Ayudantía 7

Introducción al Machine Learning

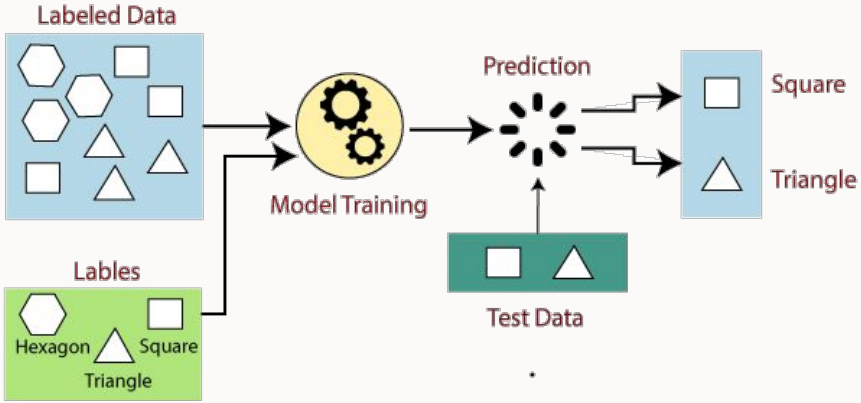
Por Martín Vial y Rodrigo Figueroa

13 de mayo 2024

Aprendizaje Supervisado

Definición

Algoritmos entrenados para **predecir** o **clasificar** datos basándose en el aprendizaje previo sobre ejemplos **etiquetados**.



Necesita un conjunto de datos de entrenamiento que consta de entradas (características) y las salidas deseadas (etiquetas).

Set de datos (pera o manzana)

id	Color	Volumen	Área	....	Peso
1	Verde	2	2		200
2	Verde	3	1		300
3	Café	4	3		250
4	Verde	5	2		210
5	Roja	2	2		280
6	Roja	3	1		350
7	Verde	4	3		100

X_train

X_test

Etiqueta
pera
manzana
pera
pera
manzana
manzana
pera

Set de entrenamiento

id	Color	Volumen	Área	....	Peso
1	Verde	2	2		200
2	Verde	3	1		300
3	Café	4	3		250
4	Verde	5	2		210

Matriz X_train

Etiqueta
pera
manzana
pera
pera

Vector y_train

Set de testeo

id	Color	Volumen	Área	....	Peso
5	Roja	2	2		280
6	Roja	3	1		350
7	Verde	4	3		100

Matriz X_test

Etiqueta
manzana
manzana
pera

Vector y_test
(oculto al algoritmo)

Preprocesamiento

Métodos, técnicas o algoritmos para **limpiar** y **ordenar** los datos antes de realizar la clasificación.

- Representación de características categóricas
- Normalización para dejar en misma escala
- Extracción de características relevantes
- Manejo de datos faltantes

Preprocesamiento (normalización)

	Muestra	A
Peras	1	13234
	:	12129
	50	11957
Manzanas	51	12911
	:	
	125	17288

Columna original

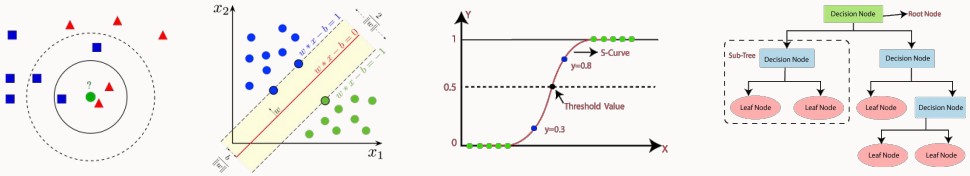
Preprocesamiento (normalización)

	Muestra	A
Peras	1	0.4981
	:	0.3681
	50	0.3479
Manzanas	51	0.4601
	:	
	125	0.9751

Columna normalizada

Diseño y entrenamiento del clasificador

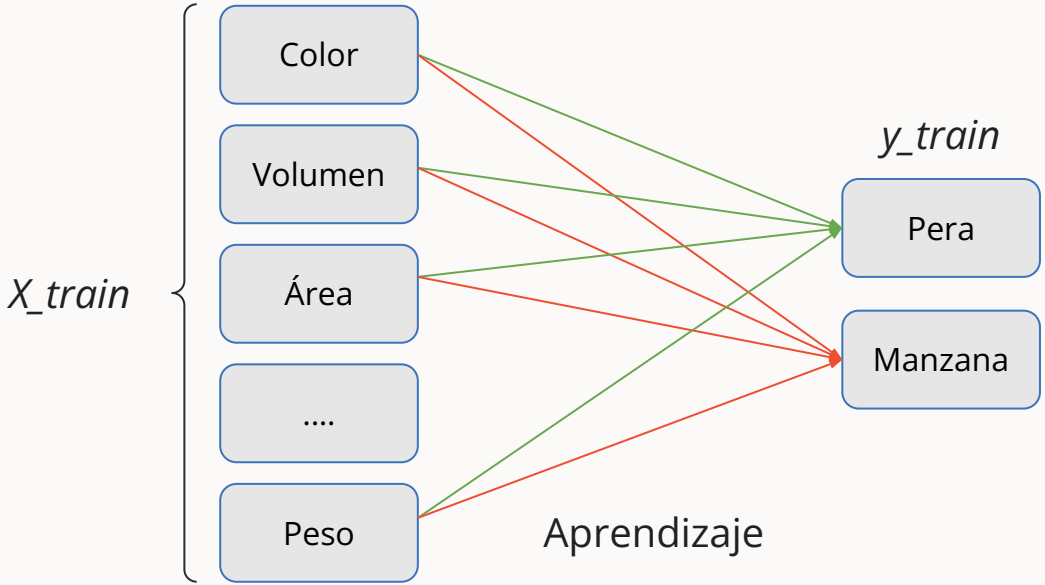
- **Selección de algoritmo:** Se elige un algoritmo de clasificación según el tipo de problema y los datos disponibles.



- **Entrenamiento del algoritmo:** El algoritmo se entrena con el set de entrenamiento y sus etiquetas para saber cómo realizar las clasificaciones.

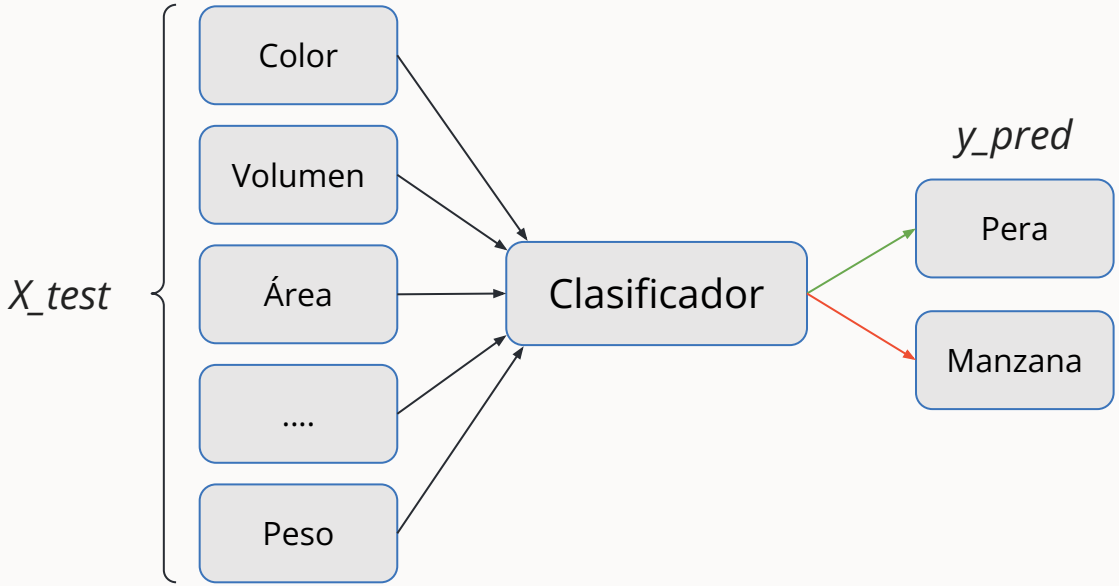
Entrenamiento del clasificador

Aprende **relaciones** y cálculos entre los **atributos** y las **etiquetas**:



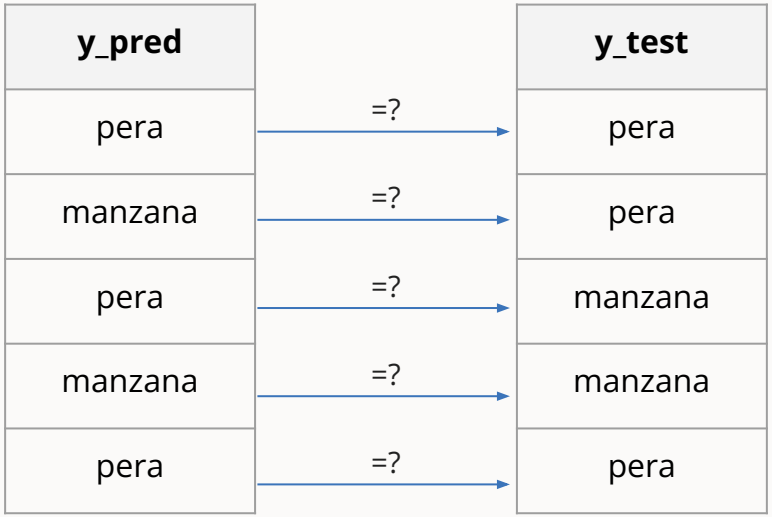
Clasificación y predicciones

Entregamos al algoritmo ya entrenado los **nuevos datos** de testeo para que haga las **predicciones** de la etiqueta de esos datos.

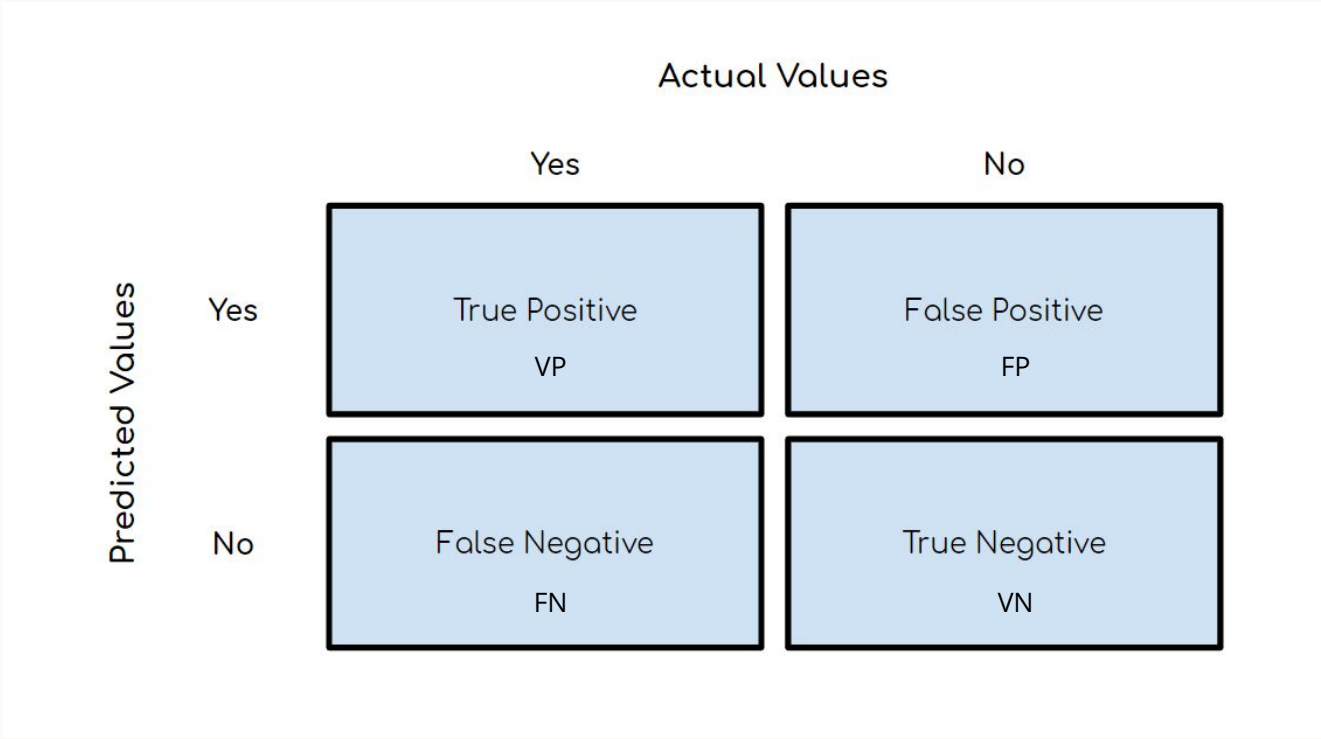


Evaluación

Comparamos la **predicción** de la clasificación del algoritmo con los **valores reales** de la etiqueta del conjunto de testeo y obtenemos distintas **métricas de rendimiento**.



Matriz de Confusión



Accuracy

Proporción de predicciones correctas que realiza un modelo en relación con el número total de predicciones.

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

Útil como una métrica general de la calidad de nuestro modelo sobre la tarea.

Precisión

Mide la **capacidad de un modelo para hacer predicciones positivas correctas** en relación con todas las predicciones positivas realizadas.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}}$$

Útil cuando se desean minimizar los falsos positivos, ya que se enfoca en la calidad de las predicciones positivas.

Recall o sensibilidad

Mide la capacidad de un modelo para **identificar todos los ejemplos positivos** en un conjunto de datos.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

Útil cuando se trata de problemas en los que los falsos negativos (omisiones) son costosos o críticos.

Clasificación multiclase

Problema de clasificación de datos con **varias clases distintas**. Le asignamos un número a cada categoría o creamos varias columnas de etiquetas y poner 1 si pertenece o 0 si no.

id	Color	Volumen	Área	....	Peso
1	Verde	2	2		200
2	Verde	3	1		300
3	Café	4	3		250
4	Verde	5	2		210
5	Roja	2	2		280
6	Roja	3	1		350
7	Verde	4	3		100

X

Etiqueta
pera
manzana
plátano
melón
melón
manzana
plátano

y

Clasificación multiclase

Problema de clasificación de datos con **varias clases distintas**. Le asignamos un número a cada categoría o creamos varias columnas de etiquetas y poner 1 si pertenece o 0 si no.

Etiqueta
pera
manzana
plátano
melón
melón
manzana
plátano

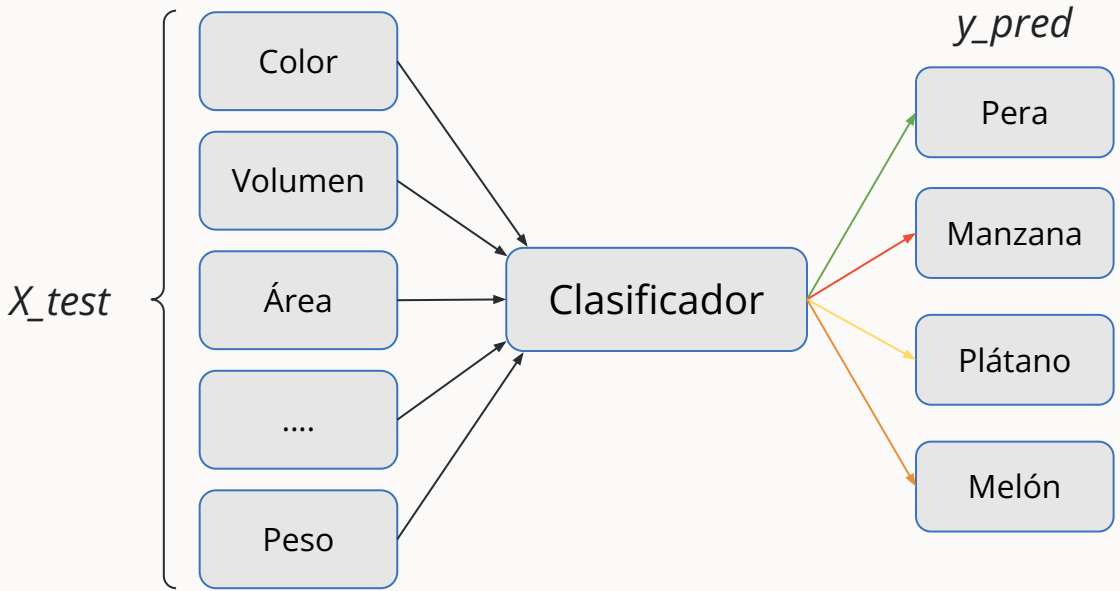
=

Etiqueta
0
1
2
3
3
1
2

=

Pera	Manzana	Plátano	Melón
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
0	0	0	1
0	1	0	0
0	0	1	0

Clasificación multiclase



Micro/Macro promedios

Los **micro-promedios** son dominados por las clases más **frecuentes**

Los **macro-promedios** pueden sobre-representar a más clases **minoritarias**