# **Selección de clasificadores con costes de mala clasificación diferentes**

### Medida del rendimiento de un clasificador.

Como es sabido, la *matriz de confusión* se expresa en el caso de que solo existan dos clases como:



Tabla : Matriz de Confusión y Métricas

Las medidas más inmediatas de evaluación del modelo que se nos ocurren son:

* Exactitud:
* Tasa de error:

Sin embargo, cuando existe un claro desequilibrio entre clases, como es el caso de los tuits de odio que pueden ser un 2 ‰ del total, es posible obtener una exactitud enorme si clasificamos todos los tuits (erróneamente) como negativos (no de odio):



Tabla : Matriz de Confusión en Conjuntos Desequilibrados

Lo que nos daría unos valores de exactitud del 99%:



Por ello, se utilizan en estos casos los indicadores clásicos en clasificación binaria:

1. Precisión (*p*recision)
2. Exhaustividad (*r*ecall)

* Precisión (p) es el porcentaje de los *tuits* clasificados como pertenecientes a la clase que verdaderamente son de dicha clase (aciertos).
* Exhaustividad (r) es el porcentaje de los *tuits* de la clase en cuestión existentes en el conjunto que han sido clasificados correctamente.

Obsérvese que en el primer caso el denominador está constituido por el número de *tuits clasificados* en la clase + y en el segundo los *realmente existentes*.

Un valor p=1 nos dice que todos los elementos recuperados como relevantes, lo son, pero no nos dice nada acerca de si hemos recuperado todos los documentos relevantes (r).

En el ejemplo de los tuits, estos valores son nulos:



Ambos indicadores se combinan equilibradamente mediante su *media armónica* en F:

## Costes de mala clasificación. ROC y AUC.

Una cuestión importante es cómo *equilibrar los falsos positivos y los falsos negativos* cuestión que, evidentemente, depende del *coste de las consecuencias de cada error*.

Para esto puede utilizarse una herramienta que se está usando cada vez más: el ***ROC[[1]](#footnote-1)*** habitual en Medicina y Biología donde las consecuencias de la detección de falsos positivos y negativos puede tener costes muy diferentes.

Ahora a la exhaustividad - - se la denomina ***sensibilidad****.* Como se ve*,* es la

y se introduce la ***especificidad****:*

De manera que

que es la *proporción de falsos positivos* ***fp****:*

Por el teorema de la probabilidad total sabemos que:

|  |  |
| --- | --- |
|  | Ec. ( 2) |

Si dibujamos el gráfico que relaciona ambas magnitudes, obtenemos la Fig. 3 en que AUCes el *área bajo la curva* que puede tomar valores entre 0 (no acierta nunca) y 1 (la predicción acierta siempre).



Fig. 3: ROC

La Tabla 1 pone de manifiesto que mientras que los indicadores usuales (*p*, *r*) usan valores de ambas filas de la matriz, los de ROC usan valores de la misma fila.

Por tanto, en el primer caso cuando cambian las proporciones de las clases estas medidas cambian aunque no lo haga el rendimiento del clasificador, cosa que no sucede en ROC como se ve en la Tabla 3 al cambiar la proporción de clases.



Tabla 3: Métrica ROC frente a Métrica Convencional

## Clasificadores discretos y por puntuación (‘scoring’).

Un clasificador mapea las *instancias* en *clases.*

Existen dos tipos de clasificadores según el tipo de salida que proporcionan:

1. discreta: proporcionan solo una etiqueta que indica la clase predicha para la instancia (p.e. árboles de decisión) y
2. continua: estimación de la *probabilidad* (entre 0 y 1) de pertenecer a una clase o un *puntaje*. La asignación a una u otra clase se hace depender de un *umbral* de corte para dicho puntaje. Los métodos *bayesianos* y *redes neuronales* pertenecen a este segundo tipo.

Un clasificador *discreto* proporciona solo *un punto* en el diagrama ROC y *una sola matriz de confusión,* mientras que otro que funcione *por puntuación* proporcionará un *punto* y una *matriz de confusión* diferente *para cada umbral*.



Fig. 13: Matrices de Confusión para Umbrales Distintos

En efecto, si llamamos tendremos que

|  |  |
| --- | --- |
|  | Ec. (3) |

|  |  |
| --- | --- |
|  | Ec. (4) |

Ecuaciones paramétricas de la curva ROC al variar los valores del umbral.

Entre Ec. (3) y Ec. (4) podemos eliminar θ y obtenemos la ecuación de la *ROC*:

Si planteamos lo anterior en términos de *contraste de hipótesis* para contrastar que una instancia pertenece a la clase +,

* *fp* sería el error tipo I y
* *tp* el error tipo II.

Un problema frecuente es el de *seleccionar el umbral de clasificación* que, cuando no se toman en cuenta costes diferentes, puede enfocarse bien maximizando *F* [Ec. (1)] o utilizando el ROC de manera que la *AUC* *sea lo mayor posible*.

*Si los costes son diferentes e identificables*, cabe construir una curva de costes a partir de la ROC y seleccionar el umbral de coste mínimo (Fig. 14).



Tabla : ROC y Curva de Costo

Para ello tendremos en cuenta que:

No obstante, es difícil definir una relación de costes explícita y, por ello, una alternativa es examinar las matrices de confusión de umbrales diferentes y seleccionar entre ellos aquel que más nos interese.

P.e. entre los resultados de un clasificador con dos umbrales distintos y con las siguientes matrices de confusión:



Tabla 5: Comparación de Umbrales

Es probable que un usuario que dé más importancia a la exhaustividad - recuperar el máximo de elementos de la clase + - que a la precisión elija el umbral *b* con mayor *exhaustividad*. Como se ve esto se hace a costa de un aumento de falsos positivos que, por consiguiente, tienen para el usuario un coste menor que los falsos negativos.

1. Receiver Operating Characteristics. [↑](#footnote-ref-1)