Grau en Estadística

(Facultat de Matemàtiques i Estadística)

Anàlisi Multivariant

Examen final

11 juny 2014

Contesteu cada pregunta a l'espai reservat

Nom Estudiant:

PREGUNTES BREUS. Seleccioneu l'opció correcta posant-hi una marca

1. La projecció d'una variable categòrica sobre un pla factorial és
    1. un centroide
    2. un vector
    3. una regió de l'espai
    4. un conjunt de punts, un per modalitat

2. La projecció d'una variable numèrica sobre un pla factorial és
    1. un punt
    2. un vector
    3. una regió de l'espai
    4. un centroide

3. Relaciona correctament els mètodes d'anàlisi multivariant en funció de l'estructura de la matriu de dades que cal analitzar

Tabla de contingencia	Análisis en componentes principales
Matriz de variables numéricas	Análisis de correspondencias múltiples
Matriz de variables categóricas	Análisis de correspondencias simples

4. Quina/es de les següents mètriques es pot utilitzar en un procés de clustering jeràrquic amb una matriu de dades que conté simultàniament variables numèriques i categòriques (pregunta de resposta múltiple)?
    1. Generalització de la mètrica de Minkowsly proposada per Ichino-Yaguchi
    2. Distància Euclídea normalitzada per la desviació típica
    3. Mètrica mixta de Gibert
    4. Coeficient de dissimilitud de Gower al quadrat
    5. Distància de $\chi 2$
    6. Coeficient de similitud de Gower
    7. Distància del valor absolut
    8. Coeficient de dissimilitud de Gower
    9. Distància de Hamming

5. Quins dels següents mètodes de clustering no requereixen indicar el número de classes que es volen trobar com a paràmetres d'entrada (resposta múltiple)
    1. Mètode de les distàncies mínimes (Single linkage)
    2. Mètode del centroide
    3. K-means
    4. Mètode de Ward
    5. DBSCAN
    6. Núvols dinàmics
    7. Mètode de classificació conceptual de Michalski
    8. Fuzzy C-Means

PREGUNTA 1 Expliqueu l'estructura del triplet $(X_{nxp}, M_{pxp}, D_{nxn})$ propi d'una Anàlisi Factorial Descriptiva

PREGUNTA 2: Per als dos escenaris que es plantegen tot seguit, expliqueu amb precissió

a) com s'identifica si una variable té una contribució alta sobre el primer eix factorial
b) com es coneix si la contribució és directa o inversa en els següents casos:
    1. Només es disposa de la imatge del pla factorial
        a)

        b)

    2. Només es disposa de la matriu de components dels vectors directors dels eixos factorials
        a)

        b)

PREGUNTA 3: Suposeu unes dades on sabem que hi ha 3 classes. Executar un k-means amb k=3 dóna garantia de trobar les 3 classes existents? Per què?

PREGUNTA 4: Es disposa d'informació sobre les preferències d'un grup de 4500 clients sobre la compra d'un producte de 4 marques diferents: A, B, C, D. La companyia que comercialitza una d'aquestes marques realitza una campanya de marqueting televisiu i recull les preferències dels mateixos clients després de la campanya.

Entre d'altra informació es disposa de l'edat del client (edat), el número d'anys que ha estudiat (estd), els ingressos anuals (ingr) i amb quantes persones viu (memb).

a) Indiqueu quin tipus d'anàlisi estadístic realitza la següent instrucció d'R

```
dcon <- data.frame (edat, memb, estd, ingr)
pc1 = prcomp(dcon, scale=T)
```

L'execució de l'anterior instrucció genera el següent resultat

```
> print(pc1)
Standard deviations:
[1] 1.1669555 1.1011225 0.9111627 0.7717037

Rotation:
            PC1        PC2        PC3        PC4
edat -0.3058089 -0.6134984  0.6547389 -0.3184612
memb -0.3860449  0.6786170  0.1620308 -0.6034851
estd -0.5068837 -0.3706227 -0.7235122 -0.2867714
ingr -0.7074738  0.1604289  0.1469459  0.6724212
```

b) Quin percentatge de la inèrcia projectada es conserva en el primer eix?

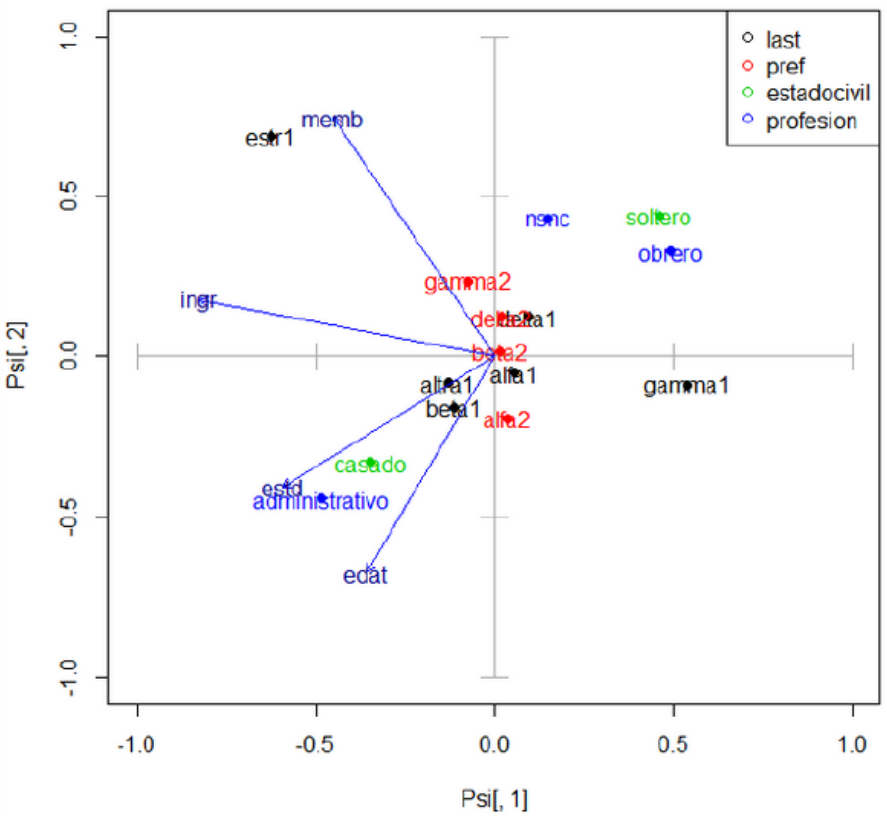
c) Per què hi ha quatre columnes de números a la matriu

d) Què significa i com s'obté el -0.3058089 de la primera cel.la de la matriu

e) A partir d'aquests resultats quina valoració feu de l'anàlisi?

f) Com es trobarien les coordenades dels punts que representen a cada client a partir d'aquesta informació i la matriu de dades inicial (X)

També es disposa d'informació sobre l'estat civil de la persona i la professió (obrero, administrativo, nsnc). Per la següent representació gràfica, expliqueu:



g) Què representa cada eix

h) Què representa la fletxa Edat, i quina relació hi ha entre les seves components i els resultats de la matriu anterior

e) Que representa el punt etiquetat com “obrero” de la gràfica i per quin procediment s’obtindrien les seves components.

f) Analitzeu la gràfica i comenteu què hi veieu

g) Quin impacte ha tingut l’anunci publicitari sobre els clients, tenint en compte que les variables preferred (alfa 1, beta1, gamma1, delta1) indiquen quina marca preferia la persona abans de la campanya i last (alfa 2, beta2, gamma2, delta2) quina compra després de l’anunci

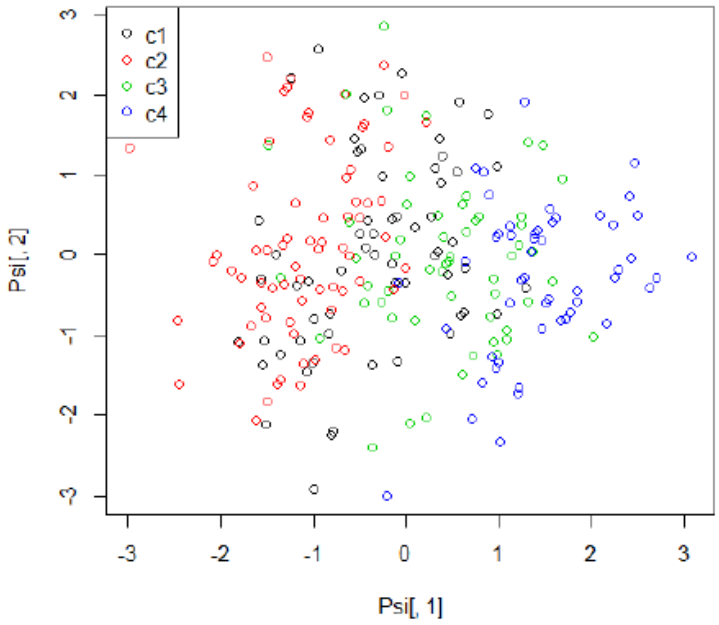
PREGUNTA 5. Per les mateixes dades del problema anterior, s'executa la següent instrucció d'R

```
h <- hclust(d)
```

a) Què és d?
b) Què fa aquest procediment d'R
c) Quin resultat genera?

d) Quina instrucció R caldria executar per a obtenir d'aquí una columna amb una classe assignada per a cada client?

Suposant que es disposa d'aquesta columna, s'ha obtingut la següent representació gràfica



Expliqueu d'on surt i què representa i valoreu el resultat de l'anàlisi.