Estadística
Clase Práctica 9
Ejercicio 1

Carlos Bermudez Porto C.BERMUDEZ@ESTUDIANTES.MATCOM.UH.CU
Grupo C412

Leynier Gutiérrez González L.GUTIERREZ@ESTUDIANTES.MATCOM.UH.CU
Grupo C412

Tony Raúl Blanco Fernández T.BLANCO@ESTUDIANTES.MATCOM.UH.CU
Grupo C411

Tutor(es):
Lic. Dalia Diaz Sistachs, *Departamento de Matemática Aplicada*

Palabras Clave: k-means, acp, clusters, d-tree
Tema: Estadística, Análisis de Componentes Principales, Técnicas de Clasificación

1. Introducción

El siguiente trabajo corresponde a la realización del ejercicio 1 de la clase práctica 9 de los autores como evaluación de la asignatura de Estadística de la carrera de Ciencia de la Computación de la Facultad de Matemática y Computación de la Universidad de La Habana.

Los datos utilizados para el trabajo son sacados de un estudio en adolecentes con desórdenes alimenticios conocidos con el objetivo de encontrar una sintomatología estandar para comportamientos de anorexia y bulimia. En cada observación las pacientes fueron valoradas en síntomas diferentes.

Se decidió eliminar la columna *number* pues en si esta es equivalente a un identificador del paciente y no aporta informacion real al problema.

En las pruebas de hipótesis se asumió un nivel de significación del 5%, por tanto, fue posible aceptar las hipótesis nulas con una probabilidad de error menor o igual a 0.05.

2. Análisis de Correlación

Se analizó la correlación entre las diferentes variables utilizando la función *symnum* (Fig. 1)).

Como se pudo observar casi todas las variables estan poco correlacionadas con el resto, debido a la abundancia de '.' y ' '. Existe solo un caso de variables muy correlacionadas, la variable *tidi* y *time*. De cualquier forma la reduccion de dimensión es factible.

3. Análisis de Componentes Principales

Se analizaron todas las variables pues no serán utilizadas para una regresión. (Fig. 2)

```
> symnum(cor_data)
       w mn fs bn v prg h fm e fr sc st sb md pre bd tm d td
weight 1
mens   . 1
fast   . .  1
binge          1
vomit          .  1
purge          ,  . 1
hyper  .    .       .   1
fami   . .  .             1
eman   . .                . .  1
frie                      .    1
school . .                .    . .  1
satt   . .                . .  ,  1
sbeh   . .  .     .       . .  .    .  .  1
mood           .                                 1
preo   . .  .        .      .  .           .  .  1
body   . .  .     .          .              .  .  .   1
time   . .  .              . .  .           .     .      1
diag                                                     1
tidi   . .  .              . .  .           .     .      B    1
attr(,"legend")
[1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```

Figure 1: Resultado de la función *symnum*

Despues de analizados los valores propios se pudo notar que usando el criterio del porcentaje o el criterio de Kaiser (para valores propios mayores que 1) era posible usar entre 5 y 6 componentes principales. Por simplicidad se utilizaron 5.

3.1 Descripción de las Componentes

- **PC1:** Para la primera componente encontramos que exite una precencia de las pacientes con alto peso, (aparece menstruacion pero no se explicarla), que ademas tienen una alta restriccion de comida, que ademas purgan(viene de la traduccion de purge en este contxto qu no domino) despues de ingerir alimentos, con hiperactividad, buenas relaciones con familiares y amigos, alta emancipacion de sus familias, (desconozco school para que sera), buena actitud y comportamiento sexual, con

```
> summary(acp)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9     PC10
Standard deviation     2.4629 1.5181 1.34396 1.14045 1.0480 0.95115 0.90961 0.82943 0.76638 0.73641
Proportion of Variance 0.3193 0.1213 0.09506 0.06845 0.0578 0.04761 0.04355 0.03621 0.03091 0.02854
Cumulative Proportion  0.3193 0.4406 0.53562 0.60408 0.6619 0.70949 0.75304 0.78925 0.82016 0.84870
                          PC11    PC12    PC13    PC14    PC15    PC16    PC17    PC18     PC19
Standard deviation     0.70641 0.68386 0.67191 0.62754 0.59065 0.53467 0.49823 0.42393 1.135e-15
Proportion of Variance 0.02626 0.02461 0.02276 0.02073 0.01836 0.01505 0.01307 0.00946 0.000e+00
Cumulative Proportion  0.87497 0.89958 0.92334 0.94407 0.96243 0.97748 0.99054 1.00000 1.000e+00
```

Figure 2: Resultado del análisis de las componentes principales

```
> acp$rotation[,1:5]
              PC1          PC2           PC3          PC4          PC5
weight 0.26018299  0.13857742  0.0300116280 -0.03962329  0.385779306
mens   0.27897801  0.17871707 -0.0461520138 -0.16651359  0.031586103
fast   0.28352595  0.07972769 -0.1533155604 -0.17391704  0.251762456
binge  0.10984013 -0.43995597 -0.1526746411  0.01352275 -0.350485591
vomit  0.14127246 -0.38984532  0.0014588249 -0.17008771 -0.006236922
purge  0.15932667 -0.44776612 -0.2083185949  0.13560922 -0.206276044
hyper  0.22657774 -0.09925471 -0.0067517113  0.37233000  0.036160583
fami   0.23290131  0.24066264  0.0679077297 -0.09992284 -0.320219937
eman   0.26454043  0.10433176  0.1860452164  0.21156779 -0.257862925
frie   0.17475192 -0.08301149  0.2766397139 -0.28692094 -0.351848308
school 0.22877371 -0.02439793  0.4683689482  0.21998375  0.097931248
satt   0.21671942 -0.04871510  0.5243756162  0.05645200  0.081793415
sbeh   0.27887530 -0.10644358  0.1619370570  0.04284165  0.092882734
mood   0.18263221 -0.15061145 -0.1344293404 -0.21549092  0.422801990
preo   0.29475877 -0.04762011 -0.2185621069 -0.26505520  0.012361748
body   0.24184929 -0.11872011 -0.0611822812 -0.33275291 -0.003438117
time   0.27756672  0.17834287 -0.3213731571  0.36379533 -0.049032165
diag   0.03096976  0.39306377 -0.0001241641 -0.36471817 -0.324594542
tidi   0.27708140  0.26696242 -0.3122954021  0.26656744 -0.124998879
```

Figure 3: Descripción de las Componentes

buen estado de animo, muy preocupadas por su alimentacion y peso corporal, con mucha percepcion de su cuerpo, (por ultimo time y tidi tambien pero no logro interpretar a que se refieren).

- **PC2:** La segunda componente esta conformada por un grupo de pacientes con caracteristicas como pocos atracones, poca precensia de vomitos, que purgan(ver anterior referencia) poco las comidas, con buenas relaciones familiares, pero que fueron diagnosticadas con la enfermedad(asumimos q alto es q sea mas grave), (tambien esta tidi pero no tengo interpretacion).

- **PC3:** La tercera componente presenta pacientes con caracteristicas como buenas relaciones con amigos, (alto school pero sin interpretacion), buena actitud sexual, (poco time y poco tidi).

- **PC4:** La cuarta componente muestra pacientes con alta hiperactividad, con emancipacion de sus familias, sin buenas realciones con sus amigos, (school alto sin interpretacion), desanimadas, con poco preocupacion por su peso y alimentacion, poca percepcion de su cuerpo, (altos time y tidi sin interpretacion), y con poca precensia de la enfermedad.

- **PC5:** La quinta componenete se caracterisa por tener presencia de pacientes con alto peso corporal, pocos atracones, malas relaciones familiares, malas relaciones con amistades, buen estado de animo y sin presencia de la enfermedad.

Ahora veamos como se agrupan estos datos usando cluster jerarquicos, kmeans y arbol de decisión.

4. Clusters

- Con 4 clusters se puede observar (Fib. 4) que se forman grandes grupos pero que aun pudiera seguir particionandoce.

- Lo haremos para 5 clusters (Fib. 5). Esto a partir de la observacion del dendrograma con 4 clusters. Además intuitivamente por la cantidad de componentes principales que tomamos.

- Con 6 clusters (Fib. 6) no se nota un gran cambio con respecto a los grupos formados para 5 clusters por eso trabajaremos con 5 como principal cantidad de conjuntos.

Lo que si es cierto es que existen una gran cantidad de subtipos y por ende muchos grupos pequeños.

5. K-means

- Comprobando para 4 (Fib. 7) la similitude entre elementos del mismo grupo es del 36%, la cual es bien baja. Las cantidades en los grupos son de 67, 19, 88 y 43.

- De igual forma usaremos 5 (Fib. 8) grupos como medidor principal. La similitud es del 41%, la cual igualmente no es tan buena. Las cantidades son 28, 85, 47, 19 y 38.

- Para 6 (Fib. 9) esta similitud alcanza el 44.3%. La diferencia no es substancial con respecto a 5. Las cantidades son 37, 71, 27, 18, 36 y 28. De esas cantidades si se puede observar que al menos los elementos estan mejor distribuidos que con las cantidades de clusteres anteriores.

6. Árbol de Decisión

Analisaremos el arbol de decision (Fig. 10) formado para la variable diag(Diagnostico) utilizando un metodo CART de clasificacion. En este caso alcanzamos un 40% de error aproximadamente. El arbol es posible de construir pero no tienve mucho valor predictivo debido al alto error en un contexto en el cual este no es permisible.

7. Referencias

- Conferencia 5 - Corelacion
- Conferencia 6 - Regresion Lineal Simple
- Conferencia 7 - Regresion Lineal Multiple
- Conferencia 8 - ANOVA
- Conferencia 10 - Cluster, Arboles de Decisión

8. Anexos

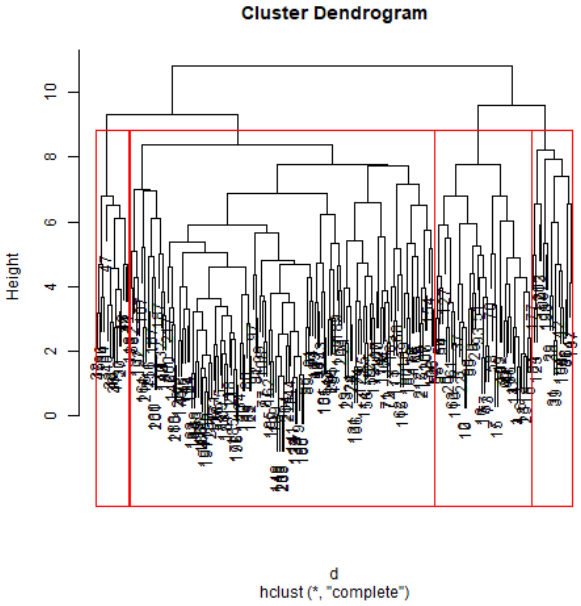

Figure 4: Análisis con 4 clusters

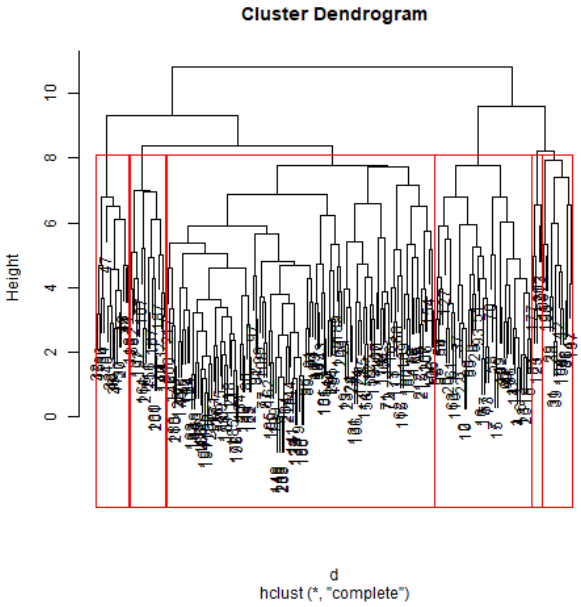

Figure 6: Análisis con 6 clusters

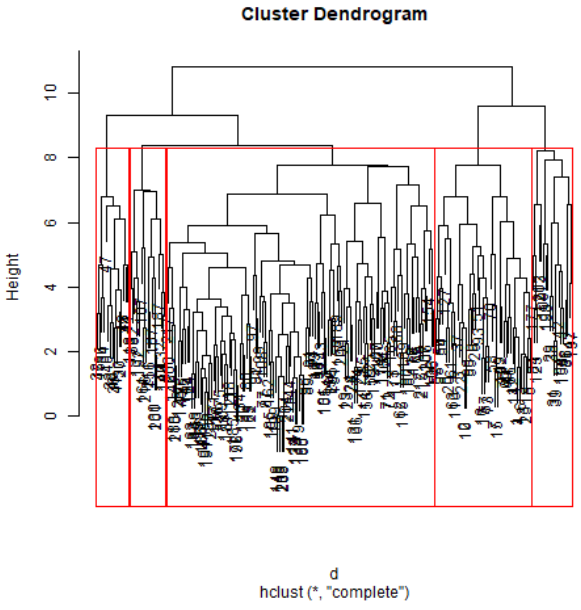

Figure 5: Análisis con 5 clusters

Figure 7: Análisis de kmeans con 4

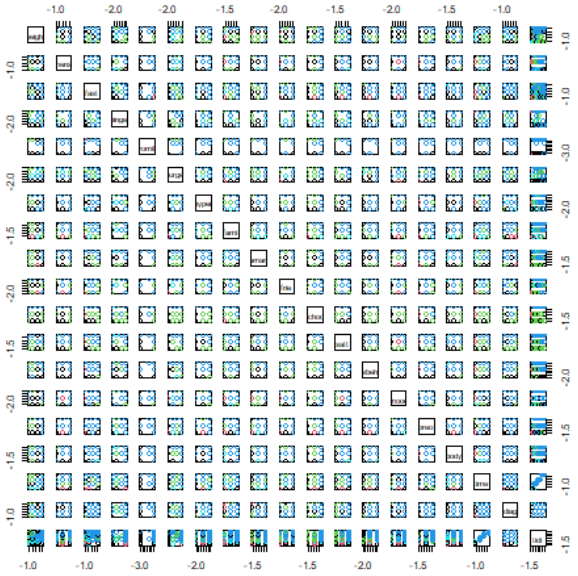

Figure 8: Análisis de kmeans con 5

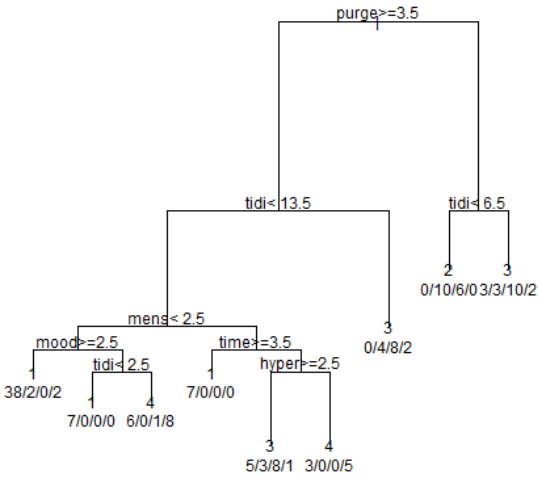

Figure 10: Árbol de Decisión

Figure 9: Análisis de kmeans con 6