Universidad de Ingeniería y Tecnología - UTEC
Departamento de Ingeniería Mecatrónica
Curso: Tópicos de Ingeniería Mecatrónica (2020 -2)

Tema 4:
Métricas de Evaluación

Prof. Oscar E. Ramos Ponce

Las métricas de evaluación son valores que permiten evaluar cuán bien se comporta un sistema al realizar la predicción, sea para una tarea de clasificación o para una tarea de regresión. Según sea el caso, existen diversas métricas cuya importancia depende de cada problema específico, ya que brindan información diferente y, en muchos casos, complementaria. Primero se presentará las métricas para clasificación.

1. Matriz de Confusión

En un problema binario de clasificación, solo existen dos clases posibles: la clase 1, llamada clase *positiva*, y la clase 0, llamada clase *negativa.* (a veces a la clase negativa se le denomina clase -1). Para este problema de clasificación, una matriz de confusión es un arreglo bidimensional que contiene la cantidad de instancias clasificadas como 0 o 1, y la cantidad de instancias que realmente son 0 y 1. Se considerará que el valor de salida real (también llamado etiqueta o *label*) de una instancia está representado por $y = \{0, 1\}$, y que el valor que se predice o estima para dicha instancia es $\hat{y} = \{0, 1\}$. Con esta notación, la matriz de confusión contiene la siguente información:

		Valor estimado		
		$\hat{y} = 1$	$\hat{y} = 0$	
Valor real	$y = 1$	TP	FN	n_P
	$y = 0$	FP	TN	n_N
		$\hat{n}_p$	$\hat{n}_N$	

La terminología usada en cada elemento de la matriz se describe a continuación.

- *TP:* son los verdaderos positivos (*True Positive*) y representan las instancias que han sido clasificadas como $\hat{y} = 1$ siendo su valor real $y = 1$.
- *TN:* son los verdaderos negativos (*True Negatives*) y representan las instancias que han sido clasificadas como $\hat{y} = 0$ siendo su valor real $y = 0$.
- *FP:* son los falsos positivos (*False Positives*) y representan las instancias que han sido clasificadas como $\hat{y} = 1$ cuando su valor real es $y = 0$. A este valor a veces también se le denomina *falsa alarma.*

- *FN:* son los falsos negativos (*False Negatives)* y representan las instancias que han sido clasificadas como $\hat{y} = 0$ cuando su valor real es $y = 1$.

Utilizando los elementos de la matriz de confusión, el número total de instancias que realmente son positivas se representará como $n_{_P}$, y el número total de instancias que realmente son negativas como $n_{_N}$, donde

$$n_{_P} = TP + FN, \quad \text{y} \quad n_{_N} = FP + TN.$$

Además, el número total de instancias clasificadas como positivas se representará como $\hat{n}_{_P}$, y el número total de instancias clasificadas como negativas como $\hat{n}_{_N}$, donde

$$\hat{n}_{_P} = TP + FP, \quad \text{y} \quad \hat{n}_{_N} = FN + TN.$$

Usando esta notación, si existen un total de n instancias, se cumple que

$$n = n_{_P} + n_{_N} = \hat{n}_{_P} + \hat{n}_{_N}.$$

Cuando se realiza una matriz de confusión de un clasificador binario, se asume que una clase es positiva (1) y la otra clase es negativa (0). El determinar qué clase se asigna como positiva y qué clase se asigna como negativa es dependiente del problema. Usualmente, la clase positiva es la clase de interés, y típicamente minoritaria, que contiene aquellos elementos que se desea clasificar del resto. Consecuentemente, la clase negativa es ese resto. Por ejemplo, en un diagnóstico médico, la clase positiva suele ser la detección de una enfermedad; en una detección de correo no deseado, la clase positiva sería el correo no deseado. A pesar de que en su forma más básica esta matriz es binaria, se puede fácilmente extender a varias clases, incrementando el número de filas y columnas. En este caso, además de especificar las clasificaciones correctas (*true*), se debe especificar a qué clases corresponden las clasificaciones incorrectas.

La matriz de confusión se puede aplicar al conjunto de entrenamiento, al conjunto de validación, o al conjunto de prueba, para tener una visión de cómo está funcionando el clasificador en cada uno de estos casos. Sin embargo, a pesar de brindar información sobre el desempeño del clasificador, por sí misma no permite una evaluación cuantitativa que pueda ser útil para la comparación de clasificadores; es decir, no brinda un "número"único para la comparación. Por este motivo se definen diversas métricas haciendo uso de los datos que contiene esta matriz.

2. Métricas de Clasificación

2.1. Métricas más usadas

Los siguientes valores son métricas comúnmente utilizadas para evaluar cuán bien se comporta un clasificador, y se basan en los valores que posee la matriz de confusión.

a. *Exactitud (A).* En inglés se denomina *accuracy*, y representa la cantidad de instancias correctamente clasificadas con respecto al total de instancias:

$$A = \frac{TP + TN}{n}$$

b. *Precisión (P).* Es un valor que indica cuántas instancias, de todas las que han sido clasificadas como positivas, realmente son positivas; es decir:

$$P = \frac{TP}{\hat{n}_{_P}} = \frac{TP}{TP + FP}$$

También se puede interpretar como $P = p(y = 1|\hat{y} = 1)$, que es la probabilidad de que una instancia sea realmente positiva dado que fue clasificada como positiva.

c. *Recall (R).* También se denomina *sensitividad*, *exhaustividad*, TPR (*True Positive Rate*), o *hit rate*. Indica cuántas instancias que realmente son positivas han sido clasificadas como positivas:

$$R = \frac{TP}{n_P} = \frac{TP}{TP + FN}$$

Se puede interpretar como $R = p(\hat{y} = 1|y = 1)$, que es la probabilidad de que una instancia sea clasificada como positiva dado que realmente es positiva.

Si bien la exactitud es una de las métricas más usadas, no brinda una información completa y puede llevar a conclusiones erróneas, sobre todo cuando se tiene clases desbalanceadas (muchos más elementos en una clase que en otra). Por ese motivo, se suele acompañar de otras métricas que brindan información adicional. Algunas de las más usadas son la precisión y el *recall*. De hecho, para algunas aplicaciones puede ser más importante tener un alto valor de precisión, mientras que para otras puede ser más importante tener un alto valor de *recall*. Los siguientes ejemplos ilustran estos casos.

Ejemplo 1. Supóngase que se tiene un sistema que diagnostica una cierta enfermedad en un conjunto de pacientes. Se considerará 1 cuando se diagnostica a la persona como enferma y 0 cuando se le diagnostica como sana. De un total de diez mil personas, se tiene la siguiente matriz de confusión:

		Diagnóstico	
		Enfermo	Sano
Realidad	Enfermo	1000	200
	Sano	800	8000

A partir de la tabla, y usando las definiciones anteriores, se puede concluir que los pacientes clasificados correctamente son 90 % (valor de exactitud A). Igualmente, de todos los pacientes diagnosticados enfermos, el 55.6 % se encuentran realmente enfermos (valor de precisión P). De todos los pacientes enfermos, el 83.3 % fue diagnosticado como enfermo (valor de recall R). En este ejemplo resulta crítico que un paciente enfermo sea diagnosticado como sano, por lo que es más importante tener un alto valor de R que un alto valor de P; es decir, importa más que si un paciente está realmente enfermo, sea diagnosticado efectivamente como enfermo.

Ejemplo 2. Supóngase que se tiene un sistema que realiza la clasificación de un correo electrónico como *spam* o no *spam*. La matriz de confusión, para un total de 1000 correos electrónicos, se muestra a continuación:

		Clasificación	
		Enviado a *Spam*	Enviado a *Inbox*
Realidad	*Spam*	100	170
	No *spam*	30	700

Con las definiciones de las métricas anteriores se obtiene que el 80 % de correos son clasificados adecuadamente (exactitud A). De todos los correos enviados a spam, el 76.9 %

son realmente spam (precisión P). De todos los correos que son realmente spam, el 37 % fueron enviados a la carpeta de spam (recall R). En este ejemplo, resulta muy malo que un correo que no es spam sea enviado a spam incorrectamente, por lo que se desea un alto valor de precisión en la clasificación (es más importante P que R). Es decir, es importante no necesariamente encontrar todo el spam, pero si se envía un correo a spam, realmente debe ser spam.

Ejemplo 3. Se tiene un clasificador binario que clasifica elementos como pertenecientes a la clase 1 o a la clase 0. El conjunto de entrenamiento tiene 14 instancias, de las cuales 8 son clasificadas como pertenecientes a la clase 1 y 6 como pertenecientes a la clase 0. Dos instancias son incorrectamente clasificadas como pertenecientes a la clase 1 cuando en realidad pertenecen a la clase 0, y 1 instancia es incorrectamente clasificada como perteneciente a la clase 0 cuando en realidad pertenece a la clase 1. A partir de estos datos, se genera la siguiente matriz de confusión:

		Predicción	
		Clase 1	Clase 0
Realidad	Clase 1	6	1
	Clase 0	2	5

Usando los datos de esta matriz, la exactitud de la clasificación es de 78.5 %. Además, de las instancias que se predicen como pertenecientes a la clase 1, 75 % son realmente de la clase 1 (precisión), y de las instancias que son realmente de la clase 1, el 85.7 % es clasificado correctamente (recall).

2.2. Valores F

Estos valores, denominados también *F-scores*, son medidas del desempeño de un clasificador que están basadas en una combinación de la precisión (P) con el *recall* (R).

Valor F_1. Este valor, llamado también F_1 *score*, evalúa cuán bien clasificadas están las instancias que fueron clasificadas como positivas. Se define como la media armónica de la precisión y el *recall*:

$$F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2PR}{P+R}$$

El valor de F_1, debido a que es una media armónica, siempre se encuentra más cerca al menor valor de P o de R, y por lo tanto permite determinar los clasificadores que no son muy buenos independientemente de si P o R es más importante para el problema en cuestión. En otras palabras, un valor alto de P, cuando R tiene un valor bajo, no genera un valor resultante muy alto; y viceversa.

En el límite, si $F_1 = 0$, entonces o P o R serían cero (o ambos) y sería el peor caso de clasificación. Por otro lado, si $F_1 = 1$, se tendría que tanto P como R son 1 y se tendría un caso perfecto de clasificación (aunque posiblemente podría haber *overfitting*).

Valor F_β. Este valor es una generalización del valor F anterior y se define, usando la precisión y el *recall*, como

$$F_\beta = (1+\beta^2)\frac{PR}{\beta^2 P + R},$$

donde $\beta \geq 0$. El valor de β permite evaluar la importancia relativa que tienen P y R para un clasificador determinado. Algunos casos de β son los siguientes:

- Cuando $\beta = 0$, se tiene $F_0 = P$. Es decir, se recupera la precisión.
- Cuando $\beta \to \infty$, se tiene $F_\infty = R$. Es decir, se recupera el valor de *recall*.
- Cuando $\beta = 1$, se tiene $F_1 = \frac{2PR}{P+R}$ y se recupera el valor de F_1.

Como se puede observar, F_1 es un caso particular cuando $\beta = 1$. De forma similar se puede dar diferentes valores a β para brindar mayor importancia a P o a R, ya que en los límites se obtiene estos valores.

2.3. Otras métricas de clasificación

Algunas otras métricas, menos utilizadas, para evaluar el desempeño de un clasificador son las siguientes.

- *Especificidad (S).* También denominado TNR (*True Negative Rate*), representa la cantidad de instancias negativas que han sido correctamente clasificadas como negativas:
$$S = \frac{TN}{n_{_N}} = \frac{TN}{FP + TN}$$
De manera probabilística se tiene $S = p(\hat{y} = 0|y = 0)$ e indica la probabilidad de que una instancia sea clasificada como negativa dado que es realmente negativa.

- *Error Tipo 1.* También se denomina FPR (*False Positive Rate*) o razón de falsa alarma, y representa la cantidad de instancias negativas que han sido incorrectamente clasificadas como positivas:
$$E_1 = \frac{FP}{n_{_N}} = \frac{FP}{FP + TN}$$
Se puede interpretar como $E_1 = p(\hat{y} = 1|y = 0)$, que es la probabilidad de que una instancia sea clasificada como positiva dado que es en realidad negativa.

- *Error Tipo 2.* También se denomina FNR (*False Negative Rate*) y representa la cantidad de instancias positivas que han sido incorrectamente clasificadas como negativas:
$$E_2 = \frac{FN}{n_{_P}} = \frac{FN}{TP + FN}$$
De forma probabilística se tiene $E_2 = p(\hat{y} = 0|y = 1)$ y representa la probabilidad de que una instancia sea clasificada como negativa dado que en realidad es positiva.

Usando propiedades básicas de probabilidades, es directo notar que $p(\hat{y} = 1|y = 1) + p(\hat{y} = 0|y = 1) = 1$, y que $p(\hat{y} = 0|y = 0) + p(\hat{y} = 1|y = 0) = 1$. Comparando estos resultados con las interpretaciones probabilísticas de las métricas, se desprende que

$$R + E_2 = 1, \quad \text{y} \quad S + E_1 = 1.$$

Debido a estas propiedades, estas cuatro últimas métricas son a veces utilizadas para la elaboración de una matriz de confusión con valores relativos, la cual queda dada por:

		Valor estimado	
		$\hat{y} = 1$	$\hat{y} = 0$
Valor real	$y = 1$	TPR	Error II
	$y = 0$	Error I	TNR

Cada elemento de esta matriz es un valor entre 0 y 1, interpretado según su definición, y la suma de los elementos de cada fila es siempre 1. Esta forma de matriz de confusión, que contiene únicamente elementos relativos, es menos usada que la forma en la que se incluye los valores absolutos.

3. Curva ROC

La curva ROC (*Receiver Operating Characteristic*) es un método visual que permite evaluar la eficiencia de la clasificación a través de un gráfico que muestra las mediciones TPR (*True Positive Rate*) en función de las mediciones TNR (*True Negative Rate*). Se utiliza para comparar el desempeño relativo de diversos clasificadores.

Esta curva se construye dando diversos valores de umbral al clasificador. Con cada valor de umbral se encuentra un par (FPR, TPR), y se ubica cada uno de estos pares en una figura, donde el eje x es FPR y el eje y representa al valor TPR. El área bajo la curva (llamado usualmente AUC de *Area Under the Curve*) mide la calidad del clasificador: un valor de 0.5 es un clasificador aleatorio, y mientras más se acerque el área a 1 es mejor.

Algoritmo Explotar la monotonicidad de clasificaciones con umbral Cualquier instancia que es clasificada como positiva con respecto a un umbral, será clasificada positiva para todos los umbrales más bajos Algoritmo Ordenar las instancias de prueba, por orden decreciente Bajar en la lista, procesando una instancia por vez Actualizar TP y FP La curva ROC se puede crear a partir de un escaneo lineal