Métricas en Aprendizaje Automático

Prof. Wílmer Pereira

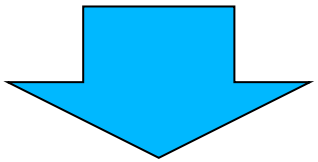
Necesidad de las métricas

Son parámetros que permiten evaluar la tasa de aprendizaje de la técnica de clasificación o regresión una vez entrenada. Para ello se usa el conjunto de prueba y se verifica el ajuste de la predicción $\hat{y}$ contra la clasificación real y

- Es muy común usar sólo el *accuracy* pero tiene ciertos problemas con la clasificación. Por ejemplo, dada dos clases (A y B) y un conjunto de entrenamiento desbalanceado y compuesto por 98% de casos de la clase A y sólo 2% de casos de la clase B. Si forzamos a que siempre responda A, el *accuracy* es del 98%

… pero …

En realidad no hay ninguna garantía de aprendizaje



Es necesario, definir diferentes tipos de métricas

Métricas de evaluación

- En el caso de la clasificación, para poder discriminar bien el aprendizaje obtenido, conviene separar la cantidad de ejemplos acertados y no acertados por clase.

Clase real

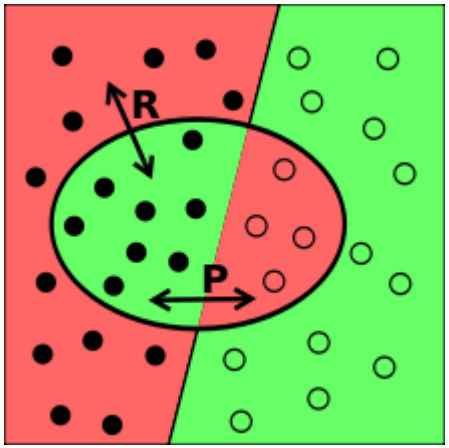
Clase predicha	**Clase**	**No Clase**
Clase	Verdadero Positivo	Falso Positivo
No Clase	Falso Negativo	Verdadero Negativo

- Un falso positivo se considera como una falsa alarma, es decir, se clasificó como perteneciente a la clase sin serlo. En contraposición el falso negativo es que clasificó como no perteneciente a la clase cuando en realidad si lo era. El vocabulario proviene de la medicina, donde falso positivo es una falsa detección de la enfermedad y un falso negativo es la no detección de la enfermedad

Métricas para la clasificación

- Para simplificar usemos las siglas en inglés:

  P: positivos reales del conjunto de prueba
  N: negativos reales del conjunto de prueba
  TP: positivos bien clasificados
  TN: negativos bien clasificados
  FP: positivos mal clasificados
  FN: negativos mal clasificados



- A la matriz que cuenta todos los ejemplos y como fueron clasificados se le denomina **matriz de confusión** o tabla de contingencia. Por ejemplo

		Valor Predicho		
		Gato	Perro	Conejo
Valor Real	Gato	5	3	0
	Perro	2	3	1
	Conejo	0	2	11

La técnica utilizada clasificó a los conejos de buena manera porque sólo se equivoco con dos ejemplos del conjunto de prueba

En cambio, perros y gatos no fueron tan bien clasificados y, un poco peor, perros

- La diagonal son las clasificaciones correctas. El resto son las equivocaciones del clasificador sobre el conjunto de prueba.

Métricas para la clasificación

P: positivos reales del conjunto de prueba
N: negativos reales del conjunto de prueba
TP: positivos bien clasificados
TN: negativos bien clasificados
FP: positivos mal clasificados
FN: negativos mal clasificados

- La métrica para clasificadores más simple es el *accuracy* (traducida como exactitud en español) que se obtiene como:

$$ACC = \frac{TP+TN}{P+N}$$

mide lo que se clasificó bien, conjuntamente con lo que pertenece o no a la clase …

- Frecuentemente es mejor discriminar por VP/FP y VN/FN cuando el clasificador es multiclase porque se verifica mejor su eficacia … existen varias parejas …

Precision/Recall: PPV/TPR — *Positive Predictive Value* / *True Positive Rate*

$$PPV = \frac{TP}{TP+FP} \quad \text{y} \quad TPR = \frac{TP}{TP+FN}$$

Sensivity/Especificity: TPR/TNR — *True Positive Rate* / *True Negative Rate*

$$TPR = \frac{TP}{TP+FN} \quad \text{y} \quad TNR = \frac{TN}{TN+FP}$$

Métricas para la clasificación

P:	positivos reales del conjunto de prueba
N:	negativos reales del conjunto de prueba
TP:	positivos bien clasificados
TN:	negativos bien clasificados
FP:	positivos mal clasificados
FN:	negativos mal clasificados

Precision/Recall (PPV/TPR) → $PPV = \frac{TP}{TP+FP}$, $TPR = \frac{TP}{TP+FN}$

Sensivity/Especificity (TPR/TNR) → $TPR = \frac{TP}{TP+FN}$, $TNR = \frac{TN}{TN+FP}$

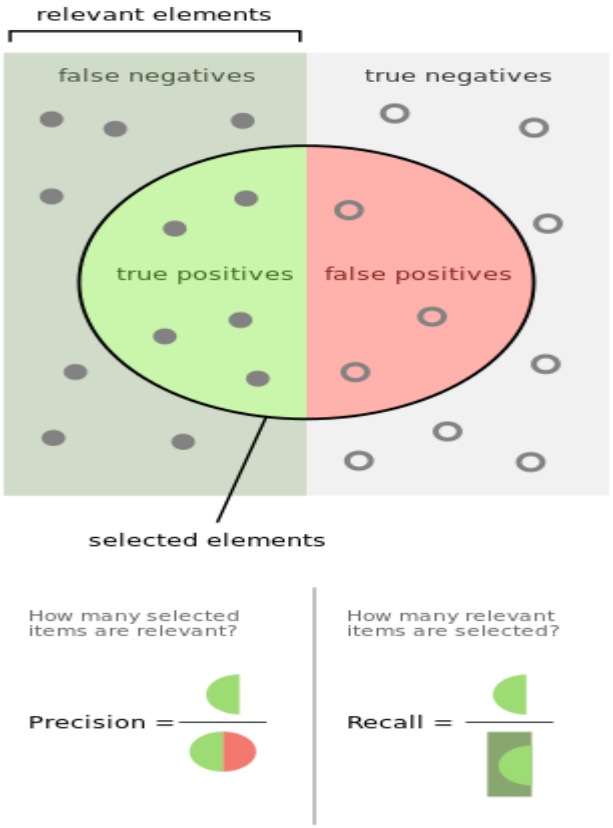
- *Precision* se puede parafrasear como: "la fracción de los ejemplos positivos que clasificó contra todos los que clasificó como positivos".
- *Recall*, en cambio, es: "la fracción de los ejemplos positivos que clasificó contra los reales positivos.

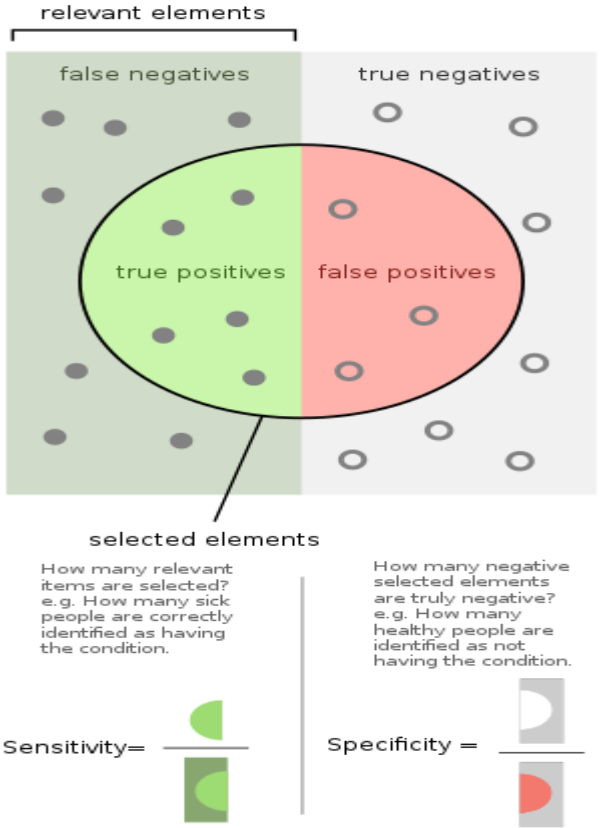
… por otro lado …

- *Sensivity* se puede parafrasear como: "la proporción de positivos identificados correctamente contra el total de los positivos clasificados bien o mal"
- *Especificity*, en cambio, es: "la proporción de negativos identificados correctamente contra el total de los negativos bien o mal".

Métricas para la clasificación

P: positivos reales del conjunto de prueba
N: negativos reales del conjunto de prueba
TP: positivos bien clasificados
TN: negativos bien clasificados
FP: positivos mal clasificados
FN: negativos mal clasificados





$$Precision = \frac{TP}{TP+FP} \quad , \quad Recall = \frac{TP}{TP+FN}$$

$$Sensivity = \frac{TP}{TP+FN} \quad , Especificity = \frac{TN}{TN+FP}$$

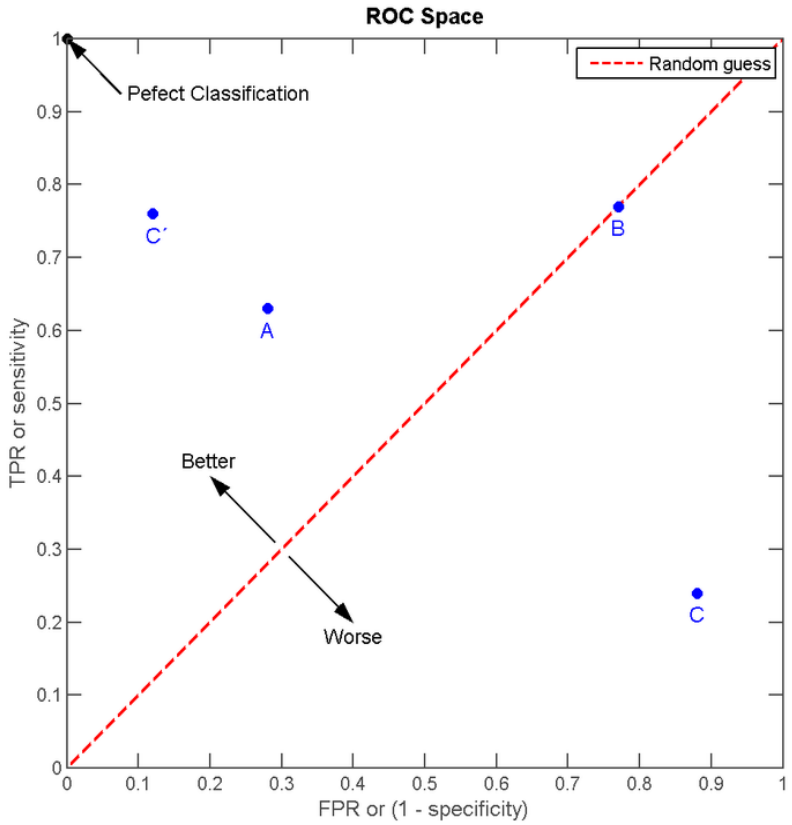
Métricas para la clasificación

P: positivos reales del conjunto de prueba
N: negativos reales del conjunto de prueba
TP: positivos bien clasificados
TN: negativos bien clasificados
FP: positivos mal clasificados
FN: negativos mal clasificados

		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

Tabla de ROC

Es una razón de la sensibilidad contra 1-especificidad para un clasificador, según se varía el umbral de discriminación

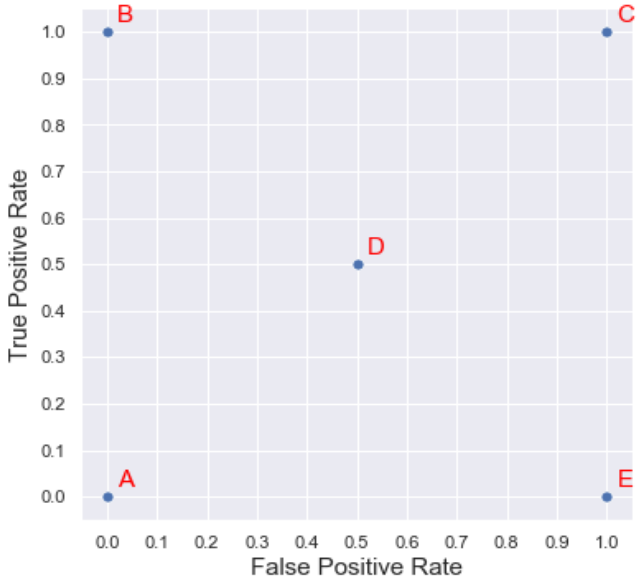


El mejor resultado sería en la esquina superior izquierda pues representaría 100% de sensibilidad (ningún falso negativo) y 100% de especificidad (ningún falso positivo)

La discriminación completamente aleatoria sería con un punto sobre la diagonal (por ejemplo, B)

En consecuencia, los puntos por debajo de la diagonal son pobre (peor que aleatorio) … Sin embargo, de un mal predictor se puede obtener uno bueno sólo invirtiéndolo …

Tabla de ROC



Así como B sería el ideal, hay otras posibilidades como A (no detectaría positivo ningún negativo) pero no identificaría ninguna muestra positiva …

El punto D significa que identificaría la mitad de las muestras positivas y negativas como erróneas y, sin duda, E es el peor caso …

Cuando se desea ajustar un clasificador, se se obtienen varios puntos y por ello una cur-curva del comportamiento del clasificador.

El área bajo la curva (AUC) se interpreta como la probabilidad de clasificar bien elementos que pertenecen a sus clases. Por ejemplo, la azul es mejor que la amarilla