Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación

Máquinas de aprendizaje

Reporte: Diagnóstico de cáncer de mama

Docente: Abraham Sánchez López

Alumno	Matrícula
Taisen Romero Bañuelos	202055209

K-NN y el cáncer de mama

Haciendo memoria, K-NN es un algoritmo que consiste en agrupar conjuntos de datos según las similitudes que existan entre ellos. Para el caso de los datos sobre el cáncer de mama contamos con el radio del tumor/bulto, su textura, perímetro, área, suavidad, compacidad, concavidad, puntos cóncavos, simetría y dimensión fractal (aunque no sepa que tienen que ver los fractales en esto). Una manera sencilla de asociar alguna de estas características con el valor categórico de "maligno" es pensar en que un tumor maligno crece y en consecuencia su tamaño aumenta, por lo que la relación tamaño y malignidad es sencilla de ver.

Metodología

Como siempre, el primer paso en esta clase de trabajos es la exploración y tratamiento de los datos. Si visualizamos los datos notaremos que incluye una columna llamada ID, que al ser simplemente un identificador de paciente no resulta relevante para nuestro análisis, así que habrá que eliminar esa columna porque además, puede alterar los resultados haciéndolos menos precisos.

```
> str(wbcd)
'data.frame':   569 obs. of  31 variables:
 $ diagnosis        : Factor w/ 2 levels "Benign","Malignant": 1 1 1 1 1 1 1 2 1 1 ...
 $ radius_mean      : num  12.3 10.6 11 11.3 15.2 ...
 $ texture_mean     : num  12.4 18.9 16.8 13.4 13.2 ...
 $ perimeter_mean   : num  78.8 69.3 70.9 73 97.7 ...
 $ area_mean        : num  464 346 373 385 712 ...
 $ smoothness_mean  : num  0.1028 0.0969 0.1077 0.1164 0.0796 ...
 $ compactness_mean : num  0.0698 0.1147 0.078 0.1136 0.0693 ...
 $ concavity_mean   : num  0.0399 0.0639 0.0305 0.0464 0.0339 ...
 $ points_mean      : num  0.037 0.0264 0.0248 0.048 0.0266 ...
 $ symmetry_mean    : num  0.196 0.192 0.171 0.177 0.172 ...
 $ dimension_mean   : num  0.0595 0.0649 0.0634 0.0607 0.0554 ...
 $ radius_se        : num  0.236 0.451 0.197 0.338 0.178 ...
 $ texture_se       : num  0.666 1.197 1.387 1.343 0.412 ...
 $ perimeter_se     : num  1.67 3.43 1.34 1.85 1.34 ...
 $ area_se          : num  17.4 27.1 13.5 26.3 17.7 ...
 $ smoothness_se    : num  0.00805 0.00747 0.00516 0.01127 0.00501 ...
 $ compactness_se   : num  0.0118 0.03581 0.00936 0.03498 0.01485 ...
 $ concavity_se     : num  0.0168 0.0335 0.0106 0.0219 0.0155 ...
 $ points_se        : num  0.01241 0.01365 0.00748 0.01965 0.00915 ...
 $ symmetry_se      : num  0.0192 0.035 0.0172 0.0158 0.0165 ...
 $ dimension_se     : num  0.00225 0.00332 0.0022 0.00344 0.00177 ...
 $ radius_worst     : num  13.5 11.9 12.4 11.9 16.2 ...
 $ texture_worst    : num  15.6 22.9 26.4 15.8 15.7 ...
 $ perimeter_worst  : num  87 78.3 79.9 76.5 104.5 ...
 $ area_worst       : num  549 425 471 434 819 ...
 $ smoothness_worst : num  0.139 0.121 0.137 0.137 0.113 ...
 $ compactness_worst: num  0.127 0.252 0.148 0.182 0.174 ...
 $ concavity_worst  : num  0.1242 0.1916 0.1067 0.0867 0.1362 ...
 $ points_worst     : num  0.0939 0.0793 0.0743 0.0861 0.0818 ...
 $ symmetry_worst   : num  0.283 0.294 0.3 0.21 0.249 ...
 $ dimension_worst  : num  0.0677 0.0759 0.0788 0.0678 0.0677 ...
>
```

Así pues, quedan sólo las columnas esenciales. Ahora el siguiente aspecto a analizar es la suavidad de una variable con respecto a otra, por ejemplo, hay datos que varían en un rango siempre menor a cero, mientras hay otros que lo hacen en un rango mayor.

```
> summary(wbcd[c("radius_mean","area_mean","smoothness_mean")])
  radius_mean       area_mean      smoothness_mean
 Min.   : 6.981   Min.   : 143.5   Min.   :0.05263
 1st Qu.:11.700   1st Qu.: 420.3   1st Qu.:0.08637
 Median :13.370   Median : 551.1   Median :0.09587
 Mean   :14.127   Mean   : 654.9   Mean   :0.09636
 3rd Qu.:15.780   3rd Qu.: 782.7   3rd Qu.:0.10530
 Max.   :28.110   Max.   :2501.0   Max.   :0.16340
```

Para esta parte tenemos la opción de normalizar nuestros datos o escalarlos con la estandarización de la puntuación z. Para no hacer extenso el reporte iré directamente a la puntuación z.

La puntuación z tiene valores estandarizados sin un mínimo ni un máximo, por lo que los valores extremos se dirigen al centro. Entonces, así quedarían nuestros datos aplicando dicho reescalado.

```
> summary(wbcd_z$area_mean)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.4532 -0.6666 -0.2949  0.0000  0.3632  5.2459
>
```

Ahora lo que toca es aplicar el algoritmo k-nn y observar qué resultados nos da si cambiamos el valor de k, pero primero, estos son los resultados si seguimos el k sugerido (la raíz cuadrada del número de instancias, en este caso, nuestro conjunto de entrenamiento tiene 469, por lo que el k recomendado es 21).

```
> #k=21 (k sugerido)
> wbcd_test_pred<-knn(train = wbcd_train, test=wbcd_test, cl=wbcd_train_labels,k=21)
> CrossTable(x=wbcd_test_labels,y=wbcd_test_pred, prop.chisq = FALSE)


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


Total Observations in Table:  100


                 | wbcd_test_pred
wbcd_test_labels |    Benign | Malignant | Row Total |
-----------------|-----------|-----------|-----------|
          Benign |        61 |         0 |        61 |
                 |     1.000 |     0.000 |     0.610 |
                 |     0.924 |     0.000 |           |
                 |     0.610 |     0.000 |           |
-----------------|-----------|-----------|-----------|
       Malignant |         5 |        34 |        39 |
                 |     0.128 |     0.872 |     0.390 |
                 |     0.076 |     1.000 |           |
                 |     0.050 |     0.340 |           |
-----------------|-----------|-----------|-----------|
    Column Total |        66 |        34 |       100 |
                 |     0.660 |     0.340 |           |
-----------------|-----------|-----------|-----------|
```

Según la tabla, el número de tumores benignos que fueron detectados como malignos fue cero, en cambio, el de tumores malignos que fueron detectados como benignos fue de cinco.

Si variamos el valor de k obtenemos los siguientes resultados:

- k=1:

	wbcd_test_pred		
wbcd_test_labels	Benign	Malignant	Row Total
Benign	59	2	61
	0.967	0.033	0.610
	0.952	0.053	
	0.590	0.020	
Malignant	3	36	39
	0.077	0.923	0.390
	0.048	0.947	
	0.030	0.360	
Column Total	62	38	100
	0.620	0.380	

- k=5:

	wbcd_test_pred		
wbcd_test_labels	Benign	Malignant	Row Total
Benign	59	2	61
	0.967	0.033	0.610
	0.952	0.053	
	0.590	0.020	
Malignant	3	36	39
	0.077	0.923	0.390
	0.048	0.947	
	0.030	0.360	
Column Total	62	38	100
	0.620	0.380	

- k=11:

	wbcd_test_pred		
wbcd_test_labels	Benign	Malignant	Row Total
Benign	59	2	61
	0.967	0.033	0.610
	0.952	0.053	
	0.590	0.020	
Malignant	3	36	39
	0.077	0.923	0.390
	0.048	0.947	
	0.030	0.360	
Column Total	62	38	100
	0.620	0.380	

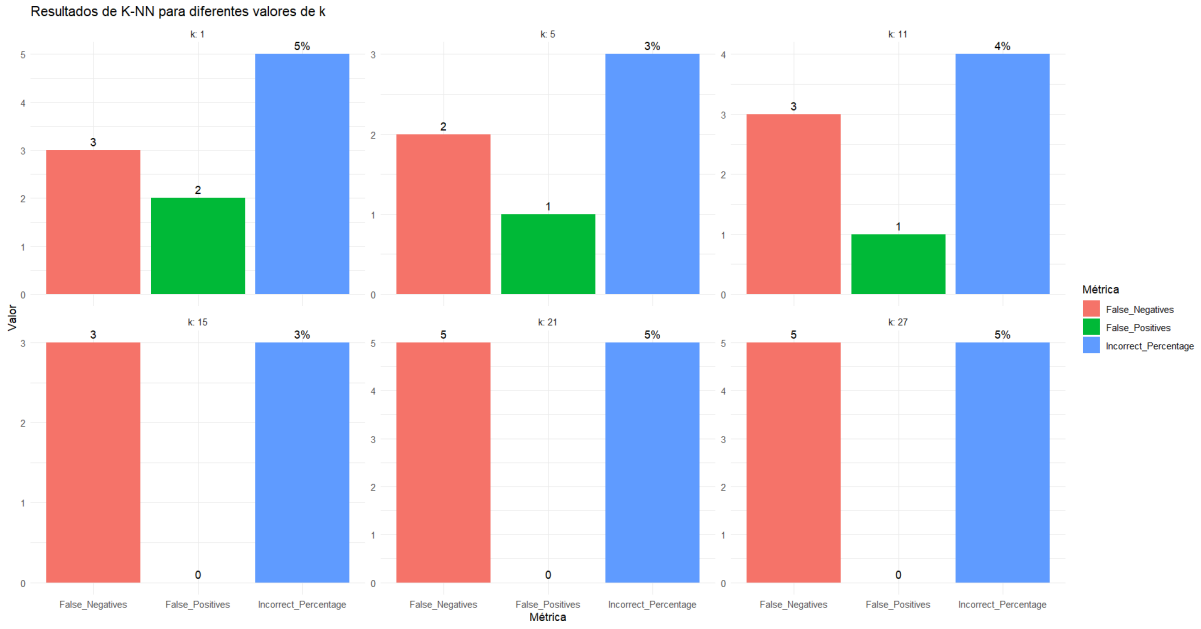
- k=15:

wbcd_test_labels	wbcd_test_pred Benign	Malignant	Row Total
Benign	61	0	61
	1.000	0.000	0.610
	0.953	0.000	
	0.610	0.000	
Malignant	3	36	39
	0.077	0.923	0.390
	0.047	1.000	
	0.030	0.360	
Column Total	64	36	100
	0.640	0.360	

- k=27:

wbcd_test_labels	wbcd_test_pred Benign	Malignant	Row Total
Benign	61	0	61
	1.000	0.000	0.610
	0.924	0.000	
	0.610	0.000	
Malignant	5	34	39
	0.128	0.872	0.390
	0.076	1.000	
	0.050	0.340	
Column Total	66	34	100
	0.660	0.340	

Ahora bien, una forma más bonita de ver los resultados es realizando un gráfico facetado como el siguiente.



Bueno, teniendo a la mano los resultados que obtuve es hora de compararlos con los del PDF.

Valor k	Falsos negativos	Falsos positivos	Porcentaje clasificado incorrectamente
1	1	3	4%
5	2	0	2%
11	3	0	3%
15	3	0	3%
21	2	0	2%
27	4	0	4%

Como podemos observar, mis resultados sólo coinciden con uno de los resultados de la tabla final del PDF.