Introducción al reconocimiento de patrones: Métricas para clasificadores

M. Sc. Saúl Calderón Ramírez
Instituto Tecnológico de Costa Rica,
Escuela de Computación, bachillerato en Ingeniería en Computación,
PAttern Recongition and MAchine Learning Group (PARMA-Group)

24 de agosto de 2018

A continuación se presentan distintoas métricas para la precisión y confiabilidad de un sistema reconocimiento de patrones.

Parte I
Métricas para clasificación

1. Falsos positivos, falsos negativos, precisión y exhaustividad

Una forma intituiva de medir la confiabilidad de un clasificador es la cantidad o porcentaje de clasificaciones correctas. Sin embargo considere un ejemplo como el diagnóstico de cáncer de cérvix. Este cáncer tiene una prevalencia de 10 % en la población. Un clasificador de histologías cervicales *perezoso,* el cual siempre clasifica como negativa la presencia de cáncer a partir de una muestra lograría un 90 % de clasificación correcta. Tal tasa de correctitud desprecia el hecho de que 10 % de las muestras fueron incorrectamente diagnosticadas, correspondiendo al caso en el que las mujeres presentan la enfermedad pero según el clasificador indicó lo contrario.

En el reconocimiento de patrones y el aprendizaje automático los conceptos de falso positivo y falso negativo son muy utilizados para medir la confiabilidad de un sistema de clasificación. Para determinarlos, es necesario utilizar las muestras del campo de verdad o groundtruth, las cuales son muestras con las etiquetas verdaderas correspondientes, usualmente manualmente asignadas. Tómese el ejemplo que para un clasificador $\tilde{t}_i = c\left(\vec{x}_i\right)$ el cual estima la etiqueta t_i para la muestra a clasificar $\vec{x}_i$. Suponga que el clasificador c fue entrenado a

		Clase real (T)		
		C_1	C_2	C_3
Clase estimada ($\tilde{T}_v$)	C_1	5	2	0
	C_2	3	**3**	**2**
	C_3	0	**1**	**11**

Cuadro 1: Matriz de confusión de ejemplo.

partir de un conjunto de n muestras de entrenamiento:

$$X_{\mathrm{e}} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$$

para el cual se dispone del conjunto de etiquetas correspondiente $T_e = \{t_1, t_2, \ldots, t_n\}$. Para la etapa de validación, se dispone de un conjunto de m muestras distintas:

$$X_v = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_m\}$$

y sus respectivas etiquetas $T_v = \{t_1, t_2, \ldots, t_m\}$. La etiqueta puede tomar un valor de natural entre 0 y k para k clases, de modo que $t_i \in [1 - k]$. Para un conjunto X_v, la salida del clasificador sería un conjunto de etiquetas estimadas $\tilde{T}_v$, de modo que:

$$\tilde{T}_v = c\,(X_v)$$

Para ejemplificar los conceptos de falso positivo y negativo, tómese $m = 27$ muestras de entrenamiento y $k = 3$ clases, de modo que la clase C_1 corresponde a la categoría «perro», la clase C_2 a la categoría «gato» y la clase C_3 a la categoría «ratón». Después de ejecutar la estimación de las etiquetas en el conjunto $\tilde{T}_v$ con el clasificador c para el conjunto de etiquetas de validación X_v se obtienen los siguientes resultados, tabulados en lo que se denomina una **matriz de confusión**.

- **Falsos y verdaderos positivos**: Un verdadero positivo para una clase C_i corresponde a la ocurrencia de una estimación de la etiqueta $\tilde{t}_j$ para una muestra $\vec{x}_j$ correcta, de modo que:

  $$\text{si} \quad t_j = i \Rightarrow i = \tilde{t}_j = c\,(\vec{x}_j)$$

  por lo que entonces un falso positivo (error de tipo 1) para una clase C_i corresponde a una estimación incorrecta de la etiqueta si $\quad t_j \neq i \Rightarrow i = \tilde{t}_j = c\,(\vec{x}_j)$. En el ejemplo de la Tabla 1, para la clase C_1existen 5 verdaderos positivos y 2 falsos positivos.

- **Falsos y verdaderos negativos**: Un verdadero negativo para una clase C_i se asocia con la estimación de la etiqueta $\tilde{t}_j$ para una muestra $\vec{x}_j$ para la cual según el conjunto de datos verdaderos T_v, la etiqueta real t_j corresponde a una clase distinta, por lo que $t_j \neq i$. Es decir, la muestra

		Clase real (T)	
		C_1	No C_1
Clase estimada ($\tilde{T}_v$)	C_1	5 Verdaderos positivos (VP)	2 Falsos positivos (FP)
	No C_1	3 Falsos negativos (FN)	17 Verdaderos negativos (VN)

Cuadro 2: Matriz de confusión de ejemplo para la clase C_1.

$\vec{x}_j$ pertenece a la *no-clase* C_i (cualquiera de las clases distintas a C_i), y el clasificador . Más formalmente, un **verdadero negativo** se da :

$$\text{si} \quad t_j \neq i \Rightarrow i \neq \tilde{t}_j = c\left(\vec{x}_j\right)$$

y un **falso negativo** (error de tipo 2) si $t_j = i \Rightarrow i \neq \tilde{t}_j = c\left(\vec{x}_j\right)$. En el caso de ejemplo, para C_1, existen 17 verdaderos negativos (en negrita), correspondientes a las muestras que pertenecen realmente a las clases C_2, C_3 y fueron estimadas como de la no-clase C_1 (clases C_2 y C_3). Los falsos negativos para la clase C_1 corresponden a las 3 muestras que correspondían realmente a la clase C_1 pero fueron clasificadas en la categoría C_2.

A partir de lo anterior, es posible, para una clase C_i construir una matriz de confusión con los falsos positivos, negativos y verdaderos positivos y negativos. Ello se ilustra para la clase C_1 en la Tabla 2.

- **Exhaustividad o sensibilidad**: Se refiere a la tasa de verdaderos positivos (TVP) o probabilidad de detección para una clase C_i, respecto a la cantidad de muestras reales (según el *groundtruth*). Mide la proporción de positivos correctamente clasificados del total de estimaciones para la clase según el ground truth, lo cual se denominan ***elementos relevantes***, y se ilustra en la Figura 1. Su fórmula está dada por:

$$\text{TVP} = \frac{\text{VP}}{\text{FN+VP}}$$

Para ilustrar mejor el concepto, tómese el ejemplo de la búsqueda de documentos en un repositorio. Los documentos relevantes se definen como los documentos que son correctamente asociados al términos de búsqueda, y los documentos recuperados son los documentos encontrados automáticamente por el algoritmo. De esta forma, los **verdaderos positivos** corresponden a la cantidad de elementos en la intersección:

$$\{\text{documentos relevantes}\} \cap \{\text{documentos recuperados}\}$$

y la cantidad total de documentos relevantes equivale a la cantidad de falsos negativos y verdaderos positivos (el círculo en la Figura 1). De esta forma, se puede reescribir la tasa de verdaderos positivos como sigue:

$$\text{TVP} = \frac{|\{\text{documentos relevantes}\} \cap \{\text{documentos recuperados}\}|}{|\{\text{documentos relevantes}\}|}$$

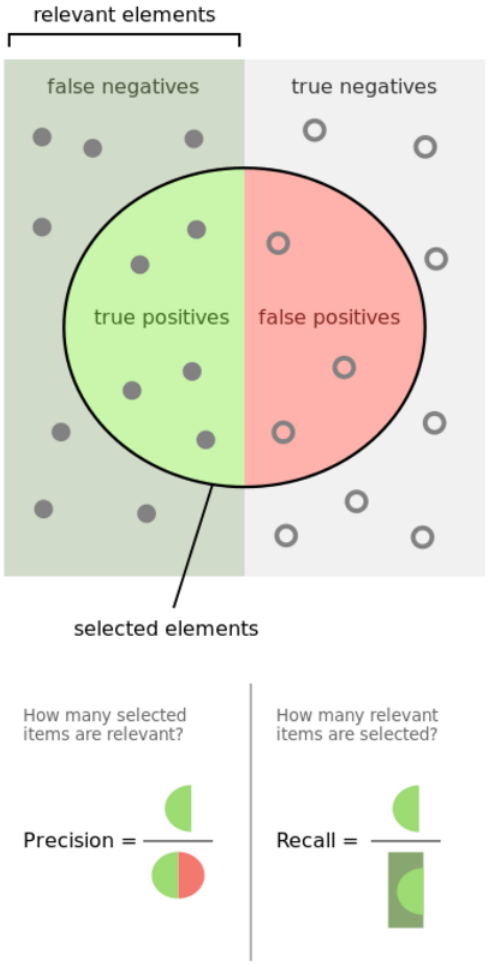

Figura 1: Exhaustividad y precisión.

		Clase real (T)		
		C_1	C_2	C_3
Clase estimada ($\tilde{T}_v$)	C_1	50	12	5
	C_2	4	32	5
	C_3	0	1	11

Cuadro 3: Matriz de confusión de ejemplo.

- **Precisión o valor predictivo positivo**: Se refiere al valor predictivo positivo para una clase C_i frente al total de estimaciones realizadas para esa clase. El total de estimaciones realizadas viene entonces dado por la suma de los falsos positivos y los verderdaderos positivos para la clase C_i, correspondiente a los ***elementos seleccionados***. Su cálculo viene dado por:

$$\text{VPP} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

  Siguiendo el ejemplo de los documentos, la precisión vendría dada por la proporción de los documentos correctamente recuperados, correspondiente a $|\{\text{documentos relevantes}\} \cap \{\text{documentos recuperados}\}|$ respecto a los documentos recuperados, por lo que entonces:

$$\text{VPP} = \frac{|\{\text{documentos relevantes}\} \cap \{\text{documentos recuperados}\}|}{|\{\text{documentos recuperados}\}|}.$$

Las métricas anteriores toman en cuenta aspectos distintos en la clasificación. La precisión de clasificación de una clase C_i se realiza respecto a las estimaciones de esa clase, y la exhaustividad respecto a las muestras correctas de tal clase. Para ponderar ambas métricas en una sola, se utiliza la *medida F* o ***F-score***. Su cálculo con peso equivalente para precisión y exhaustividad viene dado por:

$$F = \frac{2 \cdot \text{TVP} \cdot \text{VPP}}{\text{TVP} + \text{VPP}}$$

Para calcular cualquiera de las métricas, se recomienda usar un porcentaje del banco de muestras para entrenamiento, y otro para validación, lo cual se conoce como **validación cruzada**.

2. Ejemplos

Para los siguientes ejemplos, calcule: los falsos positivos, falsos negativos, verdaderos positivos, verdaderos negativos, exhaustividad, precisión y la medida F para las clases C_1 y C_2.

		Clase real (T)			
		C_1	C_2	C_3	C_4
Clase estimada ($\tilde{T}_v$)	C_1	92	5	2	10
	C_2	3	12	30	11
	C_3	0	1	20	5
	C_4	0	2	1	60

Cuadro 4: Matriz de confusión de ejemplo.

Parte II
Métricas para regresión

En el problema de regresión la medición de la precisión se realiza con métricas las cuales consideran de forma continua el ajuste de un modelo sobre un conjunto de datos, en este caso el modelo realiza la estimación de valores continuos, correspondientes a las etiquetas, por lo que entonces, para un conjunto de muestras de validación

$$X_v = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_m\}$$

se define el conjunto de etiquetas de prueba$T_v = \{t_1, t_2, \ldots, t_m\}$, donde en este caso $t_i \in \mathbb{R}$, y las etiquetas estimadas están definidas en el conjunto $\tilde{T}_v = \left\{\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_m\right\}$, resultado de la estimación del modelo $\tilde{T}_v = c\left(X_v\right)$.

Dos métricas usuales para la medición de la precisión del modelo de regresión son el error medio absoluto y la raíz del error medio cuadrado, los cuales se detallan a continuación.

El error medio absoluto o MAE en inglés se define como:

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}\left|\tilde{t}_i - t_i\right|$$

mientras que la raíz del error medio cuadrado (RMSE en inglés) se define como:

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\tilde{t}_i - t_i\right)^2}$$

Similitudes entre las métricas: Ambas métricas son similares en el sentido de que expresan el error promedio de predicción del modelo en unidades de la variable de interés. Por ejemplo, si el modelo tiene por objetivo estimar la edad de un individuo, ambas métricas tienen por unidad los años. Además, ambas métricas pueden variar de 0 a ∞ y son indiferentes a la dirección de los errores. Son puntuaciones negativamente orientadas, lo que significa que los valores más bajos son mejores.

i	$\lvert\tilde{t}_i - t_i\rvert$	$\left(\tilde{t}_i - t_i\right)^2$	$\lvert\tilde{t}_i - t_i\rvert$	$\left(\tilde{t}_i - t_i\right)^2$	$\lvert\tilde{t}_i - t_i\rvert$	$\left(\tilde{t}_i - t_i\right)^2$
1	2	4	1	1	0	0
2	2	4	1	1	0	0
3	2	4	1	1	0	0
4	2	4	1	1	0	0
5	2	4	1	1	0	0
6	2	4	3	9	0	0
7	2	4	3	9	0	0
8	2	4	3	9	0	0
9	2	4	3	9	0	0
10	2	4	3	9	20	400
$\sigma = 0$			$\sigma = 1{,}054$		$\sigma = 6{,}324$	
MAE $= 2$			MAE $= 2$		MAE $= 2$	
RMSE $= 2$			RMSE $= 2{,}236$		RMSE $= 6{,}325$	

Cuadro 5: 3 conjuntos de muestras distintos.

Diferencias entre las métricas: El cálculo de la diferencia cuadrada de los errores en la métrica RMSE conlleva ciertas implicaciones. Dado que los errores se cuadran antes de promediarlos, el RMSE otorga un peso relativamente alto a los grandes errores. Esto significa que el RMSE debería ser más útil para castigar los errores grandes, siendo particularmente indeseables. Las tres tablas de la Figura 5, muestran ejemplos donde MAE es estable y RMSE aumenta cuando existen **valores atípicos** más grandes.

Los valores atípicos grandes afectan la desviación estándar, sin embargo, no necesariamente una mayor desviación estándar de los errores aumenta el RMSE, como se muestra en la Figura 6.

Parte III

Métricas para señales binarias

En el problema de segmentación más simple, es necesario etiquetar los pixeles como de fondo o pertenecientes al objeto de interés. Para evaluar la precisión de la segmentación, las métricas usuales evalúan la diferencia de señales binarias (donde por ejemplo los pixeles etiquetados con 1 se refieren a los objetos de interés y los etiquetados con 0 al fondo). Una métrica muy utilizada para tales efectos es el coeficiente de *dice* o de Sørensen, el cual se puede usar para evaluar la similitud estructural entre dos imágenes binarias $U \in \mathbb{R}^{m \times n}$ y $V \in \mathbb{R}^{m \times n}$, y viene dado por:

$$S = \frac{2\left(\sum_j^m \sum_i^n U\left[i,j\right]\ V\left[i,j\right]\right)}{\left(\sum_j^m \sum_i^n U\left[i,j\right]\right) + \left(\sum_j^m \sum_i^n V\left[i,j\right]\right)}$$

i	$\lvert\tilde{t}_i - t_j\rvert$	$\left(\tilde{t}_i - t_i\right)^2$	$\lvert\tilde{t}_i - t_j\rvert$	$\left(\tilde{t}_i - t_i\right)^2$
1	5	25	3	9
2	5	25	3	9
3	5	25	3	9
4	5	25	3	9
5	5	25	3	9
6	0	0	4	16
7	0	0	4	16
8	0	0	4	16
9	0	0	4	16
10	0	0	4	16
$\sigma = 2{,}635$			$\sigma = 0{,}527$	
MAE $= 2{,}5$			MAE $= 3{,}5$	
RMSE $= 3{,}536$			RMSE $= 3{,}536$	

Cuadro 6: 2 conjuntos con distinta desviación estándar pero mismo RMSE.

Figura 2: *Groundtruth* a la izquierda, e imagen automáticamente segmentada a la derecha.

donde el numerador equivale conceptualmente a la cantidad de pixeles del objeto de interés intersecados en ambas imágenes U y V : $2\,|U \cap V|$(con las barras denotando la cantidad de pixeles con 1's en la matriz) y el denominador a la suma de los pixeles del objeto de interés en ambas imágenes: $|U| + |V|$, lo que entonces significa que se puede reescribir S como:

$$S = \frac{2\,|U \cap V|}{|U| + |V|}$$

La imagen U puede asociarse a la imagen segmentada por el algoritmo, y V la imagen segmentada manualmente, también conocida como imagen de *groundtruth.* De este modo, una segmentación perfecta según el *groundtruth* resulta en un coeficiente de *dice* $S = 1$, y una segmentación completamente errada un coeficiente $S = 0$. La Figura 2 ilustra la segmentación automática de una radiografía pulmonar, y el *groundtruth* de la misma.