Gestión de Empresas de Base Tecnológica

Machine Learning - Introducción

Máster en Ingeniería Informática
Universidad Complutense de Madrid, 2021-2022

Humberto Martínez Silva

Definición más aceptada de *Machine Learning*

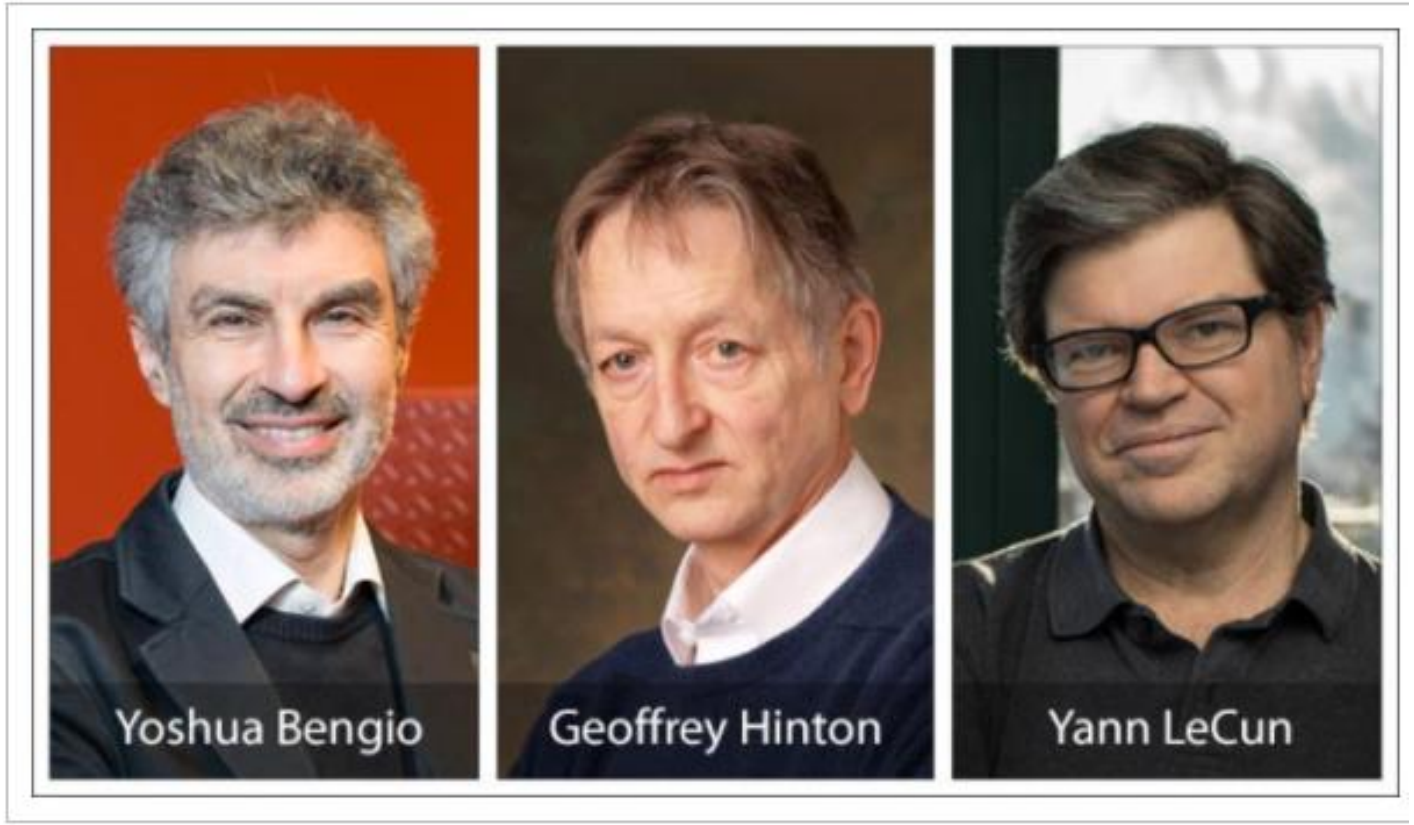
Tom Mitchell (1998): *A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.*

Tarea T: *clasificar email como spam*

*Experiencia E: entrenamos al algoritmo con muchos correos clasificados en **spam** o **no spam***

Performance P: 90% de correos clasificados correctamente, y mejora en base a la experiencia E

Los tres padres del ML



CAPTION

Yoshua Bengio, Geoffrey Hinton and Yann LeCun are the recipients of the 2018 ACM A.M. Turing Award for their contributions to deep neural networks.

https://www.globenewswire.com/news-release/2019/03/27/1773710/0/en/FATHERS-OF-THE-DEEP-LEARNING-REVOLUTION-RECEIVE-ACM-A-M-TURING-AWARD.html

Geoffrey Hinton

Backpropagation: En un artículo de 1986, "Learning Internal Representations by Error Propagation", en coautoría con David Rumelhart y Ronald Williams, Hinton demostró que su algoritmo de *backpropagation* permitía a las redes neuronales descubrir sus propias representaciones internas de datos, lo que hizo posible el uso de redes neuronales para resolver problemas que anteriormente se pensaba que estaban fuera de su alcance.

Boltzmann Machines: En 1983, con Terrence Sejnowski, Hinton inventó las Máquinas Boltzmann, una de las primeras redes neuronales capaces de aprender representaciones internas en neuronas que no formaban parte de la entrada o salida.

Mejoras en las redes neuronales convolucionales: En 2012, con sus estudiantes, Alex Krizhevsky e Ilya Sutskever, Hinton mejoró las redes neuronales convolucionales utilizando RELU y regularización por *droput* (que previene el *overfitting*). En la prominente competencia de ImageNet, Hinton y sus estudiantes redujeron casi a la mitad la tasa de error para el reconocimiento de objetos y remodelaron el campo de visión por computadora.

https://www.globenewswire.com/news-release/2019/03/27/1773710/0/en/FATHERS-OF-THE-DEEP-LEARNING-REVOLUTION-RECEIVE-ACM-A-M-TURING-AWARD.html

Yoshua Bengio

Probabilistic models of sequences: En la década de 1990, consigue la lectura de cheques escritos a mano, que se consideró un hito de la investigación de redes neuronales. Los modernos sistemas de reconocimiento de voz de aprendizaje profundo extienden estos conceptos.

High-dimensional word embeddings and attention: En 2000, Bengio fue autor del artículo histórico, "A Neural Probabilistic Language Model", que introdujo un impacto enorme en la traducción lingüística y en tareas de procesamiento del lenguaje natural.

Generative adversarial networks: Desde 2010, los artículos de Bengio sobre el aprendizaje profundo generativo, en particular las Redes Generativas Adversariales (GAN) desarrolladas con Ian Goodfellow, han generado una revolución en la visión por computadora y los gráficos por computadora.

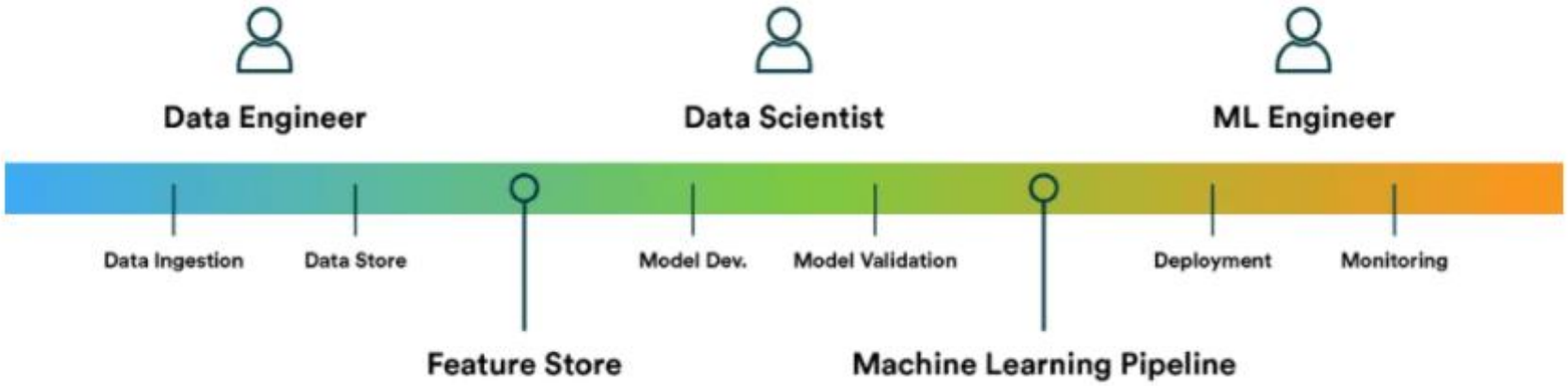
Yann LeCun

https://www.globenewswire.com/news-release/2019/03/27/1773710/0/en/FATHERS-OF-THE-DEEP-LEARNING-REVOLUTION-RECEIVE-ACM-A-M-TURING-AWARD.html

Convolutional neural networks: Gran impacto en Deep Vision.

Mejoras en el algoritmos backpropagation: Impacto en el rendimiento y performance de NN.

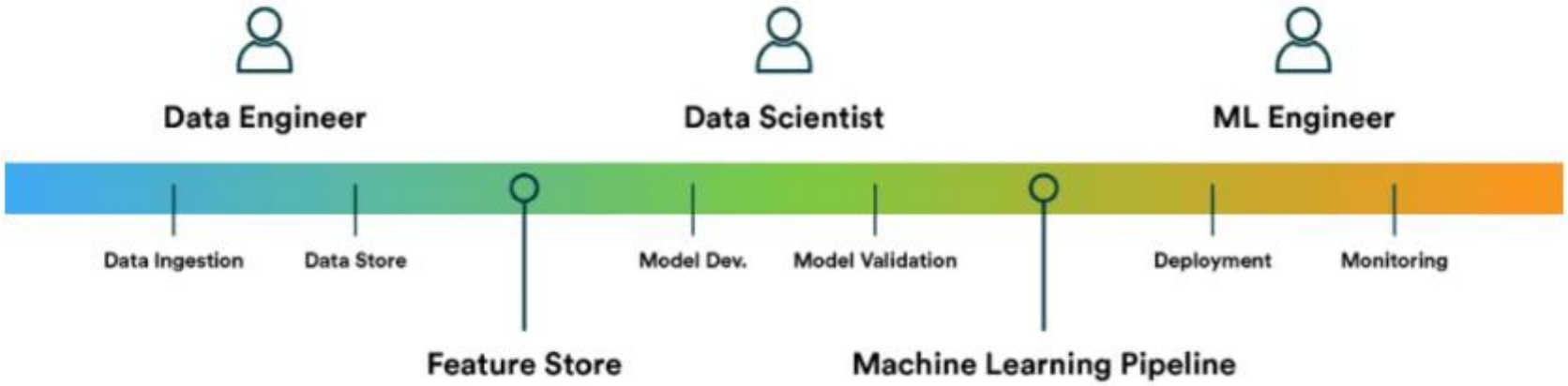
Broadening the vision of neural networks: Crea arquitecturas de NN de gran impacto sobre todo en el mundo de Deep Vision y en la manipulación de datos estructurados y gráficos.



Machine Learning Roles

Data Engineer

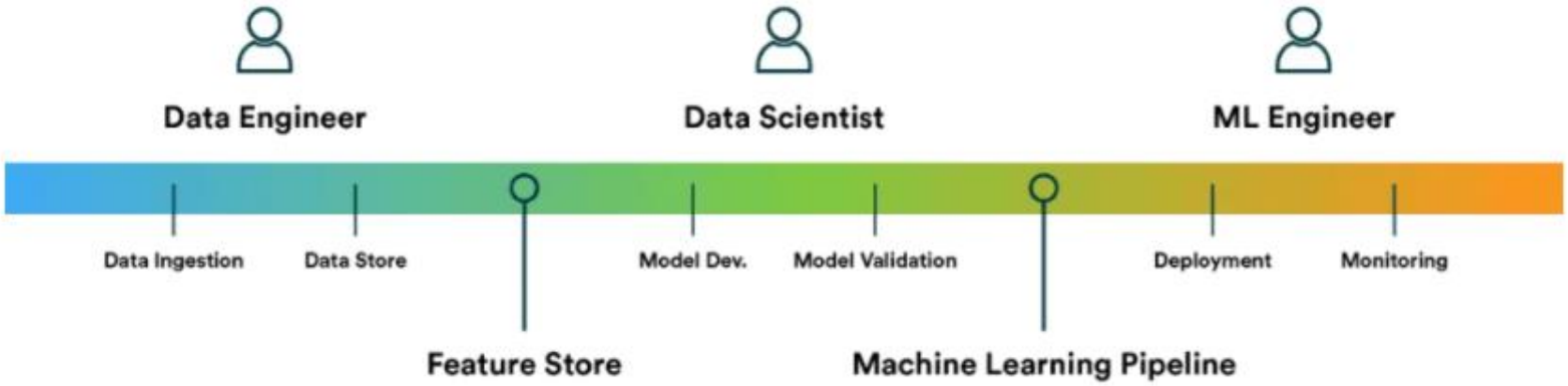
- *Recopilación de datos.*
- *Calidad de los datos.*
- *Limpieza y transformación de datos.*
- *Interpretación de datos.*



Machine Learning Roles

Data Scientist

- *Análisis estadístico.*
- *Creación de modelos de inferencia.*
- *Investigación.*
- *Poner los datos en contexto respecto al Negocio.*



Machine Learning Roles

ML Engineer

- *Intersección entre la ciencia de datos y el desarrollo de SW.*
- *Despliegue.*
- *Optimización.*
- *Eficiencia.*

Supervised vs Unsupervised learning

https://www.r-craft.org/r-news/supervised-vs-unsupervised-learning-explained/

Supervised learning

Se cuenta con un *dataset* perfectamente etiquetado, a partir del cual los algoritmos *aprenden.* Es decir, el algoritmo aprende con ejemplos que le son dados.

<u>Ejemplos:</u>

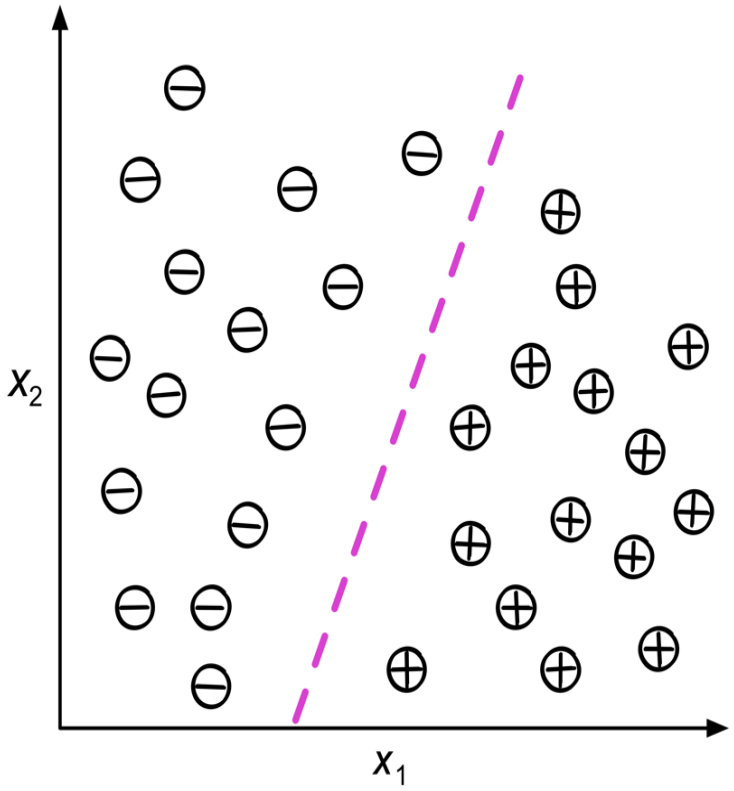
- tenemos una *muestra* de pacientes para deducir cierta enfermedad, clasificados si la padecen o no.

- tenemos una muestra de casas y sus características con sus precios.

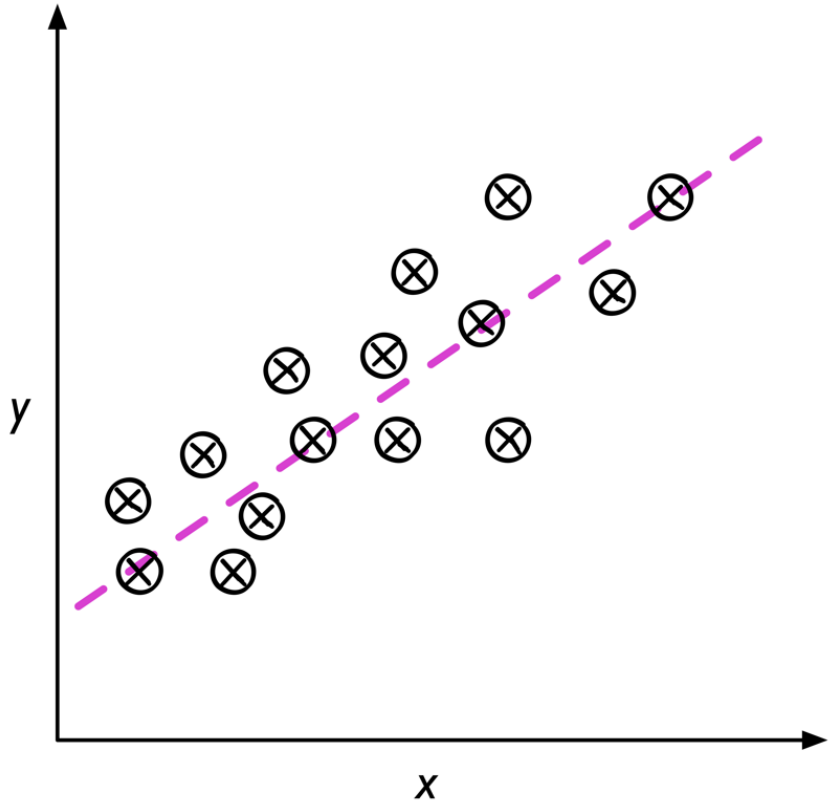
- tenemos una muestra de imágenes para clasificar objetos de distintos tipos.

Supervised learning

Problemas de clasificación



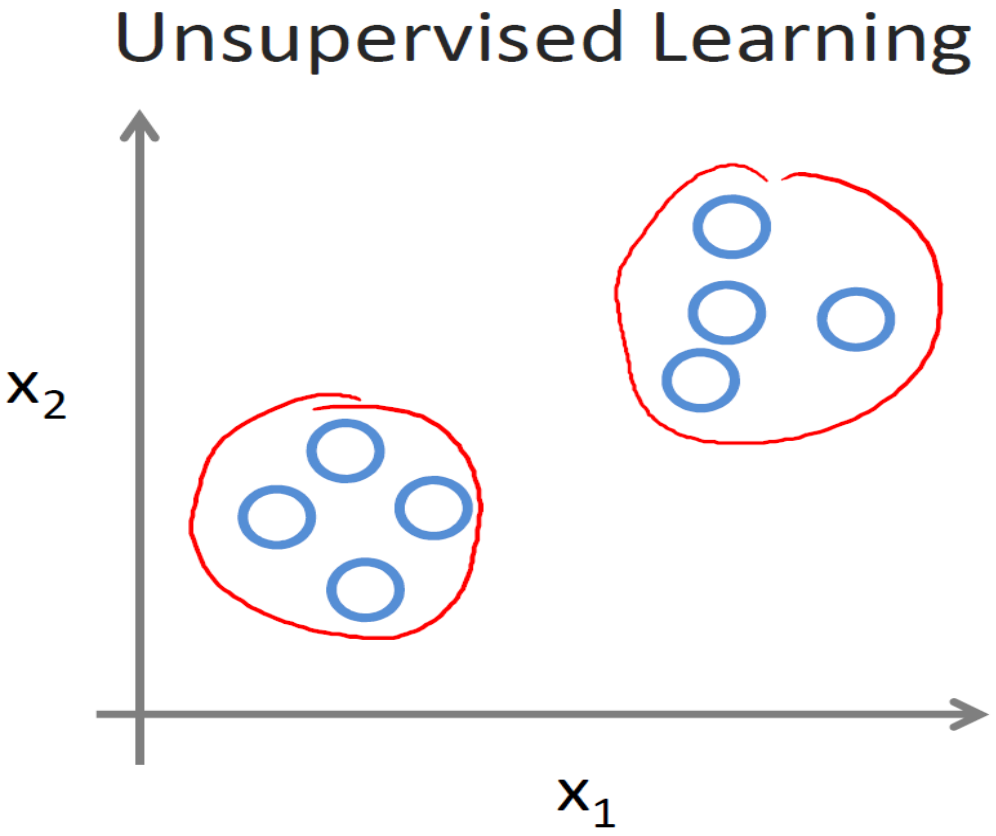
Problemas de regresión

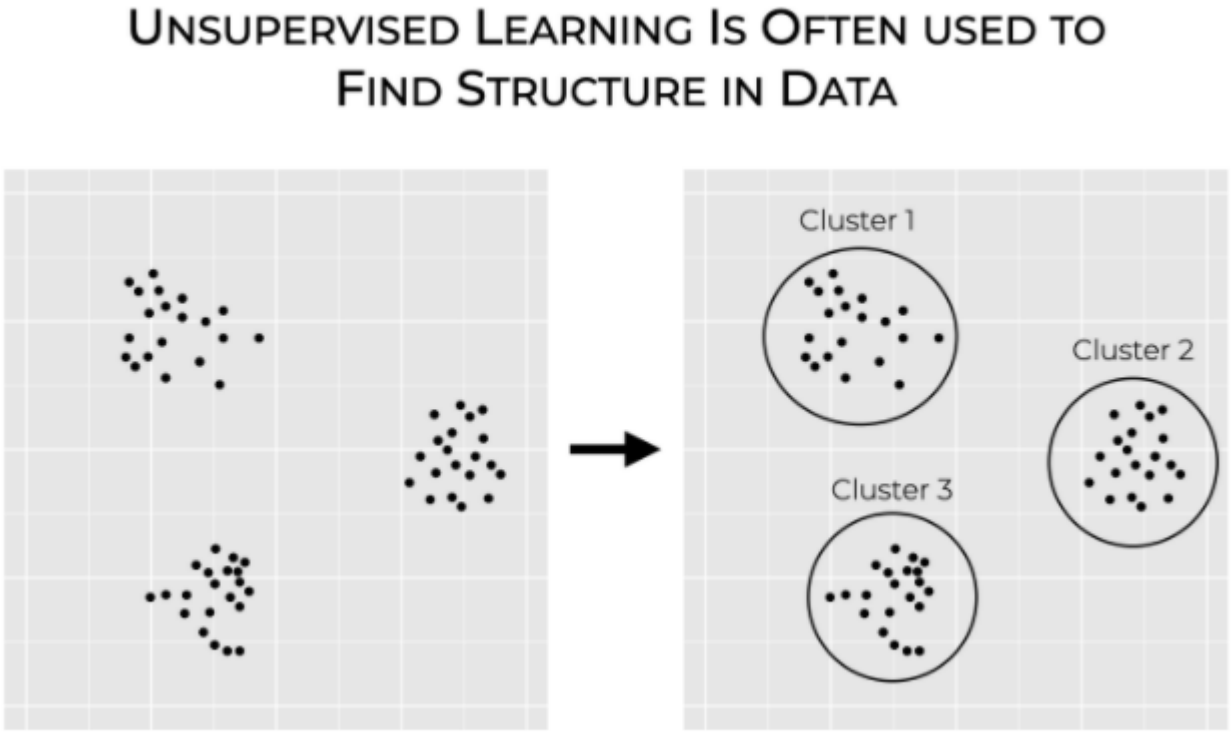


Ref: *Python Machine Learning Book (Sebastian Raschka)*

Unsupervised learning

El algoritmo deduce *grupos o clusters* de items *parecidos*.





Ref: *deeplearning.ai*

https://www.r-craft.org/r-news/supervised-vs-unsupervised-learning-explained/

Unsupervised learning

El algoritmo deduce *grupos o clusters* de items *parecidos*.

Ejemplos:

- detección de anomalías.
- sistemas de recomendación.
- segmentación de clientes.
- genética (ADN, etc).

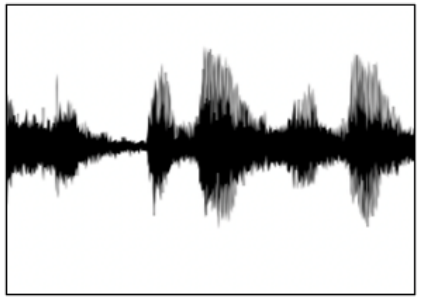
Structured vs Unstructured data

Structured Data

Size	#bedrooms	...	Price (1000$s)
2104	3		400
1600	3		330
2400	3		369
⋮	⋮		⋮
3000	4		540

User Age	Ad Id	...	Click
41	93242		1
80	93287		0
18	87312		1
⋮	⋮		⋮
27	71244		1

Unstructured Data



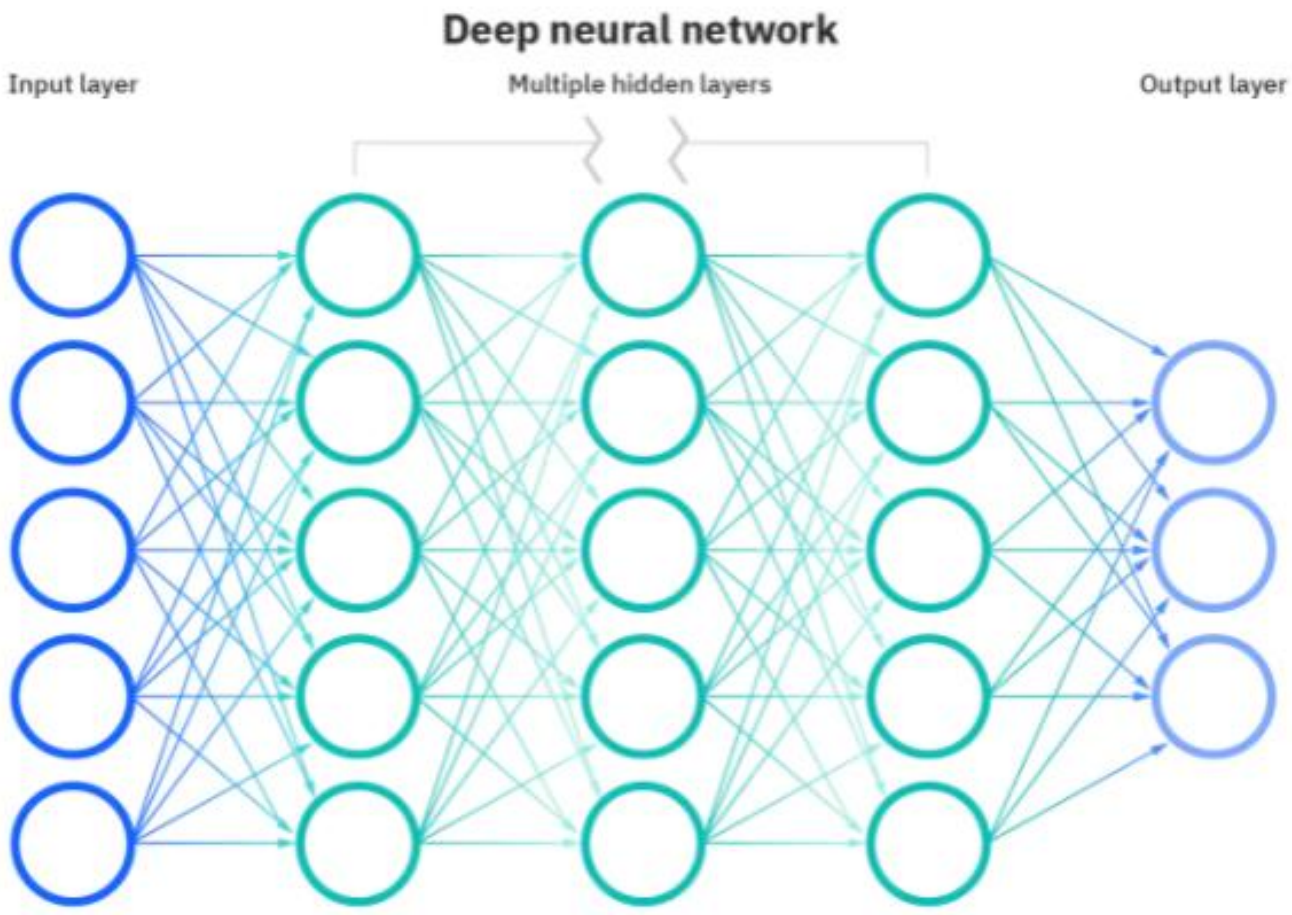
Audio

Image

Four scores and seven years ago...

Text

Ref: *deeplearning.ai*

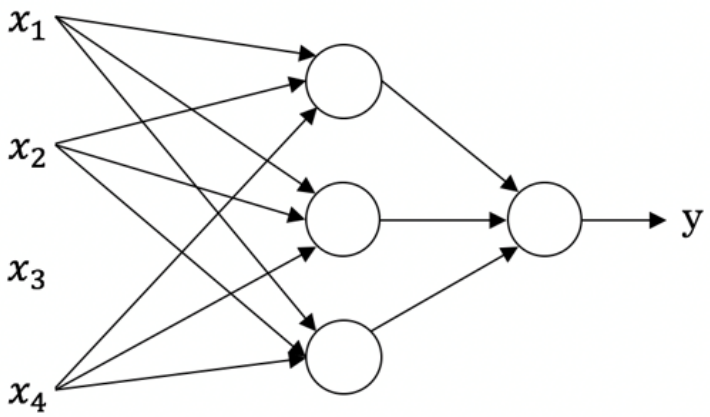


Neural Network

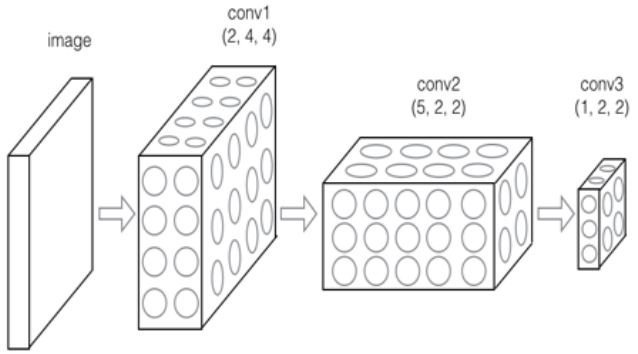
Training ML

Ref: *deeplearning.ai*

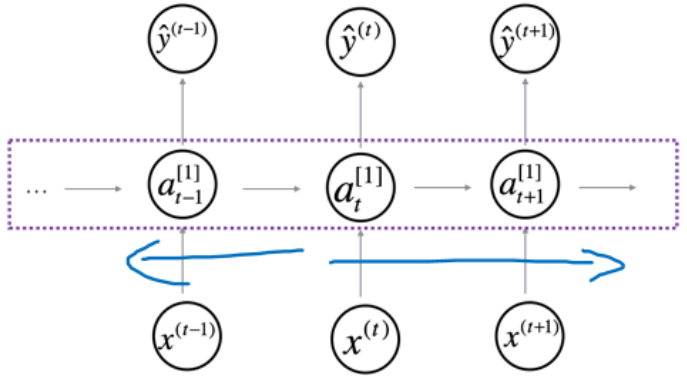
Neural Network examples



Standard NN



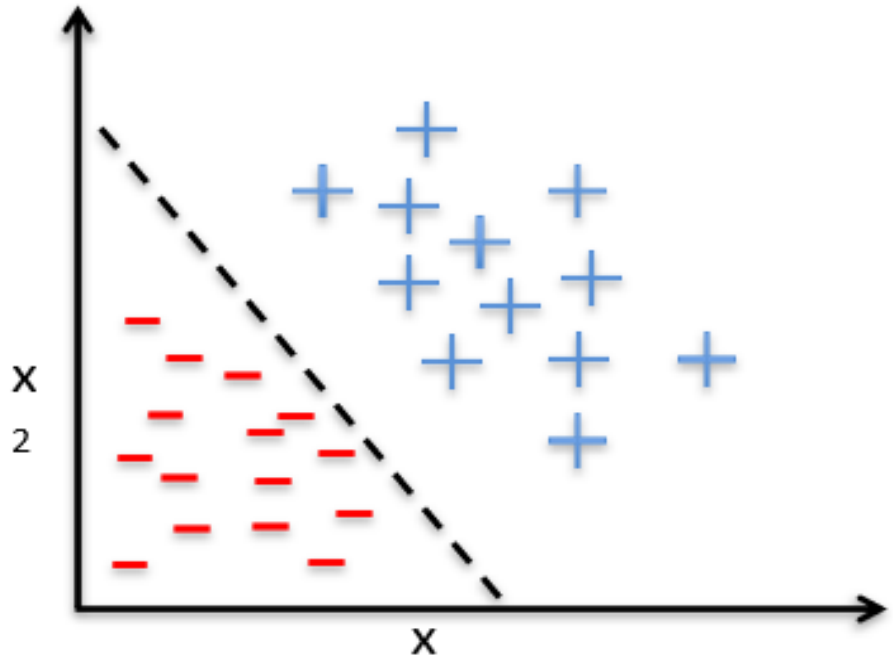
Convolutional NN



Recurrent NN

Training ML

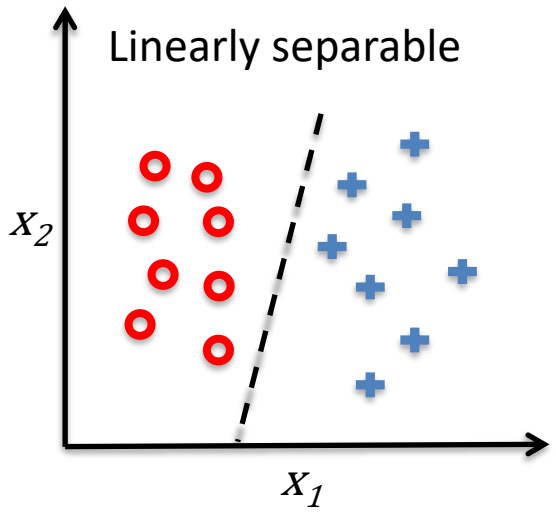
Ref: *Python Machine Learning Book (Sebastian Raschka)*

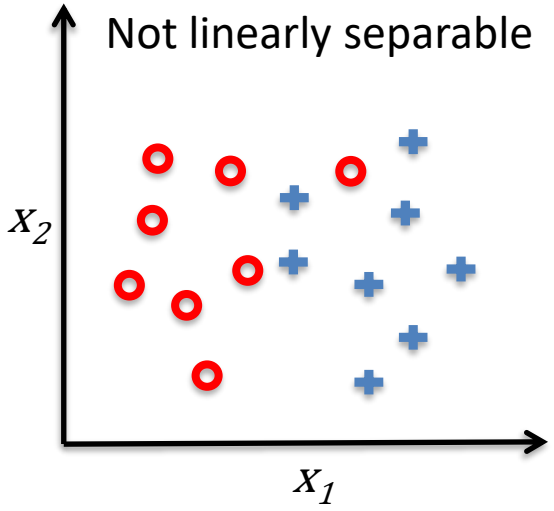


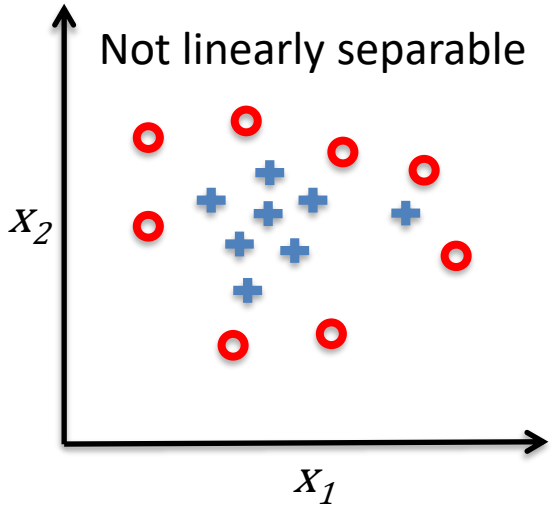
Example of a linear decision boundary for binary classification.

Training ML

Ref: *Python Machine Learning Book (Sebastian Raschka)*

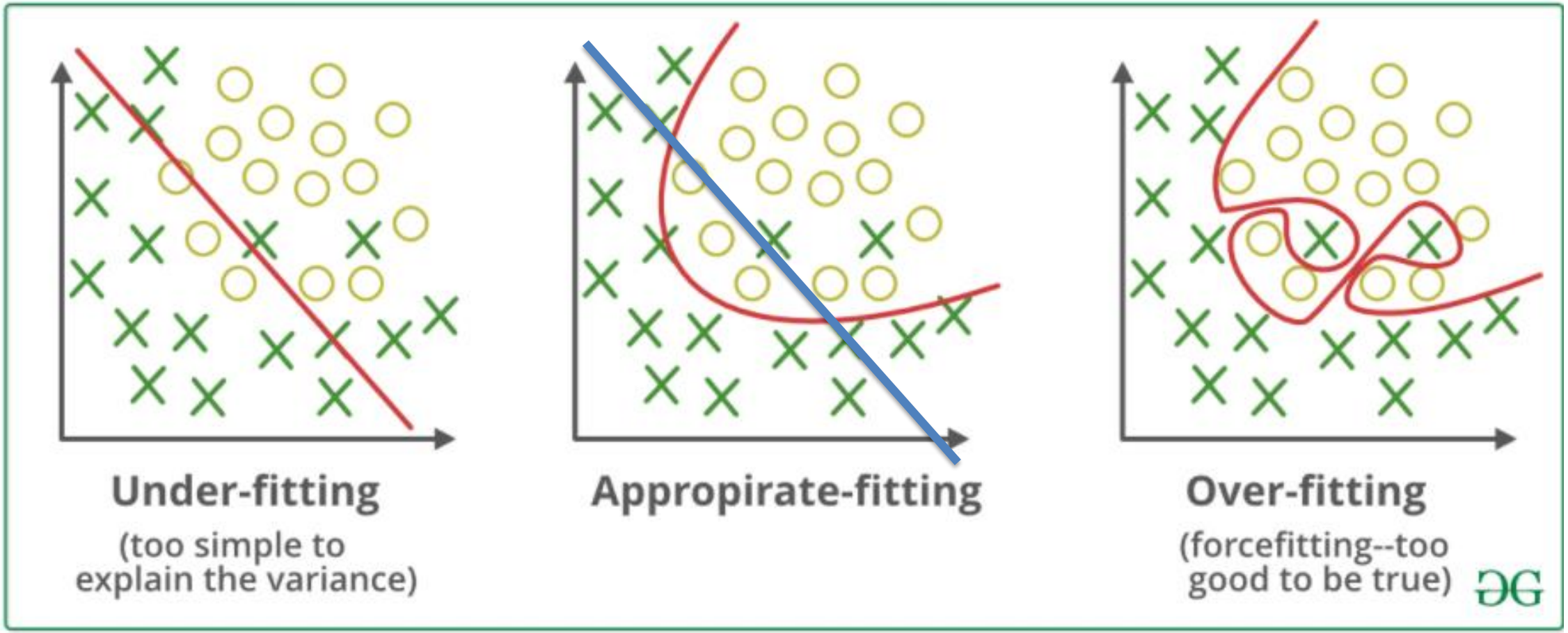






Underfitting / Overfitting

Ref: *https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/*



High bias

High variance

Underfitting

Ref: *https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/*

Técnicas para reducir *Underfitting*

Añadir features.

Aumentar la complejidad del modelo (más capas, más parámetros).

Aumentar la calidad de los datos.

Aumentar la duración del entrenamiento para obtener mejores resultados.

Overfitting

Ref: *https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/*

Técnicas para reducir *Overfitting*

Obtener más datos.

Reducir la complejidad del modelo.

Utilizar *Regularización*.

Utilizar *dropout*.

Z = wX

Logistic Cost function with regularization

$$J(W) = \sum_{i=1}^{m} [-y^{(i)} \log(\phi(z^{(i)})) - (1 - y^{(i)}) \log(1 - \phi(z^{(i)}))] + \frac{\lambda}{2} \| w \|^2$$

$$w_j = w_j + \alpha \sum (y^i - \Phi(z^i)) x_j^i + \frac{\alpha\lambda}{2} w_j$$

Generative Adversarial Networks

Neural style transfer


Content (C) Style (S)

Generated image (G)


Content (C) Style (S)

Generated image (G)

[Images generated by Justin Johnson]

Andrew Ng

Algunos de los datasets más conocidos

- MNIST
- ImageNet
- Open Images Dataset
- The Street View House Numbers
- CIFAR 10/CIFAR 100
- IMDB Reviews
- Million Song Dataset

https://www.analyticsvidhya.com/blog/2018/03/comprehensive-collection-deep-learning-datasets/#h2_5

Algunos de los datasets más conocidos

Image Datasets

MNIST

MNIST es un conjunto de datos de dígitos escritos a mano y contiene un *training set* de 60,000 ejemplos y un *test set* de 10,000 ejemplos.

Size: ~50 MB
Number of Records: 70,000 images in 10 classes

https://www.analyticsvidhya.com/blog/2018/03/comprehensive-collection-deep-learning-datasets/#h2_5

Algunos de los datasets más conocidos

ImageNet

ImageNet es un conjunto de datos de imágenes que se organizan de acuerdo con la jerarquía de WordNet. WordNet contiene aproximadamente 100,000 frases e ImageNet ha proporcionado alrededor de 1000 imágenes en promedio para ilustrar cada frase.

Size: ~150GB
Number of Records: Total number of images: ~1,500,000; each with multiple bounding boxes and respective class labels

https://www.analyticsvidhya.com/blog/2018/03/comprehensive-collection-deep-learning-datasets/#h2_5

Algunos de los datasets más conocidos

Open Images Dataset

Open Images es un conjunto de datos de casi 9 millones de URL para imágenes. Estas imágenes se han anotado con etiquetas a nivel de imagen que delimitan cuadros que abarcan miles de clases.

Size: 500 GB (Compressed)
Number of Records: 9,011,219 images with more than 5k labels

https://www.analyticsvidhya.com/blog/2018/03/comprehensive-collection-deep-learning-datasets/#h2_5

Algunos de los datasets más conocidos

The Street View House Numbers (SVHN)

Este es un conjunto de datos de imágenes del mundo real para desarrollar algoritmos de detección de objetos. Es similar al conjunto de datos MNIST mencionado en esta lista, pero tiene más datos etiquetados (más de 600,000 imágenes). Los datos se han recopilado de los números de las casas vistos en Google Street View.

Size: 2.5 GB
Number of Records: 6,30,420 images in 10 classes

https://www.analyticsvidhya.com/blog/2018/03/comprehensive-collection-deep-learning-datasets/#h2_5

Algunos de los datasets más conocidos

CIFAR-10

This dataset is another one for image classification. It consists of 60,000 images of 10 classes (each class is represented as a row in the above image). In total, there are 50,000 training images and 10,000 test images. The dataset is divided into 6 parts – 5 training batches and 1 test batch. Each batch has 10,000 images.

Size: 170 MB
Number of Records: 60,000 images in 10 classes

https://www.analyticsvidhya.com/blog/2018/03/comprehensive-collection-deep-learning-datasets/#h2_5

Algunos de los datasets más conocidos

Natural Language Processing

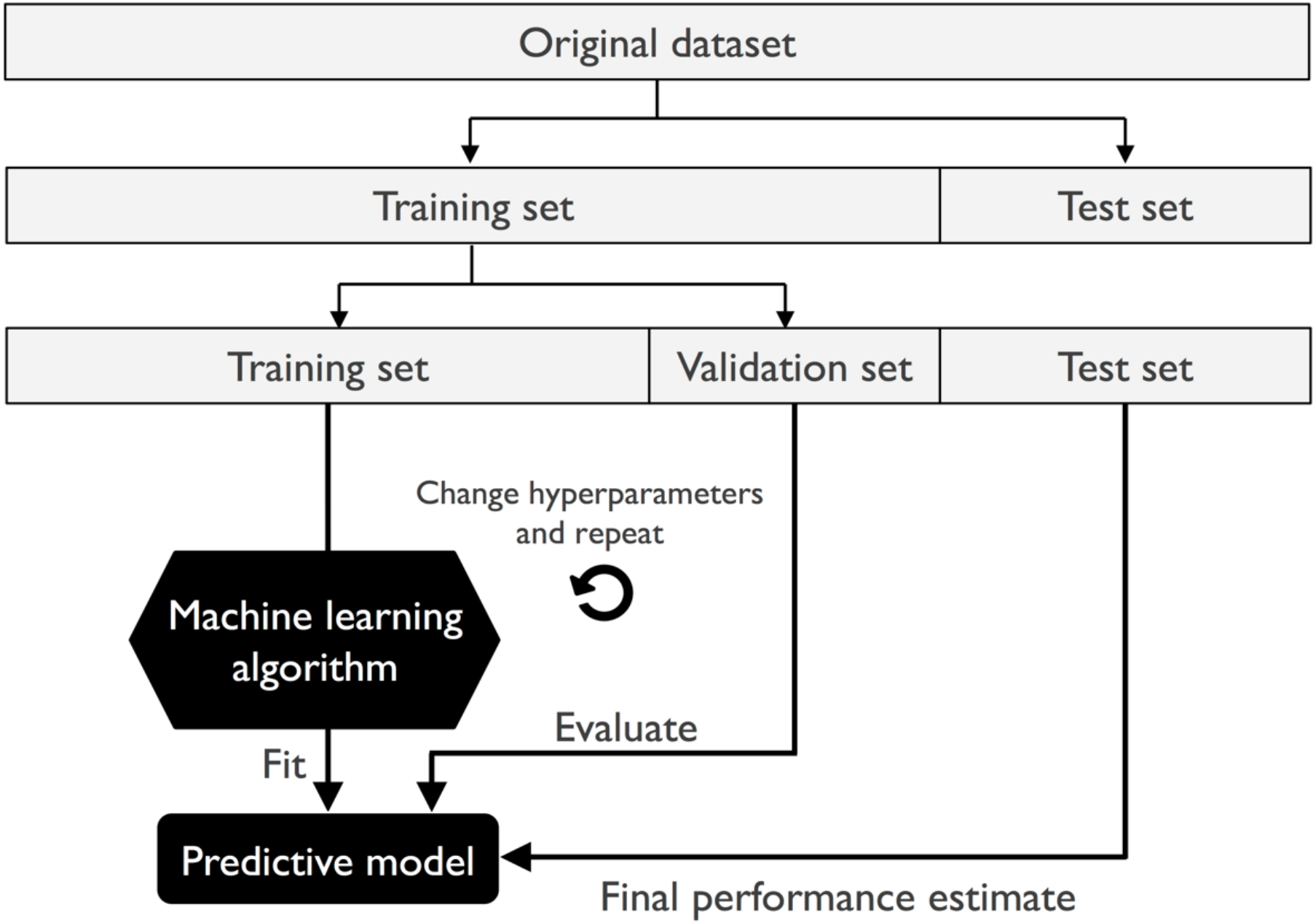
IMDB Reviews

Está destinado a la clasificación de sentimiento binario y tiene muchos más datos que cualquier conjunto de datos anterior en este campo.

Size: 80 MB
Number of Records: 25,000 highly polar movie reviews for training, and 25,000 for testing

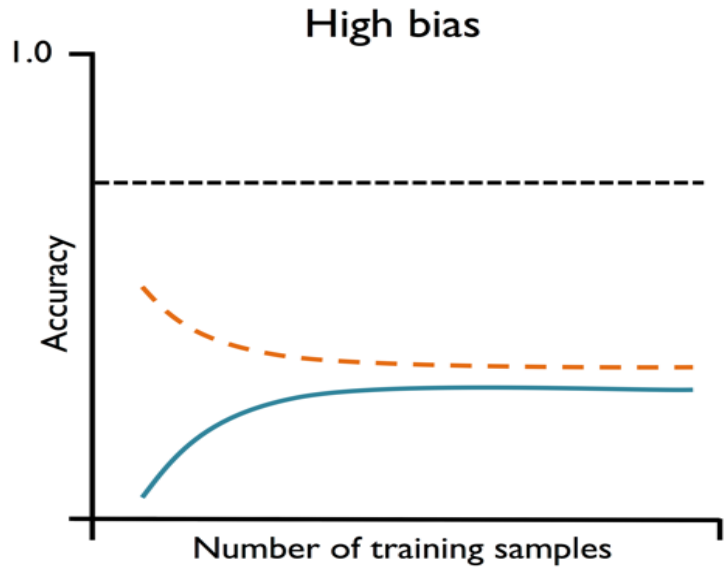
https://www.analyticsvidhya.com/blog/2018/03/comprehensive-collection-deep-learning-datasets/#h2_5

Ref: *Python ML Book (Sebastian Raschka)*

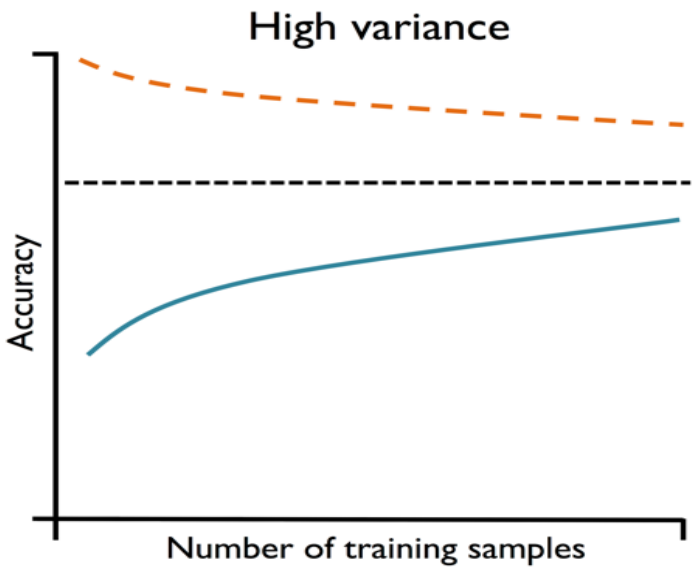


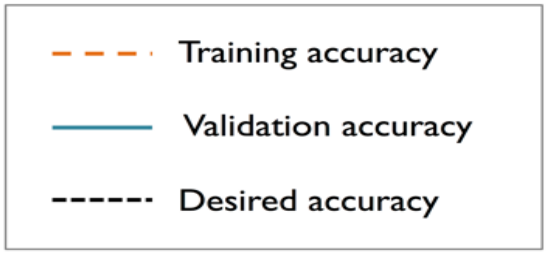
Learning curves

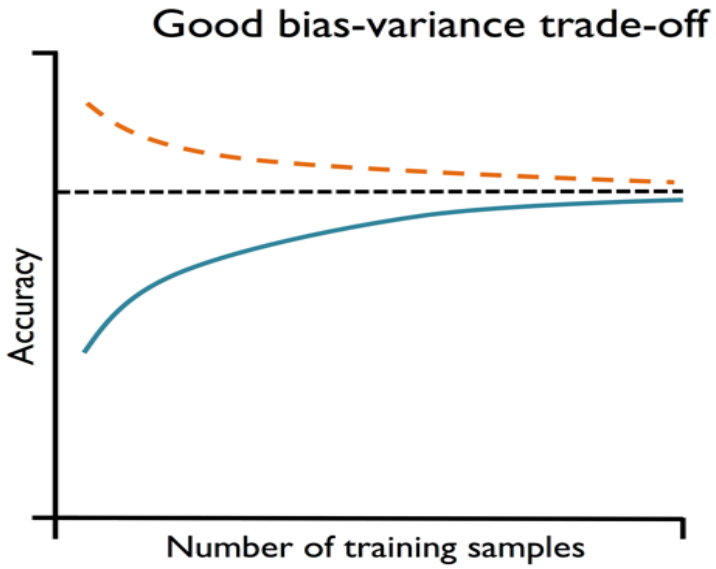
Underfitting



Overfitting







Ref: *Python ML Book (Sebastian Raschka)*

Ref: *Python ML Book (Sebastian Raschka)*