Anàlisi Multivariant de Dades
Grau Interuniversitari d'Estadística
Curs 2018-2019

Professorat: Belchin Kostov, Bea Sevilla, Karina Gibert Oliveras

Estructura de les dades

- Es recomana buscar matrius de dades NO AGREGADES
- Un mínim de 7 variables numèriques
- Un mínim de 7 variables qualitatives (d'entre elles un màxim de 2 variables binàries)
- Grandària de la Base de Dades: a menys que l'estructura del domini limiti a menys, entre 2000 i 5000 observacions.
- Evitar BD massa grans que requereixin temps de processos massa llargs perquè perdreu molt temps a processar-les i no aprendreu més.
- Feu un cop d'ull a les pràctiques de cursos anteriors i tracteu de no repetir temes
- Recordeu que queden excloses BD provinents de pàgines acadèmiques o lligades a softwares

Estructura del material a entregar

1. Portada (pag1) (data d'entrega D1)

1. Títol
2. Nom complert i el correu electrònic dels components del grup ordenats alfabèticament
3. Data d'entrega

2. Definició del projecte i assignació (data d'entrega D1)

La primera pàgina de la memòria ha de contenir:

1. Font d'on s'han obtingut les dades incloent la pàgina web d'origen (o totes les implicades si n'hi ha més d'una)
2. Un paràgraf que expliqui a què fan referència les dades
3. Estructura bàsica de la matriu de dades (quantes files i columnes té, quantes variables són numèriques i quantes categòriques, nombre de caselles missing i % que representa respecte del total de la matriu de dades, % de missing per variable (taula i/o histograma))

La següent pàgina ha de contenir l'índex de la memòria

3. Pla de treball (data d'entrega D2)

1. Descomposició de la pràctica en tasques i seqüenciació temporal en diagrama de Gantt (veure documentació sobre treball en equip al Campus Virtual)
2. Parrilla de distribució de les tasques entre els components del grup
3. Pla de riscos (que penseu que pot fer fallar el planning? Com ho podríeu combatre?)

4. Estructura de les dades i descriptiva (data d'entrega D3)

1. Un paràgraf amb la motivació del treball
2. Descripció formal de l'estructura de les dades
   - **i.** Determineu clarament les files de la matriu què contenen, i així, l'àmbit de l'estudi, explicitar seleccions de files i columnes si és el cas, indicar filtres utilitzats, criteris, etc
   - **ii.** Llistat de variables continguts a la matriu de dades amb la corresponent meta informació (assegureu-vos que no us deixeu res)
3. Anàlisi descriptiva univariant inicial de totes les variables (inclou resum numèric i gràfic i comentaris corresponents)
   - **i.** Per les variables numèriques: summary, histograma i/o boxplot (eventualment altres representacions adients)
   - **ii.** Per les variables categòriques: table, barplot i/o pie chart (eventualment altres representacions adients)
4. Descripció detallada del procés de preprocessament de dades seguit i justificació de totes les decisions preses (inclou tractament de missings, etc). Si hi ha poques variables modificades, es pot afegir la distribució de la variable depurada al costat de la de la variable original i fer comentari únic. En aquest cas, el punt 5 desapareixeria de la memòria.
5. Anàlisi descriptiva univariant de les dades preprocessades i discussió sobre l'aleatorietat de les dades mancants si n'hi ha.

5. Clustering jeràrquic (data d'entrega D3)

1. Descriure bé quin mètode d'agregació i quina distància es fa servir (discutir amb els professors però per defecte seria Mètode de Ward i distància de Gower al quadrat)
2. Mostrar l'arbre jeràrquic resultant i seleccionar el número de classes. Validar el tall de l'arbre amb els professors.
3. Profiling dels clusters: Utilitzar la variable de cluster per fer estadístiques descriptives per grups. Eventualment visualitzar la relació entre parells de variables contínues i les classes. Determinar les particularitats de cada classe i elaborar una descripció prototípica de cada classe. Si les dades són georeferenciades presentar visualitzacions en mapes de les classes.

6. ACP de les variables numèriques (data d'entrega D4)

1. Resultats de l'anàlisi, screeplot i selecció de components factorials retingudes.
2. Per cada mapa factorial: projecció dels individus; projecció comú de variables numèriques i modalitats de les variables qualitatives. Eventualment projecció de noves variables creades específicament per l'estudi en qüestió.
3. Interpretació de les relacions i oposicions entre variables i interpretació dels eixos.

7. ACM de les variables qualitatives: (data d'entrega D4 fins al final del document)

1. Resultats de l'anàlisi, screeplot i selecció de components factorials a retenir.

2. Per al primer mapa factorial: projecció de modalitats de les variables qualitatives (eventualment amb les il·lustratives qualitatives que es vulguin afegir).
3. Interpretació de relacions i oposicions entre variables. Anàlisi sobre les coincidències i discrepàncies amb l'ACP.

8. Clustering jeràrquic sobre les components factorials d'ACP (i eventualment les d'ACM)

1. Descriure bé quina matriu de dades serveix d'input al procés de clustering i quin mètode d'agregació i quina distància es fa servir. Mostrar l'arbre jeràrquic resultant i seleccionar el número de classes.
2. Profiling dels clusters: Repetir el mateix procés d'interpretació de les classes per aquesta segona classificació i comparar les dues classificacions. Representar la visualització de les classes sobre els primers plans factorials de l'ACP i de l'ACM.

9. AFM o Anàlisi Discriminant o Anàlisi Textual

1. Aplicar a les dades una de les tres metodologies en funció de la seva adequació amb els objectius. Justificar el motiu de la selecció i interpretar els resultats principals obtinguts.

10. Anàlisi comparativa: Analitzar coincidències i divergències entre l'ACP, l'ACM i el clustering

11. Conclusions generals

12. Pla de treball REAL

1. Ha de tenir la mateixa estructura que l'original però adaptat al decurs real del projecte.
2. Discussió crítica sobre les desviacions de la planificació final respecte de l'original.
3. Per facilitar la tasca s'admetrà que a l'última entrega aquesta secció repeteixi els continguts de la secció 3 també per fer la comparació més àgil.

13. Scripts d'R utilitzats per elaborar les anàlisis (opcionalment es poden incrustar les parts rellevants dels scripts al llarg dels capítols anteriors).

Terminis d'entrega i presentacions

D1: Submissió de la pràctica a aprovació dels professors. Entregar punts 1 i 2.

D2: Entregar punt 3 en format pdf. També cal entregar el fitxer amb dades originals

D3 (Pre-entrega): Memòria amb els punts 1 a 5 i presentació oral (15 minuts). Noteu que es demana que aquesta memòria incorpori des del punt 1. La memòria s'ha de presentar en el format pdf (llegir instruccions sobre les memòries) i la presentació en format ppt (llegir les instruccions sobre les presentacions). També cal entregar el fitxer amb dades processades i el script de R.

D4 (Presentació final): Document escrit total (punts 1 a 13) i ppt amb la presentació (15 minuts). La memòria s'ha de presentar en el format pdf (llegir instruccions sobre les memòries) i la

presentació en format ppt (llegir les instruccions sobre les presentacions). També cal entregar el fitxer amb dades processades i el script de R.

Nota important: Les entregues s'hauran de fer via Campus Virtual utilitzant els espais generats per penjar-les. La data límit per fer cada entrega serà la que apareix en el document d'introducció a laboratoris. En el cas de les dues presentacions, les dues entregues s'hauran de fer almenys tres dies abans de les presentacions.

Instruccions sobre les memòries

Les memòries s'han d'entregar en format pdf. Es valorarà positivament la capacitat de síntesi dels alumnes. En aquest sentit, el nombre màxim de full permès per la memòria de la pre-entrega amb tamany de lletra 11 i espai 1,5 serà 20 pàgines. De la mateixa manera, la memòria final entregada haurà de tenir com a màxim 40 pàgines. La superació dels límits indicats seran penalitzades.

Instruccions sobre les presentacions

Els membres del grup es distribuiran els 15 minuts designats per presentar el treball realitzat a tota la classe i discutir amb els professors sobre el mateix. Es valorarà la capacitat de comunicació, l'aspecte de la presentació, la capacitat de síntesi expressada a les transparències, l'ajust al temps previst de presentació, la qualitat del treball, la correctesa tècnica i metodològica, la capacitat de defensa mostrada al torn de preguntes i els coneixements demostrats durant la discussió, així com la bona sincronització i planificació del grup de treball. Les deficiències en el procediment d'entrega seran penalitzades.

Presentació en ppt de la pre-entrega (D3) ha d'incloure (màxim 12 diapositives)

1) Una transparència/portada amb els noms de tots els components del grup, el curs acadèmic i el títol de la pràctica
2) Una transparència amb el tema de la pràctica i les url d'on provenen les dades
3) Índex
4) Una transparència amb l'estructura de la BD tot descrivint les variables d'anàlisi
5) Una transparència per la descriptiva univariant
6) Una transparència pel procés de preprocessing (eventualment afegir una transparència per cada aspecte concret a comentar)
7) Una transparència amb la descripció del procés de clustering realitzat i l'arbre resultant
8) Una transparència amb els perfils de les classes finalment descrits, incloent la grandària de la classe, el percentatge que representa respecte de la grandària total de la mostra i el títol assignat a la classe

Presentació final (D4) haurà d'incloure (màxim 20 diapositives)

9) Una transparència/portada amb els noms de tots els components del grup, el curs acadèmic i el títol de la pràctica
10) Un parell de transparències per resumir els resultats més rellevants dels punts "1-8" de la pre-entrega

11) Una transparència amb el primer pla factorial (eventualment afegir-ne més pels següents plans si és rellevant)
12) Una transparència per les conclusions de l'ACP
13) Una transparència per al primer pla factorial de l'ACM
14) Una transparència per les conclusions de l'ACM
15) Una transparència pels resultats principals d'AFM, Anàlisi Discriminant o Anàlisi Textual segons l'anàlisi escollit (eventualment afegir-ne alguna més per detallar els resultats rellevants)
16) Una transparència de conclusions i l'anàlisi de coincidències i divergències entre ACP, ACM i clustering (opcionalment afegir l'anàlisi addicional escollit)
17) Una transparència amb la planificació original i final del treball