

# Estadística

## Clase Práctica 9

### Ejercicio 1

**Carlos Bermudez Porto**

*Grupo C412*

**Leynier Gutiérrez González**

*Grupo C412*

**Tony Raúl Blanco Fernández**

*Grupo C411*

C.BERMUDEZ@ESTUDIANTES.MATCOM.UH.CU

L.GUTIERREZ@ESTUDIANTES.MATCOM.UH.CU

T.BLANCO@ESTUDIANTES.MATCOM.UH.CU

**Tutor(es):**

Lic. Dalia Diaz Sistachs, *Departamento de Matemática Aplicada*

**Palabras Clave:** k-means, acp, clústeres, 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 adolescentes 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 información 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 reducción 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    . . . . . 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 existe una presencia de las pacientes con alto peso, (aparece menstruación pero no se supo explicar), que además tienen una alta restricción de comida, que purgan (vómitan los alimentos ingeridos para mantener bajo peso), con hiperactividad, buenas relaciones con familiares y amigos, alta emancipación de sus familias, buena actitud y comportamiento sexual, con buen estado de animo, muy preocupadas por su alimentación y

```
> summary(ACP)
Importance of components:
PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8    PC9    PC10
Standard deviation 2.46219 1.5181 1.34396 1.14045 1.0480 0.91115 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 0.02376 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]
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

peso corporal, con mucha percepción de su cuerpo. También tienen presencia positiva las variables *school*, *tidi* y *time*.

- **PC2:** La segunda componente esta conformada por un grupo de pacientes con características como pocos atracones, poca presencia de vómitos (que purgan poco las comidas), con buenas relaciones familiares, pero que fueron diagnosticadas con la enfermedad. También tiene presencia positiva la variable *tidi*.
- **PC3:** La tercera componente presenta pacientes con características como buenas relaciones con amigos, buena actitud sexual. También tienen presencia negativa las variables *tidi* y *time*, y positiva la variable *school*.
- **PC4:** La cuarta componente muestra pacientes con alta hiperactividad, con emancipación de sus familias, sin buenas relaciones con sus amigos, desanimadas, con poca preocupación por su peso y alimentación, poca percepción de su cuerpo, y con poca presencia de la enfermedad. Tienen también alta presencia las variables *school*, *tidi* y *time*.
- **PC5:** La quinta componenete se caracteriza 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 clúster jerárquicos, kmeans y árboles de decisión.

## 4. clústeres

- Con 4 clústeres se puede observar (Fig. 4) que se forman grandes grupos pero que aún pudiera seguir particionandoe.
- Lo haremos para 5 clústeres (Fig. 5). Esto a partir de la observación del dendrograma con 4 clústeres. Además intuitivamente por la cantidad de componentes principales que tomamos.
- Con 6 clústeres (Fig. 6) no se nota un gran cambio con respecto a los grupos formados para 5 clústeres 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 (Fig. 7) la similitud 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 (Fig. 8) para ver como se observa la partición del árbol. La similitud es del 41%, la cual igualmente no es tan buena. Las cantidades son 28, 85, 47, 19 y 38. Con 5 se logró particionar un grupo grande en la distribución con 4 grupos pero se sigue manteniendo el mayor.
- Para 6 (Fig. 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 clústeres anteriores, habiendose particionado un poco el mayor de los grupos según la distribución de 5.

## 6. Árbol de Decisión

Analizaremos el árbol de decisión (Fig. 10) formado para la variable *diag* (Diagnostico) utilizando un metodo CART de clasificación. En este caso se alcanzó un 40% de error aproximadamente. El árbol es posible de construir pero no tiene mucho valor predictivo debido al alto error en un contexto en el cual este no es permisible.

## 7. Referencias

- Conferencia 5 - Correlación
- Conferencia 6 - Regresión Lineal Simple
- Conferencia 7 - Regresión Lineal Múltiple
- Conferencia 8 - ANOVA
- Conferencia 9 - ACP
- Conferencia 10 - Clúster, Árboles de Decisión

## 8. Anexos

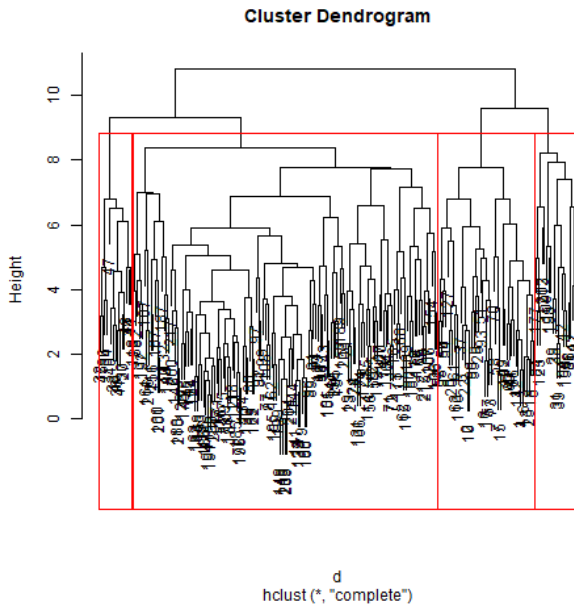


Figure 4: Análisis con 4 clústeres

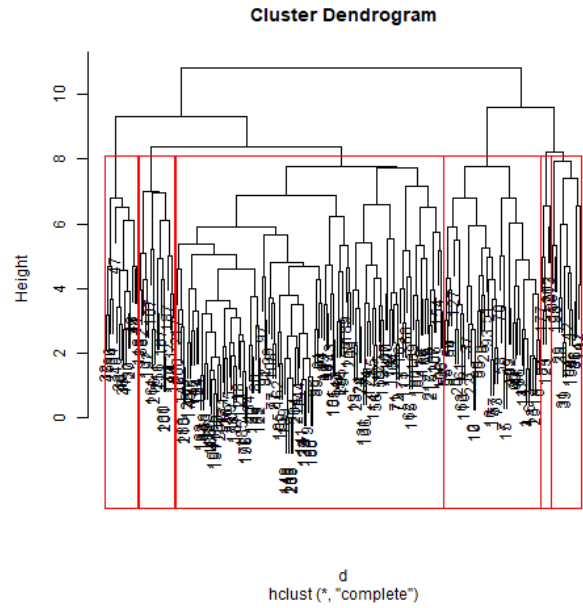


Figure 6: Análisis con 6 clústeres

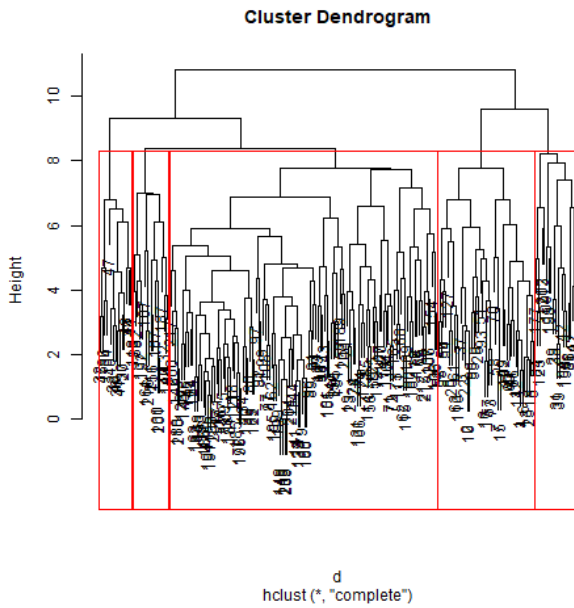


Figure 5: Análisis con 5 clústeres

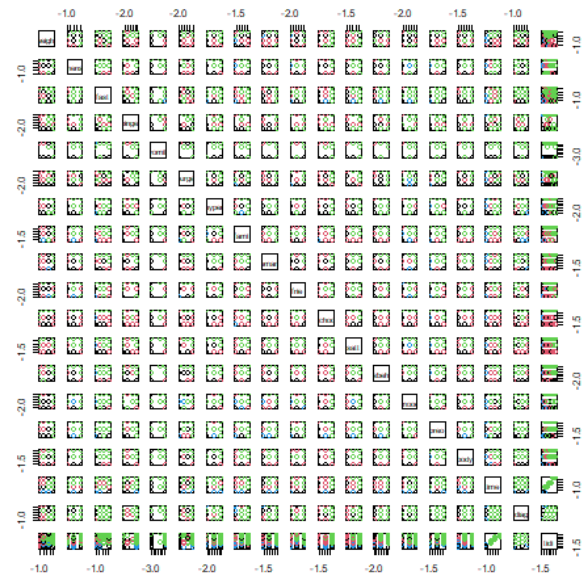


Figure 7: Análisis de kmeans con 4

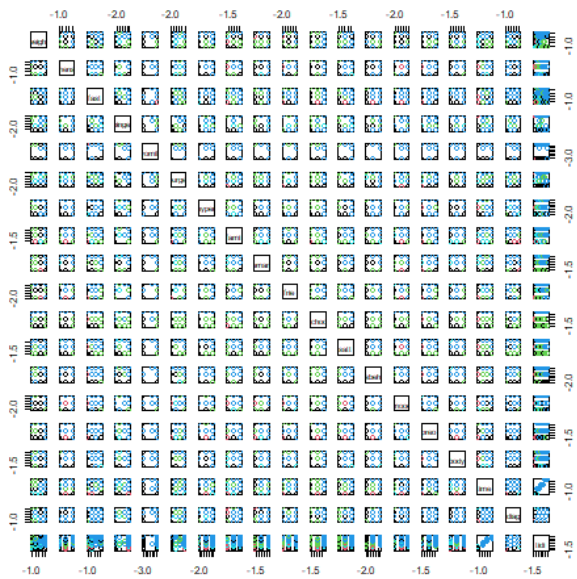


Figure 8: Análisis de kmeans con 5

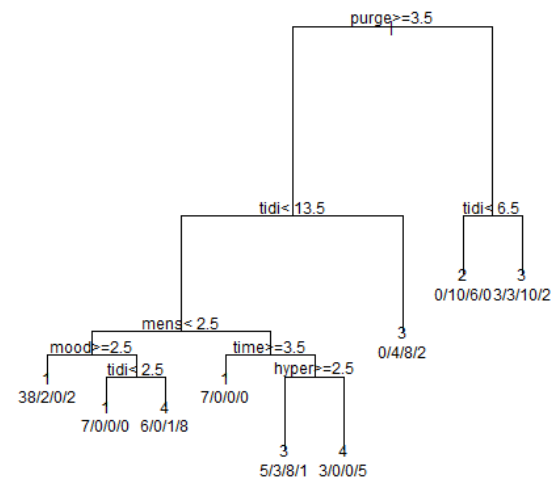


Figure 10: Árbol de Decisión

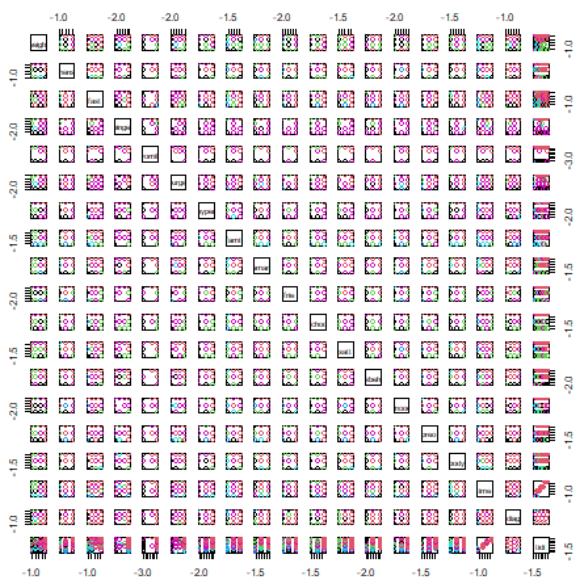


Figure 9: Análisis de kmeans con 6