



Curso DM

Clasificación

(Introducción, Framework, Evaluación)

Primavera 2023

Basado en las slides de Bárbara Poblete

Sobre la Clasificación

- Técnica utilizada en minería de datos.
- Viene del área de Machine Learning.
- Método de “aprendizaje supervisado”.

**Aprendizaje
Supervisado**

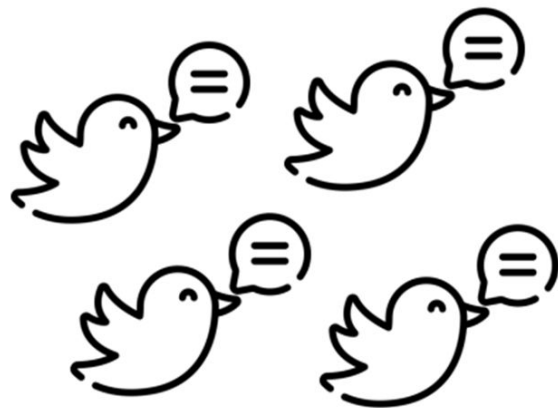


Debemos mostrar ejemplos a la máquina para que aprenda de ellos

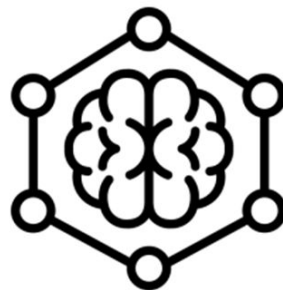
¿Qué es la Clasificación?

- Técnica que “aprende” automáticamente cómo clasificar objetos en dos o más clases determinadas.
- Este aprendizaje se basa en datos previamente etiquetados (clasificados).
- Se aplica en caso en que “etiquetar” tiene un alto costo (por ejemplo: trabajo humano experto).

Ejemplo de aplicación



Aprendizaje



Mensajes	¿Relevante para la crisis?
Ascienden 52 los muertos y 1,2 millones de afectados por terremoto en Guatemala: La cifra de muertos por el... http://t.co/YyCHBarU	Si
No sé si preocuparme porque no sentí el temblor de ahorita o qué.	No
...	

Nuevos mensajes	¿Relevante para la crisis?
Fuerte terremoto de 6,5 magnitud en #Ecuador. https://t.co/vU4N7babRb	?

¿Qué es la Clasificación?

Objetivo: Asignar objetos no vistos anteriormente a una clase dentro de un conjunto determinado de clases con la mayor precisión posible.

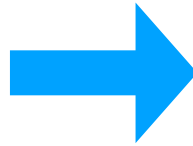
La clase usualmente es denotada por los valores 0,1 para el caso binario o bien $\{1,2,3,\dots,K\}$ para cuando se tienen K categorías.

Enfoque:

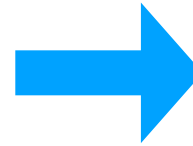
- Dada una colección de registros (conjunto de entrenamiento)
 - cada registro contiene un conjunto de atributos
 - uno de los atributos es la clase (etiqueta) que debe predecirse
- Aprender un modelo para el atributo de clase como función de los otros atributos.

Tarea de mapear set X a una clase y

Input
Conjunto de
atributos (X)



Modelo de
Clasificación



Output
Etiqueta de la
clase (y)

¿Qué es la Clasificación?

Variantes:

- Clasificación **binaria** (fraude/no fraude o verdadero/falso)
- Clasificación **multi-clase** (bajo, medio, alto)
- Clasificación **multi-etiqueta** (más de una clase por registro, por ejemplo, intereses del usuario)

Machine learning vs Data Mining

- Cuando hacemos clasificación en Machine Learning queremos **automatizar una** tarea (i.e., reconocer rostros en imágenes).
- Cuando hacemos clasificación en Data Mining queremos **encontrar un patrón** en los datos (i.e., queremos entender cómo se relaciona x con y por medio de un modelo predictivo).

Muchas veces usamos los mismos tipos de modelos pero con objetivos distintos.

Componentes principales

- Conjunto de entrenamiento
- Algoritmo de clasificación
- Conjunto de validación
- Producen un “Modelo de Clasificación”

Ejemplos de clasificación

Evaluación del riesgo crediticio

- Atributos: su edad, ingresos, deudas, ...
- Clase: ¿recibes crédito de tu banco?

Marketing

- Atributos: productos comprados anteriormente, comportamiento de navegación
- Clase: ¿es usted un cliente objetivo para un nuevo producto?

Detección de SPAM

- Atributos: palabras y campos de la cabecera de un correo electrónico
- Clase: ¿correo electrónico normal o correo basura?

Detección de sentimiento

- Atributos: palabras del mensaje.
- Clase: ¿el texto transmite un sentimiento negativo?

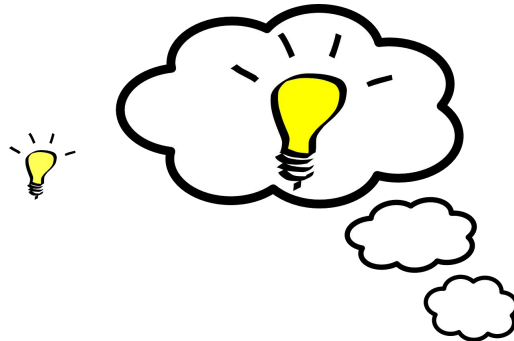
Identificación de células tumorales

- Atributos: características extraídas de radiografías o resonancias magnéticas
- Clase: células malignas o benignas

Actividad

Mencionen 3 tareas de clasificación, ¿cuáles serían los atributos y cuál la clase objetivo para cada tarea?

Reunirse y discutir en grupo (5 minutos)



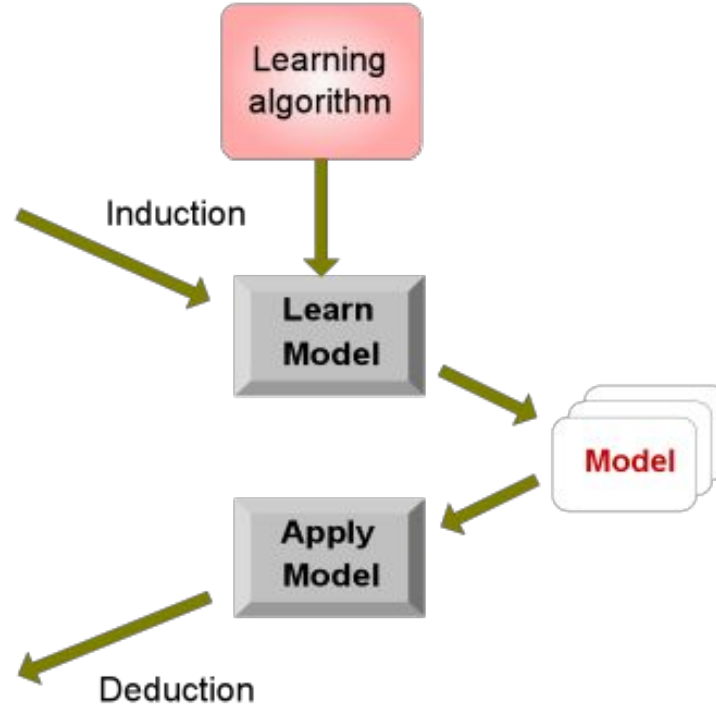
Proceso de Clasificación

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

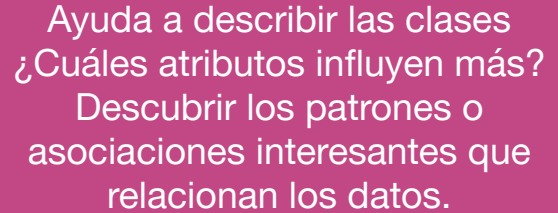
Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Usos de los modelos

Descriptivo: el modelo se utiliza como una herramienta descriptiva.



Ayuda a describir las clases
¿Cuáles atributos influyen más?
Descubrir los patrones o
asociaciones interesantes que
relacionan los datos.

Predictivo: se utiliza para predecir la clase de objetos nuevos.

Nota sobre la clasificación

- es mejor para **datos binarios y nominales**,
- **no es tan bueno para ordinales**, ya que **no consideran relación de orden** entre clases (ej. alto, mediano, bajo), también ignora información de subclases-superclases (mamíferos -> primates -> {humanos, monos}).
- Nos enfocamos en clases binarias y nominales.

Técnicas de clasificación

- Basados en Árboles de Decisión
- Métodos basados en Reglas
- Razonamiento en base a memoria
- Redes Neuronales
- Naïve Bayes y Redes de Soporte Bayesianas
- Support Vector Machines

La Clave del Éxito

- El modelo construido debe ser “generalizable”, es decir, debe aprender bien con muchos tipos de datos nuevos.

¿Cómo saber si un modelo es bueno o no?

- Lo más importante es la **capacidad predictiva del modelo**.
- Hacer predicciones correctas sobre los datos de entrenamiento no es suficiente para determinar la capacidad predictiva.
- El modelo construido debe **generalizar**, es decir, debe ser capaz de realizar predicciones correctas en datos distintos a los datos de entrenamiento.
- Otros factores importantes: interpretabilidad, eficiencia, fairness.



Lectura
recomendada

Theory In, Theory Out: The Uses of Social Theory in Machine Learning for Social Science.

¿Cómo saber si un modelo es bueno o no?

1. Resumimos la capacidad predictiva de un modelo mediante **métricas de desempeño** (performance metrics).
2. Las métricas se calculan **contrastando** los valores predichos versus los valores reales de la variable objetivo.
3. Este se hace con datos no usados durante entrenamiento.
4. Diseñamos experimentos en que comparamos las métricas de desempeño para varios modelos distintos y nos quedamos con el mejor.

Performance Metrics (métricas de desempeño)

- Basadas en **contar** datos **correcta e incorrectamente clasificados**.
- Accuracy (Exactitud): métrica más usada, o
- Error rate (Tasa de error)

Matriz de Confusión

Es una forma estándar de mostrar visualmente los resultados de la clasificación. Detalla el número de aciertos (TP, TN) y errores (FP, FN). Las métricas se pueden calcular directamente a partir de esta matriz de confusión.

		Predicted Class	
		1	0
Actual Class	1	True Positives (TP)	False Negatives (FN)
	0	False Positives (FP)	True Negatives (TN)

Accuracy (Exactitud)

		Predicted Class	
		1	0
Actual Class	1	True Positives (TP)	False Negatives (FN)
	0	False Positives (FP)	True Negatives (TN)

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

Número de casos clasificados correctamente dividido por el número total de casos evaluados.

Limitaciones del Accuracy

Consideren un problema de 2-clases

- Num. de ejemplos de la Clase 0 = 9990
- Num. de ejemplos de la Clase 1 = 10

¿Cuál es el problema?

- Modelo que clasifica todo como Clase 0, accuracy es $9990/10000 = 99.9\%$
- Pero el modelo no detecta nada de la Clase 1 podría ser una $f(x)=0$.
- No es una buena métrica cuando tenemos clases desbalanceadas.



Métricas más sensibles

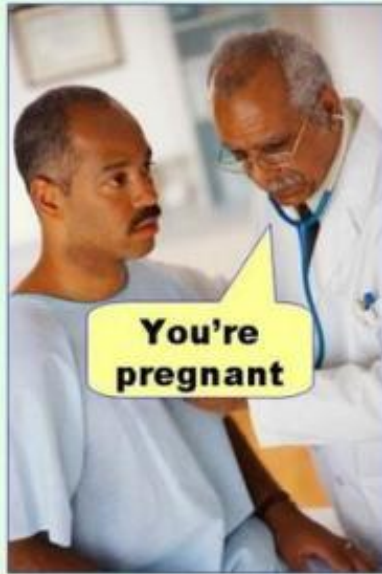
Precisión y Recall

En un problema de clasificación binaria tenemos que escoger cual es la clase positiva. Podemos pensar que clasificar algo como “positivo” es equivalente a “seleccionarlo”.

- **Precision:** % de los casos “seleccionados” que son correctos.
- **Recall:** % de los casos “positivos” que son “seleccionados”.
- Existe un trade-off entre Precision y Recall.

Falsos Positivos y Falsos Negativos

Type I error
(false positive)



Type II error
(false negative)



Precisión

		Predicted Class	
		1	0
Actual Class	1	True Positives (TP)	False Negatives (FN)
	0	False Positives (FP)	True Negatives (TN)

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

Proporción de verdaderos positivos sobre el número de casos predichos como positivos.

Determina lo bueno que es un clasificador para evitar los falsos positivos.

Recall

		Predicted Class	
		1	0
Actual Class	1	True Positives (TP)	False Negatives (FN)
	0	False Positives (FP)	True Negatives (TN)

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

Proporción de verdaderos positivos sobre el número de positivos reales.
Determina lo bueno que es un clasificador para evitar los falsos negativos.

F1-score

		Predicted Class	
		1	0
Actual Class	1	True Positives (TP)	False Negatives (FN)
	0	False Positives (FP)	True Negatives (TN)

$$F1 = 2 \cdot \frac{P \cdot R}{P + R}$$

Media armónica de la precisión y recall. Es conservadora y tiende a estar más cerca del mínimo. Generalmente usamos la F1 measure.

Ejercicio: reincidencia de cáncer

Considere 286 mujeres: 201 no tienen reincidencia de cáncer después de 5 años y 85 sí tienen. Compare los modelos:

M1: "todas reinciden"

Clase real	Clase predicha		
		+	-
		+	-
+	85	0	
-	201	0	

M2: "ninguna reincide"

Clase real	Clase predicha		
		+	-
		+	-
+	0	85	
-	0	201	

Calcular accuracy, precision, recall y F1. Reunirse y discutir en grupo (5 minutos)

Ejercicio: reincidencia de cáncer

Considere 286 mujeres: 201 no tienen reincidencia de cáncer después de 5 años y 85 sí tienen. Compare los modelos:

M1: "todas reinciden"

Clase real	Clase predicha		
		+	-
		+	-
+	+	85	0
	-	201	0

Accuracy: $85/286 = 0.3$

Precision: $85/85 = 1$

Recall: 1

F1: $2 \cdot 0.3 / (0.3 + 1) = 0.46$

M2: "ninguna reincide"

Clase real	Clase predicha		
		+	-
		+	-
+	+	0	85
	-	0	201

Accuracy: $201/286 = 0.7$

Precision: $0/0 = \text{undef}$

Recall: $0/85 = 0$

F1: undef

Matriz de Costo

A veces yo se cuales errores son más costosos y cuales aciertos son más valiosos. Puedo hacer una evaluación sensible al costo.

Clase real	Clase predicha		
	$C(i j)$	clase = +	clase = -
	clase = +	$C(+ +)$	$C(- +)$
	clase = -	$C(+ -)$	$C(- -)$

$C(i|j)$: Costo de clasificar un objeto como clase j dado que es clase i

Calculando el costo de la clasificación

A mayor costo
peor el modelo.

Matrix Costo	Clase predicha		
Clase real	C(i j)	+	-
	+	-1	100
	-	1	0

Modelo M1	Clase predicha		
Clase real		+	-
	+	150	40
	-	60	250

$$\text{Accuracy(M1)} = 400/500 = 0.8$$

$$\text{C(M1)} = -1*150 + 100*40 + 1*60 + 0*250 = 3910$$

Modelo M2	Clase predicha		
Clase real		+	-
	+	250	45
	-	5	200

$$\text{Accuracy(M2)} = 450/500 = 0.9$$

$$\text{C(M2)} = -1*250 + 100*45 + 1*5 + 0*200 = 4255$$

Clasificación Multi-clase

Cuando tenemos k etiquetas, la matriz de confusión es una matriz de $k \times k$.

Docs in test set	Assigned UK	Assigned poultry	Assigned wheat	Assigned coffee	Assigned interest	Assigned trade
True UK	95	1	13	0	1	0
True poultry	0	1	0	0	0	0
True wheat	10	90	0	1	0	0
True coffee	0	0	0	34	3	7
True interest	-	1	2	13	26	5
True trade	0	0	2	14	5	10

Métricas de desempeño por clase

Clasificador binario: One-Vs-Rest

Recall: Fracción de ejemplos de la clase i correctamente clasificados.

Precision: Fracción de ejemplos asignados a la clase i que realmente son de la clase i .

Accuracy: $(1 - \text{error rate})$. Fracción total de ejemplos correctamente clasificados.

Es posible agregarlas para tener una sola métrica que resuma el desempeño del clasificador.

Micro- vs. Macro-Averaging

Si tenemos más de una clase, ¿cómo combinamos múltiples métricas de desempeño en un solo valor?

Macroaveraging: computar métrica para cada clase y luego promediar.

Microaveraging: crear matriz de confusión binaria para cada clase, combinar las matrices y luego evaluar.

Promedio ponderado por soporte (cantidad de ejemplos) por clase

	precision	recall	f1-score
1	1.00	0.67	0.80
2	0.00	0.00	0.00
3	0.00	0.00	0.00
micro avg	1.00	0.67	0.80
macro avg	0.33	0.22	0.27
weighted avg	1.00	0.67	0.80



`sklearn.metrics.classification_report`

Micro- vs. Macro-Averaging: Ejemplo clasificación de Spam

		<i>gold labels</i>			
		urgent	normal	spam	
<i>system output</i>	urgent	8	10	1	$\text{precision}_u = \frac{8}{8+10+1}$
	normal	5	60	50	$\text{precision}_n = \frac{60}{5+60+50}$
	spam	3	30	200	$\text{precision}_s = \frac{200}{3+30+200}$
		$\text{recall}_u = \frac{8}{8+5+3}$	$\text{recall}_n = \frac{60}{10+60+30}$	$\text{recall}_s = \frac{200}{1+50+200}$	

Figure 4.5 Confusion matrix for a three-class categorization task, showing for each pair of classes (c_1, c_2) , how many documents from c_1 were (in)correctly assigned to c_2

Micro- vs. Macro-Averaging: Ejemplo clasificación de Spam

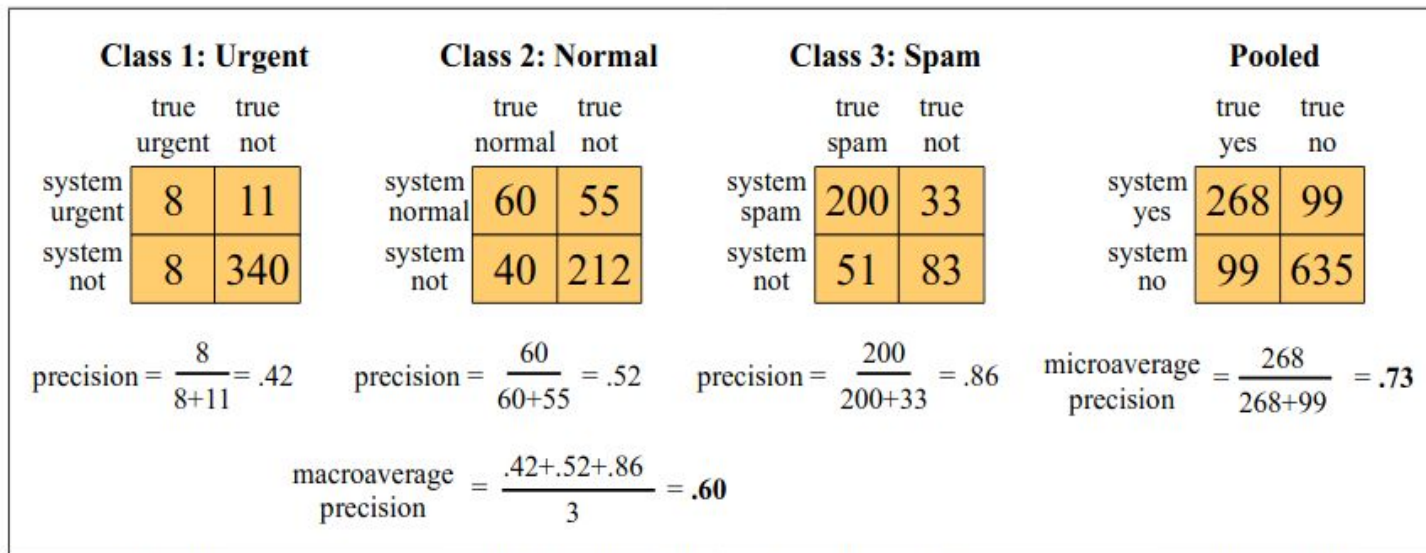


Figure 4.6 Separate contingency tables for the 3 classes from the previous figure, showing the pooled contingency table and the microaveraged and macroaveraged precision.

- Los micro-promedios son dominados por las clases más frecuentes.
- Los macro-promedios pueden sobre-representar a clases minoritarias.

En resumen

Clasificación

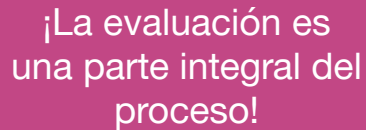
Dada una colección de objetos (set de entrenamiento)

- Cada record contiene un set de atributos, uno de los cuales es su clase.

Encontrar un modelo para el atributo clase, en base a los otros atributos.

Meta: records nuevos deben ser asignados correctamente su clase

- Un set de evaluación se utiliza para medir el desempeño del modelo.



¡La evaluación es
una parte integral del
proceso!

Próxima clase

- Continuación de evaluación del desempeño del modelo.
- Problemas prácticos en la clasificación (Overfitting, Underfitting).
- Algoritmos de clasificación.



Ejemplo: Rumores en Twitter

(presentación externa

<https://prezi.com/r6xefyatyuwg/information-credibility-on-twitter/>)



dcc

CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

www.dcc.uchile.cl