Machine Learning

Aprendizaje no supervisado

Data Science Bootcamp

The Bridge

Definición

Definición

Aprendizaje no supervisado

“Aprendizaje no supervisado es un método de Aprendizaje Automático donde un modelo se ajusta a las observaciones. Se distingue del Aprendizaje supervisado por el hecho de que no hay un conocimiento a priori, los datos no están etiquetados”

Wikipedia

Ejemplo práctico

Imagina que tienes un conjunto de imágenes de una serie de piezas de una línea de producción y quieres montar un modelo de ML que te prediga si la pieza está defectuosa.

Montaremos entonces un modelo automático que las clasifique

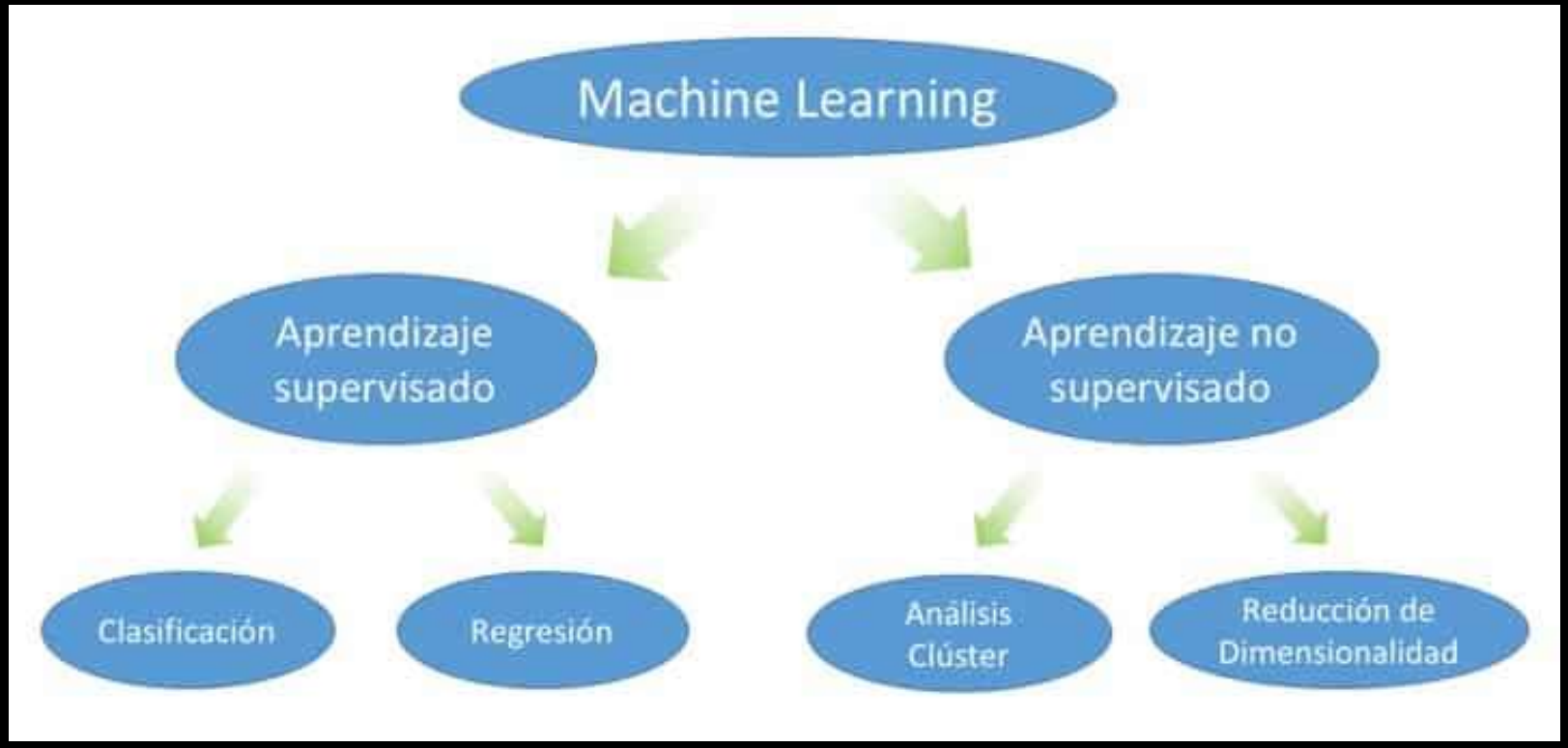
Recogemos el banco de imágenes de la BD

No tenemos los datos etiquetados!! Imposible montar un clasificador binario como los que ya conocemos.

Podemos juntar un equipo de expertos para que las vaya etiquetando manualmente. El problema es que este proceso es costoso en tiempo y dinero y cada vez que haya cambios en los productos, tendremos que volver a repetirlo

Para solucionar esto tenemos los algoritmos de aprendizaje no supervisado

Supervisado vs no supervisado

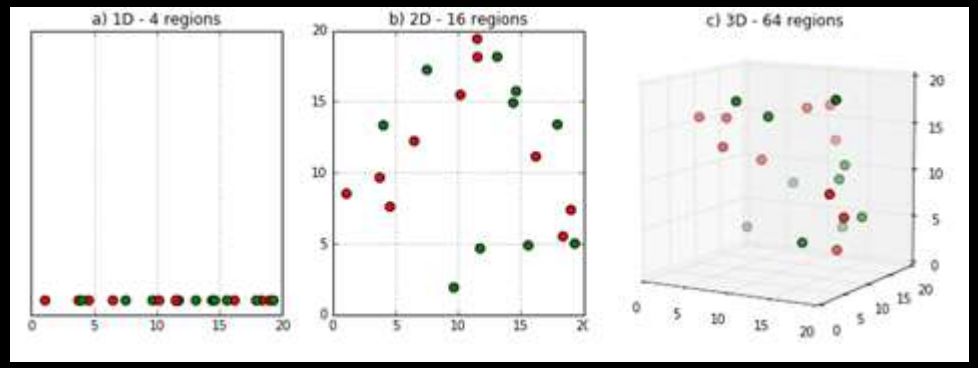


Feature Reduction

Curse of Dimensionality

El número de muestras que se necesitan para estimar una función arbitraria (un target de ML, por ejemplo) con un cierto nivel de precisión crece exponencialmente con el número de inputs/dimensiones/variables de la función.

Este fenómeno afecta mucho a la dispersión y la cercanía de los datos. Según vamos añadiendo dimensiones, se van diferenciando mejor.

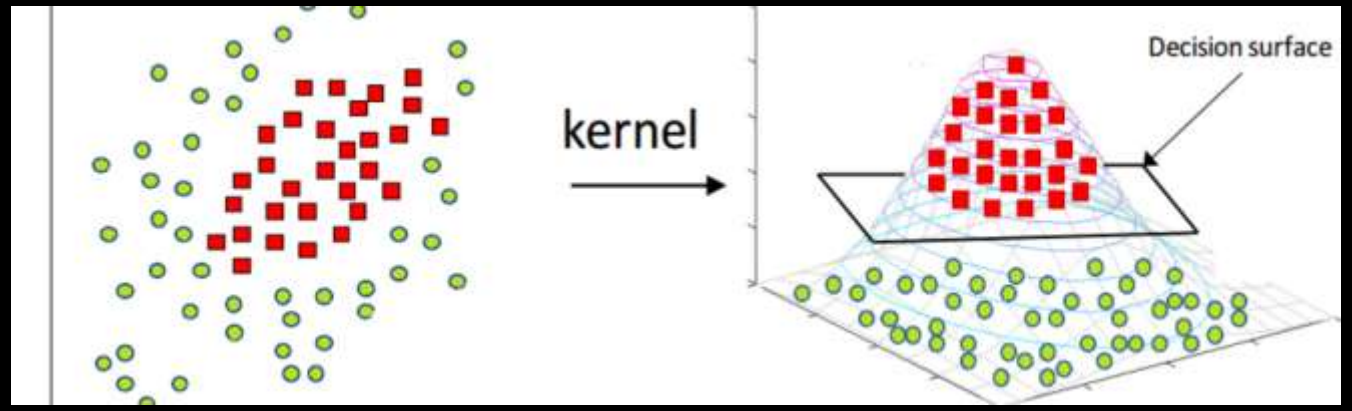


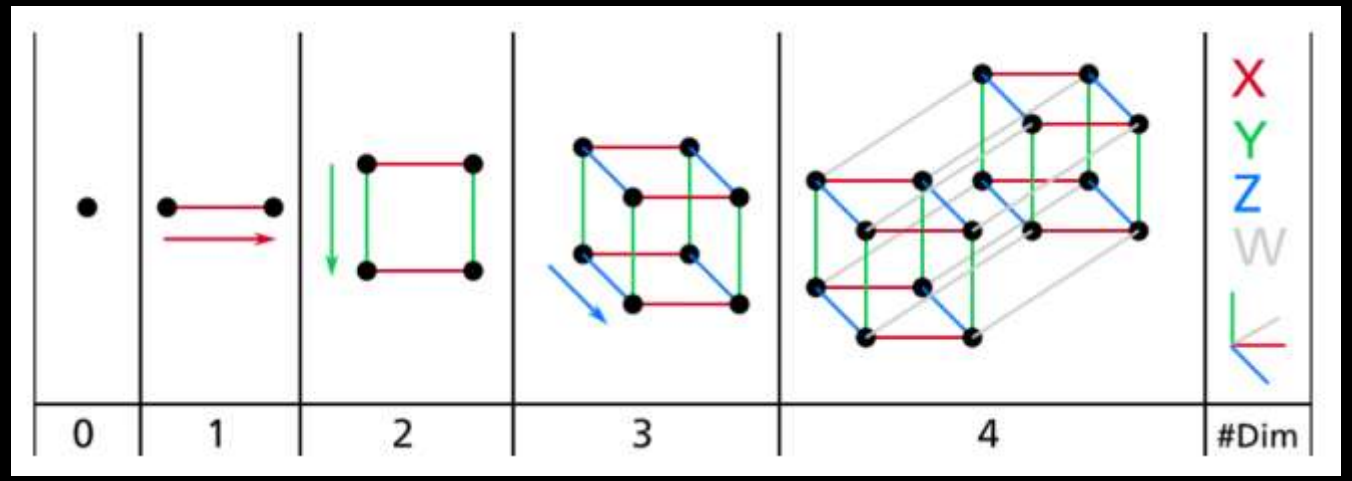
Cuando estamos ante pocas dimensiones, tenemos datos que pueden resultar muy parecidos, pero según vamos añadiendo características y dimensiones nuevas, esto cambia

Datasets con muchas dimensiones serán muy dispersos y con mucha distancia entre los puntos, lo cual es bueno para clasificar. El problema es que nuevas observaciones estarán también muy lejanas de las originales (**overfitting**), produciendo predicciones menos fiables que datasets con pocas dimensiones. **La solución sería incrementar el conjunto de train o reducir la dimensionalidad**

Aumentar vs reducir la dimensionalidad

SVM vs Feat. Reduction



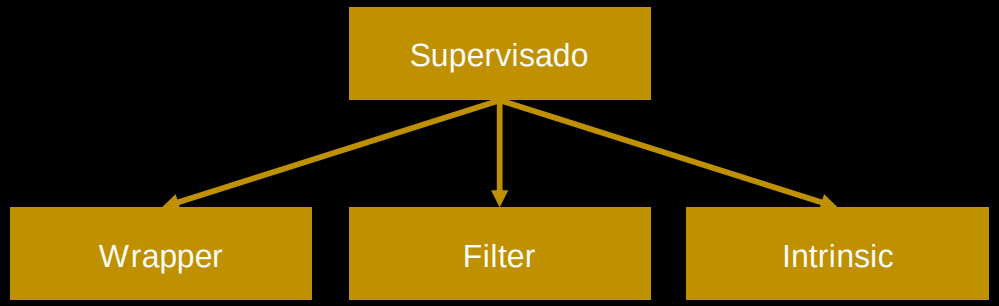


Feature Reduction

Algoritmos y aplicaciones

Aplicaciones

1. Mejora computacional
2. Detección de features discriminantes
3. Eliminación de features irrelevantes
4. Compresión de la información (imágenes)
5. Visualización (PCA)



No Supervisado

sklearn.feature_selection.VarianceThreshold

Reducción de dimensionalidad

PCA

Detalle en documento "*FEATURES SELECTION.pdf*"

Clustering

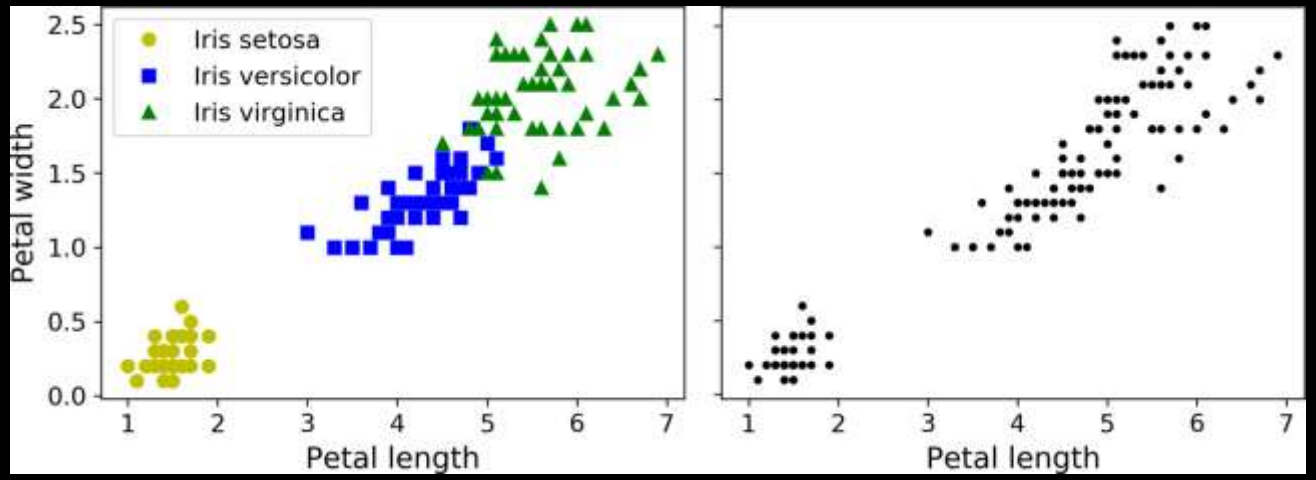
Clustering

Definición

Mediante las técnicas de clustering se pretende agrupar los datos **sin etiquetar** en diferentes grupos **(o clusters)**, de manera automática. Estos algoritmos buscan patrones y similitudes en los datos y los agrupan en k grupos, siendo k un parámetro dado.

Clasificación vs Clustering

En clasificación tenemos los datos etiquetados y por tanto acudimos a los algoritmos supervisados que ya conocemos (Regresión logística, SVC, DecisionTree…) para entrenar el modelo y predecir después las clases. En clustering no tenemos el dato etiquetado, y por tanto es el modelo el que clasifica los datos en k clases. En ningún momento sabe qué es cada clase, pero agrupa los datos según similitudes.



Clustering

Aplicaciones

Segmentación de clientes

Clasificar clientes en grupos para sistemas recomendadores de productos

Data Analysis

Resulta útil aplicar clustering cuando analicemos un dataset para descubrir posibles agrupaciones en los datos

Dimensionality Reduction

Después de aplicar clustering podemos ver la afinidad de cada observación con sus grupos y que esas k medidas sean las features.

Anomaly Detection

Toda observación con poca afinidad respecto a su grupo es susceptible de ser una anomalía o un outlier. Detección de comportamientos inusuales de os usuarios

Algoritmos de búsqueda

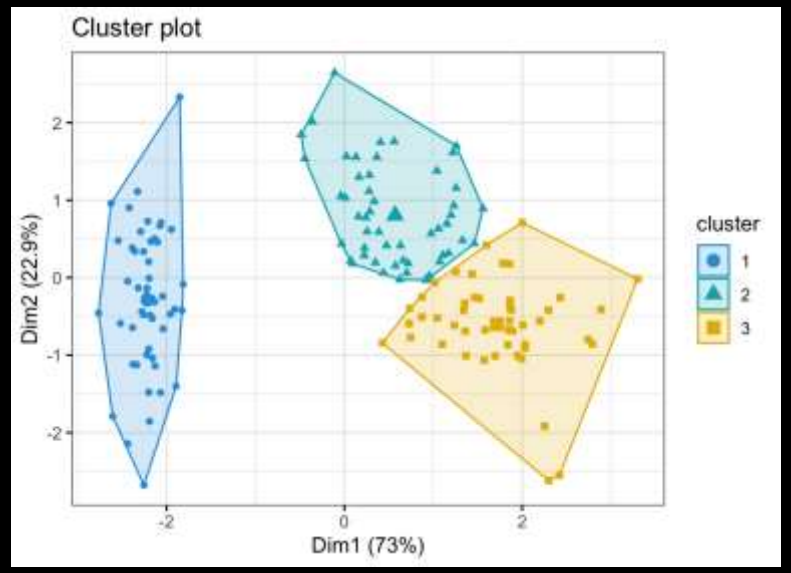
Algoritmos de búsqueda de imágenes similares a una referencia. Como si de un sistema de recomendación se tratase.

Clustering

Algoritmos de clustering

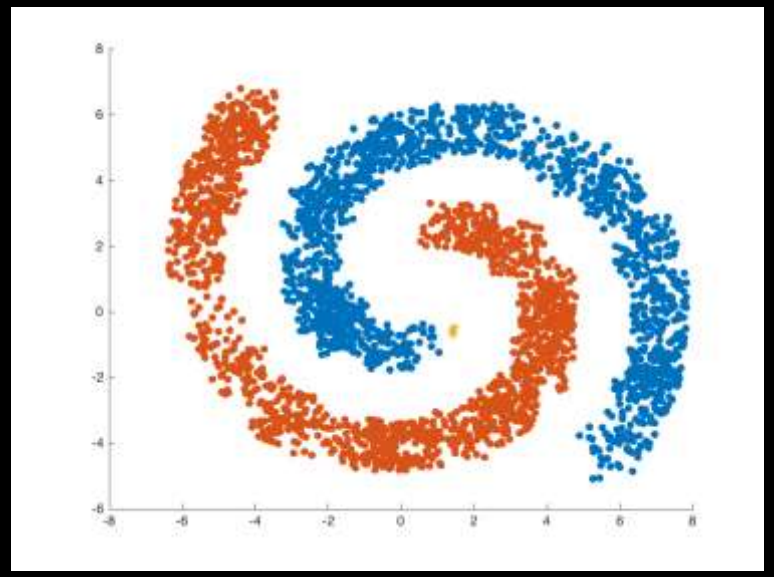
K-Means

Algoritmo de clustering basado en distancias



DBSCAN

Define los clusters como regiones continuas de alta densidad de observaciones.



Bibliografía

https://learning.oreilly.com/library/view/hands-on-machine-learning/9781492032632/ch09.html#unsupervised_learning_chapter