

Métricas

Aprendizaje Automático

Hasta ahora, en prácticas hemos utilizado para cuantificar la bondad del modelo funciones sencillas, como el MSE en problemas de regresión o la precisión en problemas de clasificación. Si bien en problemas de regresión las funciones se basan en un error calculado de una u otra manera (error medio, error cuadrático medio, etc.), en problemas de clasificación se pueden derivar otro tipo de métricas en función de cómo es el problema que se está resolviendo. Muchas de estas métricas, al menos las que se van a usar en prácticas, se basan en el cálculo de la matriz de confusión.

Una matriz de confusión es una matriz cuadrada, con tantas filas y columnas como clases, donde se muestra la distribución de los patrones en clases, y la clasificación que realiza el modelo. Habitualmente en las filas se pone cómo ha efectuado el modelo la clasificación, y en las columnas los valores de clasificación real, aunque esto pueda variar según la fuente que se consulte.

El caso más sencillo se corresponde con 2 clases, en las que una se considera “negativa” y otra “positiva”. Una matriz de confusión de dos clases sería la siguiente:

		Predicción	
		Negativo	Positivo
Real	Negativo	VN	FP
	Positivo	FN	VP

Esta matriz de confusión contiene 4 valores, que podemos dividir

- Según la salida que da el modelo: positivos o negativos.
- Según si el modelo se equivoca o no: verdaderos o falsos.

De esta forma, estos 4 valores reciben por nombre verdaderos negativos (VN), falsos positivos (FP), falsos negativos (FN) y verdaderos positivos (VP). Por ejemplo, los falsos negativos sería la cantidad de patrones que el sistema ha clasificado en negativo, y se ha equivocado porque en realidad eran positivos.

A partir de esta matriz de confusión se pueden calcular distintas métricas. Dependiendo del problema en que se esté trabajando, será más interesante seguir una u otra. Algunas de las métricas más usadas son:

- Precisión (*accuracy*, no confundir con el término inglés *precision*). Ratio de patrones en los que acierta en la predicción. Se calcula como $(VN + VP)/(VN + VP + FN + FP)$
- Tasa de error (*error rate*). Ratio de patrones en los que falla en la predicción. Se calcula como $(FN + FP)/(VN + VP + FN + FP)$
- Sensibilidad (*recall* o *sensitivity*). Indica la probabilidad de que para un caso positivo se obtenga en el clasificador un resultado positivo. Se calcula como $VP/(FN+VP)$
 - En una prueba médica, la sensibilidad de la prueba representa la probabilidad de que un sujeto enfermo (positivo) tenga un resultado positivo en la prueba.
- Especificidad (*specificity*). Indica la probabilidad de que para un caso negativo se obtenga en el clasificador un resultado negativo. Se calcula como $VN/(FP+VN)$
 - La especificidad de una prueba representa la probabilidad de que un sujeto sano (negativo) tenga un resultado negativo en la prueba.
- Valor predictivo positivo (*precision* o *positive predictive value*). Ratio de casos con valor positivo que han sido correctamente clasificados. Se calcula como $VP/(VP+FP)$.
- Valor predictivo negativo (*negative predictive value*). Ratio de casos con valor negativo que han sido correctamente clasificados. Se calcula como $VN/(VN+FN)$.
- F_1 -score, se define como la media **armónica** entre VPP y sensibilidad.

Es conveniente aclarar que estas métricas, así como otras vistas en clase de teoría (curva ROC, índice Kappa) se utilizan para valorar clasificadores ya entrenados, no para realizar el proceso de entrenamiento. Para entrenarse, cada modelo tiene su propia función para cuantificar el error o la bondad, como puede ser la función *cross-entropy* en el caso de las redes de neuronas.

Posiblemente, el valor más utilizado es el de precisión (*accuracy*), puesto que indica, de una forma sencilla, la tasa de éxito del clasificador. Sin embargo, dependiendo del problema con el que se está trabajando, podría no ser la métrica más adecuada. Por ejemplo, en un test masivo en una población sobre una enfermedad en el que se sabe que la mayoría de la

gente no padece esa enfermedad, un modelo que clasifique a todas las personas como negativo (sano) tendrá una precisión muy alta, aunque el modelo realmente no hace nada.

Por este motivo, es necesario valorar, entre todas estas métricas, cuál o cuáles son las más utilizadas, en función del problema. En muchos problemas en los que las distintas clases tienen la misma importancia, el valor de precisión puede ser suficiente, dependiendo del problema. Sin embargo, en otros problemas puede interesar más evaluar las situaciones en las que se produce o debería producir una respuesta positiva por parte del modelo, puesto que podría indicar algo crítico, como detectar una enfermedad o dar algún tipo de alarma. Por este motivo, muchas veces se tienen en cuenta, aparte de la precisión, los valores de sensibilidad y valor predictivo positivo. En los apuntes de teoría aparece una discusión más extensa sobre esto, pero una posible guía informal podría ser la siguiente:

- Si se desea minimizar el número de positivos clasificados incorrectamente como negativos (por ejemplo, maximizar el número de sujetos enfermos diagnosticados como sanos, o maximizar el número de alarmas que se dan ante situaciones de riesgo), la métrica indicada es la sensibilidad.
- Si se desea minimizar el número de positivos detectados incorrectamente (falsos positivos, es decir, clasificados como negativos, por ejemplo, sujetos sanos diagnosticados como enfermos, o situaciones en las que se debería haber dado una alarma pero no se dio), la métrica indicada es el valor predictivo positivo.

Por lo tanto, depende del problema escoger la métrica más adecuada en función de la importancia relativa de su salida y de su comportamiento. En este tipo de problemas, F_1 -score es una métrica que puede ser de más utilidad que la precisión.

Otra cuestión a tener en cuenta es el posible desbalanceo de los datos. La precisión es una métrica que ofrece una visión “global”, sin tener en cuenta que puede ser engañosa cuando la distribución en clases está desbalanceada, y en estos casos F_1 -score es una mejor métrica. El tener bases de datos desbalanceadas es algo muy común, lo cual aporta un argumento extra para utilizar F_1 -score antes que precisión.

Finalmente, si se tienen más de dos clases, es posible construir una matriz de confusión de una forma similar teniendo una fila y columna por clase, por ejemplo:

	Predicción
--	------------

		A	B	C
Real	A			
	B			
	C			

En estos casos, ya no se puede hablar de patrones positivos o negativos, puesto que hay más de dos clases, ni tomar valores de sensibilidad o valor predictivo positivo. Sin embargo, esta matriz de confusión puede ofrecer información muy interesante a la hora de comprender el funcionamiento del modelo, viendo cuáles son las clases entre las que el modelo encuentra mayor facilidad y dificultad en su separación.

- Si se ha dividido el conjunto de patrones en entrenamiento y test, ¿en cuál de los dos habría que calcular la matriz de confusión?

En esta práctica, se pide:

- Desarrollar una función llamada *confusionMatrix* que acepte dos vectores de igual longitud (igual al número de patrones), el primero con las salidas obtenidas por un modelo *outputs* y el segundo con las salidas deseadas *targets*, ambos de tipo *AbstractArray{Bool,1}* y devuelva:
 - Valor de precisión.
 - Tasa de fallo.
 - Sensibilidad.
 - Especificidad.
 - Valor predictivo positivo.
 - Valor predictivo negativo.
 - *F₁-score*.
 - Matriz de confusión, como un objeto de tipo *Array{Int64,2}* con dos filas y dos columnas.

Dado que se le están dando vectores de valores booleanos, esta función se aplicará

en problemas de dos clases (casos positivos y negativos).

En el cálculo de estos valores, es necesario tener en cuenta estos casos particulares:

- Si los verdaderos negativos constituyen la totalidad de los patrones, no se habrá podido calcular la sensibilidad ni el valor predictivo positivo. En este caso el sistema funciona correctamente, por lo que a estas dos métricas se les asignará el valor de 1.
- De forma análoga, si los verdaderos positivos constituyen la totalidad de los patrones, no se habrá podido calcular la especificidad ni el valor predictivo negativo. El sistema nuevamente funciona de forma correcta, por lo que a estas dos métricas se les asignará el valor de 1.
- Si no se ha dado ninguno de esos dos casos y alguna métrica no se ha podido calcular, esta tomará el valor de 0. Si, por ejemplo, la sensibilidad no se ha podido calcular, quiere decir que no había ningún patrón con salida deseada positiva.
- Si la sensibilidad y el VPP son ambos iguales a 0, el valor de F_1 no se puede calcular, y se asignará un valor de 0.

No utilizar bucles en el desarrollo de esta función.

- Muchos modelos (por ejemplo, RR.NN.AA.) no van a devolver una salida categórica, sino que asignarán un valor de probabilidad a la clase “positivo”. Por este motivo, se pide desarrollar una función de nombre igual que la anterior, cuyo primer parámetro, en lugar de ser un vector de valores booleanos, sea un vector de valores reales (de tipo *AbstractArray{<:Real}*), y con un tercer parámetro opcional que tenga un umbral, con un valor por defecto, y los utilice para aplicar la función anterior y devolver, por lo tanto, los mismos valores.