

Aplicaciones de Ciencia de la Computación

Aprendizaje de Máquina: Evaluación del Modelo de aprendizaje

Prof. Dr. César A. Beltrán Castañón

cbeltran@pucp.pe

Evaluación del clasificador

¿Cómo saber si nuestro modelo en verdad está aprendiendo?

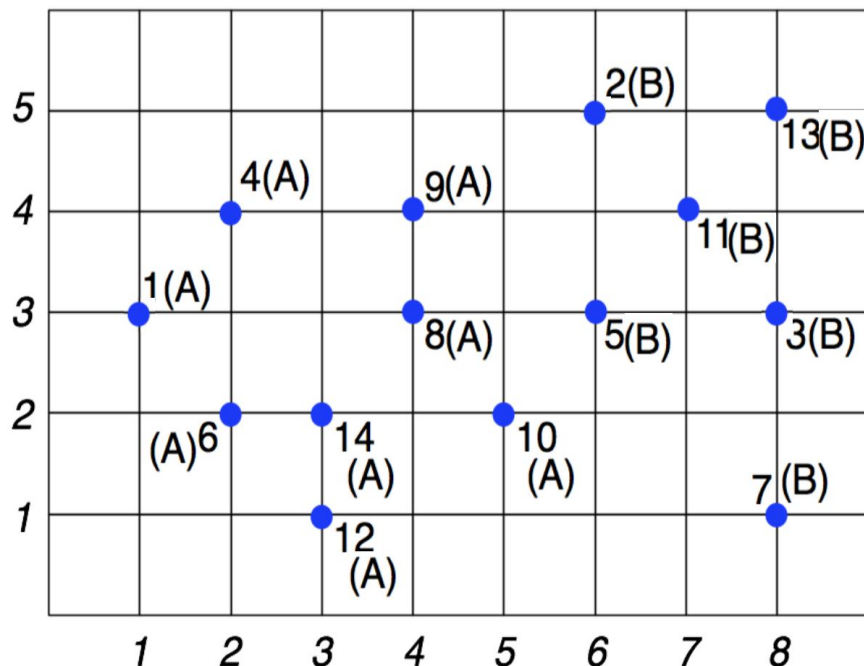


Cómo medir la eficiencia del clasificador?

- Conjunto de entrenamiento?
- Conjunto de prueba?
- En verdad mi clasificador aprende??

Ejemplo de clasificación

Dado el siguiente conjunto de entrenamiento



Indique el error de clasificar los siguientes puntos, usando kNN, con $k=3$:

Punto	Clase
(3,3)	A
(5,1)	B
(6,1)	B
(5,3)	A
(5,5)	B
(4,6)	A

Criterios para medir el desempeño del clasificador

Eficacia de la predicción: capacidad del modelo para predecir correctamente la etiqueta de una clase o de nuevos datos

Velocidad: esto se refiere a los gastos de cálculo involucrados en la generación y uso del modelo

Robustez: capacidad del modelo para hacer predicciones correctas sobre datos ruidosos o datos con valores perdidos

Escalabilidad: capacidad de construir un modelo eficiente dada una gran cantidad de datos

Interpretabilidad: nivel de comprensión y conocimiento que es proporcionado por el modelo

También podemos citar la **simplicidad**, indicadores de calidad dependientes del

Matriz de confusión

La **matriz de confusión** es una tabla de $m \times m$ de tamaño. Una entrada $CM_{i,j}$ indica el número de registros de la clase i que fueron etiquetados como clase j

Clase real /Clase predicha	<i>Clase₁</i>	<i>Clase₂</i>	...	<i>Clase_m</i>
<i>Clase₁</i>	$CM_{1,1}$	$CM_{1,2}$...	$CM_{1,m}$
<i>Clase₂</i>	$CM_{2,1}$	$CM_{2,2}$...	$CM_{2,m}$
...
<i>Clase_m</i>	$CM_{m,1}$	$CM_{m,2}$...	$CM_{m,m}$

Idealmente, la mayoría de los registros deben estar representados a lo largo de la **diagonal** de la matriz de confusión

Matriz de confusión

Para una clase binaria:

	Clasificado como...	
Clase original	Si	No
Si	Verdaderos positivos	Falsos negativos
No	Falsos positivos	Verdaderos negativos

- **Verdaderos positivos:** registros positivos correctamente etiquetados
- **Verdaderos negativos:** registros negativos correctamente etiquetados
- **Falsos positivos:** registros positivos incorrectamente etiquetados
- **Falsos negativos:** registros negativos incorrectamente etiquetados

Accuracy

La **eficacia (accuracy)** se refiere a la capacidad del modelo para **predecir correctamente** la etiqueta de una clase o de nuevos datos

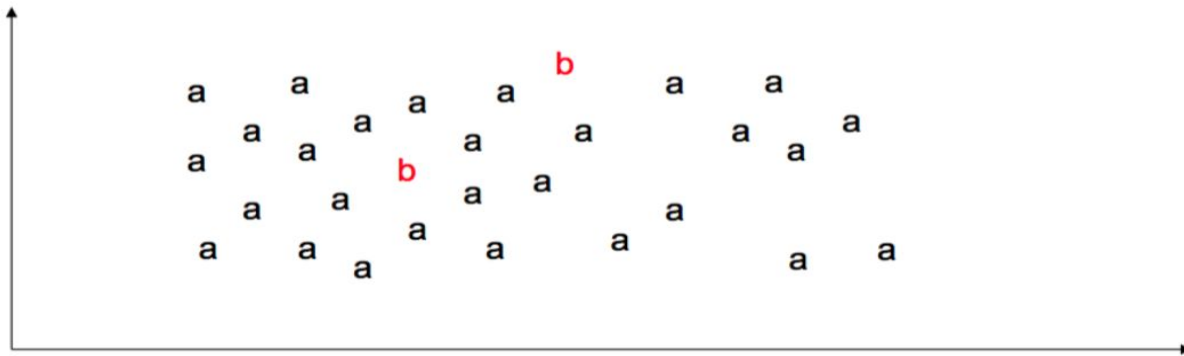
	Clasificado como...	
Clase original	Si	No
Si	Verdaderos positivos TP	Falsos negativos FN
No	Falsos positivos FP	Verdaderos negativos TN

$$accuracy = \frac{\#registrosCorrectamenteClasificados}{\#totalRegistros}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

La eficacia no es siempre suficiente

- Considere un problema de dos clases



- Si un modelo predice siempre una clase, la precisión es $28/30 = 93\%$
- La **eficacia es engañosa** porque el modelo no podría detectar ningún registro de la clase *b* (**clases desbalanceadas**)

Medidas sensibles al desbalance

	Clasificado como...	
Clase original	Si	No
Si	Verdaderos positivos TP	Falsos negativos FN
No	Falsos positivos FP	Verdaderos negativos TN

precision (P): sesgada hacia $C_{(Si|Si)}$ y $C_{(Si|No)}$. Mayor precisión, menor FP

$$precision = \frac{TP}{TP + FP}$$

También conocida como sensibilidad

recall (R): sesgada hacia $C_{(Si|Si)}$ y $C_{(No|Si)}$. Mayor recall, menor FN

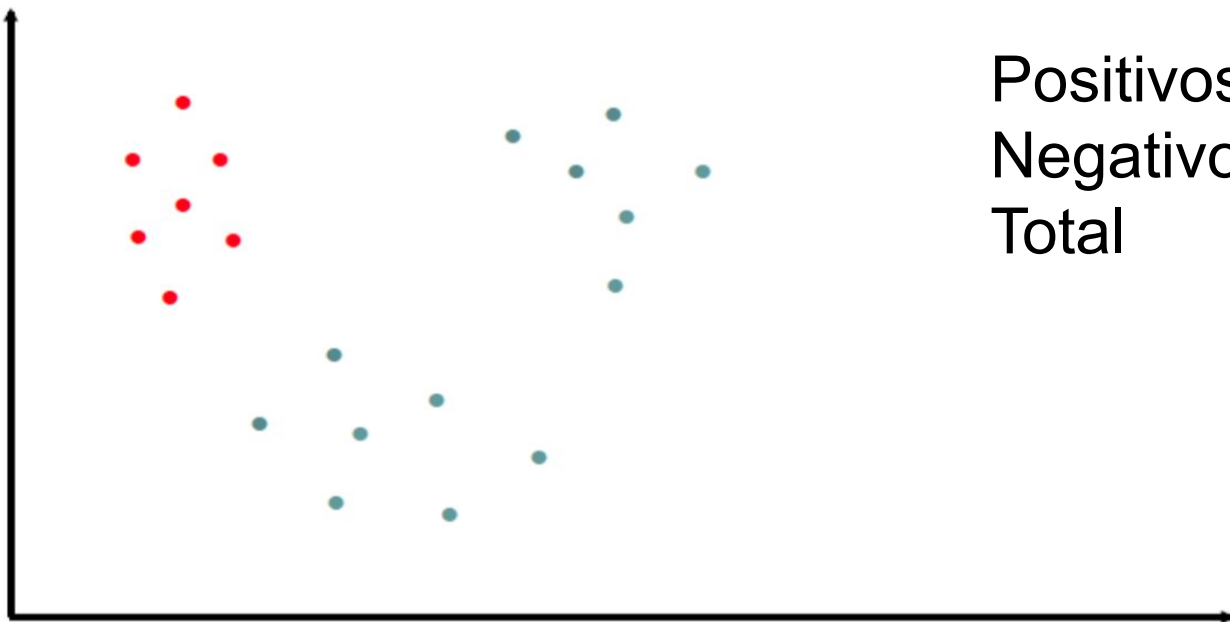
$$recall = \frac{TP}{TP + FN}$$

F1-measure (F1): sesgada hacia todo excepto al $C_{(No|No)}$. F1 es alto si el P y R son altos. Mayor F1, menores FP y FN

$$F1 - measure = \frac{2 \times R \times P}{R + P}$$

Ejemplo

Distribución original



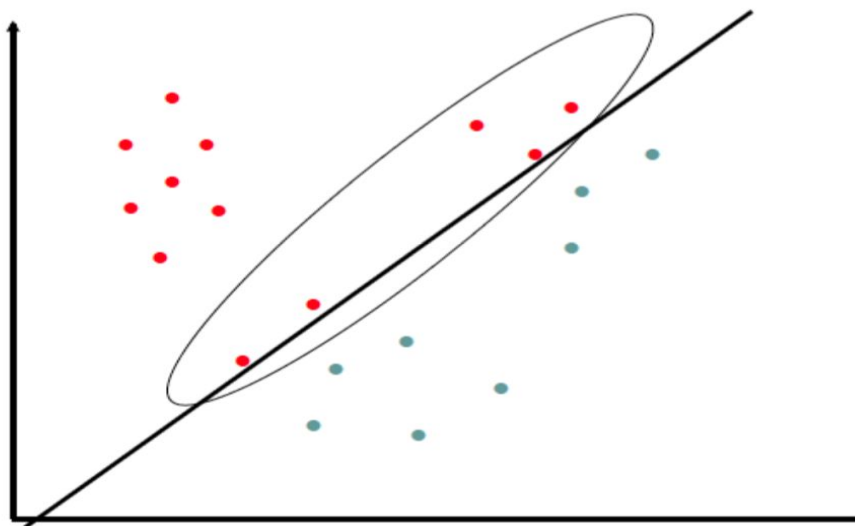
Positivos = 7

Negativos = 13

Total = 20

Ejemplo

Distribución clasificada con error



$$\text{Accuracy} = 15/20 = 0.75$$

$$\text{Recall} = 7/12 = 0.58$$

Clasificado como	Clase Original	
	Si	No
Si	7	5
No	0	8

$$\text{Precision} = 7/7 = 1.0$$

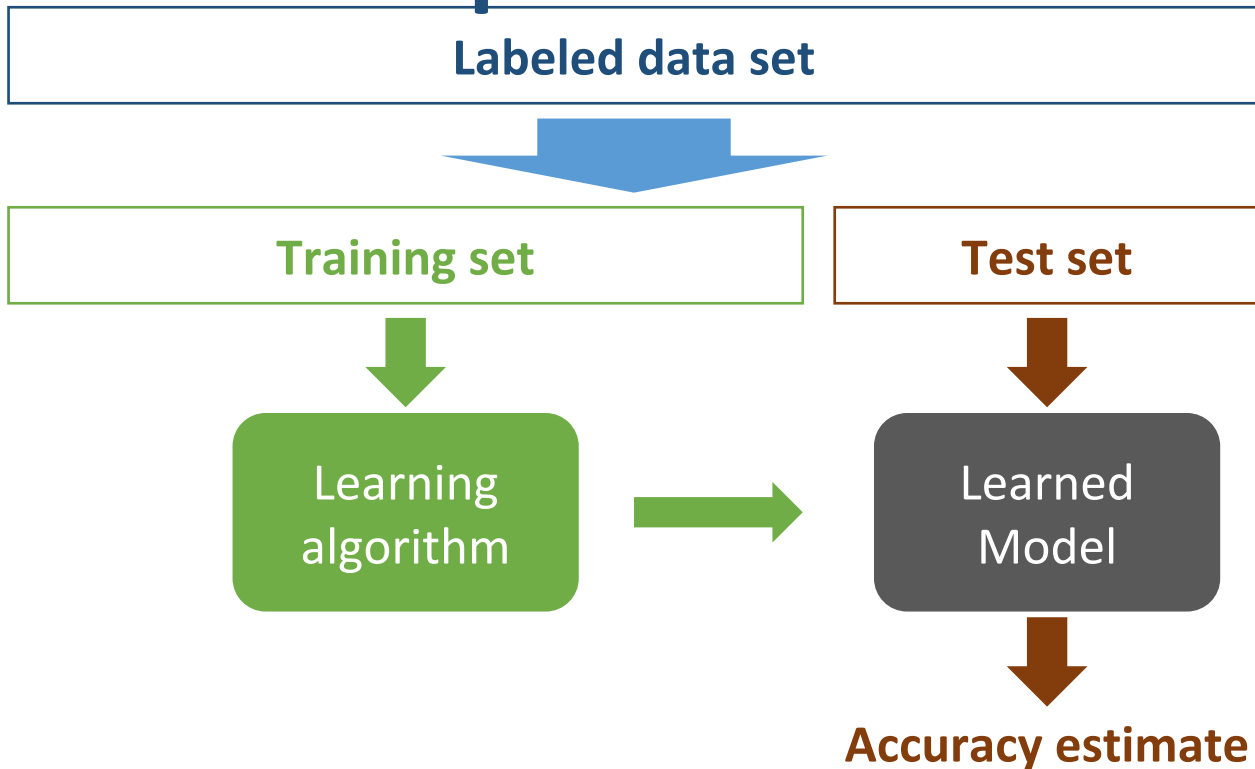
$$F1 = (2 \times P \times R)/(P + R) = 0.73$$

Matriz de confusión para múltiples clases

True positive	False positive
True negative	False negative

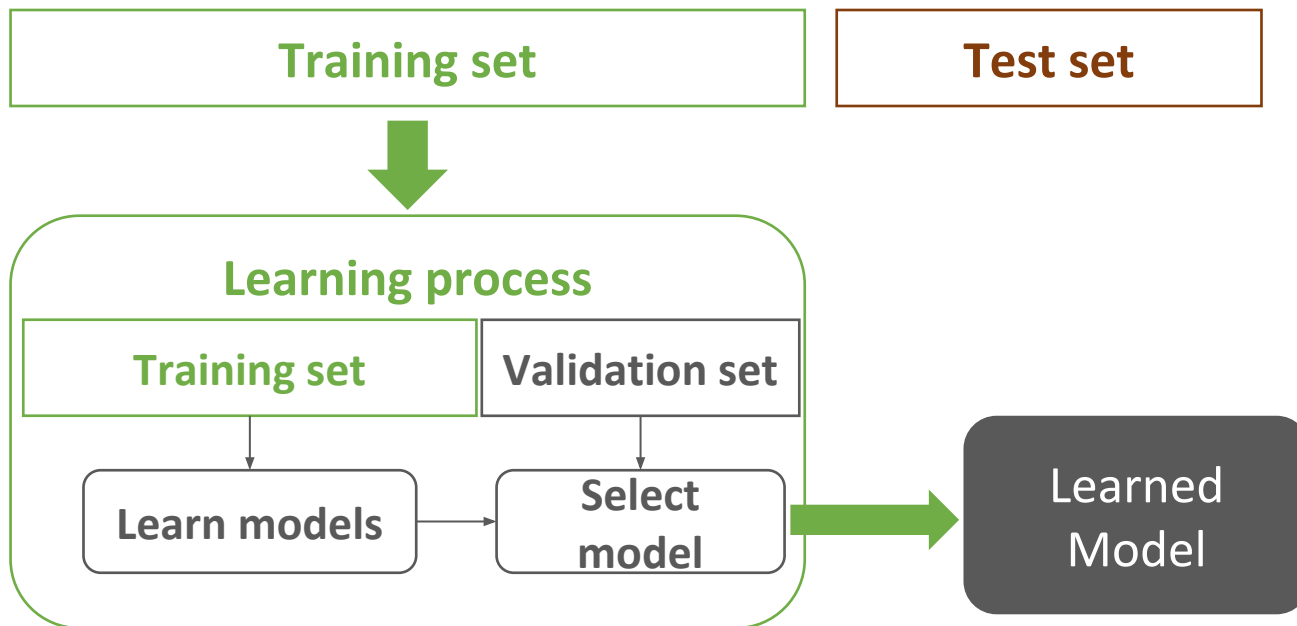
C.Matrix	1	2	3	4	5	6	ACTUAL	RECALL
1	339	15	5	0	0	0	359	94.43%
2	15	305	14	0	0	0	334	91.32%
3	6	10	242	0	0	0	258	93.80%
4	0	0	0	302	30	0	332	90.96%
5	0	0	0	15	368	0	383	96.08%
6	0	0	0	0	0	394	394	100.00%
PREDICTED	360	330	261	317	398	394	2060	94.43%
PRECISION	94.17%	92.42%	92.72%	95.27%	92.46%	100.00%	94.51%	94.66%

Conjuntos de Evaluación: Train – Test split



Conjuntos de Evaluación:

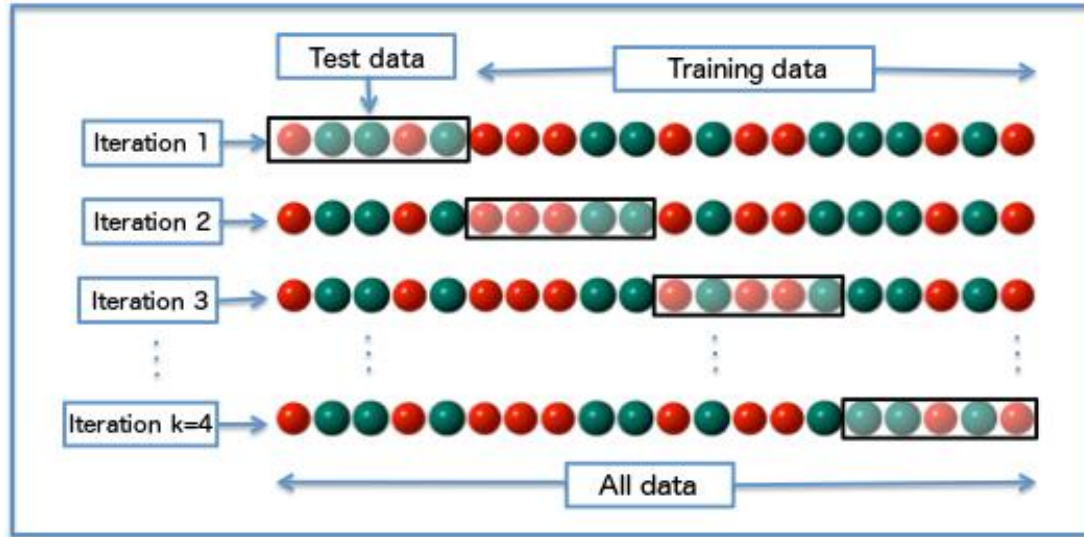
Validation set



Validación de modelos

- Objetivo: seleccionar el modelo correcto.
- Se quiere evitar las limitaciones de un simple “split” de conjuntos (¿debo darle prioridad al “train” o “test” set?)
- Métodos:
 - Random sampling
 - Stratified sampling
- Estrategias conocidas:
 - K-fold cross validation
 - Leave-one-out

Validación de modelos: *k*-fold cross validation

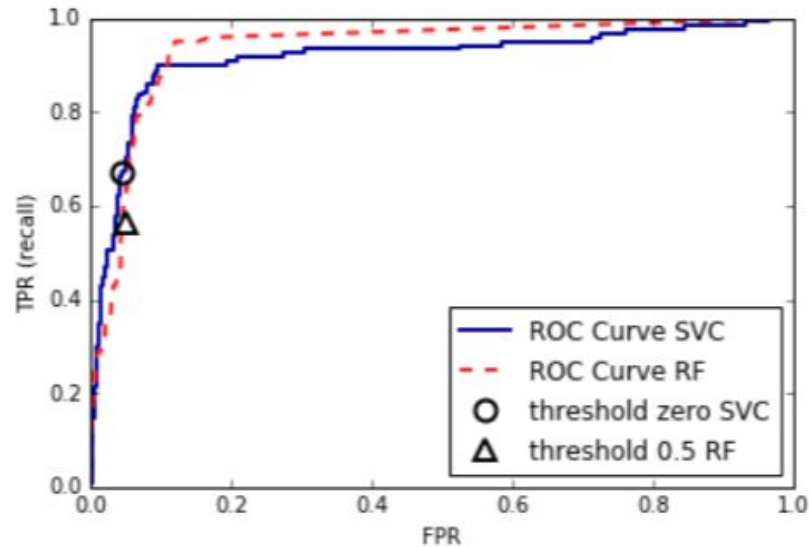


By Fabian Flöck - Own work, CC BY-SA 3.0,
<https://commons.wikimedia.org/w/index.php?curid=51562781>

Curva ROC

- La curva ROC (Receiving Operating Characteristics Curve) analiza el comportamiento de un clasificador en diversos umbrales.
- Evalúa que tan bien el clasificador asigna de manera ordenada el puntaje (ranking) al conjunto de datos
- El eje vertical es la tasa de verdaderos positivos (TPR) o exhaustividad (recall). El eje horizontal es la tasa de falsos positivos (FPR). Cada punto corresponde a un posible umbral (threshold) o punto de operación.
- La curva ideal pasa por el punto (0,1).
- El peor resultado posible es la diagonal $TPR = FPR$ (random guessing)
- Usualmente se resume en un solo valor: el área bajo la curva (AUC)

Curva ROC



∴ Comparación de la curva ROC de dos clasificadores (Mueller 2016)

Conclusión

¿En verdad aprenden los modelos?

¿Qué similitud encuentra con el aprendizaje natural de los humanos?

¿De qué depende que nuestros modelos sean mejores clasificadores?