Introducción a Machine Learning

Leonardo A. Caravaggio

Abril 2022

Big Data

Se combinan en este tiempo tres procesos que ya tienen fuerte impacto en la sociedad, y que seguirán teniendolo en los próximos años.

- Aparición de nuevos algoritmos
- Mayor generación de datos, capacidad de almacenamiento y disponibilización
- Mayor capacidad de procesamiento computacional

¿Qué es un dato?

Casi todo es o puede ser un dato.

- Números, series de números, tablas con números o palabras o categorías
- Imágenes, audios, videos, textos, patrones de comportamiento

Machine Learning

El objetivo del aprendizaje de máquina es lograr identificar un determinado patrón en un conjunto de datos.

Inferencia no causal

En economía se suele pensar en la determinación de la relación causal. En Machine Learning esto no es necesariamente cierto.

- Ejemplo Jeff Seder

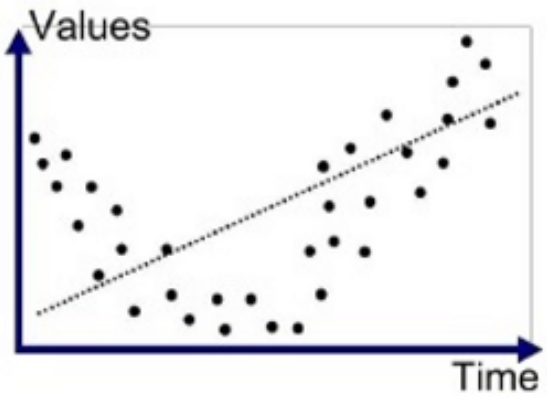
Corte transversal

En economía se usa especialmente series de tiempo. En Machine Learning es más común el uso de datos de corte transversal.

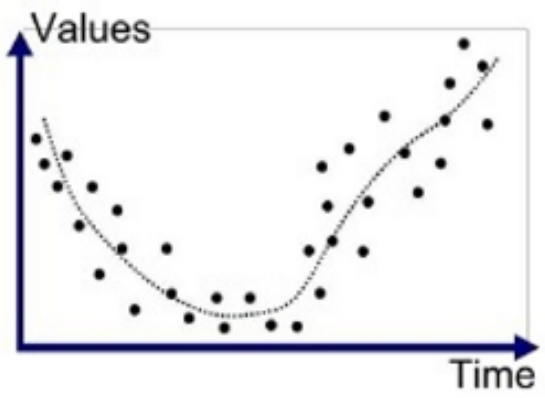
Partición Train/Test

Con el objetivo de lograr identificar si un algorítmo realmente está aprendiendo el patrón detrás del conjunto de datos es práctica común dividir el conjunto de datos en dos. La primera parte se usará para entrenar el algorítmo y la segunda para comprobar el comportamiento aprendido.

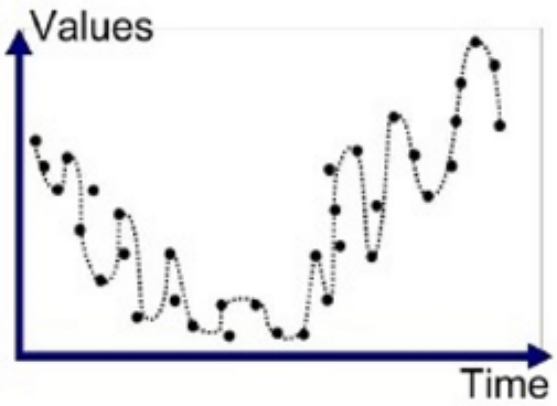
Overfitting



Underfitted



Good Fit/Robust



Overfitted

Regresión / Clasificación

En general se puede pensar en dos problemas distintos a resolver, un problema de clasificación y un problema de regresión.

Supervisado / No supervisado

Si contamos con un conjunto de datos de entrenamiento etiquetados, podemos utilizar técnicas de aprendizaje supervisado. Pero también es posible identificar patrones en datos no etiquetados, esto se conoce como aprendizaje no supervisado.

Algunos algorítmos

Estos son algunos de los algorítmos más conocidos y usados.

- Regresión lineal, regresión logística
- Naive Bayes
- Árboles de decisión
- Support Vector Machines
- K-Means, KNN
- Redes neuronales

Práctica

Veamos ahora algunos de estos algoritmos en la práctica.

Ver Ejercicio 1 en el Repositorio

Matriz de confusión

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Accuracy

La exactitud es la proporción de resultados verdaderos (tanto verdaderos positivos (VP) como verdaderos negativos (VN)) entre el número total de casos examinados (verdaderos positivos, falsos positivos, verdaderos negativos, falsos negativos).

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

Coeficiente de determinación

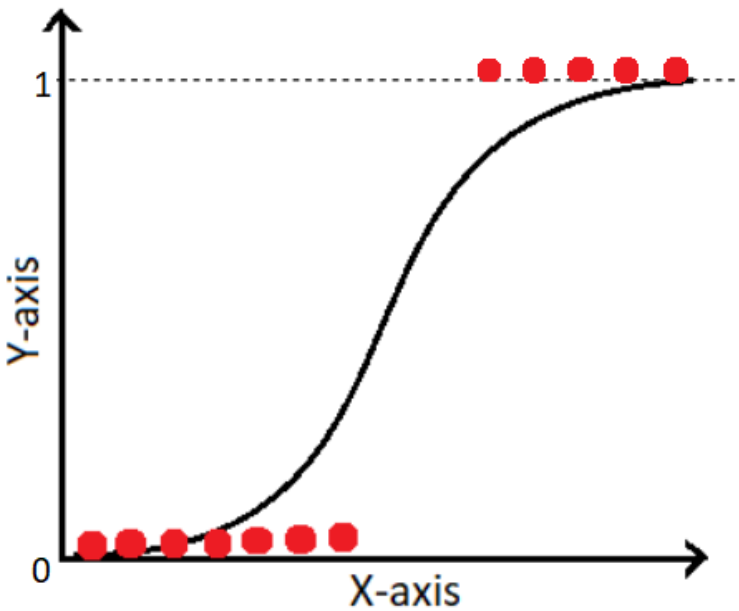
El coeficiente de determinación es la proporción de la varianza total de la variable explicada por la regresión. El coeficiente de determinación, también llamado R cuadrado, refleja la bondad del ajuste de un modelo a la variable que pretender explicar.

$$R^2 = 1 - \frac{\sum(\hat{Y}_t - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

Regresión logística

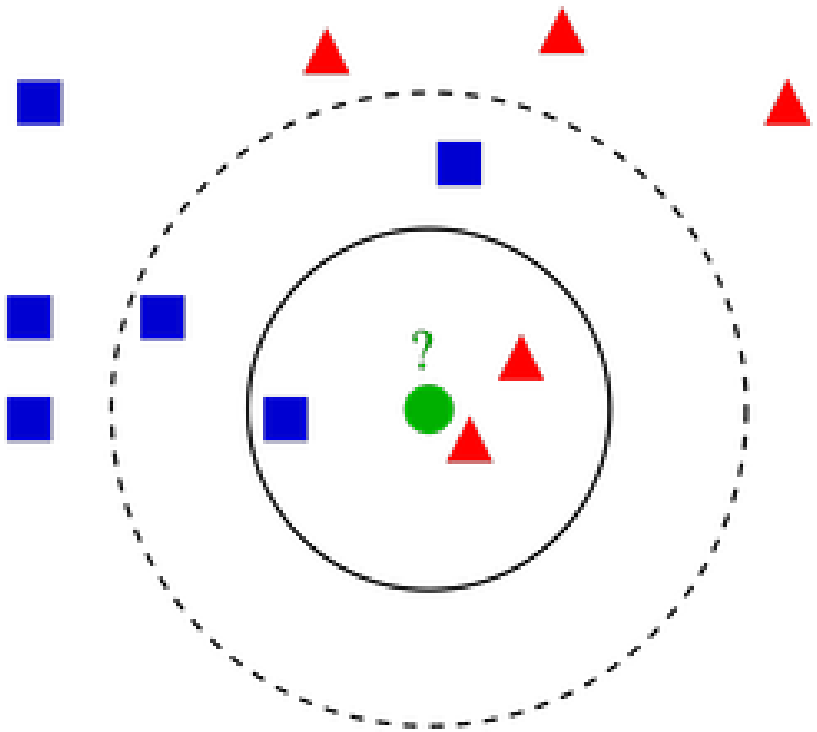
La idea de la regresión logística es clasificar los datos de acuerdo a una función logística. Veamos un ejemplo en el Ejercicio 2 del Repositorio.

$$P(t) = \frac{1}{1 + e^t}$$



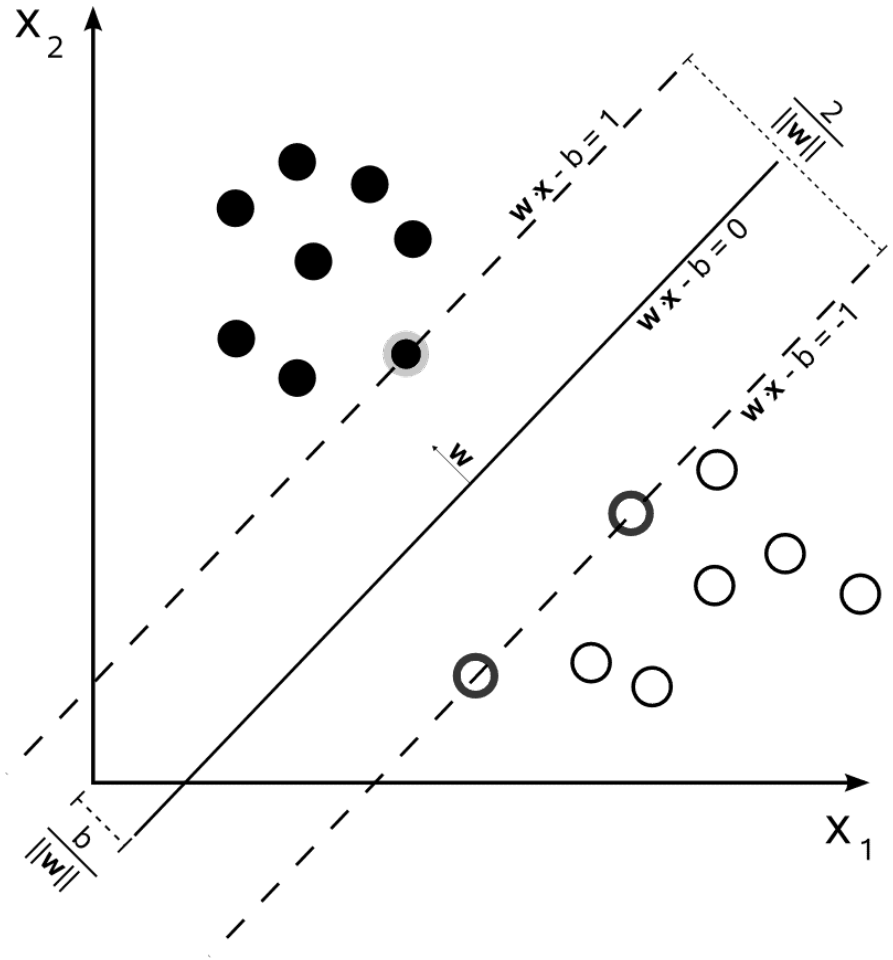
KNN

El modelo de KNN clasifica un determinado espacio de acuerdo a la aparición de K vecinos.



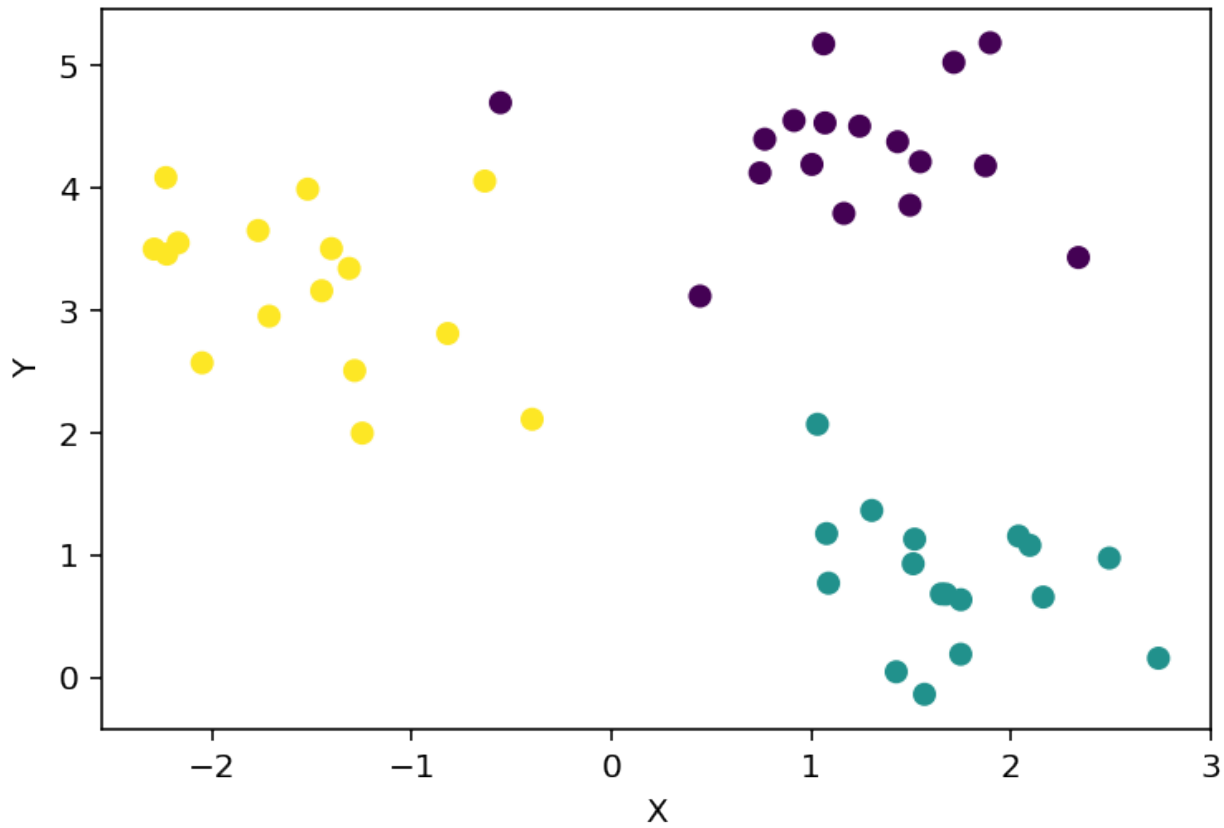
SVM

El modelo de Support Vector Machines calcula el hiperplano que mejor separe a las clases.



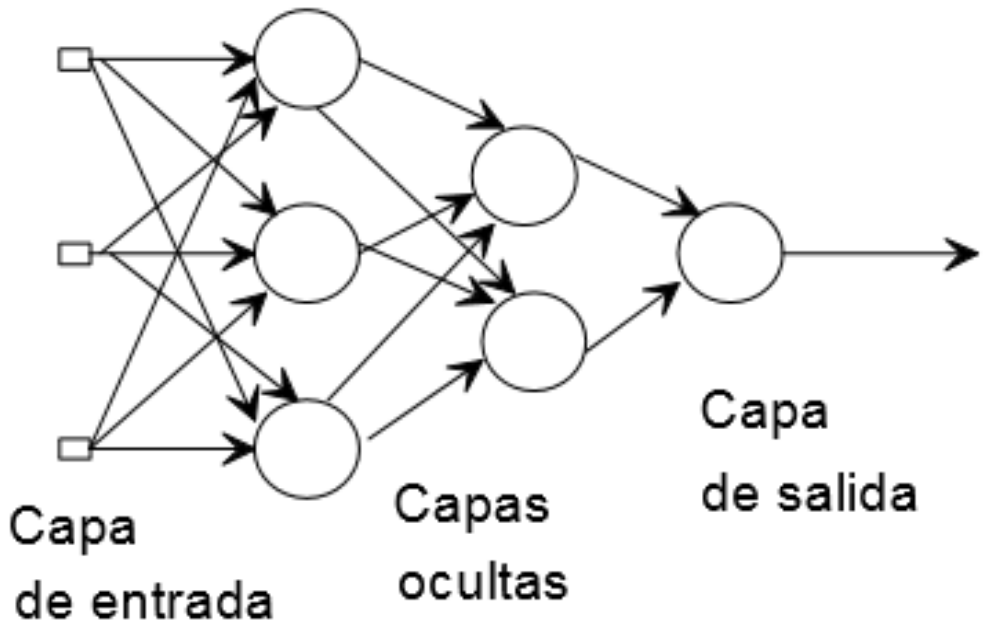
Árboles de decisión

Los árboles de decisión buscan identificar mediante preguntas binarias la mejor separación del espacio.



Redes Neuronales

Las redes neuronales identifican por back propagation la mejor estructura de pesos y relación entre pesos de cada una de las entradas para clasificar o ajustar los datos a la variable objetivo.



Práctica

Veamos código de estos modelos para clasificación usando la librería scikit-learn en el ejercicio 3 del repositorio.

Práctica

Por último podemos ver algunos ejemplos para el caso de la regresión en el Ejercicio 4 del repositorio.