## **Selección de Conjuntos de Entrenamiento**

### 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:



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):



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



Por ello, se utilizarán los indicadores clásicos en clasificación binaria:

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

Y, si es posible, el *ROC.*

* Precisión (p) es el porcentaje de los *tuits* clasificados como pertenecientes al tópico Economía que restaban asignados a dicho tópico (aciertos).
* Exhaustividad (r) es el porcentaje de los *tuits* con tópico=Economía existentes en el fichero TASS\_Test que han sido etiquetados correctamente en TASS\_Tagged\_Test.

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:



La Figura 4 muestra gráficamente estos valores.

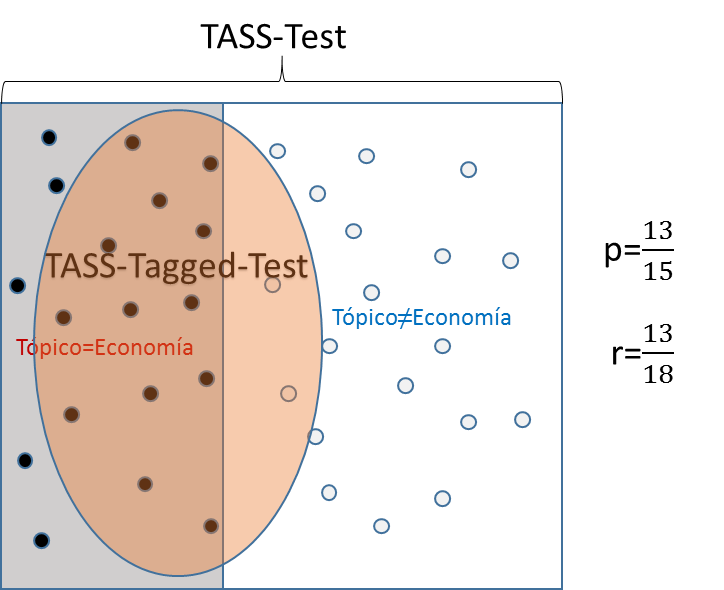


Figura 1

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

Otro indicador que se está utilizando cada vez más es el *ROC[[1]](#footnote-1)* habitual en Medicina y Biología para hablar de la detección de falsos positivos y negativos.

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

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

De manera que

Y es

Por el teorema de la probabilidad total sabemos que:

Si dibujamos el gráfico que relaciona ambas magnitudes, obtenemos la Figura 4 en que el *ROC* es el *área bajo la curva* que puede tomar valores entre 0 (no acierta nunca) y 1 (la predicción acierta siempre).

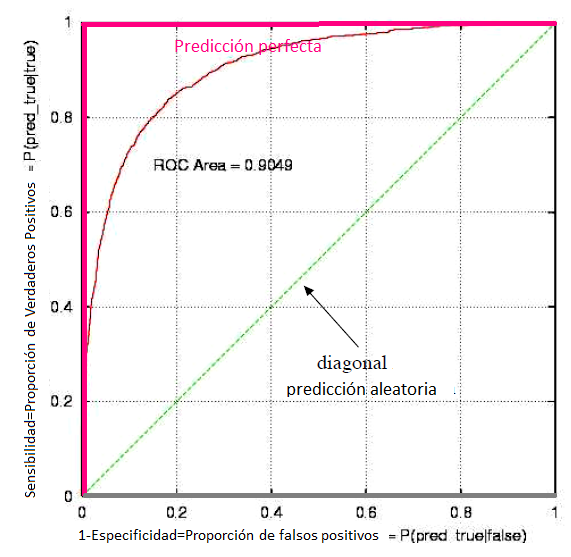


Figura 2

Lo mejor es proporcionar la ***matriz de confusión*** y cada lector del informe puede calcular el indicador que desee.

### Comparación de algoritmos.

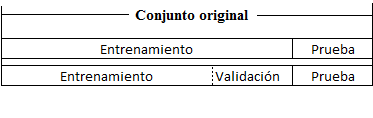
Una vez que hemos *aprendido* varios modelos y queremos seleccionar el mejor, normalmente usamos alguna de las medidas descritas anteriormente y elegimos el modelo que proporcione mejor rendimiento.

### Comparación de dos conjuntos de datos.

Si dividimos el conjunto de datos en un conjunto de *entrenamiento* y otro de *prueba* esperamos que *la distribución de los datos sea la misma en ambos*. Se usan los contrastes de Kolmogorov-Smirnov o χ2 para la clasificación.

### División de conjuntos de datos.

Lo ideal es recoger *varios conjuntos de* *datos independientes*, si ello no es posible, debemos conformarnos con un solo conjunto de datos que habremos de dividir en dos o tres conjuntos.

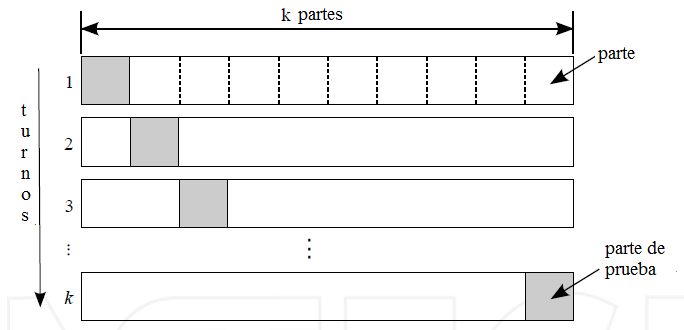


La estrategia consistente en usar

1. un *conjunto de entrenamiento* para aprender y estimar los parámetros del modelo;
2. un *conjunto de validación* para evaluar modelos y seleccionar uno de ellos y
3. un *conjunto de prueba* o test para valorar la capacidad de predicción de los modelos.

Existen múltiples métodos de división de datos, de los cuales el más simple es el *método de retención* (holdout) que consiste en dividir aleatoriamente el conjunto original en dos subconjuntos ( o para entrenamiento y el resto para prueba). Si el conjunto original no es lo bastante grande, el método es ineficiente.

El método más popular es el de *validación cruzada* p.e. en la *validación cruzada en k iteraciones* (k-fold), el conjunto original se divide en k conjuntos disjuntos del mismo tamaño. En cada uno de los *k* turnos una parte se usa para evaluación y las *k-1* restantes para aprendizaje. La exactitud resultante es un promedio de cada una de las *k* obtenidas.



### Selección de instancias.

Se trata de un procedimiento de reducir el conjunto original.

Existen dos métodos principales:

* métodos *wrapper* de los cuales el método CNN (Condensed Nearest Neighbour) es el primero en el tiempo. Se trata de un método incremental que comienza con un nuevo conjunto *R*, compuesto por una instancia por clase elegida aleatoriamente del conjunto inicial *S*. En el siguiente paso, clasifica *S* usando *R* como conjunto de entrenamiento. Después de esta clasificación, cada instancia de *S* clasificada mal (FP o FN) se agrega a R. CNN selecciona las instancias cercanas a la frontera de decisión.
* Métodos de filtrado. Se basan en algún algoritmo que selecciona instancias basadas en un vector de atributos.

### Desequilibrios de clase.

Como hemos visto, y es el caso de los tuits, las dos clases están muy desequilibradas: hay muchísimos más tuits neutros que con contenido de odio.

1. Métodos de nivel de datos. Buscan aumentar el número de instancias de la clase minoritaria (sobremuestreo) y/o reducir los de la clase mayoritaria (submuestreo).
2. Métodos de nivel de algoritmo. Se basan sobre todo en dar una sobreponderación a la clase mayoritaria.

1. Receiver Operator Characteristic. [↑](#footnote-ref-1)