Clasificación de imagenes de radiografias de tórax entre normales y con derrame

Machine Learning - PEC 1

Jorge Vallejo Ortega

10 de abril, 2020

Contents

Algoritmo k-NN

El algoritmo de los k vecinos más próximos (*k-nearest neighbours*) es un algoritmo de aprendizaje automático (*machine learning*) que se utiliza para clasificar observaciones, según si sus características las hacen más parecidas a uno u otro grupo ya establecidos.

En una primera fase de 'entrenamiento', una colección de observaciones ya clasificadas se distribuyen en un espacio n-dimensional. Cada dimensión corresponde a una de las variables medidas en las observaciones. Las nuevas observaciones, las cuales queremos clasificar, se distribuyen a su vez en ese espacio n-dimensional; y se clasifican dentro de los grupos a los que pertenezcan aquellas otras observaciones, ya clasificadas, de las que más cerca se encuentren. El número de observaciones conocidas que tenemos en cuenta para clasificar las observaciones nuevas, es ése número k.

Fortalezas y debilidades del algoritmo

Fortalezas	Debilidades
Simple y efectivo	No produce un modelo, limitando la habilidad para entender cómo las características se relacionan con la clasificación
No hace asunciones acerca de la distribución subyacente de los datos	Requiere seleccionar una k adecuada
Fase de entrenamiento rápida	Fase de clasificación lenta
	Se requiere procesado adicional para características nominales y datos incompletos

Pre-procesado de datos

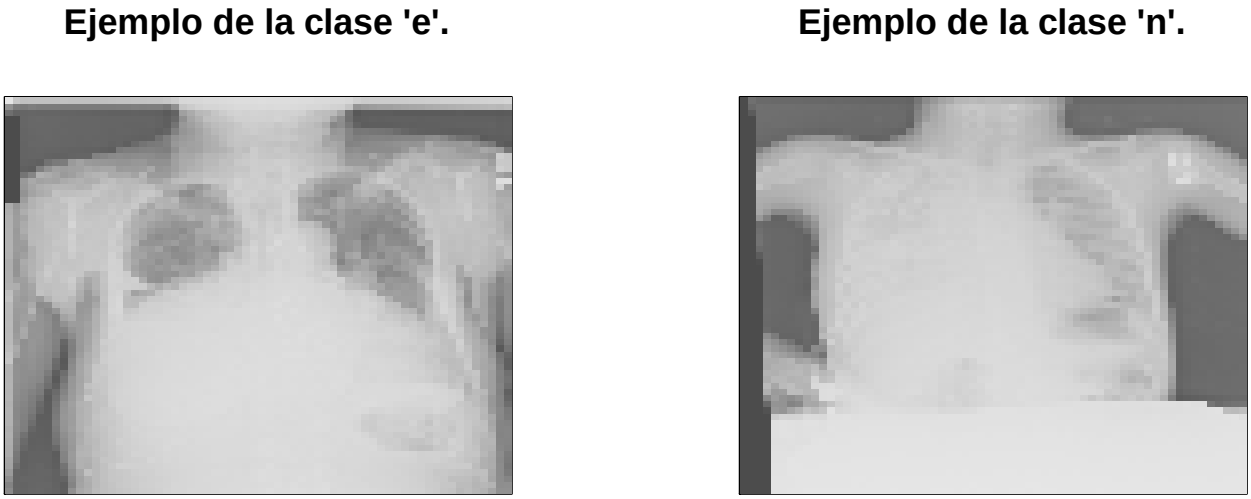
Estructura de los datos

El set de datos examinado está compuesto por 1000 observaciones.
De cada observación se han tomado 4096 variables.
El conjunto de observaciones está dividido en 2 clases (effusion y normal), codificadas como e y n.

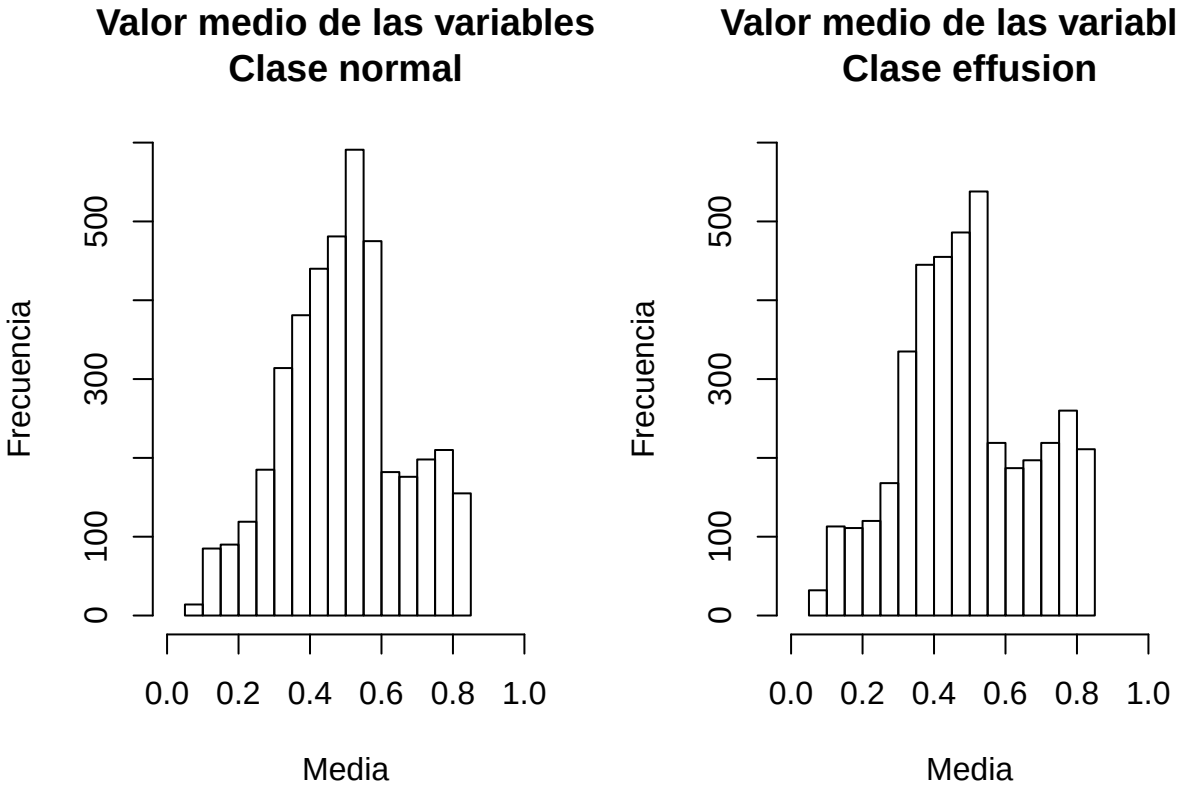
La distribución de cada clase es la siguiente:

Clase	Frecuencia
e	500
n	500

Ejemplos de observaciones de las diferentes clases

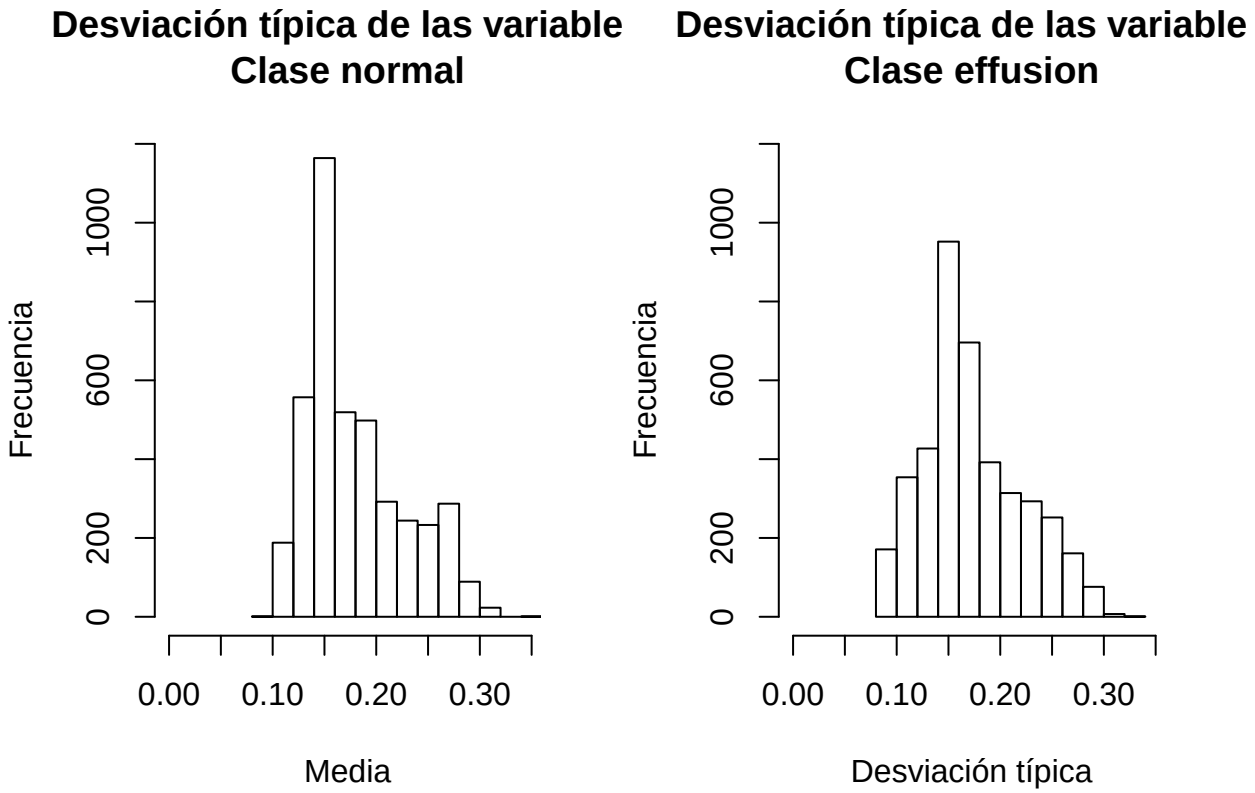


Histogramas de las medias



Los valores medios de las variables en la clase “normal” parecen estar más concentrados en el centro de la distribución, mientras que en la clase “effusion” la distribución es más achatada. Esto podría significar que las imágenes de la clase “effusion” presentan áreas más claras y más oscuras que las imágenes de la clase “normal”.

Histogramas de las desviaciones típicas



Por la forma de los histogramas, y los rangos de valores en los que se mueven, parece que las variales en las observaciones de clase "effusion" presentan valores de desviación típica mayores que las variables en la clase "normal".

Contraste de valores medios (t de Student)

Utilizando el test de la t de Student, hemos comparado los descriptores de ambos grupos de observaciones. Una vez ajustados los p-valores según el método BH, para tener en cuenta que estamos haciendo comparaciones múltiples, podemos ver en la siguiente tabla los 25 descriptores estadísticamente más significativos y la diferencia entre las medias de ambos grupos para cada descriptor.

Table 3: Tabla con los 25 descriptores de menor p-valor

	p-valores	Diferencia
V3052	5.255e-21	0.1109
V3117	7.271e-21	0.1128
V3116	1.519e-20	0.1115
V2989	1.729e-20	0.1052
V1385	6.037e-20	0.1162
V3051	6.336e-20	0.1092
V2988	6.466e-20	0.1059
V1384	6.707e-20	0.1140
V2986	6.732e-20	0.1065
V3053	6.961e-20	0.1079
V1387	1.024e-19	0.1196

	p-valores	Diferencia
V2987	1.218e-19	0.1048
V1193	1.366e-19	0.1118
V2924	1.702e-19	0.0993
V1513	2.379e-19	0.1128
V1321	3.233e-19	0.1133
V1256	3.676e-19	0.1100
V1192	3.928e-19	0.1094
V2925	4.432e-19	0.0990
V1450	4.991e-19	0.1155
V3050	5.124e-19	0.1070
V1451	6.615e-19	0.1153
V1515	7.882e-19	0.1136
V3115	8.153e-19	0.1070
V1579	9.359e-19	0.1096

Me llama la atención que en estos descriptores estadísticamente más significativos, la diferencia entre las medias del grupo “derrame”" y del grupo “normal” es de alrededor del 20%, que no parece especialmente grande.

Mapa de significatividad

A partir de los p-valores corregidos, hemos generado un gráfico en el que quedan señaladas lás áreas más significativas del conjunto de imágenes a la hora de comparar los diferentes grupos entre ellos. En escala de grises, los píxeles con un p-valor más alto (menos significativo) aparecen más oscuros; y aquellos con p-valor más bajo (más significativos) aparecen más claros:

Exploración con el algoritmo k-NN

Antes de aplicar el algoritmo hemos dividido el set de datos, al azar, en un set de entrenamiento (67% de las observaciones) y un set de prueba (33% de los datos).

El set de entrenamiento (670 observaciones en este caso) es la referencia que usa el algoritmo para realizar la clasifición de las observaciones.

El set de prueba (330 observaciones en este caso) sirve para evaluar la capacidad del algoritmo para clasificar correctamente cada observación dentro del grupo que le corresponde.

Evaluación de diferentes valores de *k*

k	Falsos Negativos	Falsos Positivos	Error Total
3	17%	21.2%	38.2%
5	14.5%	19.4%	33.9%
7	13%	22.4%	35.5%
11	13.9%	20.9%	34.8%
23	13.3%	20.6%	33.9%
45	13.6%	20.6%	34.2%
67	13.3%	21.2%	34.5%

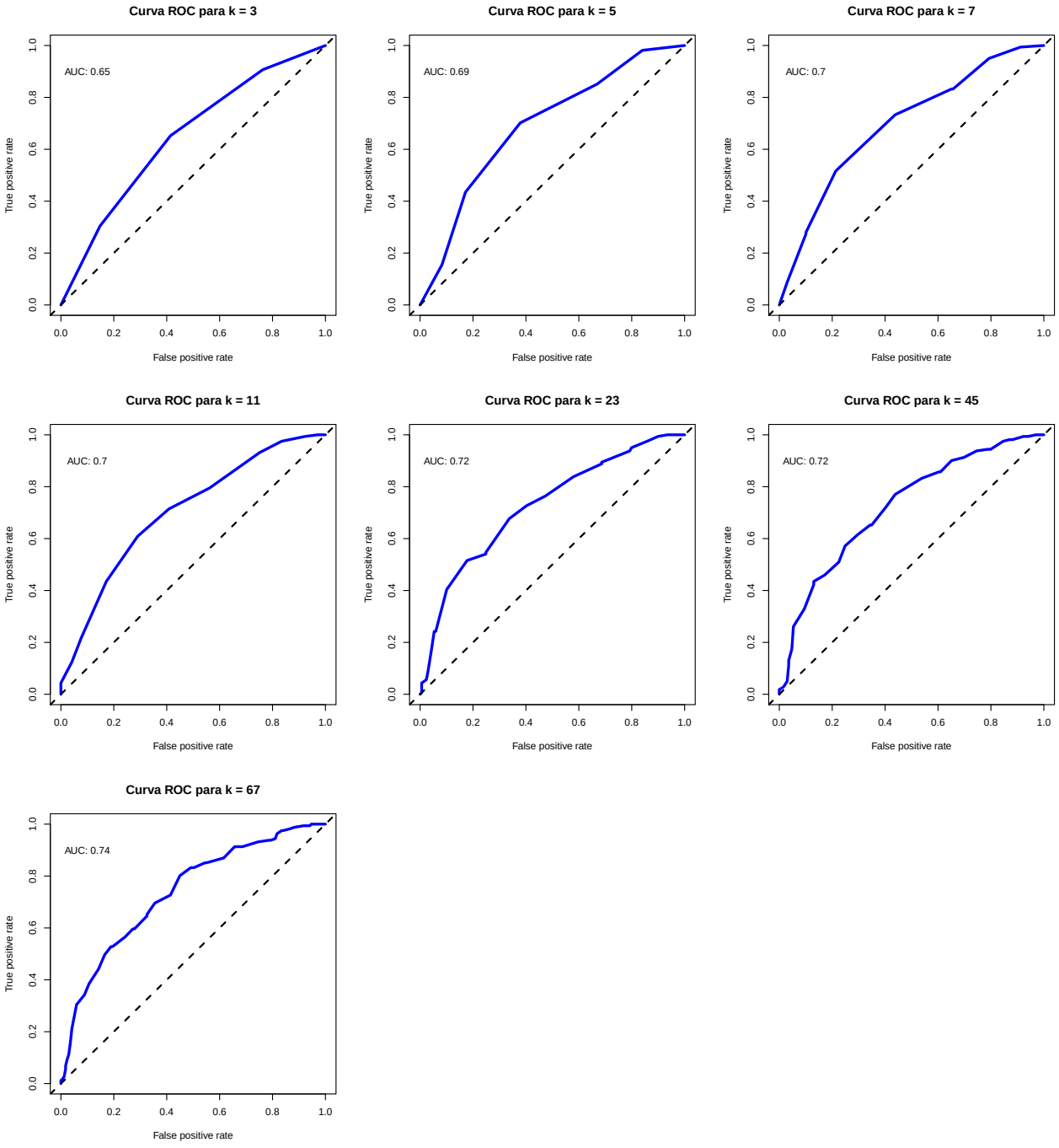
Si priorizamos los falsos negativos, que son lo que potencialmente suponen un mayor perjuicio para el paciente, el k más adecuado sería 7; que nos proporciona un valor del 13% de falsos positivos.

Análisis de rendimiento. Curvas ROC y AUC

Las curvas ROC (Receiver Operating Characteristic) son un método de análisis de rendimiento que se usa para determinar el punto de equilibrio entre la capacidad de detectar positivos auténticos, y la de evitar falsos positivos.

Un clasificador perfecto detectaría todos los positivos antes de detectar ningún falso negativo. Una forma de medir cuánto se acerca el rendimiento de nuestro clasificador al del clasificador perfecto es mediante el estadístico AUC (area under the ROC curve).

En los siguientes gráficos están representadas tanto las curvas ROC como los valores AUC para cada valor de k:



Comentario

Una vez calculadas las curvas ROC y los valores de AUC, podemos ver que el mayor AUC (0.74) corresponde al valor de k = 67.

El valor k para el mínimo de falsos negativos (13%) corresponde a k = 7.

El valor k para el mínimo de falsos positivos (19.4%) corresponde a k = 5.

El valor k para el error total mínimo (33.9%) es 5.

Dado que el error potencialmente más perjudicial para el paciente son los falsos negativos, deberíamos elegir el valor de k = 7. Este resulta en un falso negativo del 13%, falso positivo del 22.4% y AUC de 0.7.

Apéndice A: Reproducibilidad

```
sessionInfo() # For better reproducibility
```

```
## R version 3.6.3 (2020-02-29)
## Platform: i686-pc-linux-gnu (32-bit)
## Running under: Ubuntu 18.04.4 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/i386-linux-gnu/blas/libblas.so.3.7.1
## LAPACK: /usr/lib/i386-linux-gnu/lapack/liblapack.so.3.7.1
##
## locale:
##  [1] LC_CTYPE=es_ES.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=es_ES.UTF-8        LC_COLLATE=es_ES.UTF-8
##  [5] LC_MONETARY=es_ES.UTF-8    LC_MESSAGES=es_ES.UTF-8
##  [7] LC_PAPER=es_ES.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=es_ES.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ROCR_1.0-7       gplots_3.0.3     gmodels_2.18.1   class_7.3-16
## [5] OpenImageR_1.1.6 knitr_1.28
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.3         magrittr_1.5       MASS_7.3-51.5      xtable_1.8-4
##  [5] R6_2.4.1           jpeg_0.1-8.1       rlang_0.4.2        fastmap_1.0.1
##  [9] highr_0.8          stringr_1.4.0      caTools_1.17.1.3   tools_3.6.3
## [13] grid_3.6.3         xfun_0.11          png_0.1-7          KernSmooth_2.23-16
## [17] gtools_3.8.1       htmltools_0.4.0    yaml_2.2.0         digest_0.6.23
## [21] shiny_1.4.0.2      later_1.0.0        bitops_1.0-6       promises_1.1.0
## [25] evaluate_0.14      mime_0.8           rmarkdown_2.0      tiff_0.1-5
## [29] gdata_2.18.0       stringi_1.4.3      compiler_3.6.3     httpuv_1.5.2
```