Datos, métricas de desempeño y clasificación

Jocelyn Dunstan Escudero

jdunstan@uc.cl

Departamento de Ciencia de la Computación
& Instituto de Matemática Computacional
Pontificia Universidad Católica de Chile

Santiago, Chile

Objetivos

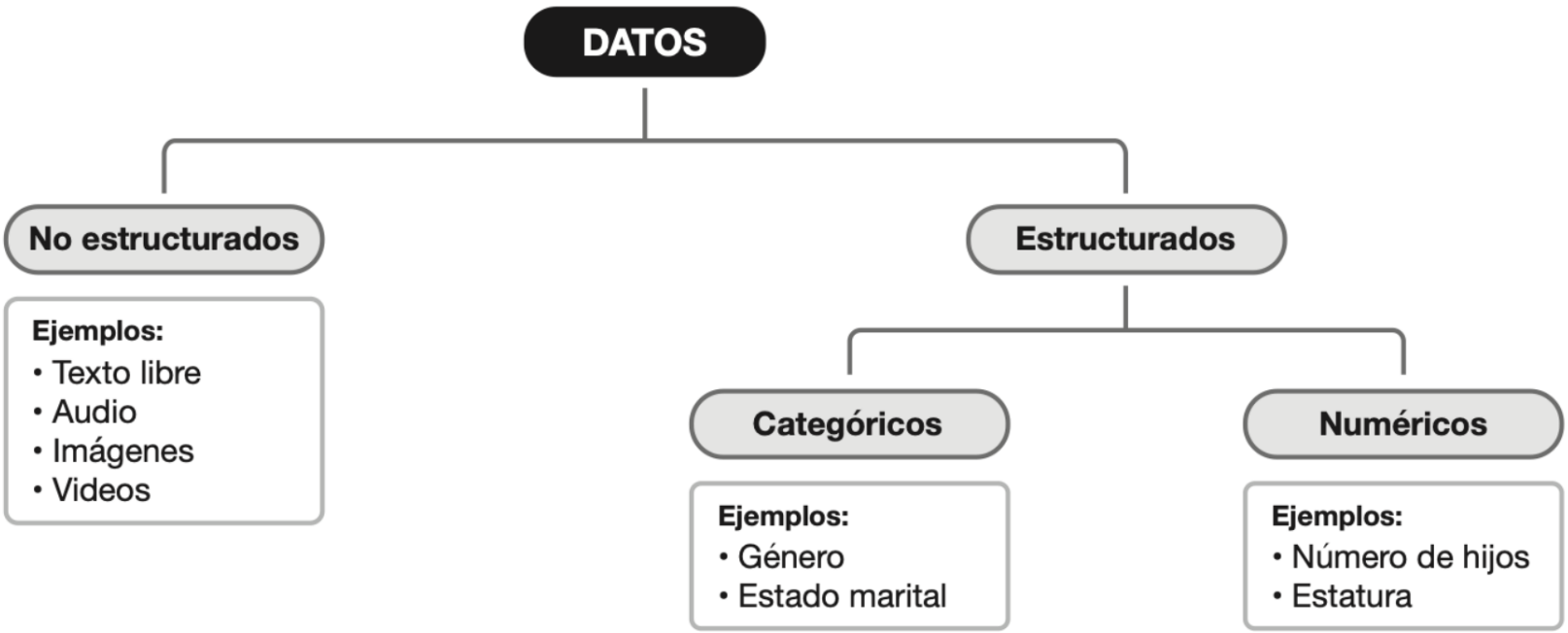
- Definir el concepto de dato y comprender su importancia en el contexto de aprendizaje de máquinas.
- Entender cuándo se trata de un problema supervisado, no-supervisado, semi-supervisado o reforzado.
- Explorar métricas de evaluación de desempeño.
- Comprender los problemas de clasificación binarios y multiclase.

¿Qué es un dato?

Definición

1. Según la OECD: Los datos son características o información, generalmente numéricos, que se recogen mediante la observación.
2. El diccionario de la Universidad de Cambridge establece que los datos son "información, especialmente hechos o números, recopilados para ser examinados, considerados y utilizados para ayudar en la toma de decisiones; o información en forma electrónica que puede ser almacenada y utilizada por un computador"

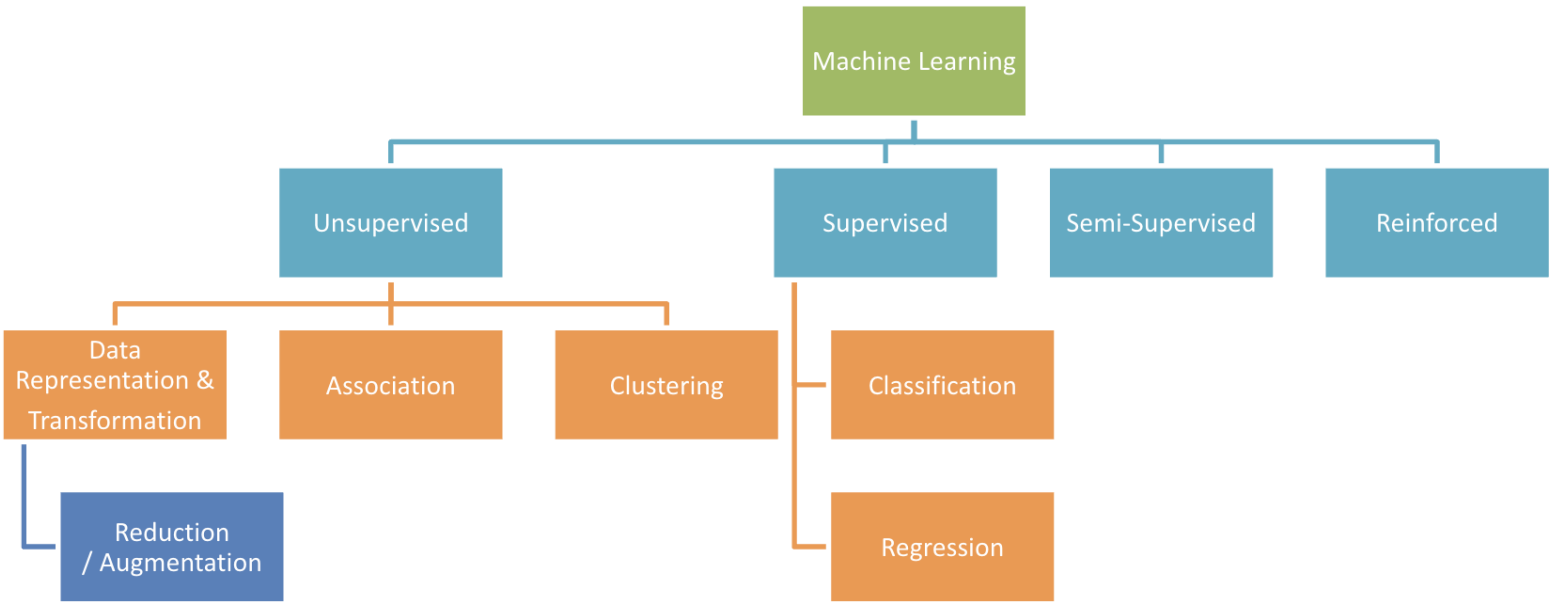
Tipos de datos



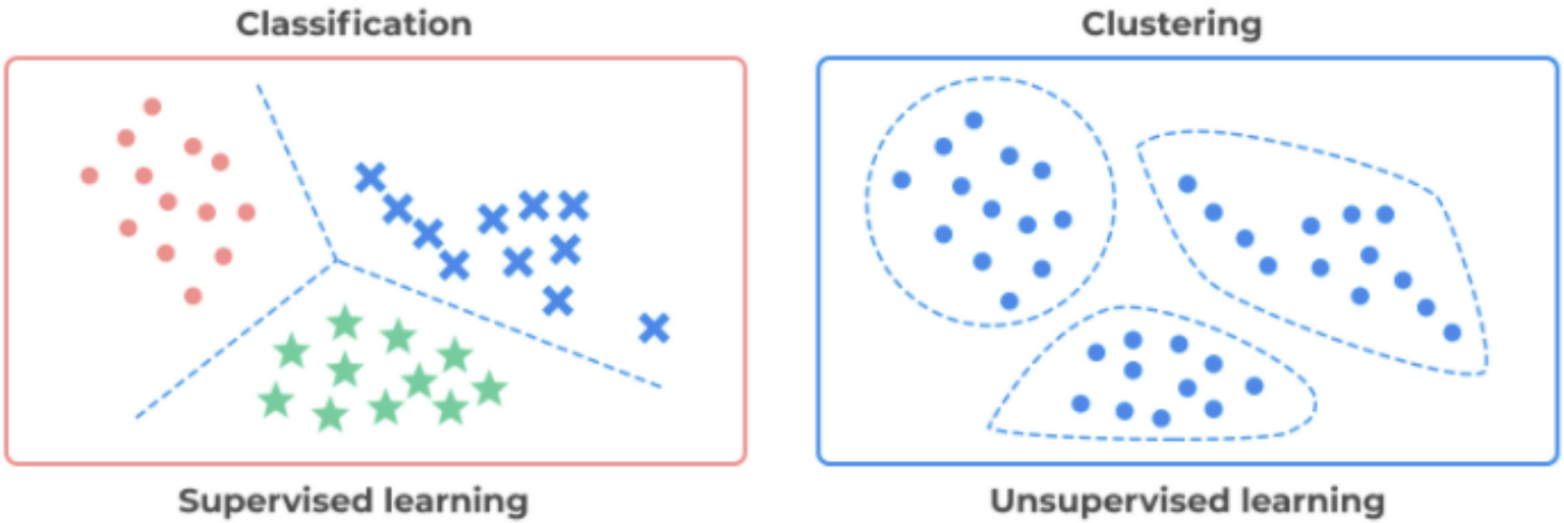
Ruta del dato

Aprendizaje de máquinas

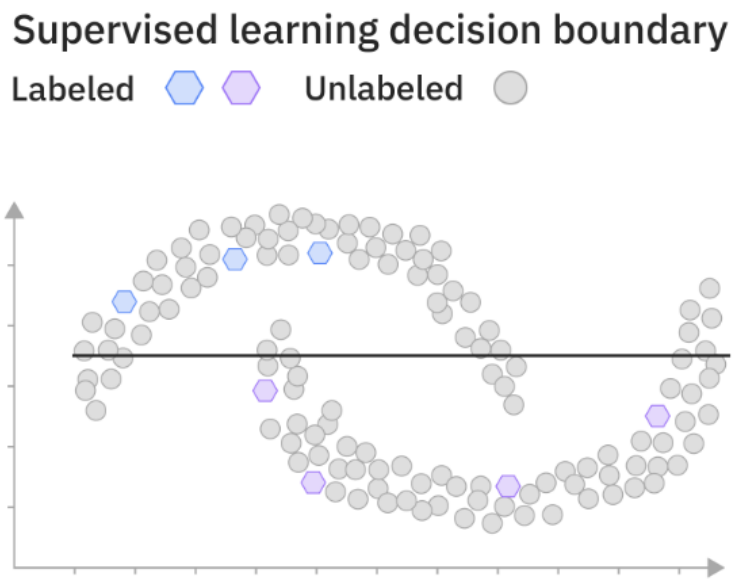
Aprendizaje de máquinas

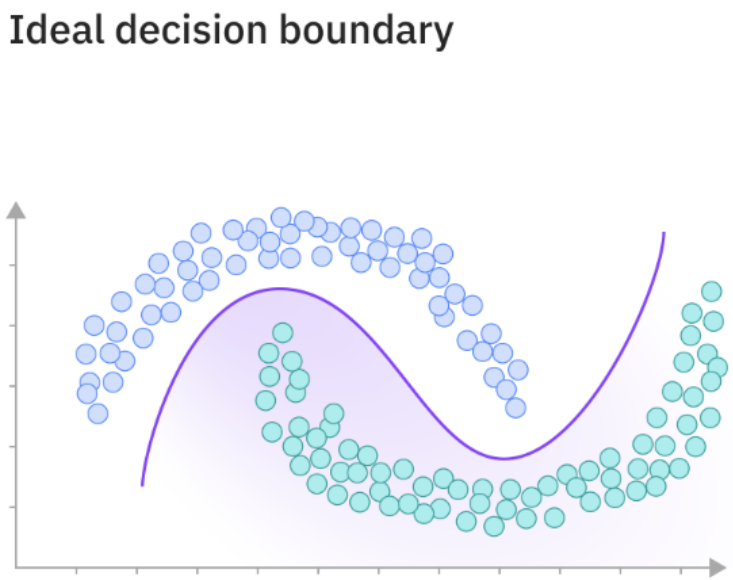


Aprendizaje supervisado vs. no-supervisado



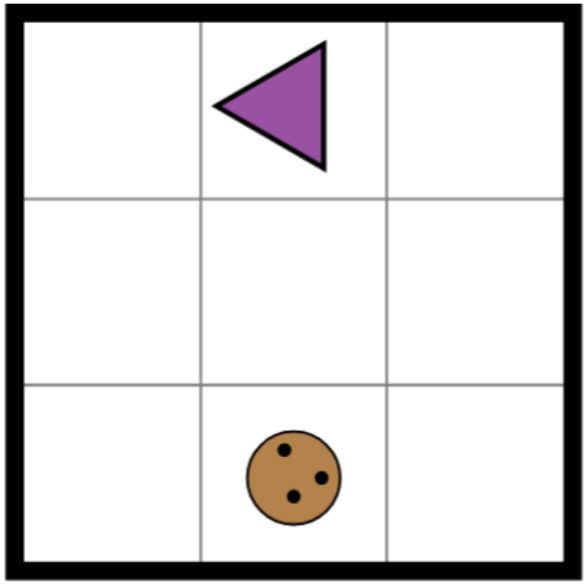
Aprendizaje semi-supervisado





https://www.ibm.com/es-es/think/topics/semi-supervised-learning

Aprendizaje reforzado: Curso de Rodrigo Toro



(0 Reward)

Definición del problema de clasificación

La máquina aprende la relación entre las características de entrada para poder predecir la clase o el valor en el conjunto de entrenamiento.

$$Y = h(X, \theta)$$

Y : target label or value
X : known features
h : relation
θ : set of parameters

Clasificación es un problema supervisado

En machine learning, a la clasificación se le considera como un enfoque de aprendizaje supervisado, pues requiere datos etiquetados.

¿Cómo saber si un modelo es bueno o no?

- Lo más importante es la capacidad predictiva del modelo.
- Pero 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 y eficiencia.

¿Cómo saber si un modelo es bueno o no?

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

Clasificación en la práctica

Proceso de clasificación (pensarlo en el contexto de la medicina)

- Conseguir datos de entrenamiento etiquetados (confiables)
- Entrenar varios modelos de clasificación.
- Evaluar en un dataset de validación.
- Poner el modelo de clasificación en producción.

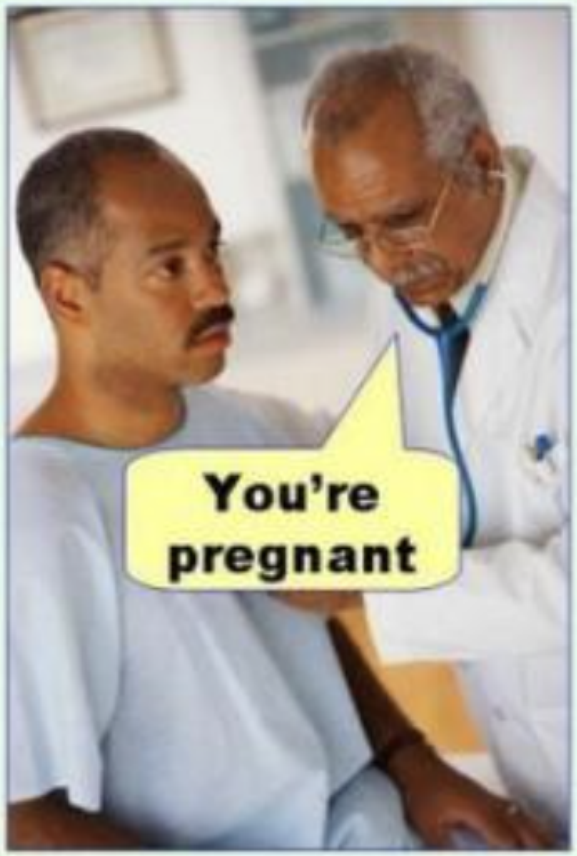
Métricas de evaluación

Matriz de confunsión

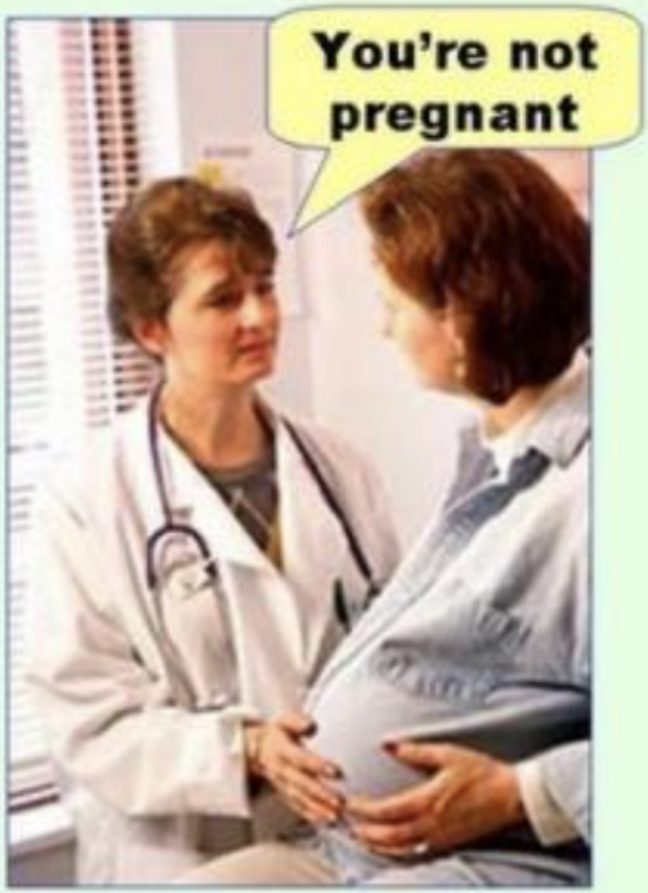
https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

En medicina

Type I error
(false positive)



Type II error
(false negative)



Exactitud (o *accuracy*)

- Basadas en contar datos correcta e incorrectamente clasificados
- Buscamos maximizar la exactitud y minimizar la tasa de error.

$$\text{Exactitud} = \frac{\text{predicciones correctas}}{\text{total de predicciones}}$$

$$\text{Tasa de error} = \frac{\text{predicciones incorrectas}}{\text{total de predicciones}}$$

Otras métricas de evaluación de desempeño

$$\text{Precision} = \frac{VP}{VP + FP}$$

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

Precision y *recall* (o sensibilidad) se calculan para cada una de las clases. Ej: si la clasificación es entre *perro* y *gato*, habrá una precisión asociado a detectar perros, y un recall de esa clase. Una medida muy utilizada es el score F_1, que es la medida armónica entre precision y recall:

$$F_1 = \frac{2(precision \cdot recall)}{precision + recall}$$

Últimos comentarios acerca de evaluación de clasificación

- Es muy normal que exista desbalance de clases y esto tiene que ser tratado con cuidado. Tanto a la hora de entrenar/testear modelos, como cuando se quiere evaluar desempeño.
- Cuando son más de dos clases se puede calcular el micro-F1 (dominado por clases más frecuentes) y el macro-F1 (todas las clases pesan igual).

Clasificación Multiclase

Matriz de confunsión multiclase

Si tenemos k etiquetas, la matriz de confusión es de kxk

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

Basado en material preparado por Felipe Bravo y Bárbara Poblete

Métricas de desempeño por clase

- **Recall:** Fracción de ejemplos de la clase i correctamente clasificado.
- **Precisión:** Fracción de ejemplos asignados a la clase i que realmente son de la clase i.
- **Exactitud:** Fracción total de ejemplos correctamente clasificados.

Micro- vs. Macro-averaging

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

- Macro-averaging: computar métrica para cada clase y luego promediar.
- Micro-averaging: crear matriz de confusión binaria para cada clase, combinar las matrices y luego evaluar.

Ejemplo de clasificación de spam con 3 clases

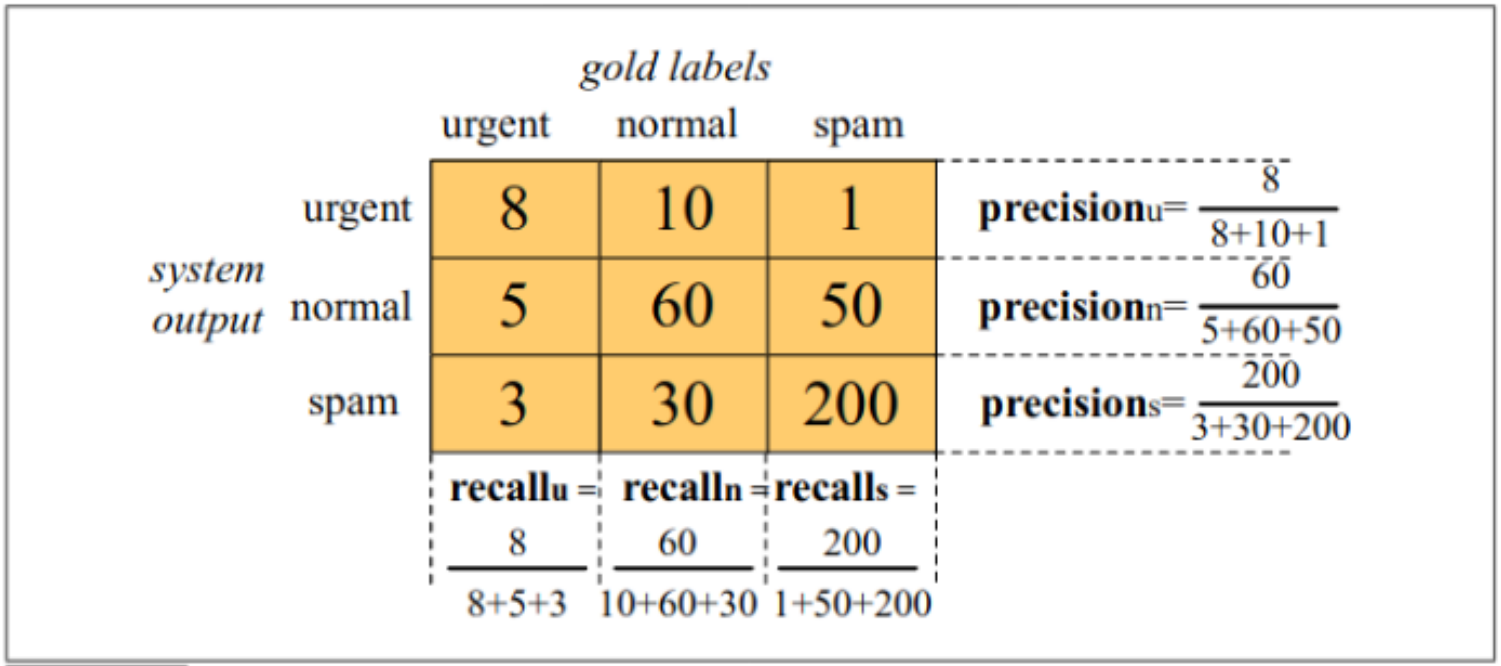


Figure 4.5 Confusion matrix for a three-class categorization task, showing for each pair of classes (c_1, c_2), how many documents from c_1 were (in)correctly assigned to c_2

https://web.stanford.edu/~jurafsky/slp3/4.pdf

Ejemplo de clasificación de spam con 3 clases

Class 1: Urgent

	true urgent	true not
system urgent	8	11
system not	8	340

$$\text{precision} = \frac{8}{8+11} = .42$$

Class 2: Normal

	true normal	true not
system normal	60	55
system not	40	212

$$\text{precision} = \frac{60}{60+55} = .52$$

Class 3: Spam

	true spam	true not
system spam	200	33
system not	51	83

$$\text{precision} = \frac{200}{200+33} = .86$$

Pooled

	true yes	true no
system yes	268	99
system no	99	635

$$\text{microaverage precision} = \frac{268}{268+99} = \mathbf{.73}$$

$$\text{macroaverage precision} = \frac{.42+.52+.86}{3} = \mathbf{.60}$$

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.

`https://web.stanford.edu/~jurafsky/slp3/4.pdf`

Objetivos

- Definir el concepto de dato y comprender su importancia en el contexto de aprendizaje de máquinas.
- Entender cuándo se trata de un problema supervisado, no-supervisado, semi-supervisado o reforzado.
- Explorar métricas de evaluación de desempeño.
- Comprender los problemas de clasificación binarios y multiclase.