(Hitz Research Center), Departamento de LSI, EHU

Sist. de Ayuda a la Decisión:
Evaluación

AitZiber AtutXa
aitziber.atucha@ehu.eus

febrero de 2023

Índice

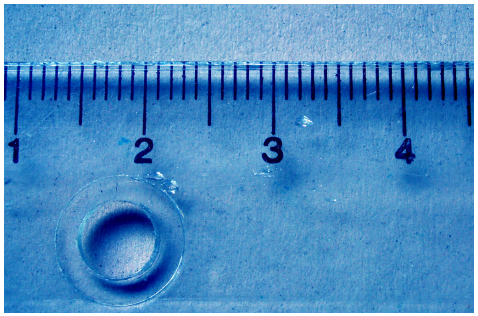
Objetivos de Aprendizaje

- Ser capaces de estimar la calidad de un modelo
- Conocer los esquemas de evaluación
  - no-honesta
  - hold-out
  - k-fold-cross validation

¿Cómo validamos un modelo de ML?

Proceso de aprendizaje automático:

- Recogemos Datos
- Preprocesamos los datos
- Entrenamos nuestro algoritmo/modelo
- Y la gran pregunta.....



Pregunta.

¿Cómo sabemos si ha aprendido lo que se suponía que tenía que aprender?

¿Cómo sabemos si un modelo es suficientemente bueno?

¿Cómo medir la bonanza del modelo?

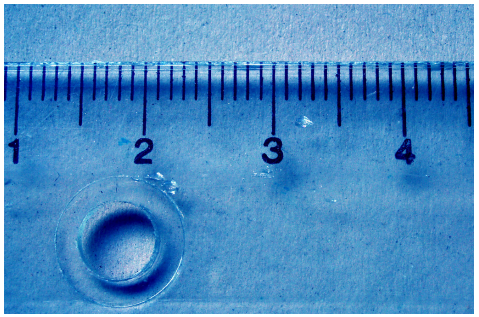
□

imagen: regla"by -Merce- CC

¿Cómo validamos un modelo de ML?

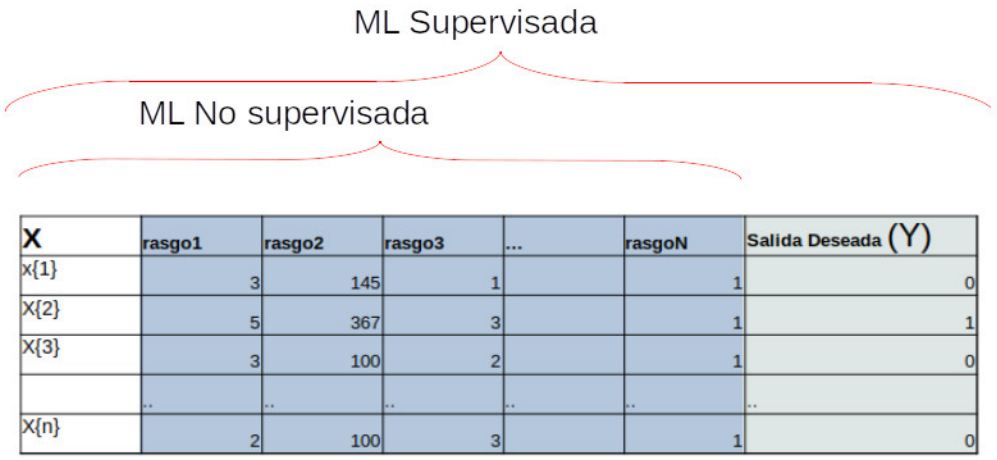
Debemos encontrar una medida objetiva:

- ¿los **errores** qué comete? (minimizarlos)
- ¿los premios que obtenemos? (maximizarlo)
- ¿lo bien que se amolda? (maximizarlo)



¿Cómo validamos un modelo de ML?

- Partiendo de los datos (recordad, supervidado vs. no-sup.)



X	rasgo1	rasgo2	rasgo3	...	rasgoN	Salida Deseada (Y)
x{1}	3	145	1		1	0
X{2}	5	367	3		1	1
X{3}	3	100	2		1	0
	..	..	..	..	..	..
X{n}	2	100	3		1	0

La muestra

$$X = \{x^{\{1\}}, x^{\{2\}}, x^{\{3\}} \ldots x^{\{n\}}\}$$

n *es tamaño de la muestra*
(i.e. núm. instancias)

$$Y = \{y^{d\{1\}}, y^{d\{2\}}, y^{d\{3\}} \ldots y^{d\{n\}}\}$$ (¡Si supervisada! i.e. deseado)

¿Cómo validamos un modelo de ML?

Las diferentes hipótesis (i.e. los diferentes modelos)

$H = \{h_1, h_2, h_3 ... h_j\}$

j núm. de hipotesis

Con cada hipótesis h_i obtengo una predicción
$Y_i^* = \{y^*\{1\}, y^*\{2\}, y^*\{3\} ... y^*\{n\}\}$

El error de la hipótesis h_i

$$\sum_{k=1}^{n} (y^*\{k\} - y^d\{k\})^2$$

n: tamaño de la muestra

$y^*\{k\}$: predicción mi modelo para la instancia k

$y^d\{k\}$: salida deseada para la instancia k

¿Cómo validamos un modelo de ML?

Las diferentes hipótesis (i.e. los diferentes modelos)

$H = \{h_1, h_2, h_3 ... h_j\}$

j núm. de hipotesis

Con cada hipótesis h_i obtengo una predicción
$Y_i^* = \{y^*\{1\}, y^*\{2\}, y^*\{3\} ... y^*\{n\}\}$

El error de la hipótesis h_i

$$\sum_{k=1}^{n} (y^{*\{k\}} - y^{d\{k\}})^2$$

n: tamaño de la muestra

$y^{*\{k\}}$: predicción mi modelo para la instancia k

$y^{d\{k\}}$: salida deseada para la instancia k

Fase de Entrenamiento: Obtención hipótesis

Ejercicio: ¿Qué son los epochs?

Supongamos que pongo el límite aleatoriamente en 8.0. Así toda nota que sea >=8 considero que da lugar a un aprobado. Si no acierto, actualizo el limite con -2.0 (learning rate).

Nombre	**Nota**	**Y deseada (0/1)**
Asier Arrutia	10	Aprobado (1)
Xabier Axular	8	Aprobado (1)
Unai Lopetegi	4	Suspenso (0)
Jorge Zelaia	6	Aprobado (1)

□

Fase de Entrenamiento: Obtención hipótesis

Ejercicio: ¿Qué son los epochs?

Supongamos que pongo el límite aleatoriamente en 8. Así toda nota que sea >=8 considero que da lugar a un aprobado. Si no acierto, actualizo el límite con -2 (*learning rate*).
Epoch 1 (o vuelta completa sobre los datos):

Nb.	**Nota**	Y^d **(0/1)**	Y^* **(0/1)**	**Error**	**Lím.**
A. A.	10	Aprob. (1)	10 >= 8 $\rightarrow Aprob.(1)$	$(1-1)^2$ = 0	8
X. A.	8	Aprob. (1)	8 >= 8 $\rightarrow Aprob.(1)$	$(1-1)^2$ = 0	8
U. L.	4	Susp. (0)	4 < 8 $\rightarrow Susp.(0)$	$(0-0)^2$= 0	8
J. Z.	6	Aprob. (1)	6 < 8 $\rightarrow Susp.(0)$	$(0-1)^2$= 1	6

□

Fase de Entrenamiento: Obtención hipótesis

Ejercicio: ¿Qué son los epochs?

Epoch 2 (o vuelta completa sobre los datos):

Nb.	**Nota**	Y^d **(0/1)**	Y^* **(0/1)**	**Error**	**Lím.**
A. A.	10	Aprob. (1)	10 >= 6 $\rightarrow Aprob.(1)$	$(1-1)^2$ = 0	6
X. A.	8	Aprob. (1)	8 >= 6 $\rightarrow Aprob.(1)$	$(1-1)^2$ = 0	6
U. L.	4	Susp. (0)	4 < 6 $\rightarrow Susp.(0)$	$(0-0)^2$= 0	6
J. Z.	6	Aprob. (1)	6 >= 6 $\rightarrow Aprob.(1)$	$(1-1)^2$= 0	6

¡¡¡¡BRAVO!!! ¡¡¡CONVERGE!!! ¡¡Num.Errores 0!! □

Fase de Entrenamiento: Obtención hipótesis

Ejercicio: ¿Qué son los epochs?

Supongamos ahora que mi muestra hubiese sido otra, y el límite es 8. Si no acierto, actualizo el límite con -2 (*learning rate*).

Epoch 1 (o vuelta completa sobre los datos):

Nb.	**Nota**	Y^d **(0/1)**	Y^* **(0/1)**	**Error**	**Lím.**
A. A.	10	Aprob. (1)			8
X. A.	8	Aprob. (1)			
M. L.	5	Aprob. (1)			
U. L.	4	Susp. (0)			
J. Z.	6	Aprob. (1)			

□

Fase de Entrenamiento: Obtención hipótesis

Ejercicio: ¿Qué son los epochs?

Epoch 1 (o vuelta completa sobre los datos):

Nb.	**Nota**	Y^d **(0/1)**	Y^* **(0/1)**	**Error**	**Lím.**
A. A.	10	Aprob. (1)	10 >= 8 $\rightarrow Aprob.(1)$	$(1-1)^2$ = 0	8
X. A.	8	Aprob. (1)	8 >= 8 $\rightarrow Aprob.(1)$	$(1-1)^2$ = 0	8
M. L.	5	Aprob. (1)	5 < 8 $\rightarrow Susp.(0)$	$(0-1)^2$= 1	6
U. L.	4	Susp. (0)	4 < 6 $\rightarrow Susp.(0)$	$(0-0)^2$= 0	6
J. Z.	6	Aprob. (1)	6 >= 6 $\rightarrow Aprob.(1)$	$(1-1)^2$= 0	6

Número de errores 1. □

Fase de Entrenamiento: Obtención hipótesis

Ejercicio: ¿Qué son los epochs?

Epoch 2 (o vuelta completa sobre los datos):

Nb.	**Nota**	Y^d **(0/1)**	Y^* **(0/1)**	**Error**	**Lím.**
A. A.	10	Aprob. (1)			6
X. A.	8	Aprob. (1)			
M. L.	5	Aprob. (1)			
U. L.	4	Susp. (0)			
J. Z.	6	Aprob. (1)			

□

Fase de Entrenamiento: Obtención hipótesis

Ejercicio: ¿Qué son los epochs?

Supongamos ahora que mi muestra es otra, y el límite seleccionado aleatoriamente es 6. Si no acierto, actualizo el límite con -2 (*learning rate*).

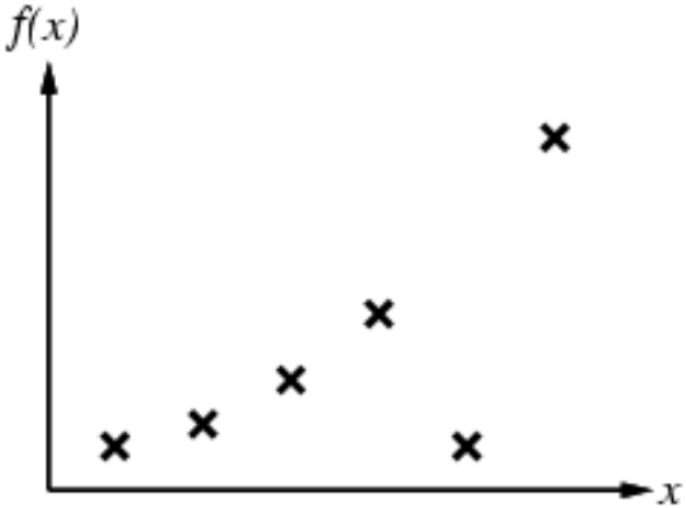
Epoch 2 (o vuelta completa sobre los datos):

Nb.	**Nota**	Y^d **(0/1)**	Y^* **(0/1)**	**Error**	**Lím.**
A. A.	10	Aprob. (1)	10 >= 6 $\rightarrow Aprob.(1)$	$(1-1)^2$ = 0	6
X. A.	8	Aprob. (1)	8 >= 6 $\rightarrow Aprob.(1)$	$(1-1)^2$ = 0	6
M. L.	5	Aprob. (1)	5 < 6 $\rightarrow Susp.(0)$	$(0-1)^2$= 1	4
U. L.	4	Susp. (0)	4 >= 4 $\rightarrow Aprob.(1)$	$(1-0)^2$= 1	2
J. Z.	6	Aprob. (1)	6 >= 2 $\rightarrow Aprob.(1)$	$(1-1)^2$= 0	2

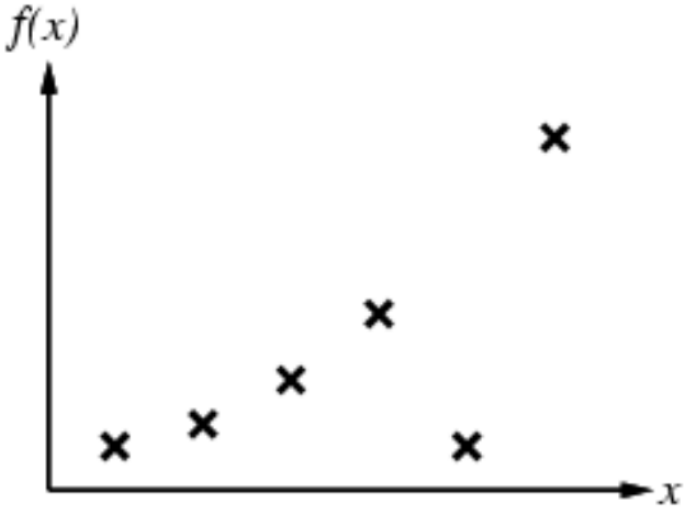
¡¡¡NO CONVERGE, Y NO CONVERGIRÁ!! Learn. rate alto

Número de errores 2. ¡¡Crece!! □

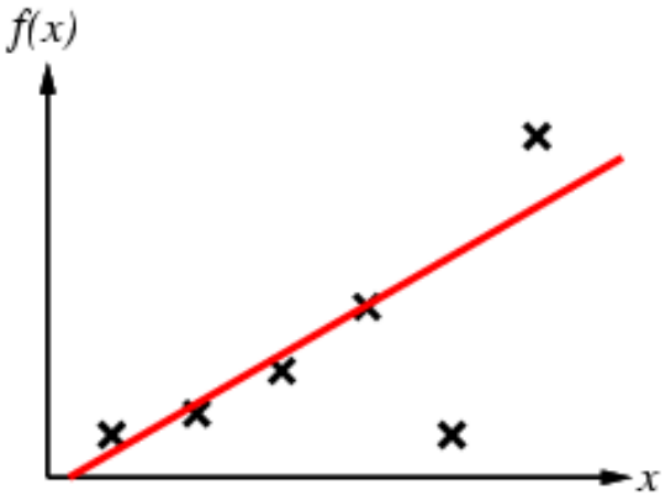
Fase de Evaluación: Intuición



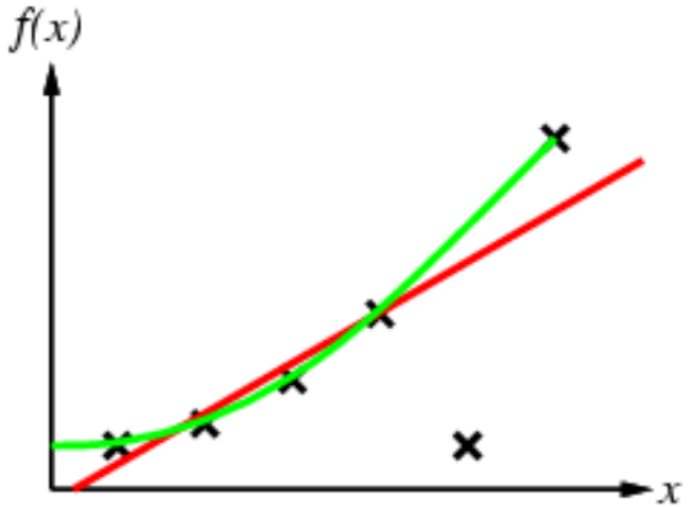
Fase de Evaluación: Intuición



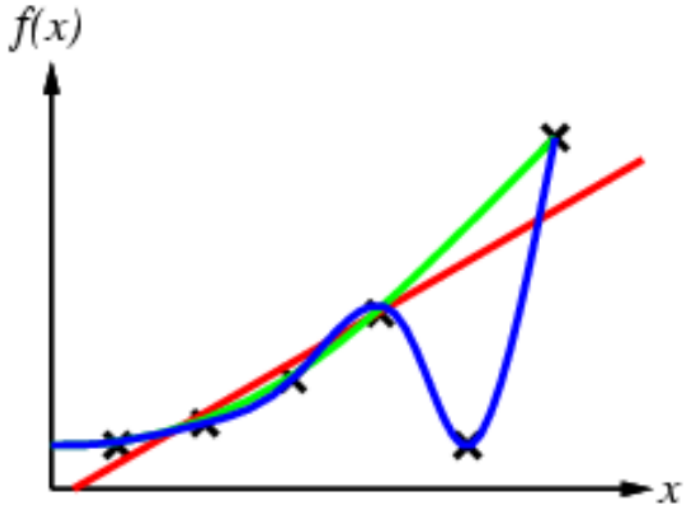
Fase de Evaluación: Intuición



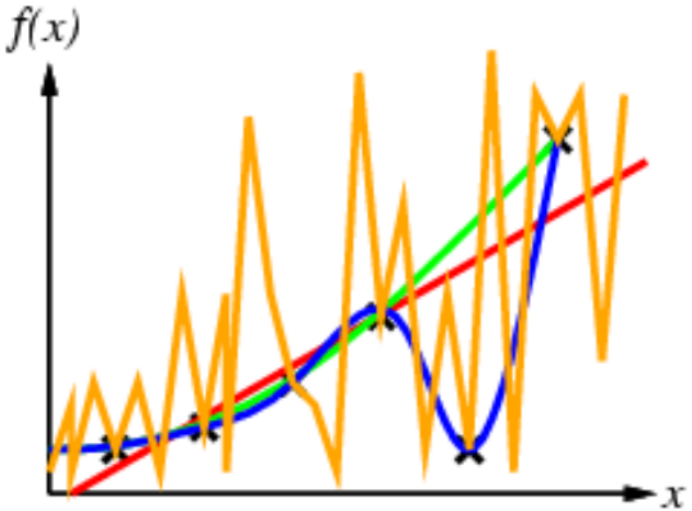
Fase de Evaluación: Intuición



Fase de Evaluación: Intuición



Fase de Evaluación: Intuición



Fase de Evaluación: Esquemas

Una vez entrenado el modelo con suerte hasta convergencia (o error < ϵ). Lo *congelamos*

Esquemas de Evaluación

- Evaluación no-honesta: Sobre la muestra empleada para entrenar.
- **Evaluación honesta:** Sobre otra muestra no empleada en el entrenamiento. Partir la muestra en:
  - Test
  - Desarrollo o development (para probar distintos hiperparámetros (learn. rate 1,2,3)
- Evaluación K-fold: Cuando nuestra muestra no es muy grande

Fase de Evaluación: Generación Muestras

Particiones de entrenamiento, desarrollo y test

- De forma aleatoria
- Estratificada
- Datos muy desbalanceados: Técnicas de under sampling u oversampling

Fase de Evaluación: Métricas

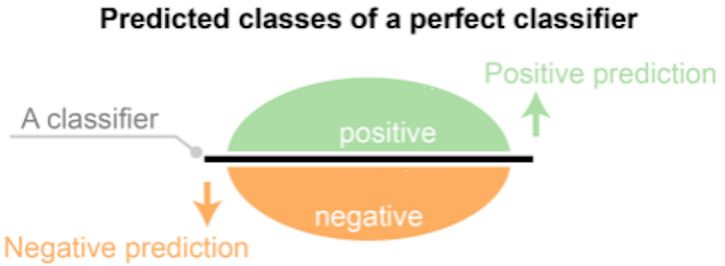
Una vez entrenado el modelo con suerte hasta convergencia (o error < ϵ). Lo *congelamos* y obtenermos la Mátriz de Confusión sobre los datos de desarrollo (o test al final de todo el proceso).

Métricas

- Accuracy.
- Precisión (versión Average con multiclass)
- Recall o sensibilidad (versión Average con multiclass)
- Especificidad o TNR (Tasa negativa real)
- F-Score
- Área bajo la curva de funcionamiento del receptor (ROC) (AUC)
- Pérdida logarítmica o cross-entropy

Fase de Evaluación: TP, FP, TN, FN

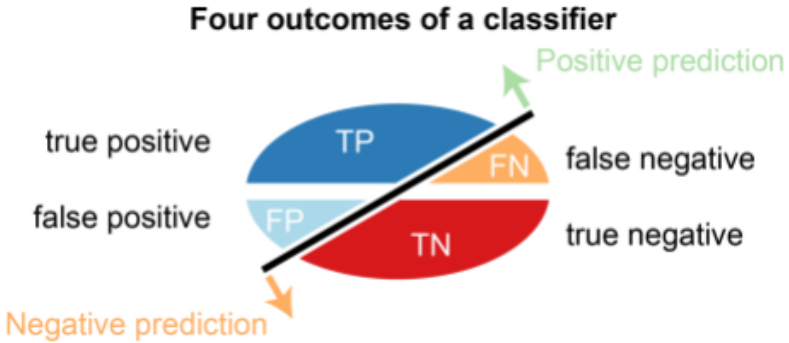
- True Positives TP (Verdaderos Positivos)
- False Positives FP (Falsos Positivos). En este caso 0
- True Negatives TN (Verdaderos Negativos)
- False Negatives FN (Falsos Negativos). En este caso 0



fuente:

Fase de Evaluación: TP, FP, TN, FN

- True Positives TP (Verdaderos Positivos)
- False Positives FP (Falsos Positivos). En este caso NO es 0
- True Negatives TN (Verdaderos Negativos)
- False Negatives FN (Falsos Negativos). En este caso NO es 0

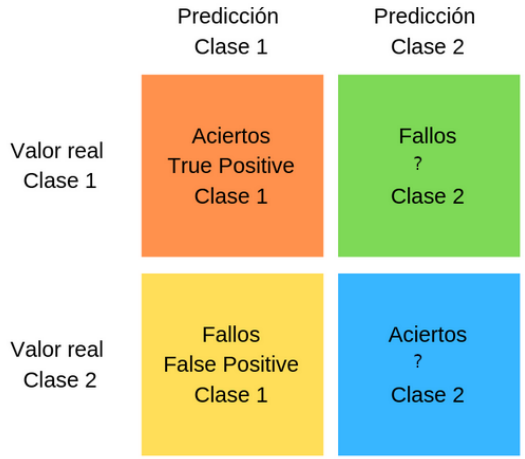


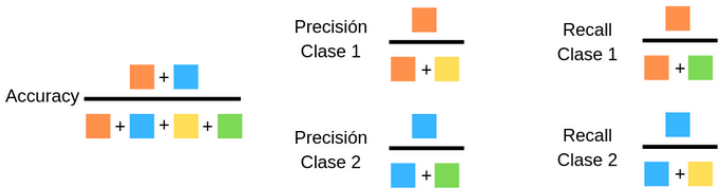
fuente: Saito et al.

Mátriz de Confusión

Proceso de aprendizaje automático (respecto a la clase 1):

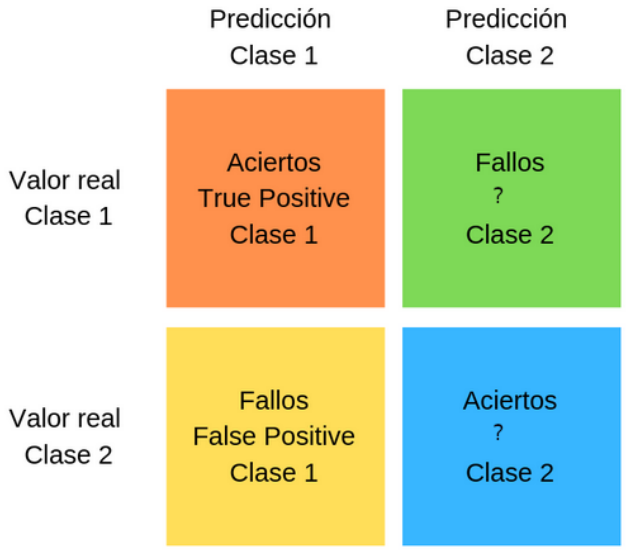
- Verdadero Positivo (TP): Predicho Verdadero y Verdadero en realidad.
- Verdadero Negativo (TN): Predicho Falso y Falso en realidad.
- Falso Positivo (FP): Predicción de verdadero y falso en la realidad.
- Falso Negativo (FN): Predicción de falso y verdadero en la realidad.

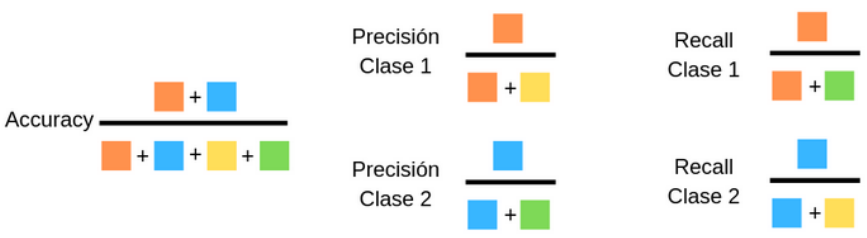




Mátriz de Confusión

- Precision: $\frac{TP}{TP+FP}$
- Recall: $\frac{TP}{TP+FN}$
- F-Score: $\frac{2*Prec*Recall}{Prec+Recall}$





fuente: Juan Ignacio Bagnato

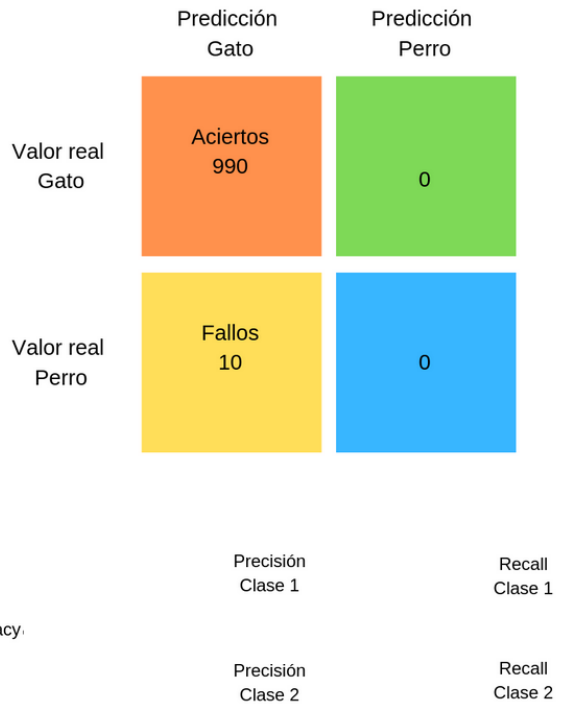
Mátriz de Confusión

¿Soy capaz de...

Calcular las métricas siguientes con respecto a gato? □

- Precision: $\frac{TP}{TP+FP}$
- Recall: $\frac{TP}{TP+FN}$
- F-Score: $\frac{2*Prec*Recall}{Prec+Recall}$

Recomendación: Visionar el vídeo https://www.youtube.con de Andrew Ng para entender la diferencia entre Precisión y Recall



 fuente: Juan Ignacio Bagnato

Evaluación Multiclass

Evaluación multiclase:

Micro and Macro Average Precision, Recall

- Macro-average Precisión: Media de la precisiones de cada clase. Rdo: **Todas clases igual importancia**. Rdo: el impacto de las predominantes se diluye $\frac{PrecCl1+PrecCl2+..+PrecCln}{n}$
- Micro-average Precision: Suma la **contribución de todas la instancias** y computa una media ponderada. Rdo: la precisión de la clase predominante (si en precisión la predicción predominante, si es recall la real predominante) tendrá mayor impacto. $\frac{TPCl1+TPCl2+..+TPCln}{(TPCl1+FPCl1)+(TPCl2+FPCl2)+..+(TPCln+FPCln)}$

Micro and Macro Average Precision, Recall

¿Soy capaz de ...

¿calcular la micro y macro average precision (precision media micro y macro) disponiendo de la siguiente matriz de confusión?

□

Matriz De Confusión Multiclase

Class	TP	FP	FN	Precision	Recall
A	5	2	1	0.71	0.83
B	10	90	7	0.1	0.58
C	15	11	2	0.57	0.88

Figura: Mátriz de Confusión multiclase

Micro and Macro Average Precision, Recall

Matriz De Confusión Multiclase

$$\frac{TP_A+TP_B+TP_C}{TP_A+TP_B+TP_C+FP_A+FP_B+FP_C} \quad \frac{5+10+15}{5+10+15+2+90+11} \quad = 0.22$$

$$\frac{Pre_A+Pre_B+Pre_C}{3} \quad \frac{.71+0.1+.57}{3} \quad = 0.46$$

Micro and Macro Average Precision, Recall

¿Soy capaz de ...

¿obtener la fórmula el Macro, Micro-average Recall?

□

Evaluación Multiclass

Evaluación multiclase:

Micro and Macro Average Recall

- Macro-average Recall (Rdo): el impacto de las predominantes se diluye $\frac{RecCl1+RecCl2+..+RecCln}{n}$
- Micro-average Recall (Rdo): la recall de la clase predominante mayor impacto. $\frac{TPCl1+TPCl2+..+TPCln}{(TPCl1+FNCl1)+(TPCl2+FNCl2)+..+(TPCln+FNCln)}$

Mátriz de Confusión

¿Soy capaz de...

¿Soy capaz de a partir de las predicciones de un modelo

- generar la Matriz de confusión?
- calcular la accuracy, precisión, recall (sensitivity), especificity y F-score?

Bajarse el enunciado asociado con este tema de eGela □

Lecturas y vídeos recomendados

- Capítulo 5 del libro de Data Mining que podréis encontrar en eGela (pg de las 163 a 177)
- Visionado del vídeo de Andrew Ng `https://www.youtube.com/watch?v=W5meQnGACGo` para entender la intuición tras la precisión y el recall
- Visionado del vídeo de *Hablando en Data* para saber como calcular la curva ROC y la AUC `https://www.youtube.com/watch?v=TmhzUdPpVPQ`

Bibliografía

- Data Mining: Practical Machine Learning Tools and Techniques Morgan Kaufmann Publishers Inc. (The Morgan Kaufmann Series in Data Management Systems) Ian H. Witten, Eibe Frank, Mark A. Hall

- Takaya Saito and Marc Rehmsmeier (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 10(3):e011843