Machine Learning en Spark con MLlib

Rafael Caballero Roldán

+ Spark ML lib

Machine Learning Basics

Una breve introducción a Machine Learning

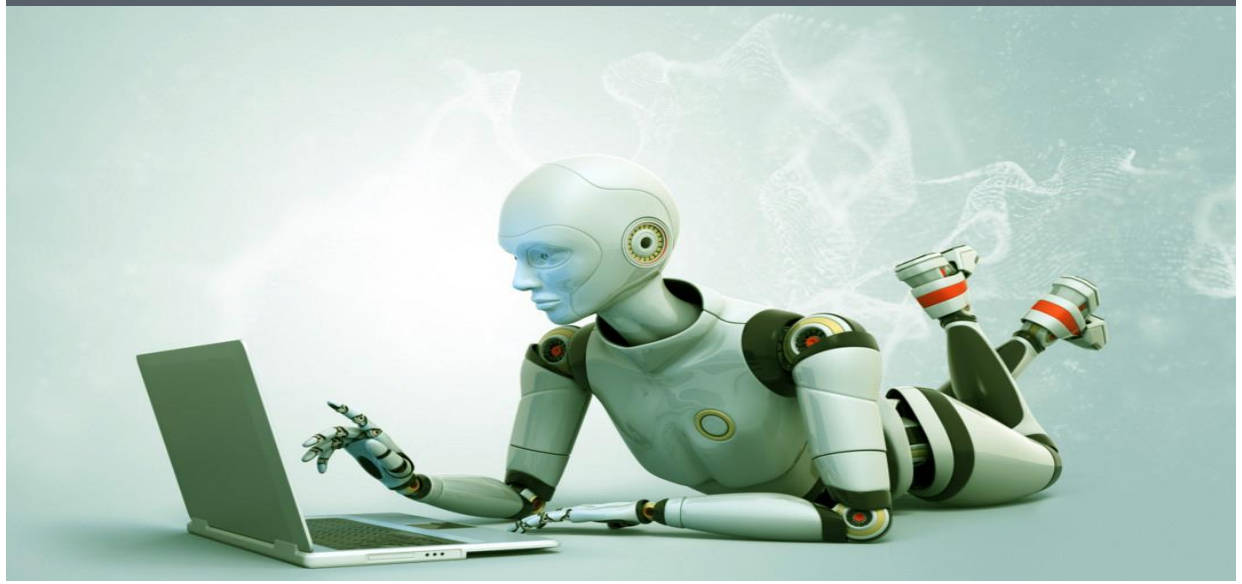
¿Qué es Machine Learning?

Aprendizaje supervisado

Aprendizaje no supervisado

Problemas comunes

Introducción a Spark



+ Spark ML lib

Machine Learning Basics

Una breve introducción a Machine Learning

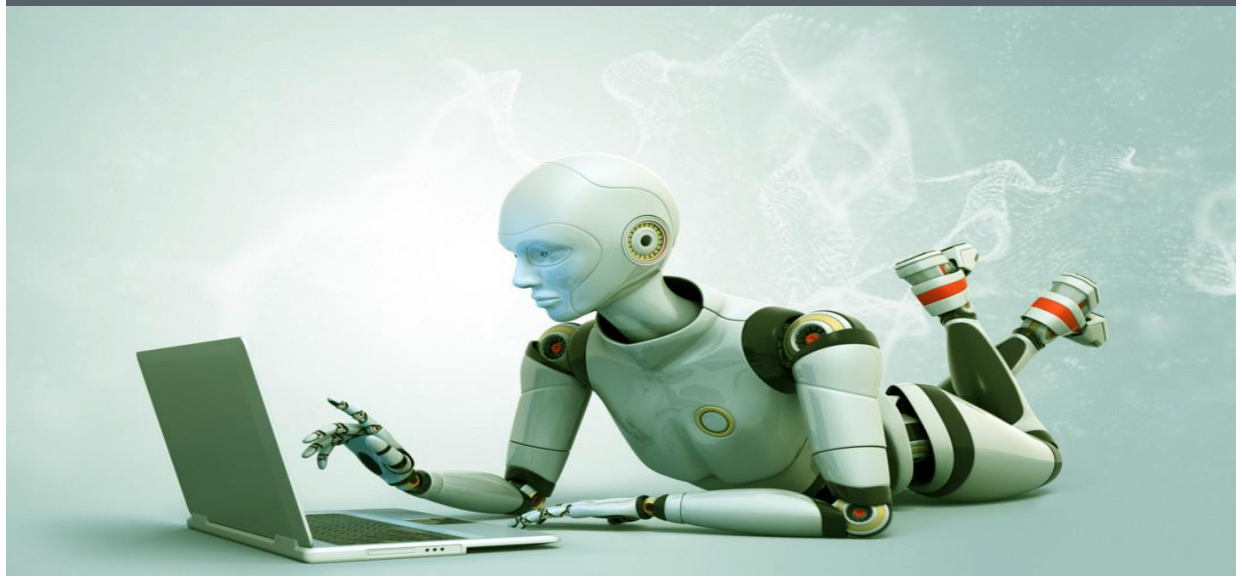
¿Qué es Machine Learning?

Aprendizaje supervisado

Aprendizaje no supervisado

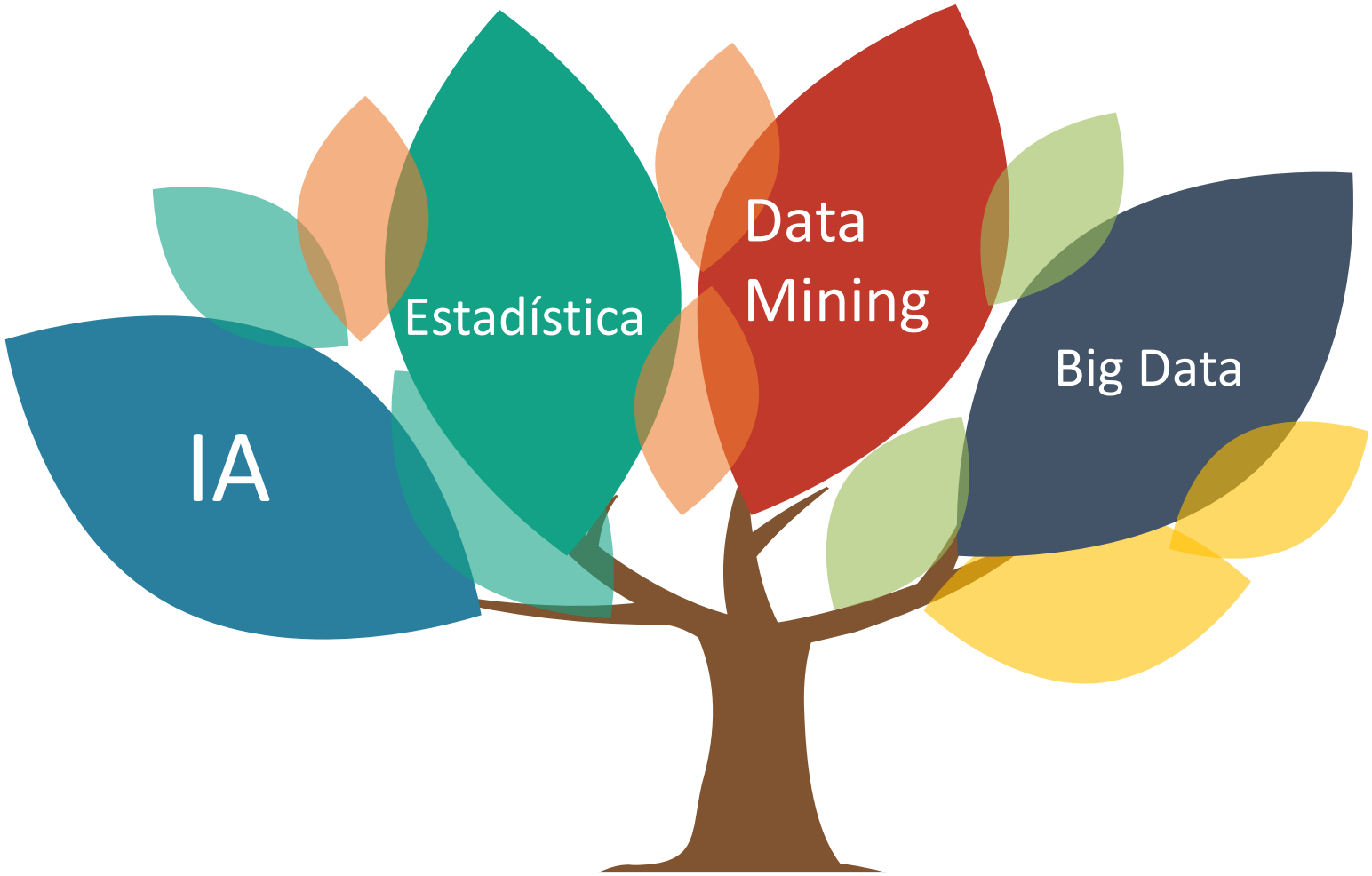
Problemas comunes

Introducción a Spark



Machine Learning

Amigos y parientes



Machine Learning

Un poco de historia

El origen

80's: AI se dedicaba a los sistemas expertos:

- Sin Estadística
- Poca atención a las redes neuronales
- Grupo de investigadores aislados continuaban en ANN: BackPropagation (86)

90's: se empieza a hablar de ML → resolver problemas prácticos. Uso de métodos estadísticos

Gente distinta, definiciones distintas

Varias definiciones, a ver si entre todas…

1959

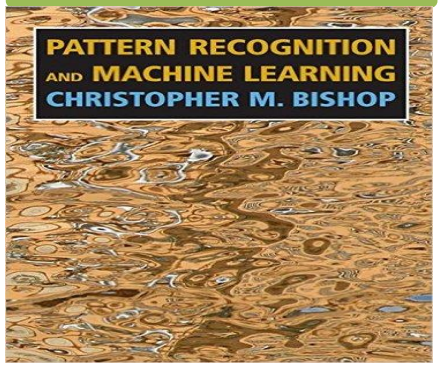
Arthur L. Samuel

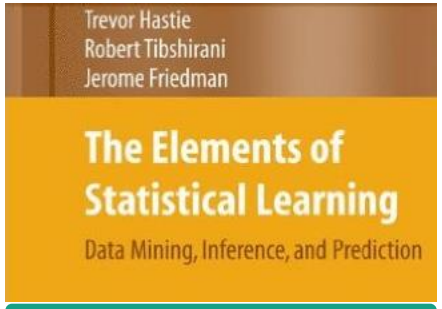
Campo de estudio que proporciona a los ordenadores la posibilidad de aprender sin ser explícitamente programados

Chistopher M. Bishop

El origen del reconocimiento de patrones está en la ingeniería, mientras que el de ML en la informática. Sin embargo, pueden verse como dos facetas de la misma tarea

2006





2008

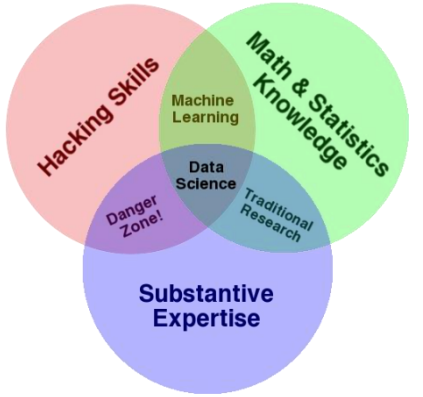
Hastie, Tibshirani, Friedman

Se están generando grandes cantidades de datos, y la tarea del estadístico es descubrir patrones y tendencias, entender lo que "dicen los datos". A eso le llamamos ML

Drew Conway

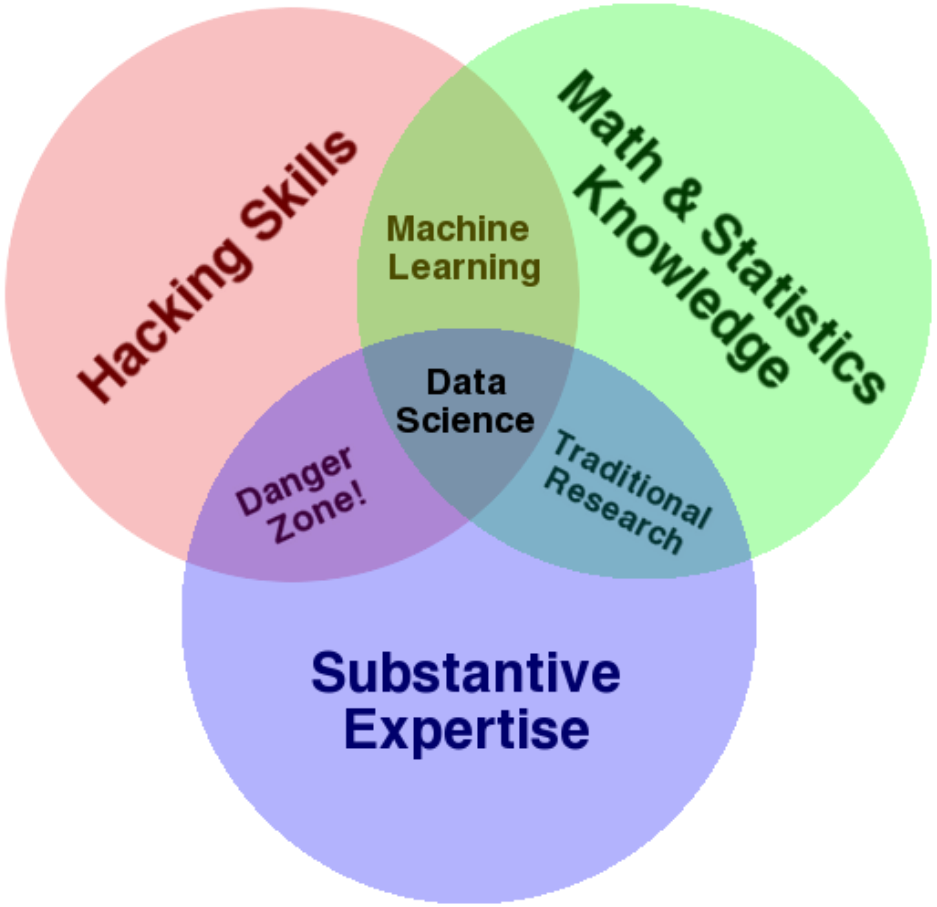
Relaciona la habilidad para codificar, el conocimiento de estadística y el grado de experiencia

2010



El diagrama de Drew Conway

El peligro de quedarse solo con una parte



La definición de Mitchell

La más formal y la más usada

Decimos que:

Un programa aprende de la experiencia E con respecto a:

a) Una clase de tareas T , y

b) Una medida de eficiencia P

Si la eficiencia en las tareas en T, medida según P, aumenta con la experiencia E

Ejemplo: recomendador

Como instancia de la definición general

Def. general

Un programa aprende de la experiencia E con respecto a:

a) Una clase de tareas T , y

b) Una medida de eficiencia P

Si la eficiencia en las tareas en T, medida según P, aumenta con la experiencia E

Instancia

Experiencia E: Conjunto de valoraciones de productos de los clientes

Tareas T : Recomendación de un producto a un usuario

Eficiencia P: Minimizar error de los mínimos cuadrados (diferencia entre la valoración estimada y la que realmente se ha dado) para un conjunto test subconjunto E

Aprendizaje automático

¿Qué tienen de especial los algoritmos basados en estas técnicas?

- ✓ No se programa una solución específica para el problema (¿cómo se escribe un programa que distingue entre perros y gatos?)
- ✓ Se usa un repertorio de técnicas, que conviene conocer
- ✓ La misma técnica sirve para distintos problemas, aprendiendo a base de ejemplos → se desplaza el foco hacia los datos

+ Spark ML lib

Machine Learning Basics

Una breve introducción a Machine Learning

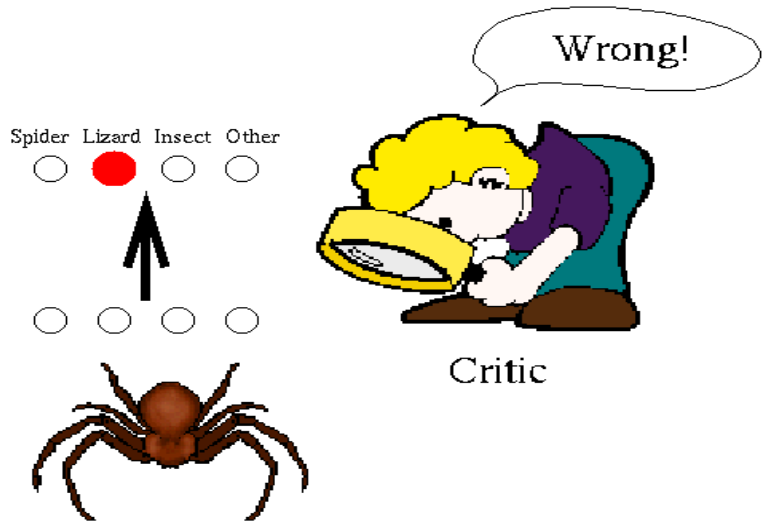
¿Qué es Machine Learning?

Aprendizaje supervisado

Aprendizaje no supervisado

Problemas comunes

Introducción a Spark



Aprendizaje supervisado

El conjunto de entrenamiento incluye resultados esperados para cada dato

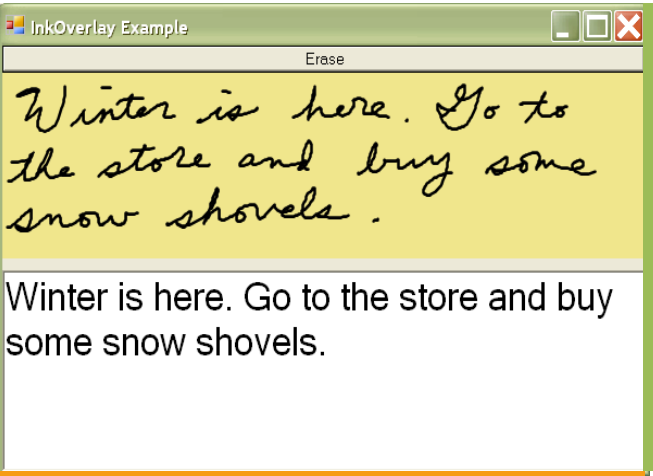
Tarea: Dado un dato, encontrar otro valor asociado
Ejemplo: Dado un dato de colesterol en sangre indicar la probabilidad de infarto

Experiencia E: conjunto de datos de entrenamiento, que incluyen para cada dato un valor asociado útil para calcular el que deseamos
Ejemplo: Personas con edad > 80 años, su nivel medio de colesterol entre los 50 y los 80 y el valor true si han sufrido algún infarto y false en otro caso.

Se dice supervisado porque tenemos una idea de los valores que se deben obtener para el conjunto de entrenamiento → útil para la eficiencia P
Ejemplo: Tras el "aprendizaje" a las personas que han sufrido un infarto les debe corresponder una probabilidad más alta que a las que no

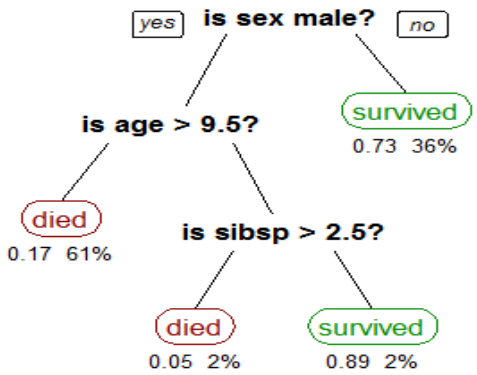
Tipos de aprendizaje supervisado

Solo unos pocos tipos



Regresión

Buscar una función que aproxime la relación de dependencia entre varias variables

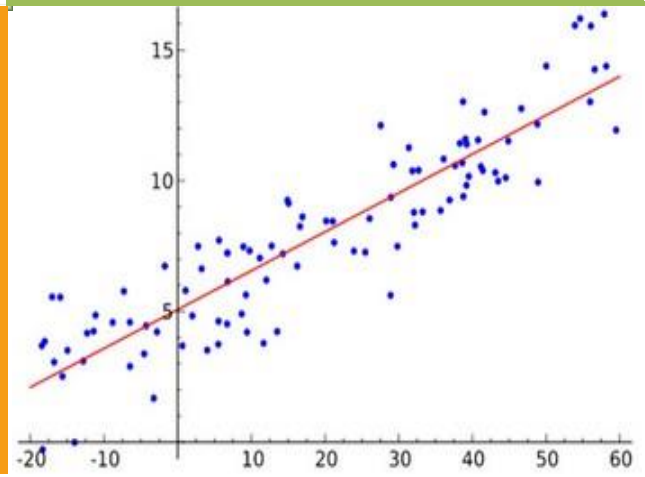


Modelos Bayesianos

Determinan los parámetros de un modelo o determinan qué modelo se ajusta mejor a unos datos

Redes neuronales

Determinar funciones que dependen de una gran cantidad de entradas con retropropagación



Árboles de decisión

Clasificación de elementos dada un serie de test consecutivos

Aprendizaje supervisado: ejemplo

Aprendiendo a conocer las frutas: https://dataaspirant.com/2014/09/19/supervised-and-unsupervised-learning/

E: conjunto de frutas; para cada una observamos su tamaño y color, y además nos dicen su nombre
T: Determinar el nombre de nuevas frutas
P: Número de aciertos con frutas nuevas

No.	SIZE	COLOR	SHAPE	FRUIT NAME
1	Big	Red	Rounded shape with a depression at the top	Apple
2	Small	Red	Heart-shaped to nearly globular	Cherry
3	Big	Green	Long curving cylinder	Banana
4	Small	Green	Round to oval,Bunch shape Cylindrical	Grape

Llega una fruta que es determinada como “Big”, “Green” y “Rounded shape with a depression at the top”
¿Qué hacemos?

Aprendizaje supervisado: ejemplo

Aprendiendo a conocer las frutas: https://dataaspirant.com/2014/09/19/supervised-and-unsupervised-learning/

	Características			Variable de decisión
No.	SIZE	COLOR	SHAPE	FRUIT NAME
1	Big	Red	Rounded shape with a depression at the top	Apple
2	Small	Red	Heart-shaped to nearly globular	Cherry
3	Big	Green	Long curving cylinder	Banana
4	Small	Green	Round to oval,Bunch shape Cylindrical	Grape

+ Spark ML lib

Machine Learning Basics

Una breve introducción a Machine Learning

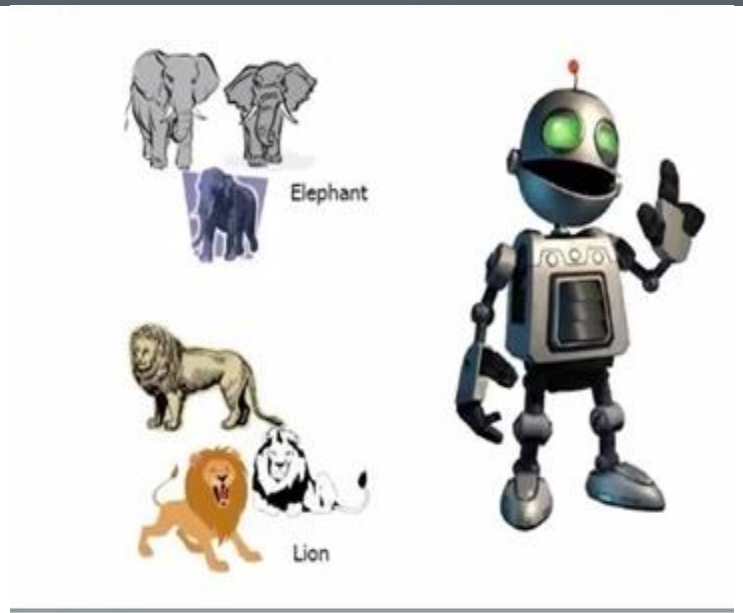
¿Qué es Machine Learning?

Aprendizaje supervisado

Aprendizaje no supervisado

Problemas comunes

Introducción a Spark



Aprendizaje no supervisado

El conjunto de entrenamiento incluye resultados esperados para cada dato

Se trata de elaborar un modelo a partir de unas observaciones (Experiencia E)

No hay conocimiento a priori y se trata de descubrir una estructura interna oculta en los datos

La técnica más común es clustering: agrupación de elementos con características similares

Otras técnicas: compleción de matrices, búsqueda de estructuras de grafos

Aprendizaje no supervisado: ejemplo

Ejemplo de Ciro Donalek

¿Cómo los agruparías?

Aprendizaje no supervisado: ejemplo

Ejemplo de Ciro Donalek





Aprendizaje no supervisado: ejemplo

Ejemplo de Ciro Donalek

Aprendizaje no supervisado: ejemplo

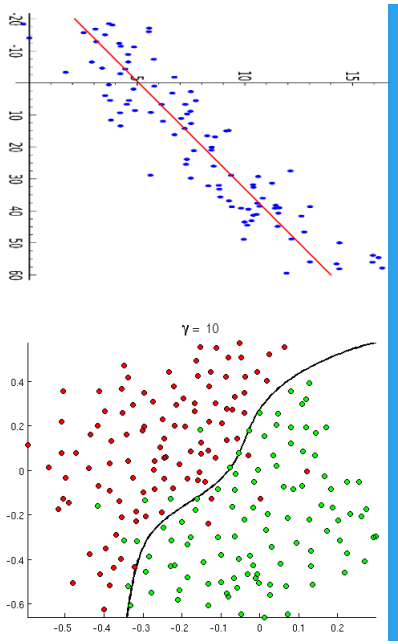
Ejemplo de Ciro Donalek

Aprendizaje no supervisado: ejemplo

Ejemplo de Ciro Donalek

Resumen: aprendizaje supervisado y no supervisado

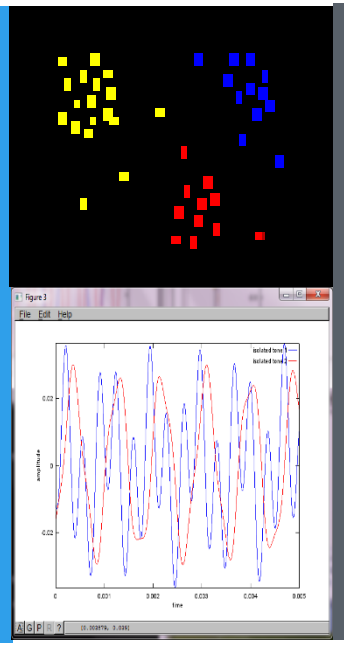
Recordamos las diferencias entre estas dos formas de aprendizaje



Supervisado

Se parte de unos cuantos valores conocidos y se trata de “adivinar” nuevos valores con la mayor precisión posible

- Calcular el tamaño de zapato a partir de la edad y el sexo de un niño (regresión)
- Determinar si un cliente debe recibir un crédito (clasificación)
- Naive Bayes
- Árboles de decisión



No supervisado

Se trata de decidir la estructura de un conjunto de datos sin clasificar.

- Distinguir segmentos de mercado (clustering)
- Separar señales acústicas (varias voces) obtenidas por distintos micrófonos

+ Spark ML lib

Machine Learning Basics

Una breve introducción a Machine Learning

¿Qué es Machine Learning?

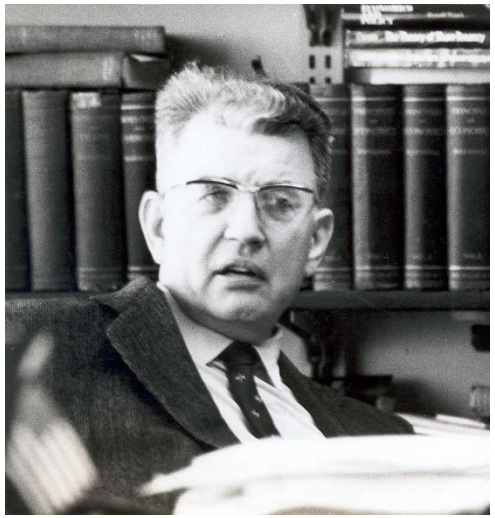
Aprendizaje supervisado

Aprendizaje no supervisado

Problemas comunes

Introducción a Spark

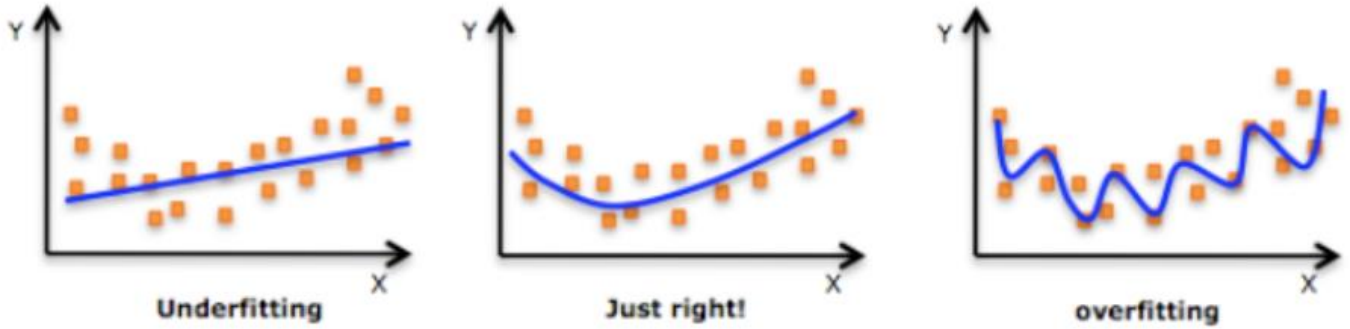
If you torture the data long enough, it will confess



Ronald H. Coase (1910-2013)

Buen ajuste

Hay que tener cuidado con ajustar demasiado o demasiado poco



Overfitting: El sistema "se aprende" los datos en lugar de buscar una función para predecir

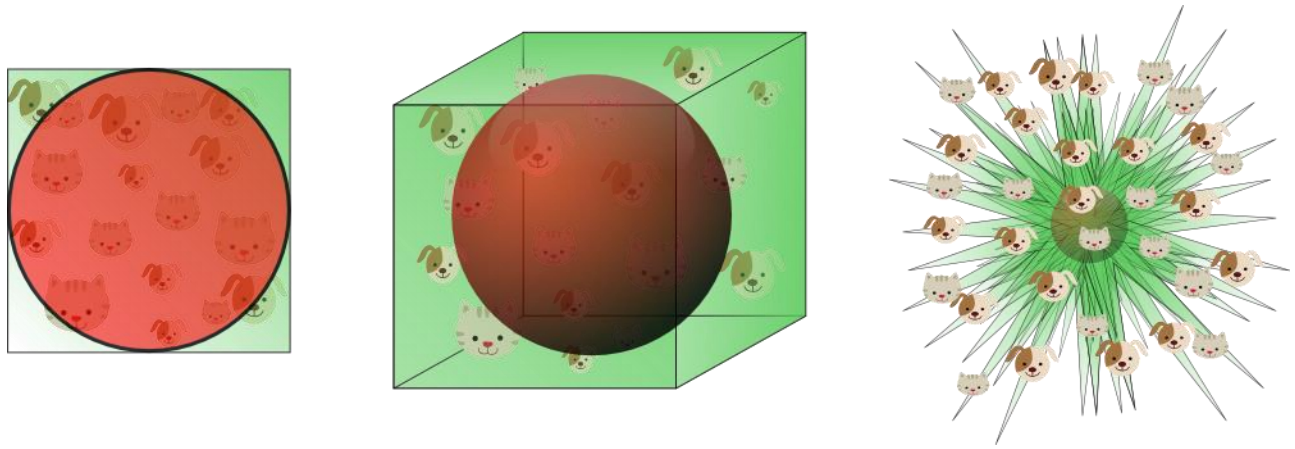
- ✓ Predicción perfecta para el conjunto de entrenamiento, pésima para otros datos
- ✓ ¿Cómo se detecta? Con conjuntos de test independientes (obtenidos quizás diviendo **E** en dos)
- ✓ ¿Cómo se evita?
  - ✓ Factores de regularización: atenúan ("quitan importancia") haciendo que la función no sea exacta
  - ✓ No demasiadas iteraciones en los algoritmos iterativos (descenso de gradiente, etc.)

La maldición de las dimensiones

Demasiadas columnas en los datos de entrada puede ser un problema

Surge en conjuntos de datos de entrada con muchas características (variables independientes, columnas)

- ✓ Un gran número de dimensiones hace que cada haya una gran cantidad de filas únicas
- ✓ Esto complica la detección de regularidades, y hace que se incremente enormemente el tamaño del conjunto de entrenamiento
- ✓ ¿Cómo se evita?: seleccionando un subconjunto relevante de dimensiones o combinaciones (PCA, SVD)



Para saber más: http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/

¿Para Qué?

Debemos tener en cuenta el uso que se hará de los resultados

El destino que vayamos a dar a los resultados puede determinar el método y la tecnología

- ✓ ¿Respuesta rápida? ¿Llegada de datos constante? ¿Recálculo del modelo?
- ✓ ¿Nivel de error permitido?
- ✓ ¿Se prefieren falsos positivos o falsos negativos?

Test

En los siguientes casos, ¿es preferible un falso negativo o un falso positivo?

A. Test de una enfermedad en un programa de screening
B. Sugerencia de un producto para comprar en una web
C. Concesión de una hipoteca evaluando posible impago
D. Detección policial de posibles terroristas

A)P B)? C) И D) P

Selección de datos de entrenamiento

Buena parte de los malos resultados en ML se debe a una mala selección de datos

- ✓ Los datos de entrenamiento deben constituir un conjunto completo (todos los casos representados, y en la proporción adecuada)
- ✓ Se deben hacer una limpieza de datos previa (ojo con las decisiones sobre missing data)
- ✓ Extracción de dimensiones (para evitar la maldición)
- ✓ Correctamente etiquetado; parece que no tiene importancia pero una mala elección dificulta el ver de un vistazo problemas/oportunidades

Papel de la estadística

Sin un mínimo de estadística no hay ciencia de datos

- ✓ Usaremos estadísticas descriptivas para “conocer” y “comprender” los datos: valores mínimos, máximos
- ✓ Muchas de las técnicas tienen que relación con estadística e internamente están definidas a partir de conceptos estadísticos
- ✓ Usaremos la estadística para medir lo bien que ha aprendido el programa