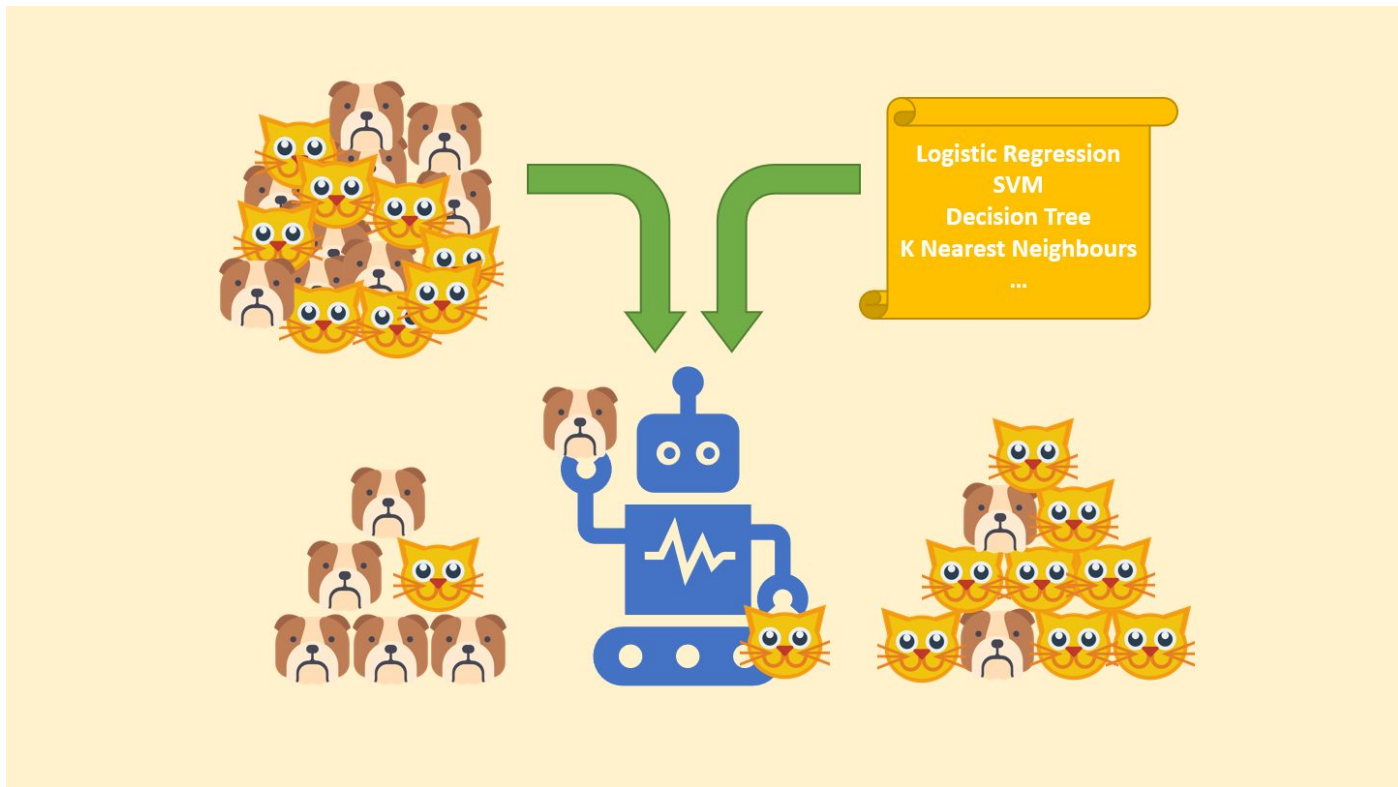




Centro de Investigación
en Computación
Instituto Politécnico Nacional

Clasificación

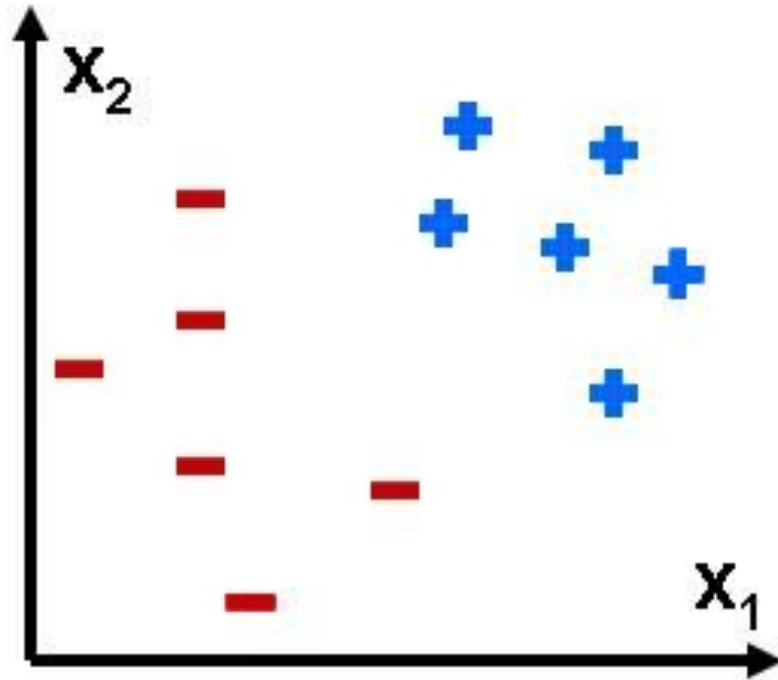
¿Qué es la clasificación?



Tipos de clasificación

Binaria	Multiclase
Multilabel	Multi-output

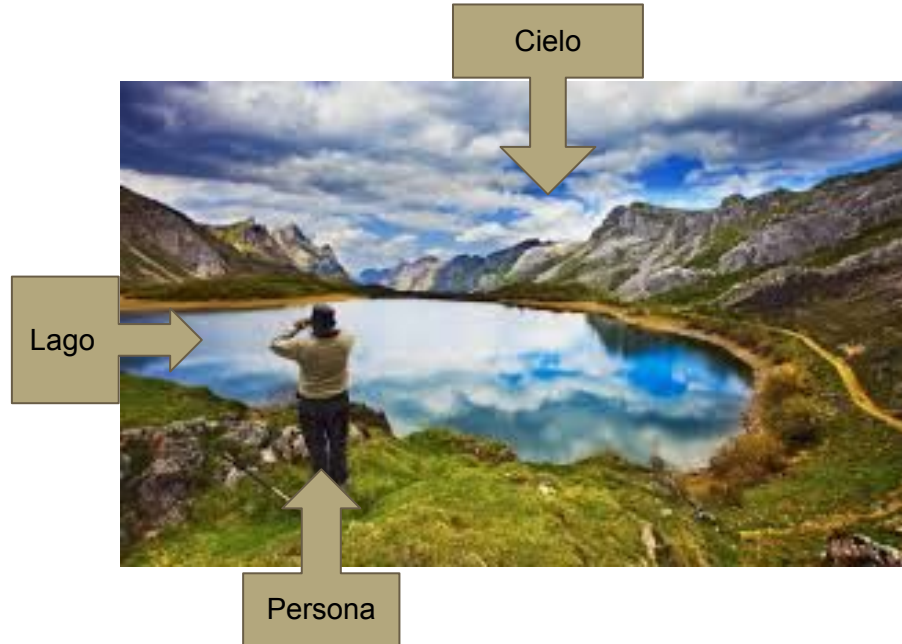
Clasificación binaria



Clasificación multiclase



Clasificación multilabel



Clasificación multi-output



Medidas de desempeño

Matriz de confusión

		Predicción	
		Positivos	Negativos
		Verdaderos Positivos (TP)	Falsos Negativos (FN)
Observación	Positivos	Verdaderos Positivos (TP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FN)	Verdaderos Negativos (TN)

Precisión y Sensibilidad (Recall)

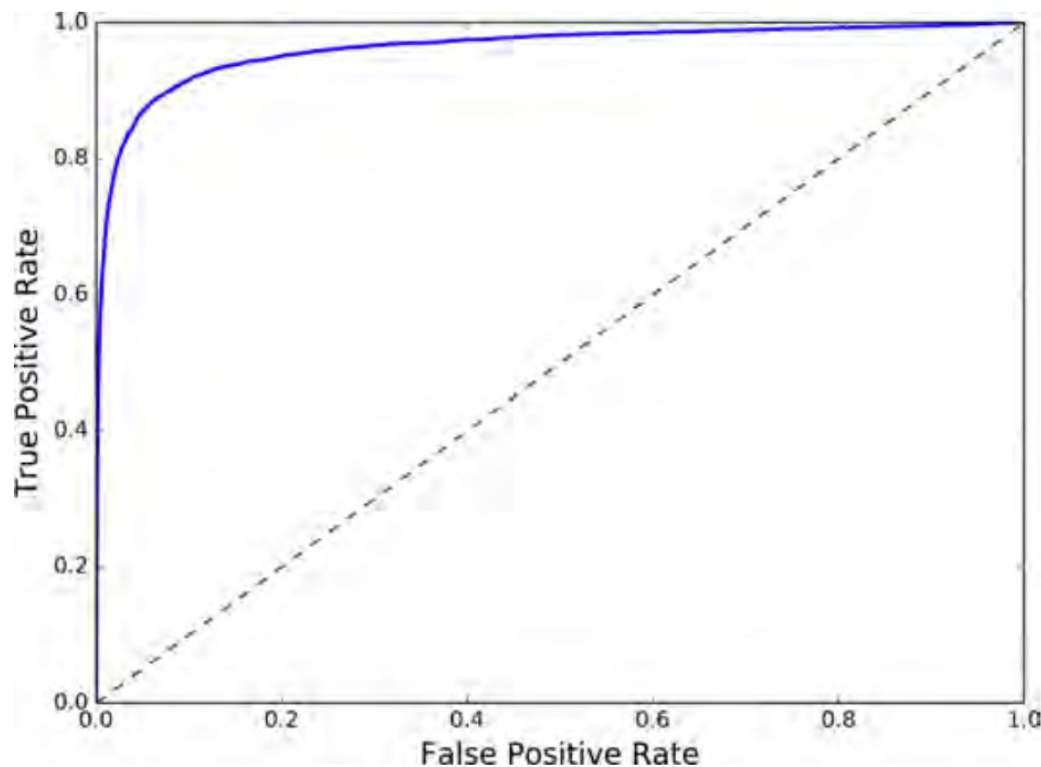
Precisión

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

Recall

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

La curva ROC (Receiver Operating Characteristic)

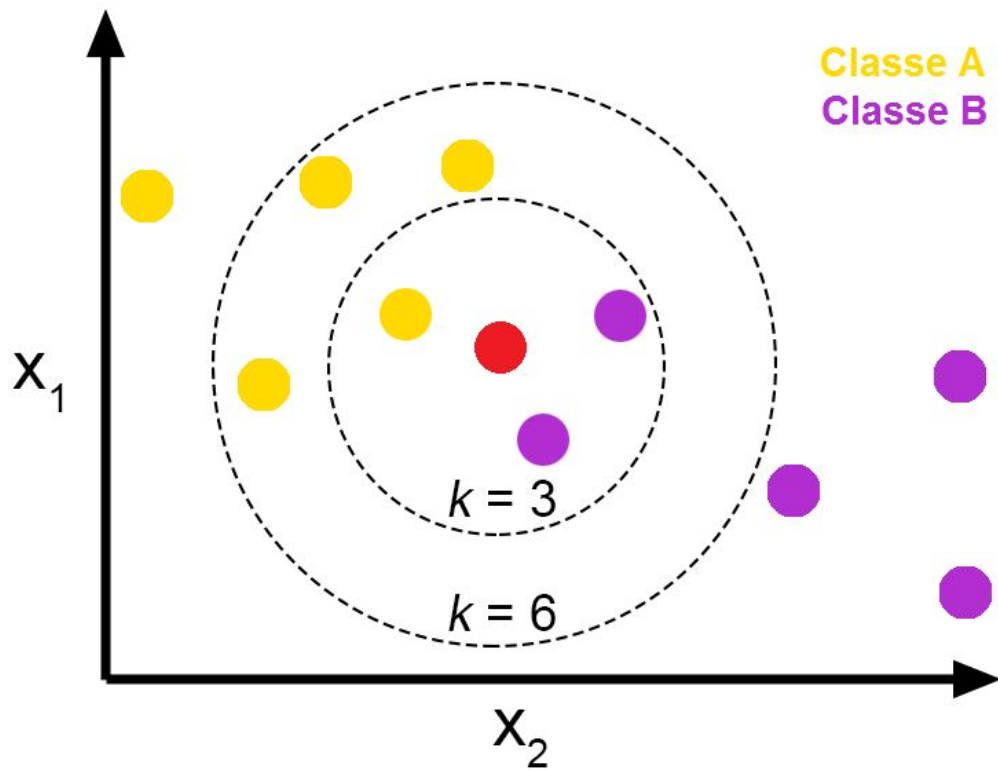


Algoritmos de clasificación

K Nearest Neighbours (KNN)

1. Llega una instancia sin clasificar
2. Calculamos la DISTANCIA entre la nueva instancia y TODAS las instancias de entrenamiento
3. Elegimos los K vecinos más cercanos
4. Calculamos la clase que aparece más entre los K vecinos más cercanos

KNN



KNN

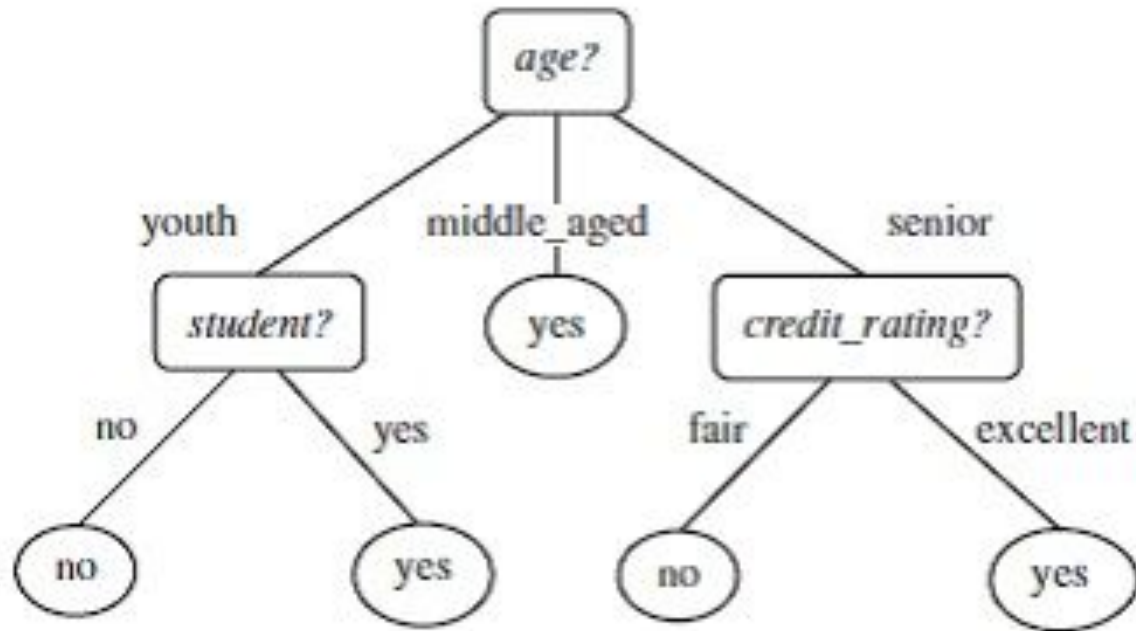
Pros

- Simple y efectivo
- No asume nada sobre la distribución de los datos
- Fase de entrenamiento rápida

Contras

- No crea un modelo
- Fase de clasificación lenta

Árboles de decisión



Árboles de decisión

Pros

- Fáciles de entender e interpretar
- Podemos sacar reglas del modelo
- Pueden usar datos numéricos y categóricos

Contras

- Overfitting
- Si las clases no están balanceadas hay sesgos
- Encontrar un árbol de decisión óptimo es NP

Naive Bayes

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

Naive Bayes

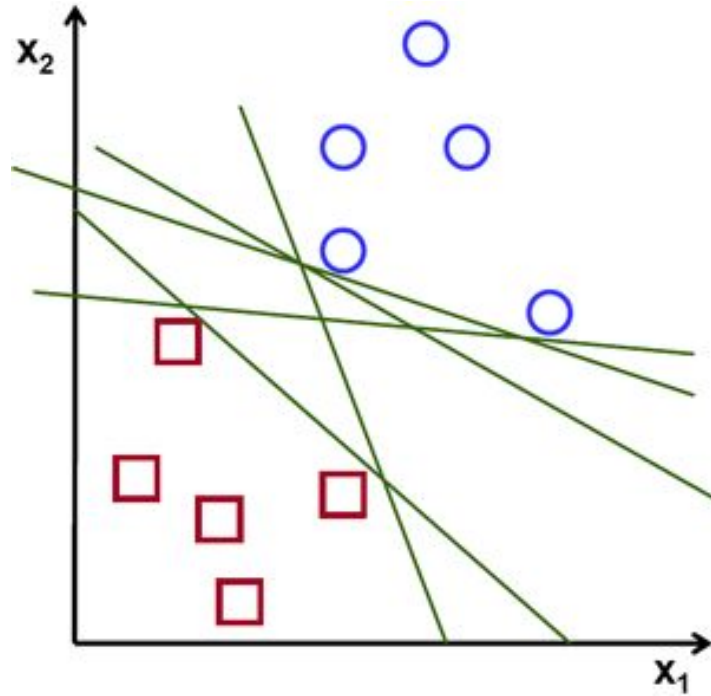
Pros

- Es barato computacionalmente y simple
- Funciona bien con pocos datos
- Es muy usado para clasificar texto

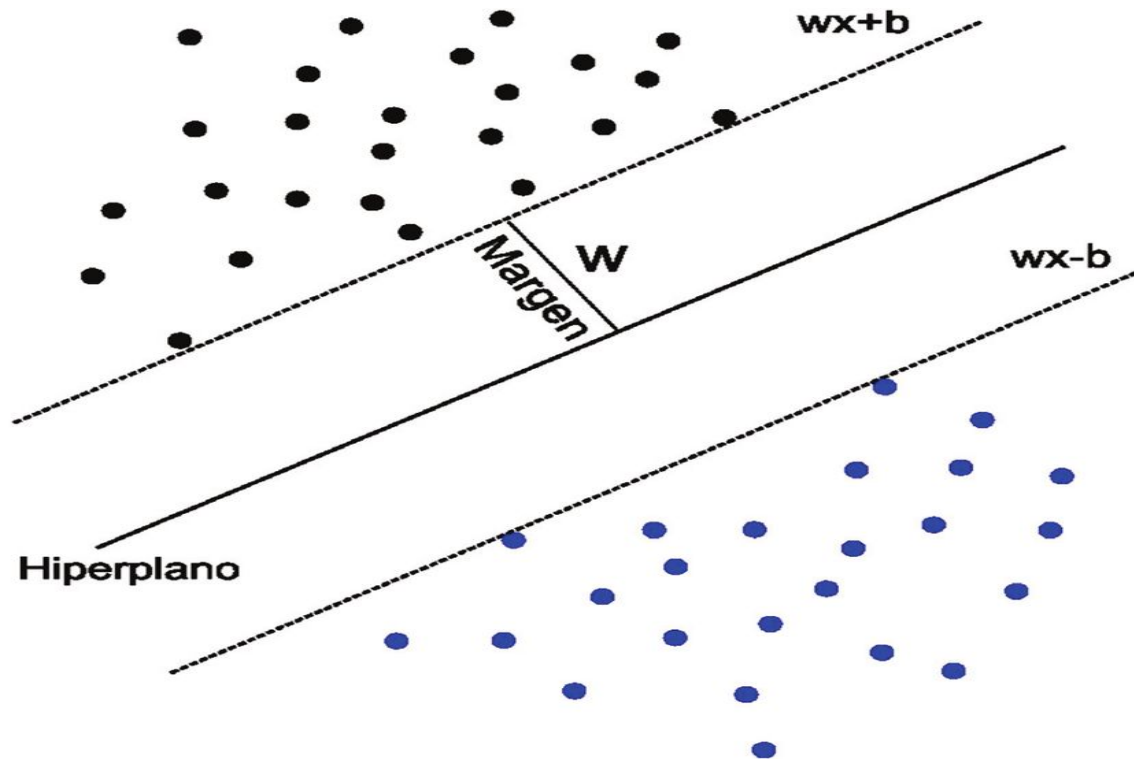
Contras

- Asume que las características son independientes, lo que muy rara vez ocurre con datos reales

Support Vector Machines (SVM)



SVM



SVM

Pros

- Funcionan con altas dimensiones (incluso con más dimensiones que instancias)
- Versátiles
- Eficientes en memoria

Contras

- Overfitting

Funciones Kernel

