Diagnóstico de cáncer de mama

Eudald Romo y Laura Santulario Verd?

9 de mayo de 2018

Introducción

Los datos en los que se basa este estudio ha sido obtenidos de un análisis cronológico (que tuvo lugar entre 1989 y 1991) sobre el diagnóstico del cáncer de mama en una muestra de 367 mujeres que se encontraban bajo seguimiento médico en hospitales de Wisconsin. Para hacer este análisis se estudiaron los cambios morfológicos en el tejido de la mama afectada por medio de citologías.

Nosotros planteamos un nuevo estudio sobre este data set con dos objetivos principales. En primer lugar, queremos realizar un análisis de corroboración. Es decir, a partir de los datos sin diagnóstico, vamos a realizar un análisis estructural para separar las mujeres en grupos (previesible y esperablemente en dos grupos, tumor benigno y maligno). Una vez realizado el análisis estructural, vamos a proponer un discriminante para los datos, vamos a corroborar que los dos grupos son distintos y vamos a comparar los resultados obtenidos con el estudio inicial.

Consideramos que lo más útil desde un punto de vista médico seria generar un estadístico discriminante durante la primera sesión (que es la que tiene más muestras con diferencia) e intentar predecir el diagnóstico de los pacientes de las sesiones posteriores.

En segundo lugar, aprovechando que tenemos datos de las pacientes en diferentes puntos temporales, haremos un estudio de seguimiento y observaremos la evolución de los tumores malignos en una selección de mujeres. El estudio de seguimiento serviría (teóricamente) para prevenir la aparición de cánceres malignos y empezar su tratamiento cuando aún estan en desarrollo. Es decir, si una mujer diera indicios de evolucionar en dirección a cáncer maligno, se podria empezar el tratamiento con valores más pequeños del estadístico discriminante: si el cambio de benigno a maligno estuviera en $F(X) = 5$ y una paciente empezara con un valor de 1 y fuera subiendo consistentemente, se podria empezar el tratamiento mucho antes. Somos conscientes que decidir el valor en el que empezar el tratamiento a partir de un estudio de seguimiento escapa el ámbito de la asignatura. Por lo tanto nos vamos a limitar a exponer los valores del seguimiento y a valorarlos cualitativamente. Sin embargo creemos que el seguimiento aporta valor al dataset estudiado y que sería útil en un entorno médico, calculando adecuadamente el valor de corte.

Las variables que se tienen en cuenta en el siguiente estudio son:

Thickness : el espesor de la masa tumoral (las células cancerosas tienden a agruparse formando multicapas)

Size: tamaño uniforme de las células tumorales (las células cancerosas tienden a tener un tamaño muy variable)

Shape: formas homogéneas de las células tumorales (las células cancerosas tienden a tener formas muy heterogéneas)

Adhesion: modo en el que se adhieren las células tumorales unas con otras (las células cancerígenas tienen menos adhesión que las células sanas)

SingleCellSize: el grosor de una capa de epitelio que forma la masa (las células cancerosas tienden a tener capas más gruesas)

Nuclei: núcleo (en las células cancerosas resulta difícil identificar el núcleo)

Chromatin: cromatina (en células cancerosos la cromatina suele ser muy gruesa)

Nucleoli: nucleolos (en las células cancerosas resulta fácil identificarlos, mientras que en las células sanas es muy difícil observarlos)

mitoses: mitosis atípicas

Class: la clasificación del tumor en benigno o maligno

Antes de empezar el estudio mencionado anteriormente, analizaremos las muestras y las distintas variables.

Análisis previo

Los datos de este estudio se han obtenido del hospital universitario de Wisconsin, y el conjunto de datos empleado en este informe es el *breast-cancer-wisconsin.data*, disponible en *https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29*. Esta base de datos cuenta con 698 observaciones divididas en 8 grupos (correspondientes a los diferentes periodos en los que se analizaron las muestras citológicas):

- Grupo 1: 367 casos (Enero 1989)
- Grupo 2: 70 casos (Octubre 1989)
- Grupo 3: 31 casos (Febrero 1990)
- Grupo 4: 17 casos (Abril 1990)
- Grupo 5: 48 casos (Agosto 1990)
- Grupo 6: 49 casos (Enero 1991)
- Grupo 7: 31 casos (Junio 1991)
- Grupo 8: 86 casos (Noviembre 1991)

En algunos periodos, se toma más de una muestra para ciertos pacientes. Al no saber el motivo por el cual se repitieron esas muestras, en cada periodo temporal descartaremos las muestras de personas que repitieron la citologia. Además, nuestro dataset presenta *missings values* en 16 registros de la variable *Nuclei*. Tenemos muestras suficientes para los análisis que vamos a realizar (como ya comentaremos más adelante) por lo que sencillamente descartaremos tambien muestras con valor de *Nuclei* inválido, quedándonos entonces con los siguientes datos:

- Grupo 1: 347 casos (Enero 1989)
- Grupo 2: 65 casos (Octubre 1989)
- Grupo 3: 30 casos (Febrero 1990)
- Grupo 4: 17 casos (Abril 1990)
- Grupo 5: 46 casos (Agosto 1990)
- Grupo 6: 47 casos (Enero 1991)
- Grupo 7: 31 casos (Junio 1991)
- Grupo 8: 79 casos (Noviembre 1991)

Cada muestra tiene 10 variables: 9 son las que hemos descrito anteriormente y representan los cambios morfológicos de las células mamarias (representadas por valores discretos entre 1 y 10) y una variable

dicotómica que caracteriza si el tumor es benigno o maligno (0 - benigno; 1 - maligno).

Un primer análisis del estudio en el que se basa nuestro data set, nos permite intuir (muy cualitativamente) que valores bajos de cada variable corresponden a tumores benignos y valores altos a malignos. Debajo de este párrafo se puede observar la media de las variables para tumores benignos y malignos.

___ Media Tumores Benignos___

```
##      Thickness           Size          Shape       Adhesion SingleCellSize
##           2.58           1.37           1.48           1.34           2.27
##         Nuclei      Chromatin       Nucleoli        Mitoses
##           1.64           2.60           1.48           1.08
```

Media Tumores Malignos

```
##      Thickness           Size          Shape       Adhesion SingleCellSize
##           7.32           6.09           6.25           5.31           5.41
##         Nuclei      Chromatin       Nucleoli        Mitoses
##           7.61           5.41           5.90           2.86
```

Como se puede observar el la siguiente tabla en nuestro estudio hay aproximadamente la misma proporción de mujeres con tumores de mama benignos (53.31%) que malignos (46.69%), con lo que esperamos ver 2 clusters de aproximadamente el mismo tamaño.

Clase	Frecuencia	Prop
0	185	53.31
1	162	46.69

Análisis

Para realizar el estudio estructural, es necesaria una medida de similitud/distancia entre las variables. Como lo que se busca es caracterizar diferencias, no definir perfiles ni dependencia/independencia, sólo se han barajado dos opciones, la distancia Euclídea y la de Mahalanobis. La matriz de correlaciones (tabla inferior) y indica cierta relación entre las variables, pero al realizar nuestro estudio sobre uno de previo, creemos razonable mantener estas correlaciones, ya que podrían implicar dar ciertos pesos a metavariables (como la forma, la actividad biológica de la célula, ...) que los investigadores del estudio previo introdujeron conscientemente. Por este motivo hemos elegido la distancia Euclídea. Cabe destacar que la distancia Euclídea se usa generalmente para variables reales. En este caso, las variables son enteras, pero a diferencia de ciertas variables categóricas, éstas representan claramente un valor cuantitativo. Por este motivo, el hecho que sean variables enteras se puede entender cómo un problema de precisión (solo tenemos una cifra significativa) y la distancia Euclídea puede ser usada tranquilamente.

```
##                Thickness Size Shape Adhesion SingleCellSize Nuclei
## Thickness           1.00 0.66  0.69     0.46           0.48   0.60
## Size                0.66 1.00  0.89     0.66           0.72   0.64
## Shape               0.69 0.89  1.00     0.64           0.67   0.68
## Adhesion            0.46 0.66  0.64     1.00           0.55   0.64
## SingleCellSize      0.48 0.72  0.67     0.55           1.00   0.53
## Nuclei              0.60 0.64  0.68     0.64           0.53   1.00
## Chromatin           0.55 0.66  0.65     0.57           0.54   0.60
## Nucleoli            0.53 0.71  0.72     0.57           0.60   0.55
## Mitoses             0.37 0.48  0.45     0.41           0.47   0.33
##                Chromatin Nucleoli Mitoses
## Thickness           0.55     0.53    0.37
## Size                0.66     0.71    0.48
```

```
## Shape              0.65     0.72    0.45
## Adhesion           0.57     0.57    0.41
## SingleCellSize     0.54     0.60    0.47
## Nuclei             0.60     0.55    0.33
## Chromatin          1.00     0.60    0.32
## Nucleoli           0.60     1.00    0.41
## Mitoses            0.32     0.41    1.00
```

A continuación se realizará un análisis jerárquico con el método Ward.D2 (por su similitud al método no jerárquico de k-means) de las muestras del primer dia que se usará para decidir un rango de número de clusters a analizar mediante un análisis no jerárquico.

Este análisis no jerárquico, usará el método *k-means* para calcular los indicadores *Delta TESS*, *pseudoF* y *average silhouette*. Se elegirá un número de clusters y se obtendrá la clasificación mediante k-means.

Los grupos obtenidos serán contrastados por inferencia (usando el test **Hotelling T2 de permutaciones** ya que, como se verá más adelante uno de los grupos no sigue una normal multivariante).

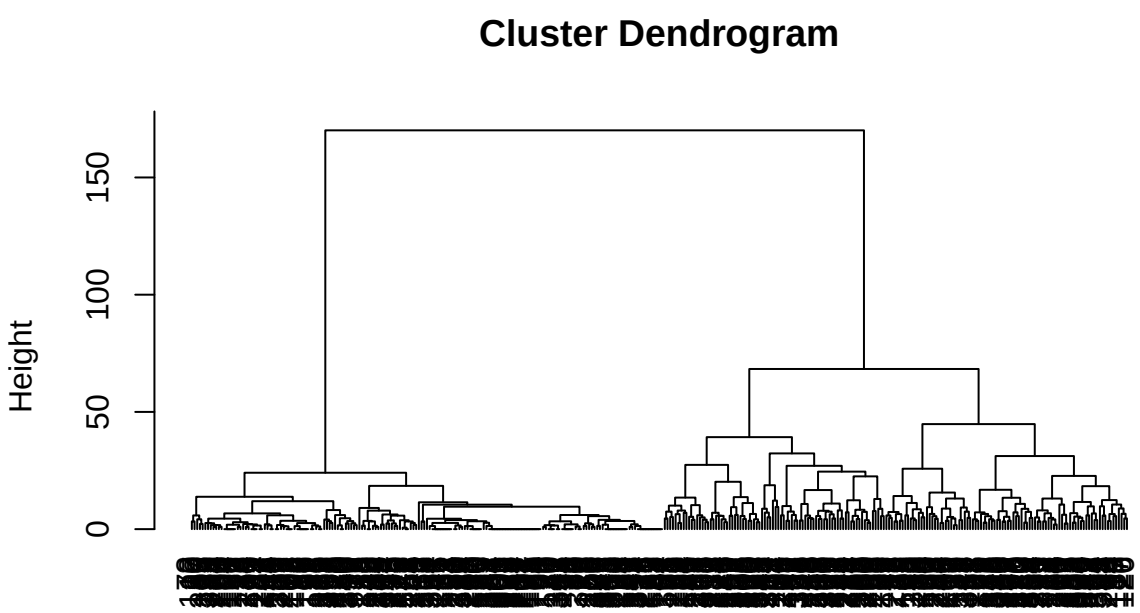
A partir de los grupos se obtendrá un discriminador (lineal o cuadrático), se comprovará su rendimiento con Leave-One-Out Crossvalidation y se compararan las predicciones obtenidas por nuestro discriminante con las obtenidas en el diagnóstico.

Finalmente, se compararán las predicciones de nuestro discriminador tanto con las variables dicotómicas que definen si la paciente tiene o no un tumor maligno. Esta comparación se hará tanto en el primer día (para comparar con las muestras de entrenamiento) como en el conjunto de los otros dias (muestras de comprovación).

Por otro lado, se usará el mismo discriminador para realizar un seguimiento a lo largo de tres sesiones a dos pacientes distintos y se comentarán los resultados obtenidos.

Resultados

```
## ---
## biotools version 3.1
```



Del análisis jerárquico podemos ver cualitativamente que el número de clusters posiblemente va a ser 2 y que dificilmente va a haber mas de 5 grupos. Por este motivo calcularemos los índices de calidad no jerárquicos para un número de grupos entre 1 y 5.

```
##   K    TOTSS  WITHINSS  PseudoF   AvgSilh  DeltaTESS
## 1 1 27411.89 27411.890       NA        NA         NA
## 2 2 27411.89 12466.375 413.6088 0.5094171 0.54522014
## 3 3 27411.89 10154.633 292.3048 0.4348668 0.18543822
## 4 4 27411.89  9203.751 226.1901 0.4506189 0.09364022
## 5 5 27411.89  8446.316 191.9839 0.3950531 0.08229637
```

Todos los indicadores coinciden en que el número de clusters es 2, cosa que corresponde tanto con el análisis jerárquico como con el resultado que esperábamos desde el inicio. Debajo de este parrafo se calcula la correlación entre los clusters calculados en este análisis y el diagnóstico real. Hay una alta correlación positiva, con lo que el cluster 2 se corresponde con el valor 1 de la variable dicotómica (tumor maligno) y el cluster 2 corresponde a tumores benignos.

```
## [1] 0.8669589
```

Para poder caracterizar estos dos grupos se analizan los representantes (figura bajo el párrafo). El representante del grupo 2 se correspondería con muestras de tumores malignos (**Class** ~ 1), al tener valores altos en todas las variables quantitativas, mientras que el representante del grupo 1 se corresponde con tumores benignos.

```
##   Thickness     Size    Shape Adhesion SingleCellSize   Nuclei Chromatin
## 1  2.726316 1.310526 1.447368 1.300000       2.184211 1.552895  2.594737
## 2  7.292994 6.312102 6.433121 5.484076       5.617834 7.895860  5.509554
##   Nucleoli  Mitoses
## 1 1.400000 1.173684
## 2 6.140127 2.796178
```

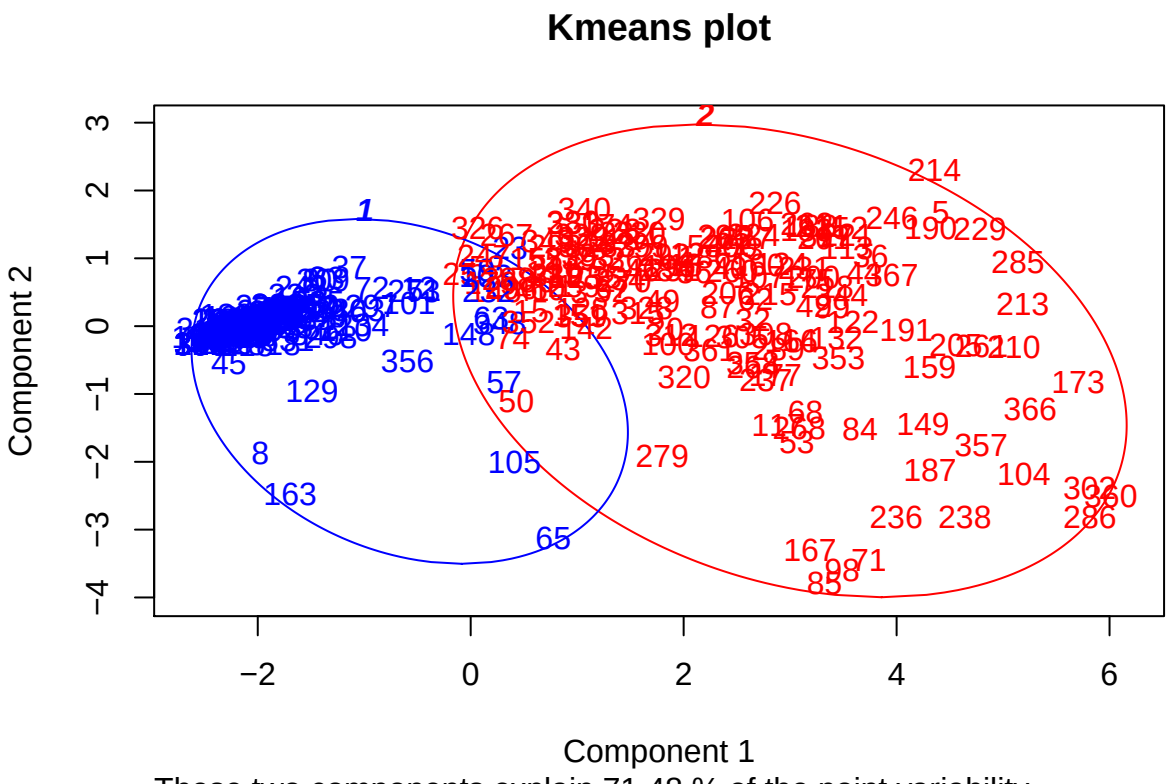
Para interpretar mejor los clusters representaremos nuestras observaciones en un espacio de dimensión reducida mediante PCA. Elegiremos un número de ejes principales de interés usando como indicador de calidad la vaiabilidad explicade e interpretaremos cada eje principal con los vectores propios del PCA.

Con dos componentes se puede explicar el 72% de la variabilidad (ver resultados debajo de este párrafo), que posiblemente será suficiente para interpretar los clusters obtenidos. La variabilidad explicada por la primera componente principal es mucho mayor que la de las otras, y añadir una tercera dimension complicaría la representación de los datos (deberíamos dar varias perspectivas elegidas adecuadamente para visualizarlos correctamente en este informe) para ganar sólo un 7% de variabilidad. Es por este motivo que hemos decidido mostrar los valores en 2 dimensiones

```
##     Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6
## 0.65470018 0.08185602 0.06805218 0.04948418 0.04110746 0.03825770
##     Comp.7     Comp.8     Comp.9
## 0.02899784 0.02584952 0.01169491
```

```
##
## Loadings:
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## Thickness       0.334         0.781  0.113  0.387  0.152  0.236  0.182
## Size            0.393 -0.246         0.197 -0.132 -0.369 -0.295
## Shape           0.388 -0.153  0.117        -0.134 -0.329 -0.461
## Adhesion        0.330  0.160 -0.562  0.326  0.630                0.190
## SingleCellSize  0.266 -0.235 -0.138  0.251 -0.507         0.651  0.309
## Nuclei          0.436  0.773        -0.203 -0.347  0.179
## Chromatin       0.229               -0.120  0.108 -0.202  0.407 -0.842
## Nucleoli        0.373 -0.435 -0.173 -0.717  0.125  0.301         0.142
## Mitoses         0.152 -0.211         0.454 -0.120  0.749 -0.212 -0.319
##                Comp.9
## Thickness
## Size            0.708
## Shape          -0.685
## Adhesion
## SingleCellSize -0.117
## Nuclei
## Chromatin
## Nucleoli
## Mitoses
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.111  0.111  0.111  0.111  0.111  0.111  0.111  0.111
## Cumulative Var  0.111  0.222  0.333  0.444  0.556  0.667  0.778  0.889
##                Comp.9
## SS loadings     1.000
## Proportion Var  0.111
## Cumulative Var  1.000
```

Se puede observar que la PC1 está influenciada por todas las variables. Crece al crecer cualquiera de las vari-



ables y el peso de

```
##             Anderson-Darling test for Multivariate Normality
##
##   data : g1
##
##   AD              : Inf
##   p-value         : 9.999e-05
##
##   Result  : Data are not multivariate normal (sig.level = 0.05)
```

```
##             Anderson-Darling test for Multivariate Normality
##
##   data : g2
##
##   AD              : 0.6655054
##   p-value         : 0.2762724
##
##   Result  : Data are multivariate normal (sig.level = 0.05)
```

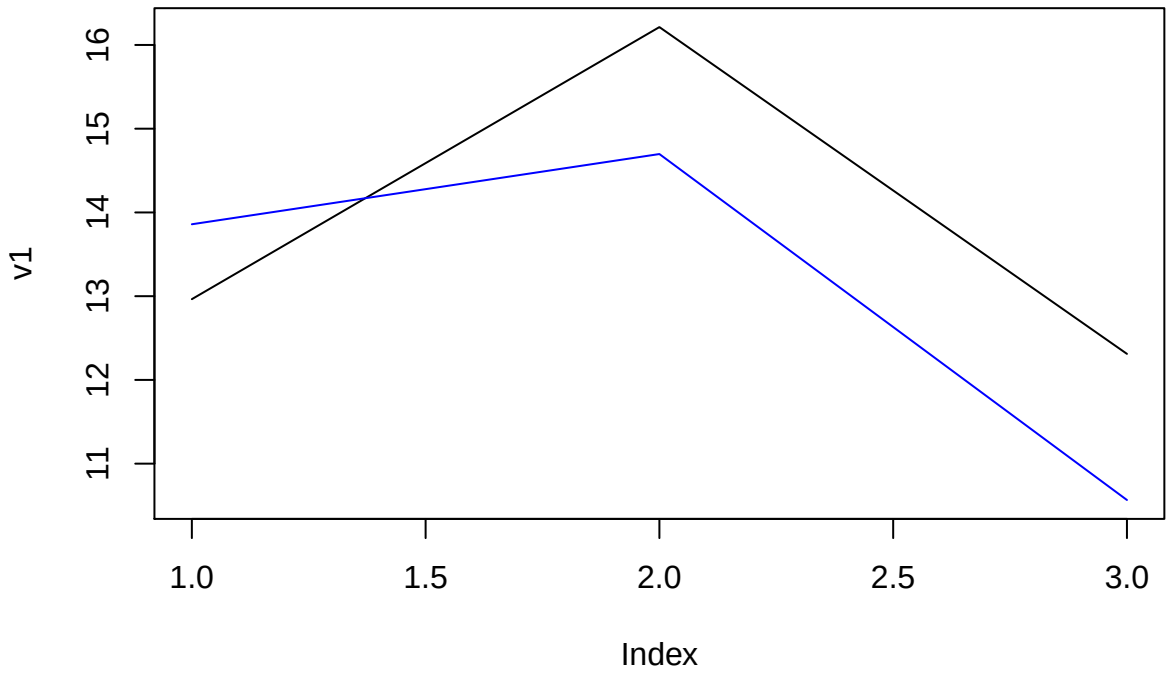
Para validar que, en realidad, estos grupos son distintos se ha aplicado el test **Hotelling T2 de permutaciones**:

$$H_0 : Grupo_1 = Grupo_2$$
$$H_1 : Grupo_1 \neq Grupo_2$$

Obteniendo un p-valor negligible (se muestra como 0 en el resultado del test) y por tanto podemos afirmar que existe evidencia estadísticamente significativa para rechazar H_0 y aceptar la hipótesis alternativa que nos dice que los grupos son diferentes.

Llegados a este punto se decide hacer un discriminante lineal para separar nuestras muestras. Hemos comprobado que este discriminante es un buen clasificador pues la tasa de error es muy baja, 0.0720461

Usando este mismo estimador para predecir el valor de los datos en los siguientes dias conseguimos una tasa de exito de 99




```
##
  |
  |                                                                      |   0%
  |
  |..                                                                    |   3%
  |
  |...                                                                   |   5%
  |
  |.....                                                                 |   8%
  |
  |.......                                                               |  11%
  |
  |.........                                                             |  13%
  |
  |...........                                                           |  16%
  |
  |.............                                                         |  18%
  |
  |...............                                                       |  21%
  |
  |.................                                                     |  24%
```

```
|
|.................                                                  |  26%
|
|...................                                                |  29%
|
|.....................                                              |  32%
|
|.......................                                            |  34%
|
|.........................                                          |  37%
|
|...........................                                        |  39%
|
|.............................                                      |  42%
|
|...............................                                    |  45%
|
|.................................                                  |  47%
|
|...................................                                |  50%
|
|.....................................                              |  53%
|
|.......................................                            |  55%
|
|.........................................                          |  58%
|
|...........................................                        |  61%
|
|.............................................                      |  63%
|
|...............................................                    |  66%
|
|.................................................                  |  68%
|
|...................................................                |  71%
|
|.....................................................              |  74%
|
|.......................................................            |  76%
|
|.........................................................          |  79%
|
|...........................................................        |  82%
|
|.............................................................      |  84%
|
|...............................................................    |  87%
|
|.................................................................  |  89%
|
|...................................................................|  92%
|
|.....................................................................|  95%
```

```
  |
  |.....................................................  |  97%
  |
  |.....................................................| 100%
## [1] "FINAL_WORK - BREAST_CANCER.R"
```