PEC 1

Miguel Díaz Martín

2024-11-04

Índice

Abstract **1**

Objetivos del estudio **1**

Materiales y métodos **2**
Datos . . . 2
Herramientas . . . 2
Métodos . . . 2

Resultados **2**
Descripción de los datos . . . 2
Obtención y procesamiento de los datos . . . 2
Uso de la clase SummarizedExperiment . . . 3
Primera parte: Análisis de cluster . . . 3
Segunda parte: Predicción de la enfermedad según los biomarcadoes . . . 9

Conclusión **13**
Análisis de cluster . . . 13
K-NN . . . 13

Limitaciones del estudio **13**

Repositorio **13**

Abstract

En este estudio se ha hecho un análisis de cluster a los biomarcadores del conjunto de datos GastricCancer_NMR, consiguiendo dividirlos en tres grupos bien definidos.

Además, se ha creado un clasificador con el objetivo de predecir el estado de salud de nuevos individuos a partir de sus biomarcadores.

Objetivos del estudio

Se va a trabajar con un dataset que contiene diferentes biomarcadores junto con el estado de salud, relacionado con el cáncer gástrico, de diferentes individuos.

El estudio se va a dividir en dos partes:

- Análisis de cluster, donde se pretende explorar la división de todos los datos en diferentes grupos y ver cómo se relacionan estos grupos creados con el estado de salud de los individuos con el algoritmo K-Means.

- Creación de un clasificador que genere predicciones del estado de salud de los individuos a partir de sus biomarcadores con el algoritmo K-NN, con el objetivo de conseguir un sistema de detección de cáncer gástrico.

Además, en las dos partes se intentará optimizar los datos para obtener los mejores resultados posibles.

Materiales y métodos

Datos

Los datos consisten en 149 diferentes biomarcadores obtenidos a partir de muestras de orina de 140 pacientes. Se puede ver más información sobre los datos en DOI:10.21228/M8B10B.

Herramientas

Para hacer el estudio se ha usado principalmente el lenguaje de programación R, junto con RMarkdown.

Métodos

Para hacer el análisis de cluster y el clasificador, se han usado los algoritmos K-Means y K-NN respectivamente.

Además, se han hecho transformaciones logarítmicas y estandarización para normalizar los datos.

Para evaluar los resultados del clasificador se han usado tablas cruzadas.

Resultados

Descripción de los datos

Para este trabajo hemos decidido que vamos a usar el dataset GastricCancer_NMR, que se encuentra en el repositorio metaboData.

Este dataset consiste en una recopilación de biomarcadores recopilados para cada individuo, relacionándolos con su estado en relación al cáncer gástrico.

Hay tres diferentes estados:

- `GC`: El individuo tiene cáncer gástrico
- `BN`: El individuo tiene una enfermedad gástrica benigna.
- `HE`: El individuo está sano.

Estos diferentes estados vienen relacionados con hasta 149 biomarcadores.

En total hay 140 individuos en el dataset.

Obtención y procesamiento de los datos

Para descargar los datos usamos el siguiente comando:

$ wget -O GastricCancer_NMR.xlsx https://github.com/nutrimetabolomics/metaboData/raw/refs/heads/main/Datasets/2023-CIMCBTutorial/GastricCancer_NMR.xlsx

Este fichero contiene dos diferentes hojas de datos: `Data`, donde están definidas todas las filas del dataset, y `Peak`, donde se definen los nombres de las columnas junto con otros datos relevantes.

Cargamos un dataframe con los datos que hay en el fichero y asignamos los nombres de las columnas que tenemos en la hoja `Peak`:

```
gastric_cancer <- read_excel("GastricCancer_NMR.xlsx", sheet="Data")
peak <- read_excel("GastricCancer_NMR.xlsx", sheet="Peak")

gastric_cancer <- rename_at(gastric_cancer, peak$Name, ~ peak$Label)
```

También vamos a descartar las 3 primeras columnas y los individuos que se encuentran en la clase de control de calidad, ya que no nos van a ser de utilidad en este estudio:

```
gastric_cancer <- gastric_cancer[c(-1, -2, -3)]
gastric_cancer <- gastric_cancer[gastric_cancer$Class != "QC",]
```

Preparamos un factor para la clase, que es la columna que nos indica el estado de cada individuo:

```
gastric_cancer$Class <- factor(gastric_cancer$Class, levels=c("GC", "BN", "HE"), labels=c("Gastric canc
```

Para terminar, asignamos el valor arbitrario 1 a todos los valores ausentes para poder hacer los análisis sin problemas:

```
gastric_cancer[is.na(gastric_cancer)] <- 1
```

Uso de la clase SummarizedExperiment

Para el estudio vamos a añadir los datos datos de los biomarcadores a una instancia de Summarized experiment, separándolos de la columna donde se encuentra el estado de los individuos:

```
gastric_cancer_classes <- gastric_cancer$Class
gastric_cancer <- gastric_cancer[-1]
gastric_cancer_se <- SummarizedExperiment(colData=colnames(gastric_cancer), assays=gastric_cancer)
```

Primera parte: Análisis de cluster

Vamos a empezar preparando una función para representar los clusters que obtengamos. En la gráfica aparecerá cada cluster generado marcado con un color diferente, mientras que la clase de cada individuo vendrá marcada con una forma.

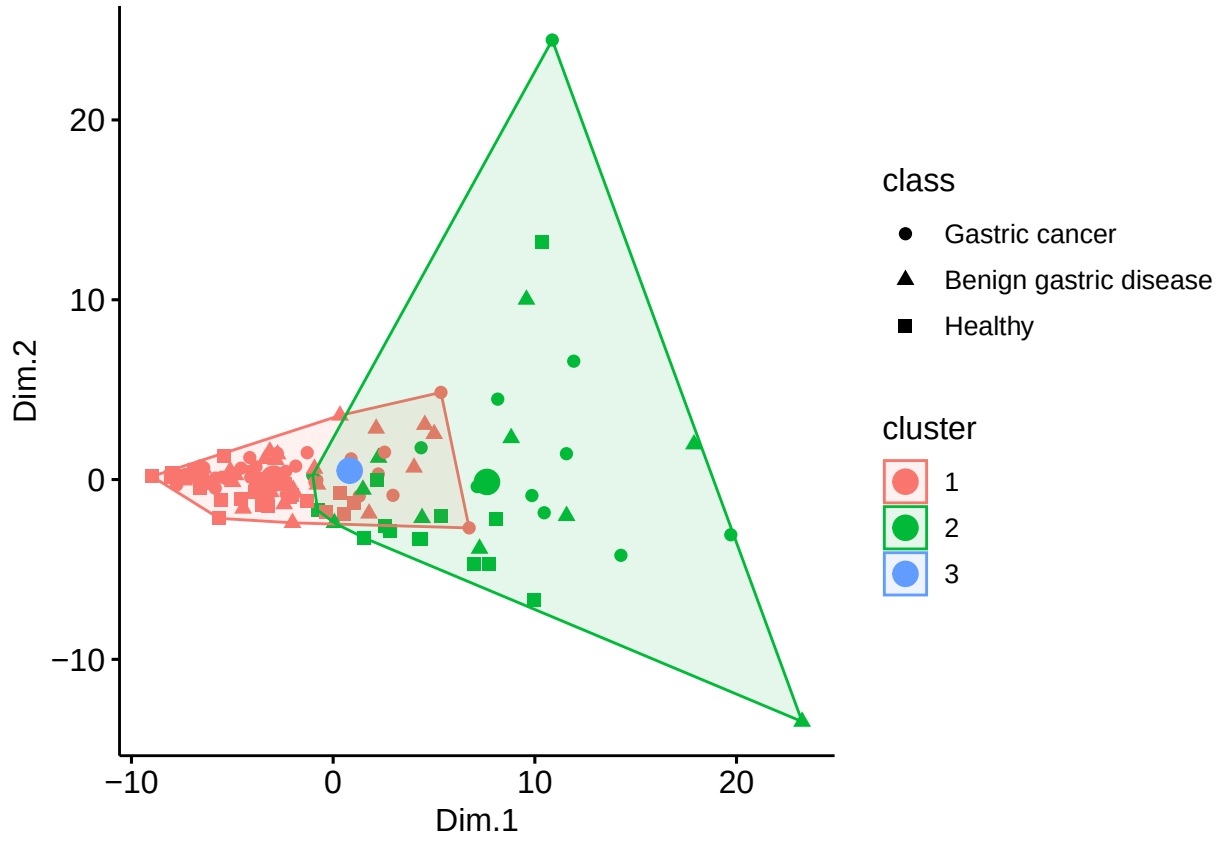
```
plot_cluster_with_class <- function (clust, gc_data, gc_class) {
  pca <- prcomp(gc_data, scale=T)
  coord <- as.data.frame(get_pca_ind(pca)$coord)
  coord$cluster <- factor(clust$cluster)
  coord$class <- gc_class

return (
  ggscatter(coord, x="Dim.1", y="Dim.2", color="cluster", shape="class", legend="right",
            ellipse=T, ellipse.type="convex") + stat_mean(aes(color=cluster), size=4)
  )
}
```

Queremos ver cómo se dividen los datos en 3 grupos, que es el número de clases con el que estamos trabajando.

Generamos el cluster con k-means y representamos el resultado:

```
km <- kmeans(assay(gastric_cancer_se), centers=3)
plot_cluster_with_class(km, assay(gastric_cancer_se), gastric_cancer_classes)
```



No parece que haya una buena división entre 3 grupos, uno de los clusters es mucho más pequeño que los otros dos.

Para ver si podemos mejorar los resultados, vamos a comparar los rangos en los que se encuentra cada columna de nuestro dataset:

```
for(c in colnames(assay(gastric_cancer_se))) {
  print(
    paste0("Rango: (", min(assay(gastric_cancer_se)[c]), ", ", max(assay(gastric_cancer_se)[c]), ")")
  )
}
```

```
## [1] "Rango: (0.4, 909.9)"
## [1] "Rango: (1, 26195.8)"
## [1] "Rango: (0.1, 862.5)"
## [1] "Rango: (0.1, 242.5)"
## [1] "Rango: (1, 2503)"
## [1] "Rango: (0.2, 339.4)"
## [1] "Rango: (1, 492.6)"
## [1] "Rango: (9.3, 525)"
## [1] "Rango: (0.7, 612.1)"
## [1] "Rango: (0.1, 2026.8)"
## [1] "Rango: (0.3, 2676.3)"
## [1] "Rango: (1.7, 576.2)"
## [1] "Rango: (17.2, 10712.7)"
## [1] "Rango: (0.2, 437.6)"
## [1] "Rango: (7.9, 212.3)"
```

```
## [1] "Rango: (1, 665.9)"
## [1] "Rango: (0.4, 187.6)"
## [1] "Rango: (1, 1236.5)"
## [1] "Rango: (0.4, 217.7)"
## [1] "Rango: (0.5, 341.9)"
## [1] "Rango: (0.1, 997.2)"
## [1] "Rango: (0.3, 713.5)"
## [1] "Rango: (1, 2499.1)"
## [1] "Rango: (0.3, 446.4)"
## [1] "Rango: (0.4, 171.8)"
## [1] "Rango: (1, 374.6)"
## [1] "Rango: (0.4, 1062.5)"
## [1] "Rango: (0.1, 14787.1)"
## [1] "Rango: (1, 4719)"
## [1] "Rango: (1, 1156.8)"
## [1] "Rango: (0.2, 1336.6)"
## [1] "Rango: (0.6, 874.2)"
## [1] "Rango: (26.8, 1127.6)"
## [1] "Rango: (0.1, 1605.4)"
## [1] "Rango: (0.1, 6596.8)"
## [1] "Rango: (0.1, 305.9)"
## [1] "Rango: (4.3, 1026.5)"
## [1] "Rango: (1, 1632.5)"
## [1] "Rango: (0.5, 913.9)"
## [1] "Rango: (0.8, 204.3)"
## [1] "Rango: (0.1, 746)"
## [1] "Rango: (1, 810.1)"
## [1] "Rango: (1, 191.7)"
## [1] "Rango: (0.3, 454.1)"
## [1] "Rango: (49.9, 16673.9)"
## [1] "Rango: (1, 3749.4)"
## [1] "Rango: (0.1, 1771.3)"
## [1] "Rango: (988, 33766.6)"
## [1] "Rango: (9.6, 1547.1)"
## [1] "Rango: (0.2, 17082.2)"
## [1] "Rango: (1, 5732.7)"
## [1] "Rango: (9.6, 3337.3)"
## [1] "Rango: (40.2, 6413.9)"
## [1] "Rango: (0.2, 82.9)"
## [1] "Rango: (0.4, 19704.1)"
## [1] "Rango: (1, 367.9)"
## [1] "Rango: (0.1, 16077.3)"
## [1] "Rango: (1, 6371.7)"
## [1] "Rango: (1, 10448.2)"
## [1] "Rango: (0.2, 160844.7)"
## [1] "Rango: (1, 4920.9)"
## [1] "Rango: (0.3, 1897.9)"
## [1] "Rango: (1, 15297.4)"
## [1] "Rango: (0.1, 5168.5)"
## [1] "Rango: (25.6, 2432.1)"
## [1] "Rango: (34.6, 16544.5)"
## [1] "Rango: (1, 686.9)"
## [1] "Rango: (0.1, 1639)"
## [1] "Rango: (0.1, 546.5)"
```

```
## [1] "Rango: (1, 598.3)"
## [1] "Rango: (4.4, 490.4)"
## [1] "Rango: (0.3, 629)"
## [1] "Rango: (10.1, 366.3)"
## [1] "Rango: (0.1, 171)"
## [1] "Rango: (21.6, 1225.3)"
## [1] "Rango: (0.1, 3230.2)"
## [1] "Rango: (0.1, 1751.5)"
## [1] "Rango: (0.1, 241.6)"
## [1] "Rango: (0.4, 3504.4)"
## [1] "Rango: (0.4, 27945.1)"
## [1] "Rango: (1.2, 3849.4)"
## [1] "Rango: (0.1, 1249.7)"
## [1] "Rango: (0.1, 8918)"
## [1] "Rango: (0.1, 813)"
## [1] "Rango: (0.1, 376)"
## [1] "Rango: (4.7, 959.3)"
## [1] "Rango: (1, 213.7)"
## [1] "Rango: (0.5, 798.8)"
## [1] "Rango: (47.4, 6317.1)"
## [1] "Rango: (2.8, 748.7)"
## [1] "Rango: (1, 337.5)"
## [1] "Rango: (0.1, 169.4)"
## [1] "Rango: (5.6, 505.6)"
## [1] "Rango: (1, 2624.2)"
## [1] "Rango: (0.7, 305.5)"
## [1] "Rango: (0.1, 32.6)"
## [1] "Rango: (0.7, 82)"
## [1] "Rango: (0.6, 1257.3)"
## [1] "Rango: (0.8, 737.2)"
## [1] "Rango: (0.1, 5188.2)"
## [1] "Rango: (0.1, 428.8)"
## [1] "Rango: (0.1, 636.2)"
## [1] "Rango: (1, 1482.8)"
## [1] "Rango: (0.2, 1579.2)"
## [1] "Rango: (0.1, 2182.2)"
## [1] "Rango: (1, 598.7)"
## [1] "Rango: (0.4, 2073.8)"
## [1] "Rango: (0.5, 1397.2)"
## [1] "Rango: (0.1, 1045.7)"
## [1] "Rango: (1, 359.1)"
## [1] "Rango: (0.2, 6373.7)"
## [1] "Rango: (18.8, 1579.1)"
## [1] "Rango: (0.2, 2078.8)"
## [1] "Rango: (2.6, 1143.4)"
## [1] "Rango: (1, 2134.5)"
## [1] "Rango: (0.5, 318.7)"
## [1] "Rango: (1, 1613.1)"
## [1] "Rango: (1, 1434.2)"
## [1] "Rango: (10.3, 526.8)"
## [1] "Rango: (0.2, 934.9)"
## [1] "Rango: (0.5, 217.2)"
## [1] "Rango: (1, 202.2)"
## [1] "Rango: (1, 1418.2)"
```

```
## [1] "Rango: (1, 3789.7)"
## [1] "Rango: (12.4, 1619.1)"
## [1] "Rango: (1, 609.4)"
## [1] "Rango: (0.1, 786.2)"
## [1] "Rango: (0.1, 5959.6)"
## [1] "Rango: (133.3, 8038.2)"
## [1] "Rango: (0.2, 1188.7)"
## [1] "Rango: (76.6, 8348.6)"
## [1] "Rango: (21.4, 3155.4)"
## [1] "Rango: (4.1, 1894.5)"
## [1] "Rango: (1, 8567.8)"
## [1] "Rango: (0.9, 53432)"
## [1] "Rango: (0.1, 5014.9)"
## [1] "Rango: (0.4, 12900.8)"
## [1] "Rango: (1, 6344.9)"
## [1] "Rango: (0.1, 5569.5)"
## [1] "Rango: (0.1, 6960.3)"
## [1] "Rango: (1, 6173.9)"
## [1] "Rango: (0.1, 282.9)"
## [1] "Rango: (0.2, 8413.2)"
## [1] "Rango: (17, 1251.4)"
## [1] "Rango: (0.1, 5479.2)"
## [1] "Rango: (0.1, 4791.5)"
## [1] "Rango: (1, 840.2)"
## [1] "Rango: (1, 2560.3)"
## [1] "Rango: (22.1, 502.5)"
```

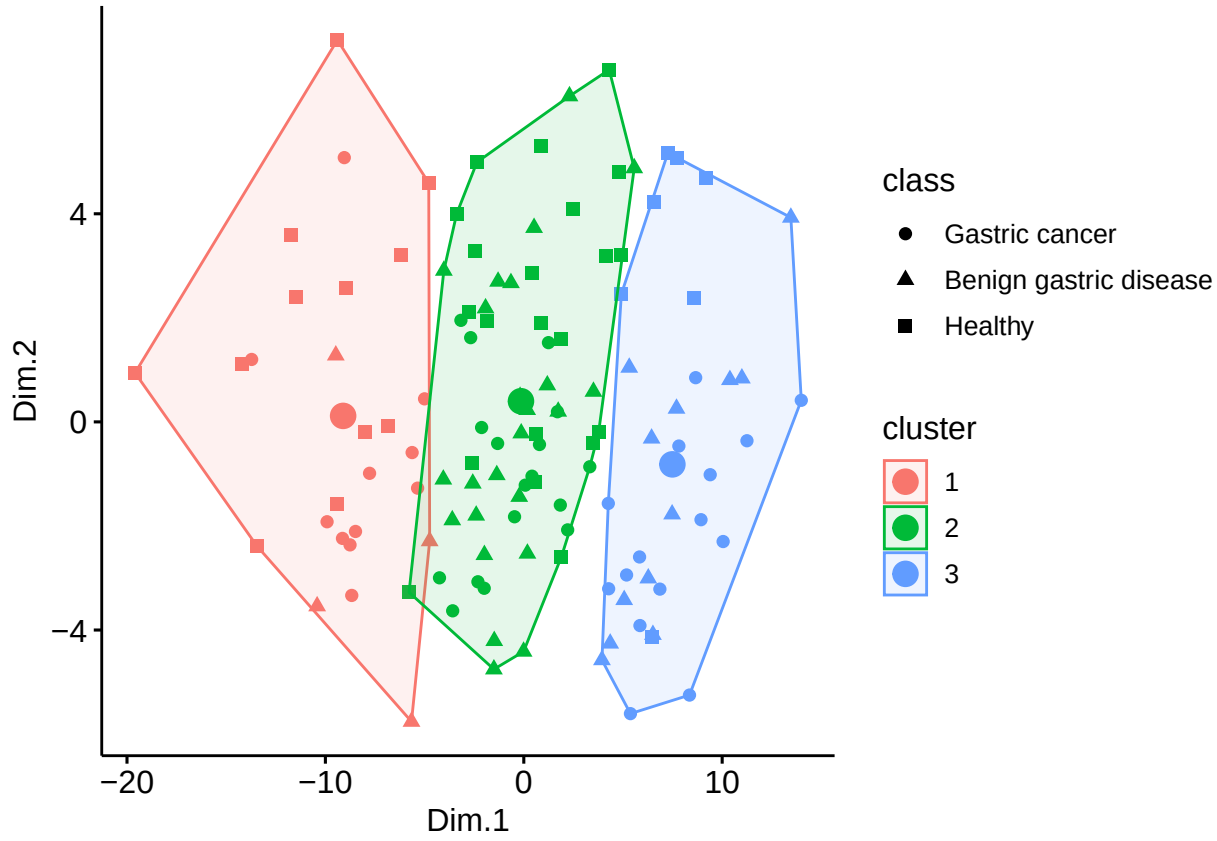
Algunos de los rangos empiezan en 1, que es el número que asignamos anteriormente a los valores ausentes. Sin embargo, esta salida nos sirve de sobra para llegar a la conclusión de que los rangos de las columnas son muy distintas entre sí, por lo que pueden estar afectando unas más que otras al cálculo del cluster.

Para solucionar esto vamos a normalizar los datos, de forma que todas las columnas se muevan en el mismo rango, y ver cómo afecta esta transformación a los cluster obtenidos.

Normalización logarítmica

Una forma de normalizar los datos es calculando el logaritmo:

```
gastric_cancer_log <- log(assay(gastric_cancer_se))
km <- kmeans(gastric_cancer_log, centers=3)
plot_cluster_with_class(km, gastric_cancer_log, gastric_cancer_classes)
```



Podemos ver que con esta transformación conseguimos tres grupos mucho mejor definidos, y parece que hay diferentes proporciones de las clases en cada grupo.

Estandarización

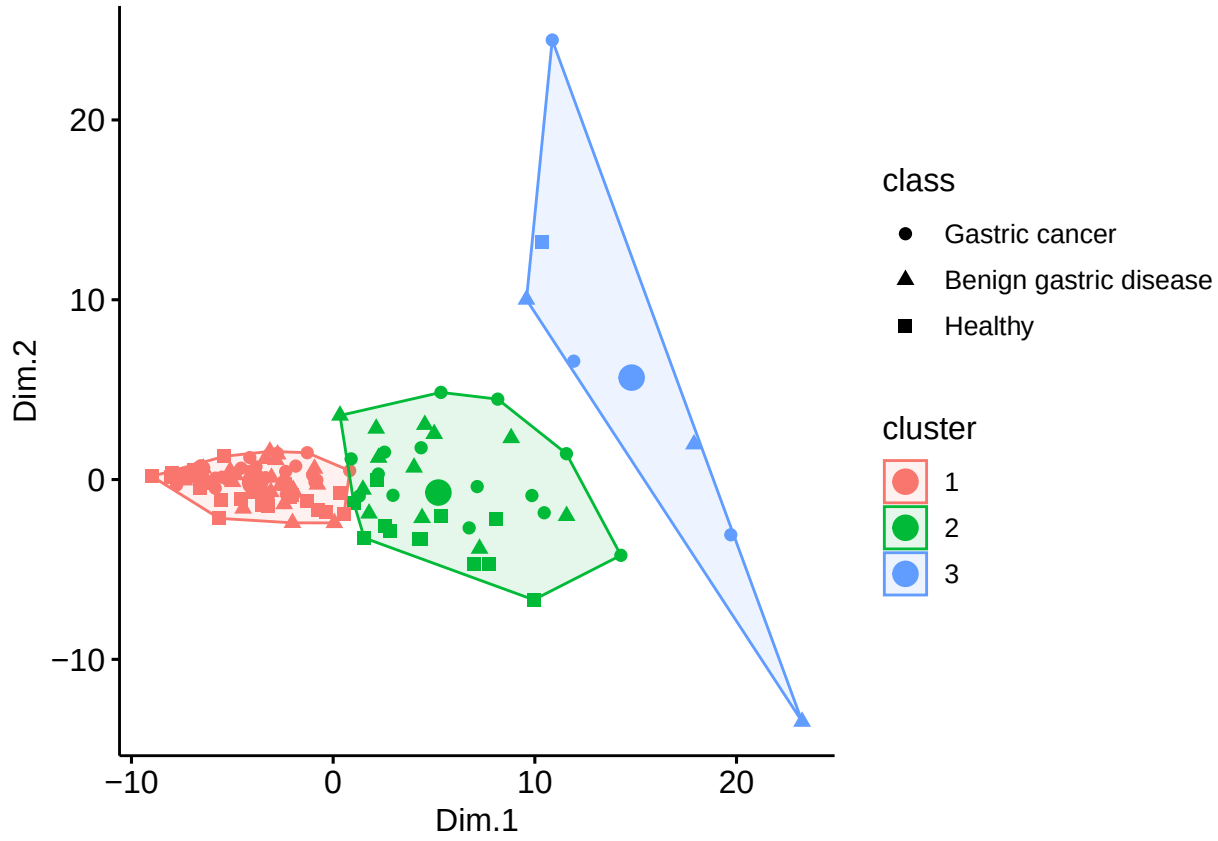
También podemos probar estandarizando las muestras, que es un proceso que se hace aplicando la siguiente fórmula a cada columna:

$$\frac{X - \mu}{\sigma}$$

Siendo μ la media y σ la desviación estándar.

```
standarize <- function(l) {
  return((l - mean(l)) / sd(l))
}

gastric_cancer_standarized <- apply(assay(gastric_cancer_se), 2, standarize)
km <- kmeans(gastric_cancer_standarized, centers=3)
plot_cluster_with_class(km, gastric_cancer_standarized, gastric_cancer_classes)
```



Con esta transformación también conseguimos tres grupos bien definidos, con diferentes proporciones de las clases.

Segunda parte: Predicción de la enfermedad según los biomarcadoes

En esta parte pretendemos entrenar un clasificador que pueda predecir el estado de un individuo a partir de sus biomarcadores, usando el algoritmo K-NN. Este algoritmo funciona calculando la distancia de todas las variables del individuo con las variables de otros individuos, y asignándole la clase más comun en su k vecinos más cercanos, siendo k un número que asignamos de forma arbitraria.

El algoritmo K-NN puede ser util para usar con los datos que estamos analizando porque funciona bien con datasets pequeños.

Para empezar, vamos a separar nuestro dataset en dos partes, una de entrenamiento para preparar el algoritmo y otra de test para evaluar los resultados. La parte de entrenamiento serán dos tercios del dataset, y la parte de test será un tercio. Vamos a usar los datos que hemos estandarizado anteriormente:

```
gc_train <- gastric_cancer_standarized[1:82,]
classes_train <- gastric_cancer_classes[1:82]

gc_test <- gastric_cancer_standarized[83:123,]
classes_test <- gastric_cancer_classes[83:123]
```

A la hora de elegir el valor de k hay que tener en cuenta que, cuanto más grande sea, menor será el impacto del ruido en los datos, pero será más dificil tener en cuenta patrones que se representen en pocos individuos. Vamos a empezar con k = 7:

```
preds <- knn(train=gc_train, test=gc_test, cl=classes_train, k=7)
```

Para evaluar los resultados vamos a usar una tabla cruzada, que dispone las predicciones junto con las clases reales, de forma que podemos ver la proporción de los fallos en las predicciones:

```
CrossTable(x=classes_test, y=preds, prop.chisq=F)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  41
##
##
##                       | preds
##           classes_test |         Gastric cancer | Benign gastric disease |                Healthy |
## -----------------------|------------------------|------------------------|------------------------|--
##         Gastric cancer |                      5 |                      5 |                      4 |
##                        |                  0.357 |                  0.357 |                  0.286 |
##                        |                  0.833 |                  0.278 |                  0.235 |
##                        |                  0.122 |                  0.122 |                  0.098 |
## -----------------------|------------------------|------------------------|------------------------|--
## Benign gastric disease |                      1 |                      7 |                      5 |
##                        |                  0.077 |                  0.538 |                  0.385 |
##                        |                  0.167 |                  0.389 |                  0.294 |
##                        |                  0.024 |                  0.171 |                  0.122 |
## -----------------------|------------------------|------------------------|------------------------|--
##                Healthy |                      0 |                      6 |                      8 |
##                        |                  0.000 |                  0.429 |                  0.571 |
##                        |                  0.000 |                  0.333 |                  0.471 |
##                        |                  0.000 |                  0.146 |                  0.195 |
## -----------------------|------------------------|------------------------|------------------------|--
##           Column Total |                      6 |                     18 |                     17 |
##                        |                  0.146 |                  0.439 |                  0.415 |
## -----------------------|------------------------|------------------------|------------------------|--
##
##
```

En la tabla podemos ver el número de precicciones correctas, que serían las que podemos encontrar en la diagonal de la tabla. Como podemos ver, ha hecho unas buenas predicciones al deducir el cáncer, pero no han sido tan buenas para las otras clases.

Vamos a ver la calidad de las predicciones que obtenemos con otros valores de k.

Para k = 5:

```
preds <- knn(train=gc_train, test=gc_test, cl=classes_train, k=5)
CrossTable(x=classes_test, y=preds, prop.chisq=F)
```

```
##
```

```
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  41
##
##
##                          | preds
##           classes_test |         Gastric cancer | Benign gastric disease |                Healthy |
## -----------------------|------------------------|------------------------|------------------------|-
##         Gastric cancer |                      9 |                      3 |                      2 |
##                        |                  0.643 |                  0.214 |                  0.143 |
##                        |                  0.750 |                  0.214 |                  0.133 |
##                        |                  0.220 |                  0.073 |                  0.049 |
## -----------------------|------------------------|------------------------|------------------------|-
## Benign gastric disease |                      3 |                      3 |                      7 |
##                        |                  0.231 |                  0.231 |                  0.538 |
##                        |                  0.250 |                  0.214 |                  0.467 |
##                        |                  0.073 |                  0.073 |                  0.171 |
## -----------------------|------------------------|------------------------|------------------------|-
##                Healthy |                      0 |                      8 |                      6 |
##                        |                  0.000 |                  0.571 |                  0.429 |
##                        |                  0.000 |                  0.571 |                  0.400 |
##                        |                  0.000 |                  0.195 |                  0.146 |
## -----------------------|------------------------|------------------------|------------------------|-
##           Column Total |                     12 |                     14 |                     15 |
##                        |                  0.293 |                  0.341 |                  0.366 |
## -----------------------|------------------------|------------------------|------------------------|-
##
##
```

Estos resultados parecen más deseables que los anteriores, ya que hay más diagnósticos de cáncer gástrico y menos diagnósticos de que los individuos están sanos cuando en realidad tienen la enfermedad.

Para k = 9:

```
preds <- knn(train=gc_train, test=gc_test, cl=classes_train, k=9)
CrossTable(x=classes_test, y=preds, prop.chisq=F)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
```

```
## Total Observations in Table:  41
##
##
##                         | preds
##           classes_test |         Gastric cancer | Benign gastric disease |                Healthy |
## ----------------------|------------------------|------------------------|------------------------|--
##         Gastric cancer |                      6 |                      5 |                      3 |
##                        |                  0.429 |                  0.357 |                  0.214 |
##                        |                  0.600 |                  0.312 |                  0.200 |
##                        |                  0.146 |                  0.122 |                  0.073 |
## ----------------------|------------------------|------------------------|------------------------|--
## Benign gastric disease |                      2 |                      6 |                      5 |
##                        |                  0.154 |                  0.462 |                  0.385 |
##                        |                  0.200 |                  0.375 |                  0.333 |
##                        |                  0.049 |                  0.146 |                  0.122 |
## ----------------------|------------------------|------------------------|------------------------|--
##                Healthy |                      2 |                      5 |                      7 |
##                        |                  0.143 |                  0.357 |                  0.500 |
##                        |                  0.200 |                  0.312 |                  0.467 |
##                        |                  0.049 |                  0.122 |                  0.171 |
## ----------------------|------------------------|------------------------|------------------------|--
##           Column Total |                     10 |                     16 |                     15 |
##                        |                  0.244 |                  0.390 |                  0.366 |
## ----------------------|------------------------|------------------------|------------------------|--
##
##
```

No parecen unos buenos resultados, hay muchos fallos en las predicciones.

Por último, vamos a probar k=3:

```
preds <- knn(train=gc_train, test=gc_test, cl=classes_train, k=3)
CrossTable(x=classes_test, y=preds, prop.chisq=F)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  41
##
##
##                         | preds
##           classes_test |         Gastric cancer | Benign gastric disease |                Healthy |
## ----------------------|------------------------|------------------------|------------------------|--
##         Gastric cancer |                      7 |                      5 |                      2 |
##                        |                  0.500 |                  0.357 |                  0.143 |
##                        |                  0.636 |                  0.333 |                  0.133 |
##                        |                  0.171 |                  0.122 |                  0.049 |
## ----------------------|------------------------|------------------------|------------------------|--
```

```
## Benign gastric disease |                   4 |                   4 |                   5 |
##                        |               0.308 |               0.308 |               0.385 |
##                        |               0.364 |               0.267 |               0.333 |
##                        |               0.098 |               0.098 |               0.122 |
## ----------------------|---------------------|---------------------|---------------------|--
##                Healthy |                   0 |                   6 |                   8 |
##                        |               0.000 |               0.429 |               0.571 |
##                        |               0.000 |               0.400 |               0.533 |
##                        |               0.000 |               0.146 |               0.195 |
## ----------------------|---------------------|---------------------|---------------------|--
##           Column Total |                  11 |                  15 |                  15 |
##                        |               0.268 |               0.366 |               0.366 |
## ----------------------|---------------------|---------------------|---------------------|--
##
##
```

No parece que los resultados sean mejores que los obtenidos con k= 5, por lo que parece que el mejor número de vecinos que podemos usar para la predicción de la salud de los pacientes es de 5, usando una normalización estándar de los biomarcadores.

Conclusión

Análisis de cluster

Al comienzo parecía que no se podían dividir los datos en tres grupos, pero hemos podido ver que, tras normalizarlos, sí que se puede distinguir la división bien definida.

K-NN

Se ha podido crear un clasificador que, aunque tenga dificultades prediciendo qué pacientes tienen una enfermedad benigna y cuáles están sanos, es capaz de predecir con un 80% de precisión si un paciente está enfermo de cáncer gástrico.

Limitaciones del estudio

Algo que falta por hacer en este estudio sería un análisis para identificar qué diferentes biomarcadores del conjunto de datos afectan con más o menos importancia a la clasificación de la enfermedad de los individuos. De esta forma se podría hacer una mejor división de grupos en el análisis de cluster y se podría conseguir un mejor clasificador para predecir el estado de nuevos individuos.

Repositorio

Link al repositorio de este trabajo: Diaz-Martin-Miguel-PEC1