `Ws` tal i com s'explica a la Secció 2.2 del mòdul. Anomeneu `CWs` a la matriu de covariàncies. Quin parell de variables (diferents) presenten la covariància més gran, en valor absolut? Representeu en un diagrama de dispersió aquestes dues variables, amb les mateixes opcions gràfiques que en les preguntes anteriors.

8. [10%] Quin parell de variables (diferents) presenten la covariància més petita, en valor absolut? Representeu en un diagrama de dispersió aquestes dues variables, amb les mateixes opcions gràfiques que en les preguntes anteriors.

9. [10%] Calculeu, sense fer servir la comanda `prcomp`, els valors i els vectors propis de la matriu de covariàncies `CWs`. Anomeneu `P` a la matriu que conté, per columnes, els vectors propis. Aneu al qüestionari associat a la pràctica i mireu quin és el valor de `N` que us ha estat assignat. Quina és la variabilitat retinguda per les primeres `N` components principals?

10. [10%] La primera component principal, `PC1`, és un vector format per la combinació lineal de les quatre variables originals:

$$\texttt{PC1} = \alpha_1 \times \texttt{Ozone} + \alpha_2 \times \texttt{Solar.R} + \alpha_3 \times \texttt{Wind} + \alpha_4 \times \texttt{Temp}$$

    Els coeficients α_i, $i = 1, 2, 3, 4$, en valor absolut, són també una mesura de la importància de cada variable original amb l'objectiu de la classificació de les dades originals. Des d'aquest punt de vista, quines són les dues variables més importants en la primera component principal? I en la segona component principal? Justifiqueu la vostra resposta.

11. [10%] Representeu gràficament els 153 punts formats per la projecció de les dades originals normalitzades sobre la primera i la segona components principals. Com abans, feu servir la instrucció `plot` amb les opcions `type = "p"` (punts) i `col = label`. Què podeu dir ara mateix sobre les classes (mesos de l'any)? Se superposen o es poden separar clarament? És possible que, tot i haver fet l'anàlisi de components principals, encara no se separin els punts. Aquest fet està explicat a l'article:

Pozo, Francesc; Vidal, Yolanda. 2016. "Wind Turbine Fault Detection Through Principal Component Analysis and Statistical Hypothesis Testing". Energies 9, no. 1: 3. https://doi.org/10.3390/en9010003

on es demostra que PCA no és suficient per separar i cal fer un pas més que, en aquest cas, és l'ús del contrast d'hipòtesi.

12. [10%] Representeu gràficament **només** els 31 punts corresponents al mes de maig i els 31 punts corresponents al mes de juliol formats per la projecció de les dades originals normalitzades sobre la primera i la segona components principals. Com abans, feu servir la instrucció `plot` amb les opcions `type = "p"` (punts) i `col = label[c(1:31,62:92)]`. Què podeu dir ara? Som capaços de distingir el mes de maig i el mes de juliol?

13. [10%] Per facilitar la correcció i la localització dels possibles errors de la programació del vostre codi en `R`, completeu la taula que trobareu al **qüestionari** *REPTE 4 - Taula resum de la Pràctica 1* en relació a les variables creades durant la resolució d'aquesta pràctica (per al vostre valor de `N` del mateix qüestionari):

	valor
Dades que falten (pregunta 2)	
Dades que falten de la variable `Ozone` (pregunta 3)	
`W[5,2]` (pregunta 4)	
Quin és el segon argument de la instrucció `apply(Ws,?,mean)`? (pregunta 6)	
`Ws[3,2]` (pregunta 6)	
`sum(CWs)` (pregunta 7)	
Variables amb la màxima covariància (pregunta 7)	
Variables amb la mínima covariància (pregunta 8)	
Variabilitat retinguda per les primeres `N` components principals (pregunta 9)	
Variables més importants en la primera component (pregunta 10)	
Variables més importants en la segona component (pregunta 10)	