Inteligencia de Negocio: Práctica 1
Resolución de problemas de clasificación y análisis experimental

David Cabezas Berrido

Grupo 2: Viernes

dxabezas@correo.ugr.es

21 de octubre de 2020

Índice

1. Introducción

En esta primera práctica abordaremos el problema de decidir si un tumor en una mamografía es benigno o maligno, por lo que deducimos que se trata de un problema de clasificación binaria (dos clases: benigno y maligno). Disponemos para ello de datos de 961 pacientes, para cada uno de ellos se han medido 6 atributos entre cualitativos y cuantitativos.

Sobre este problema real, pondremos a prueba los distintos algoritmos estudiados en teoría y las herramientas de prácticas. Trataremos de mejorar los resultados mediante un procesado básico de los datos y probando distintos hiperparámetros en los algoritmos. Finalmente compararemos los algoritmos entre sí y discutiremos cuál es el más adecuado para el problema.

Para la experimentación usaremos validación cruzada de 5 particiones. La matriz de confusión se calculará como la suma de las matrices de confusión de cada partición, de esta forma la matriz resultante tendrá en cuenta cada instancia una sola vez. Como usamos validación cruzada, nos ahorramos tener que separar un conjunto de test para pruebas y aprovechamos cada instancia tanto para entrenamiento como para evaluación.

Empezamos analizando el dataset. Como ya hemos comentado, consta de 961 instancias y cuenta con 6 atributos:

1. Código BI-RADS: se trata de un número entero entre 0 y 6 (ambos incluidos) asignado por un radiólogo tras interpretar la mamografía. Un mayor valor significa una mayor probabilidad de malignidad, a excepción del valor 0, que indica que la información de la radiografía es insuficiente. Tendremos que tener esto en cuenta, ya que puede "confundir" a algunos algoritmos como el KNN, que interpretaría que un código 0 está más próximo a un código 1 que a un código 4, lo cual no tiene sentido a priori (sin atender a más características).
2. Edad del paciente: entero positivo.
3. Forma del tumor: nominal, 4 posibles formas distintas (R, O, L, I) y N para indicar que la forma no está definida.
4. Margen de la masa: nominal, 5 posibles valores (del 1 al 5).
5. Densidad de la masa: entero entre 1 y 4 (ambos inclusive), un menor valor indica mayor densidad.
6. Severidad: variable objetivo a predecir, benigno o maligno.

Hay algunos datos perdidos en el dataset:

BI-RADS	2
Age	5
Shape	0
Margin	48
Density	76
Severity	0

En principio, no hay los suficientes datos perdidos como para dejar de considerar alguna variable, pero durante el procesado de datos discutiremos si eliminar alguna o cómo imputar los datos perdidos.

De los 961 instancias, 445 pertenecen a la clase maligno (46.3 %) y 516 a benigno (53.7 %). Las clases están bastante balanceadas, lo que en genral convierte a la accuracy en una métrica bastante adecuada para la bondad de los algoritmos de clasificación. Sin embargo, en este problema concreto, parece mucho más grave un falso negativo (predecimos benigno y enviamos a un paciente a casa que debería empezar a tratarse) que un falso positivo (predecimos maligno y el paciente irá a análisis posteriores donde probablemente se percaten del error); por lo que puede ser interesante considerar otras métricas de evaluación.

2. Procesado de datos

3. Configuración de algoritmos

4. Resultados obtenidos

5. Análisis de resultados

6. Interpretación de resultados

7. Contenido adicional

8. Bibliografía

http://faculty.marshall.usc.edu/gareth-james/ISL