

Aprendizaje Automático

Introducción a aprendizaje automático

Master Executive Big Data y Business Analytics

Edición 2015

Temario

- ❖ **Sesión 1: Introducción a aprendizaje automático**
- ❖ Sesión 2: Modelos de clasificación y clustering
- ❖ Sesión 3: Aprendizaje profundo, PCA y aplicaciones

Tabla de contenido

- ❖ Aprendizaje automático y minería de datos
- ❖ scikit-learn
- ❖ Reglas de asociación (APriori)
- ❖ Regresión lineal
- ❖ Overfitting y validación cruzada

¿Qué es el Aprendizaje Automático?

- ❖ **Aprendizaje Automático (AA) o Machine Learning (ML)** es una rama de la **Inteligencia Artificial (IA)**.
- ❖ Agrupa diferentes técnicas que permiten a las computadoras **aprender patrones** a partir de conjunto de datos de ejemplo.
- ❖ En el entorno de Big Data se utiliza el AA para extraer los patrones que se encuentran en los datos para realizar predicciones y obtener hallazgos.
- ❖ El AA tiene una amplia gama de aplicaciones:
 - motores de búsqueda,
 - detección de fraude,
 - análisis del mercado de valores,
 - detección de intrusiones,
 - reconocimiento del habla y del lenguaje escrito, etc.

¿Por qué Aprendizaje Automático?

- ❖ Las técnicas de aprendizaje automático permiten **crear modelos sobre** los datos recogidos, que pueden utilizarse para cosas como:
 - **Hacer predicciones:** ¿qué clientes se cambiarán de compañía?
 - **Buscar patrones interesantes:** ¿qué productos compran juntos habitualmente mis clientes?
 - **Buscar grupos en los datos:** ¿qué segmentos de clientes tengo?
- ❖ El aprendizaje automático es “el ingrediente clave” en la analítica de negocio, que permite aprovechar el esfuerzo hecho en organizar la información (data warehousing) y en analizarla (bases de datos analíticas).

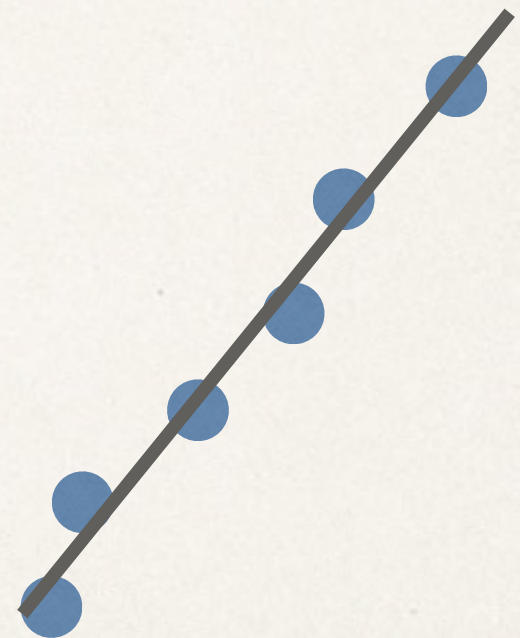
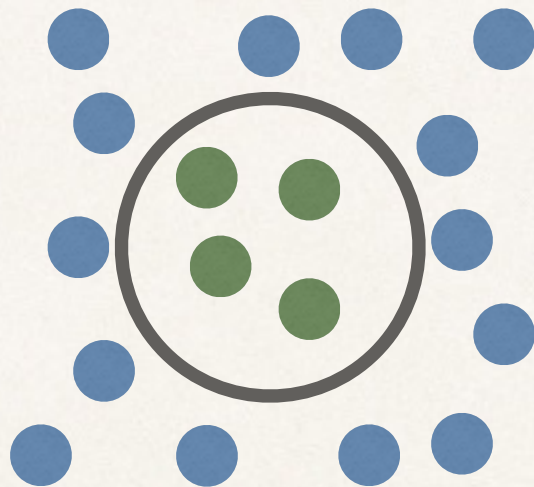
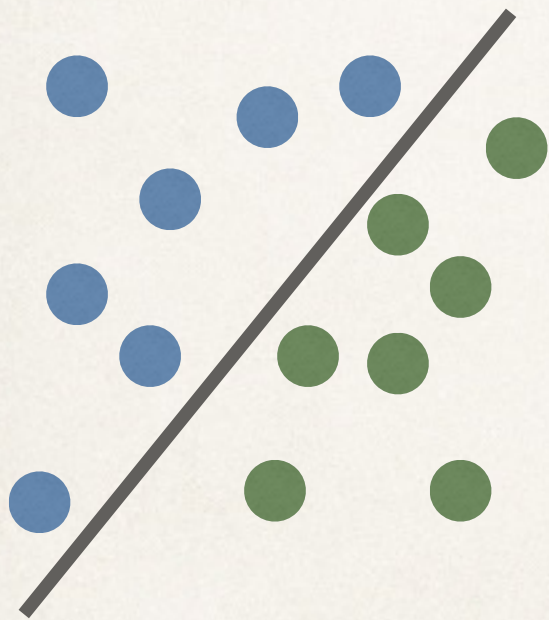
¿Qué es un modelo?

Los algoritmos utilizados en AA se utilizan para la creación de modelos que permiten realizar predicciones.

- ❖ Los algoritmos de AA toman un conjunto de datos para “entrenarse”.
- ❖ Un modelo es una abstracción de los datos que se han utilizado.
 - Puede ser una fórmula matemática, varias de ellas, un conjunto de reglas, etc.
 - En algunos casos, son valores en una estructura de conexiones (redes neuronales).
- ❖ Un caso particular de modelo es una “memorización” completa de los datos.

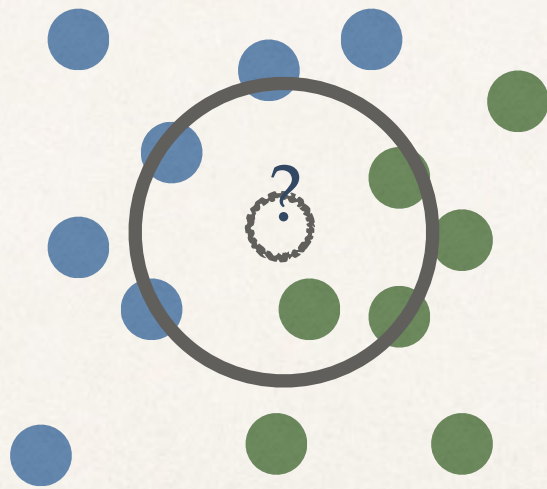
Los modelos dependen de los datos

Los modelos se han seleccionar en función de los datos y el problema a resolver.



Modelos y memorización: una excepción

- ❖ Algunos algoritmos como los clasificadores de K- vecinos (K-nn) memorizan los datos, no generalizan.
- ❖ No hay coste en el entrenamiento, sí en la predicción.
- ❖ Útiles en problemas con fronteras irregulares

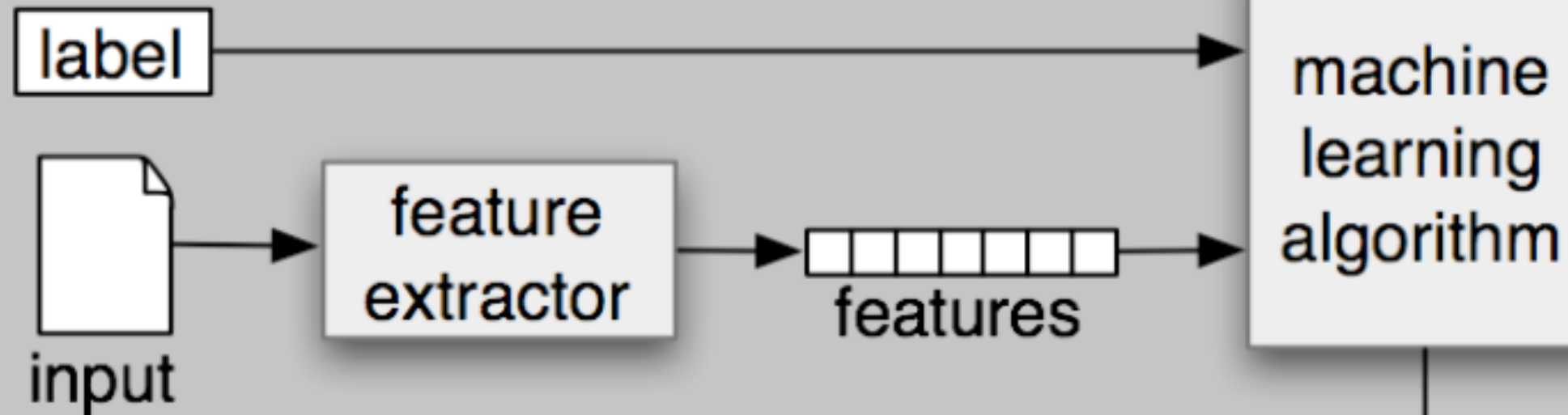


Ciclo de vida de los modelos

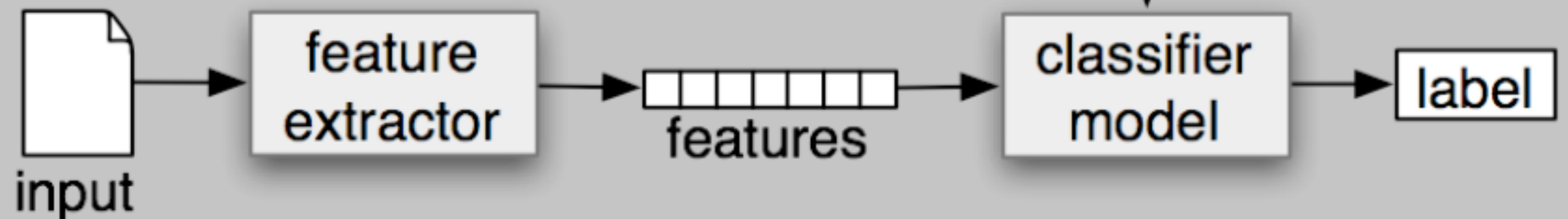
- ❖ Una vez un modelo está evaluado como bueno, pasa a producción.
 - Es decir, a utilizarse para las necesidades de negocio, por ejemplo, predecir fraude en nuevos clientes.
- ❖ El entrenamiento suele ser costoso en recursos de máquina, iterativo y requiere experiencia y conocimiento.
 - Es parte del proceso en que el científico de datos (data scientist) hace su trabajo.
- ❖ Una vez el modelo está en producción, su ejecución suele ser eficiente.
 - Lo utilizan los usuarios sin habitualmente conocer de dónde sale y cómo funciona internamente.

Modelos en producción

(a) Training



(b) Prediction





“Essentially, all models are wrong,
but some are useful.”

– *George E. P. Box*

Tipos de algoritmos en Aprendizaje Automático

Los tipos de algoritmos utilizados en Aprendizaje Automático se suelen dividir en dos categorías:

- ❖ **Aprendizaje supervisado:** los algoritmos utilizados en este caso requieren que cada una de las instancias de los datos se encuentre etiquetados con un atributo objetivo. Este atributo puede ser una clasificar los datos en una o más clases o ser un valor continuo.
- ❖ **Aprendizaje no supervisado:** los algoritmos utilizados en este caso no requiere que los datos se encuentren etiquetados con el atributo objetivo.

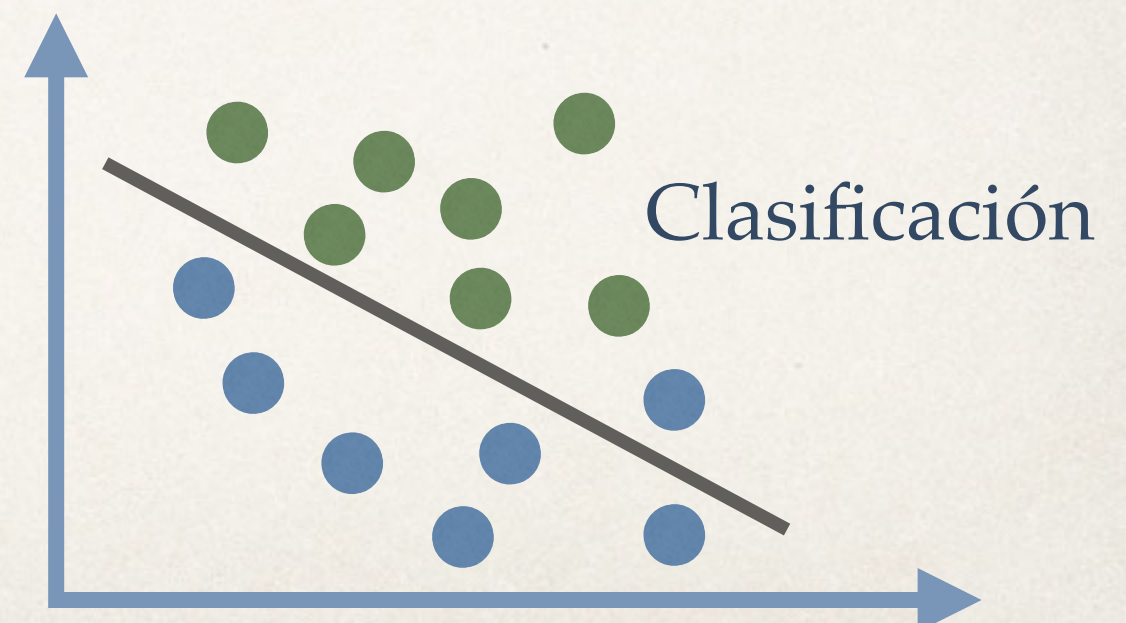
Aprendizaje supervisado y no supervisado

- ❖ **Supervisado:** si tenemos un conjunto de datos de hipotecas, podríamos querer tener un modelo del impago de las mismas. Entonces tendríamos:
 - Atributos de entrada: podrían ser en el ejemplo el principal de la hipoteca, el tipo de interés, la edad del cliente, la profesión, etc.
 - Un atributo que es la "salida", etiqueta u objetivo, que en este caso sería el campo que indicase si ese cliente impago o no la hipoteca.
- ❖ **No supervisado:** Podríamos querer simplemente buscar segmentos de clientes, es decir, buscar grupos de clientes homogéneos para crear productos personalizados, pero no sabemos a priori cuántos grupos habrá ni sus características.
 - Este es un ejemplo de "agrupamiento" (clustering).

Problemas en Aprendizaje Automático

Los tipos de problemas más habituales que se resuelven con técnicas de Aprendizaje Automático son:

- ❖ **Regresión:** En problemas de regresión los algoritmos aprendan a predecir el valor de una variable continua a partir de una o más variables explicativas.
- ❖ **Clasificación:** En los problemas de clasificación se busca que los algoritmos aprendan a predecir valores discretos a partir de una o más variables explicativas.

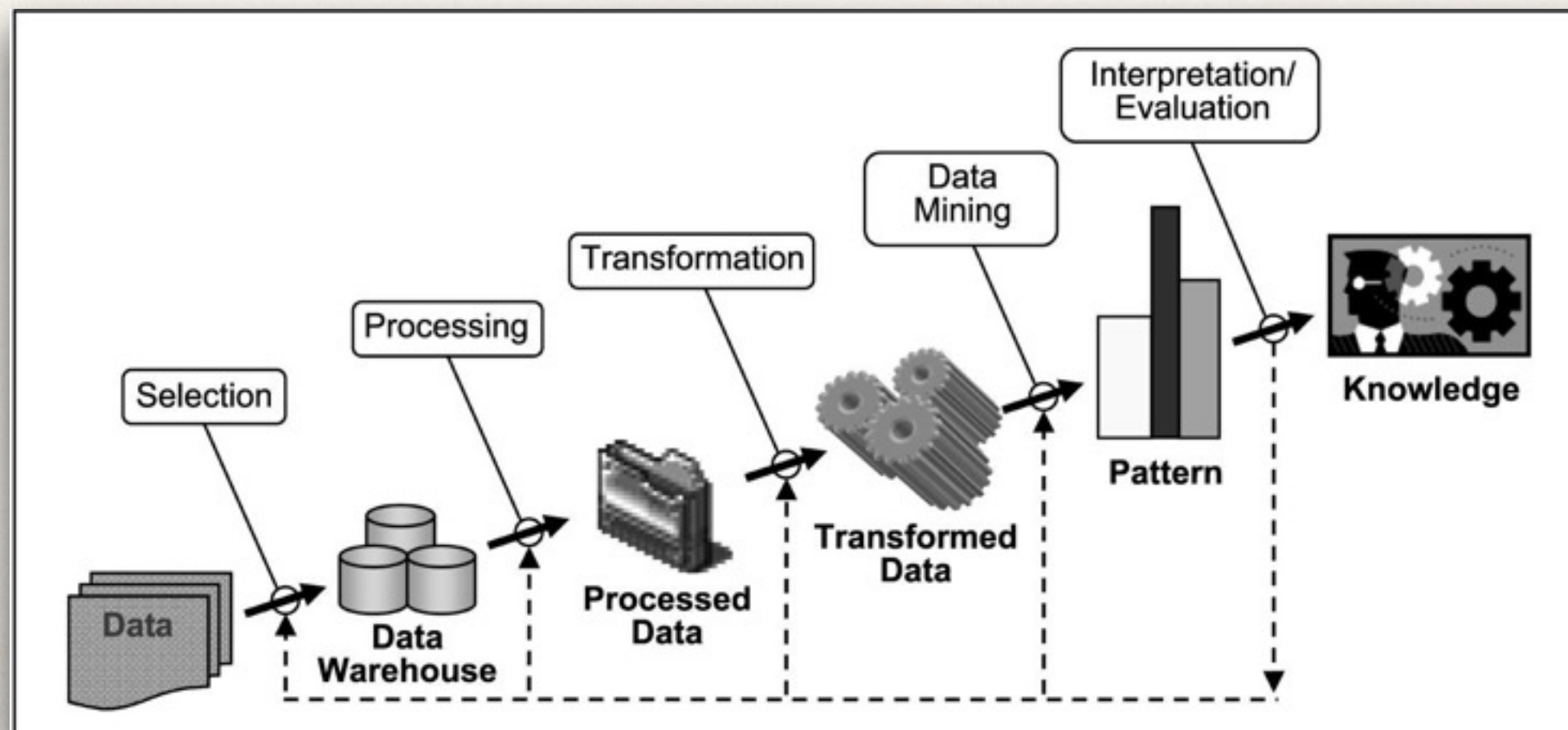


La minería de datos

- ❖ La minería comienza con algún tipo de objetivo o hipótesis sobre la existencia de minerales valiosos en una zona.
 - En la minería de datos, también hay algunas ideas de negocio o hipótesis iniciales.
 - La minería la hacen los mineros, pero se sirven de diferentes herramientas, cada vez más sofisticadas.
 - En la minería de datos, hay muchas de estas herramientas, incluyendo las de aprendizaje automático.
- ❖ La minería a veces tiene que excavar túneles alternativos.
 - En la minería de datos, también hay un proceso de ensayo y error.
- ❖ En la minería es necesario extraer muestras de mineral y analizar su calidad.
 - En la minería de datos, también es fundamental evaluar la calidad del conocimiento extraído.

El proceso de KDD

- ❖ **Knowledge Discovery in Databases (KDD)** es el proceso iterativo que incluye de manera amplia a la minería de datos y las actividades relacionadas.
- Algunos autores incluyen también: limpieza e integración al principio, y visualización y representación al final.



¿KDD = Data Mining?

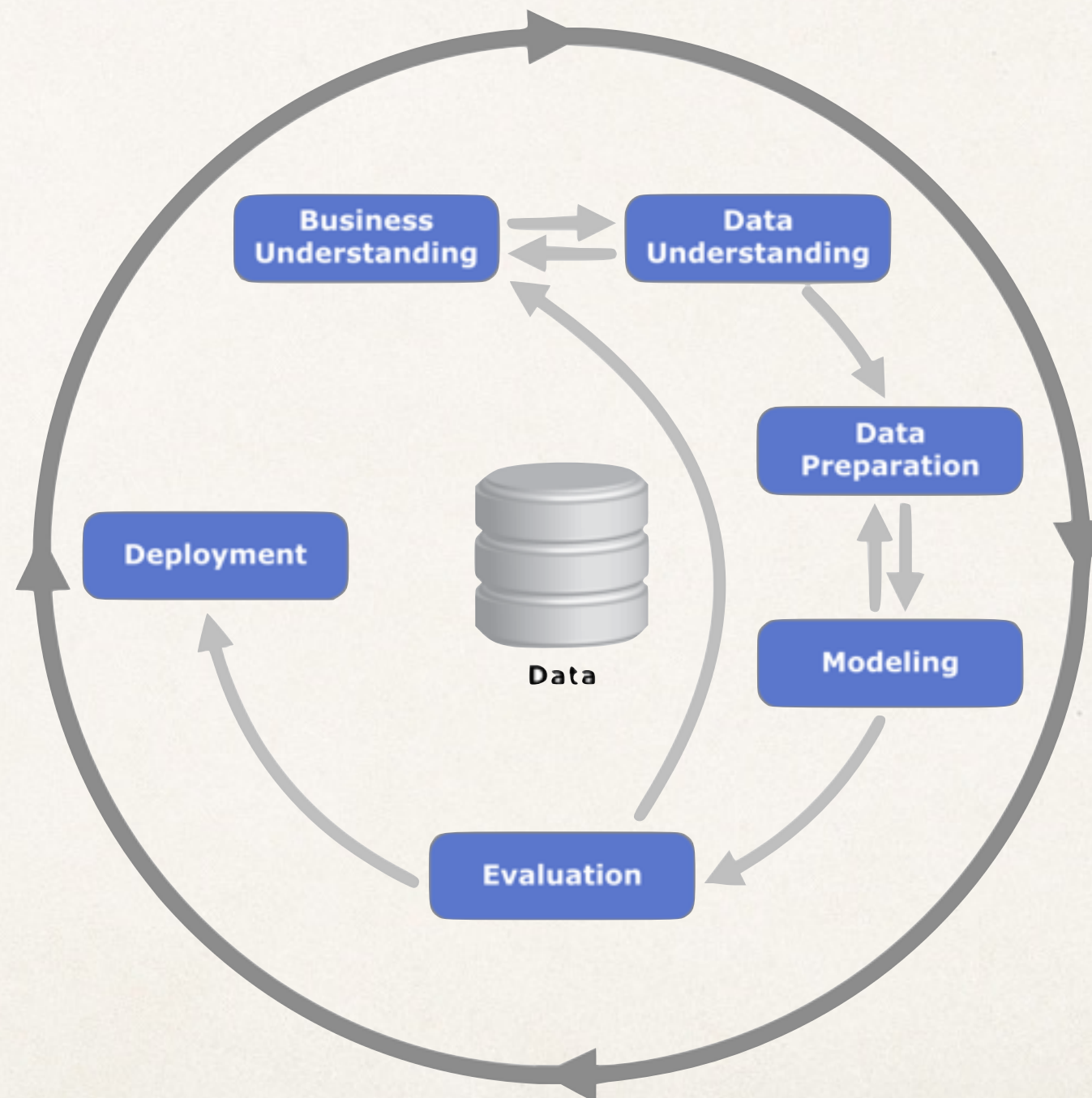
- ❖ “Data mining, also popularly referred to as knowledge discovery from data (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories or data streams.”

Data mining: concepts and techniques, J. Han & M. Kamber

- ❖ “KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data.”

From data mining to KDD. AI Magazine (1996). U. Fayad et al.

CRISP-DM Process Model



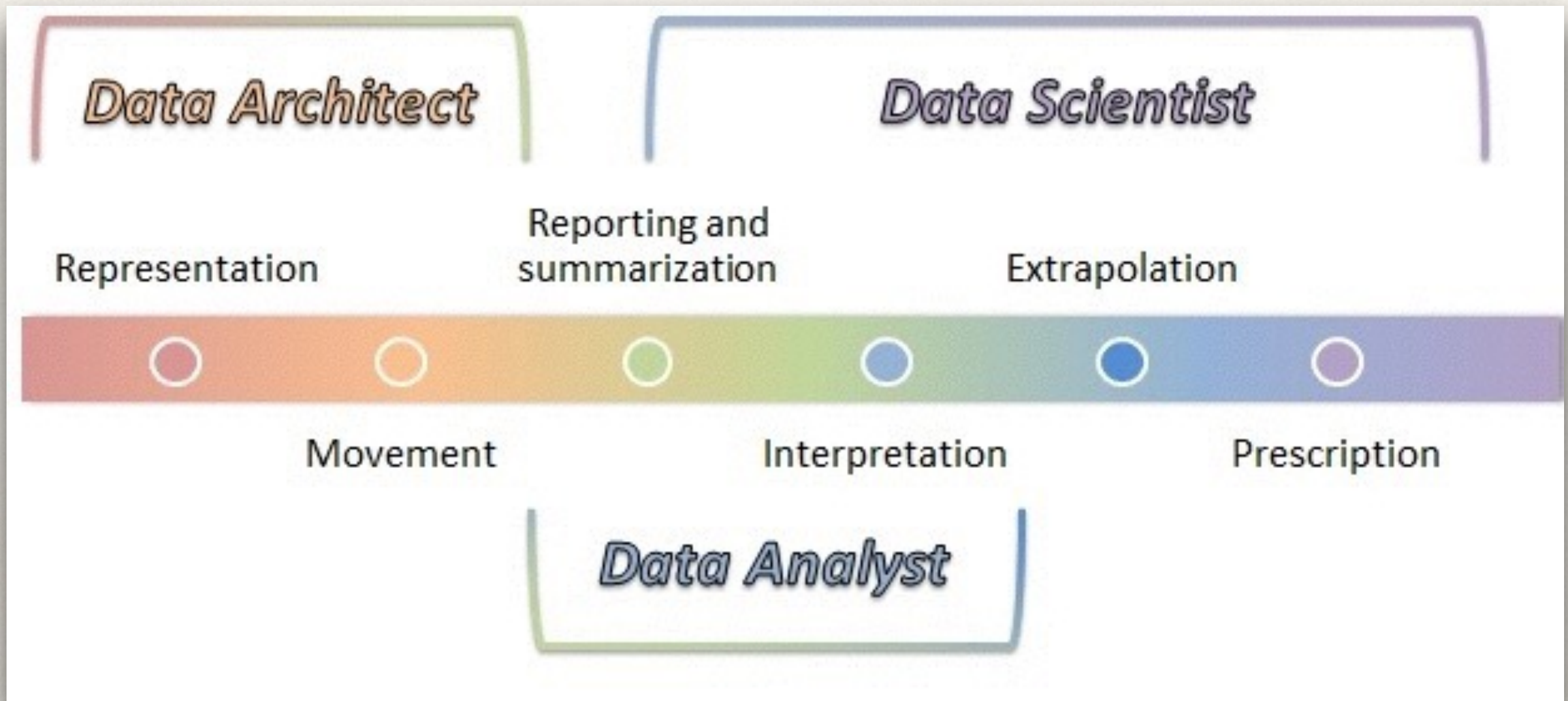
Cross Industry Standard Process for Data Mining

Las seis fases del proceso CRISP-DM

- ❖ **Comprensión del negocio:** En la fase inicial se busca comprender los objetivos y requisitos de los proyectos desde una perspectiva empresarial, y luego convertir este conocimiento en una definición del problema de minería de datos.
- ❖ **Comprensión de datos:** En esta fase se realiza una recolección inicial de datos y procesos con actividades con el objetivo de familiarizarse con los datos, identificar la calidad de los problemas, para descubrir las primeras señales dentro de los datos y detectar temas interesantes para poder formular hipótesis
- ❖ **Preparación de datos:** En esta fase se cubren las actividades para construir el conjunto de datos.
- ❖ **Modelado:** En esta fase se seleccionan y aplican varias técnicas de modelado y se calibran los parámetros para obtener los mejores resultados. Hay varias técnicas que tienen requerimientos específicos para la forma de los datos, por lo que frecuentemente es necesario volver a la fase de preparación de datos.
- ❖ **Evaluación:** En esta etapa del proyecto se ha construido un modelo (o modelos) que parece tener gran calidad, desde una perspectiva de análisis de datos.
- ❖ **Despliegue:** Esta fase se realiza el despliegue del modelo. Esto puede ser, dependiendo de los requisitos del proyecto, desde un informe como la implementación de un proceso más complejo.

Data Scientist vs Data Analyst

Diferentes roles, mismo marco de actividades



Tipos de estudios

De los más simples a los mas complejos, concretamente los siguientes:

- ❖ **Descriptivos:** descripción cualitativa de las principales características de los datos.
- ❖ **Exploratorios:** análisis de los datos en búsqueda de relaciones desconocidas.
- ❖ **Inferenciales:** evaluación de la validez de teorías en muestras de los datos.
- ❖ **Predictivos:** análisis de los datos presentes para obtener predicciones para eventos futuros:
- ❖ **Causales:** estudia lo que sucede en una variable al cambiar otra.
- ❖ **Mecanísticos:** comprensión de los cambios que se produce en las variables que producen cambios en otras variables.

Tabla de contenido

- ❖ Aprendizaje automático y minería de datos
- ❖ **scikit-learn**
- ❖ Reglas de asociación (APriori)
- ❖ Regresión lineal
- ❖ Overfitting y validación cruzada

scikit-learn

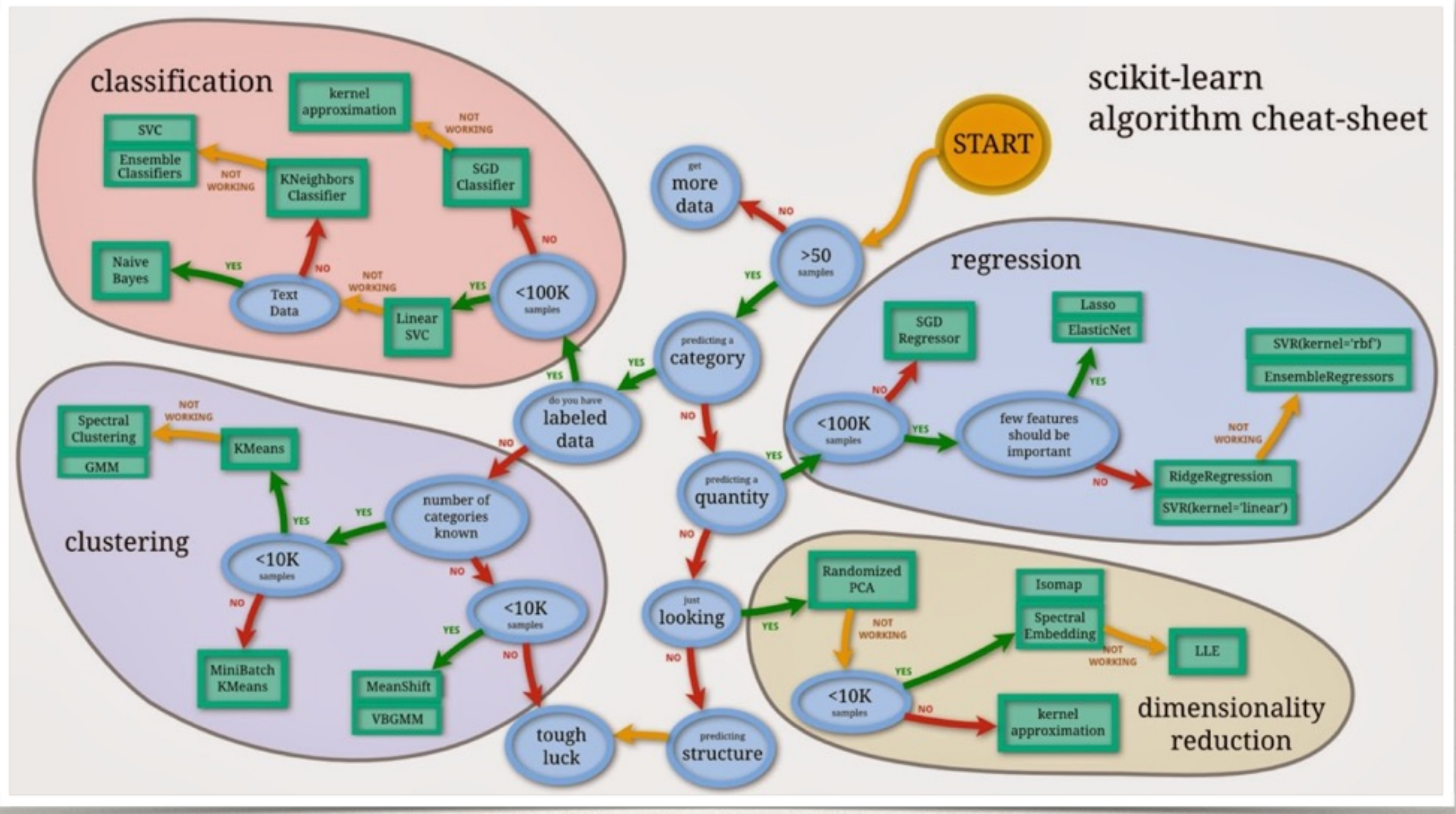
scikit-learn es una librería de open source para el lenguaje de programación Python que implemente diferentes métodos de Aprendizaje Automático.



scikit-learn cuenta con diversos algoritmos para:

- ❖ Clasificación,
- ❖ Regresión,
- ❖ Clustering,
- ❖ Reducción de la dimensionalidad

scikit-learn

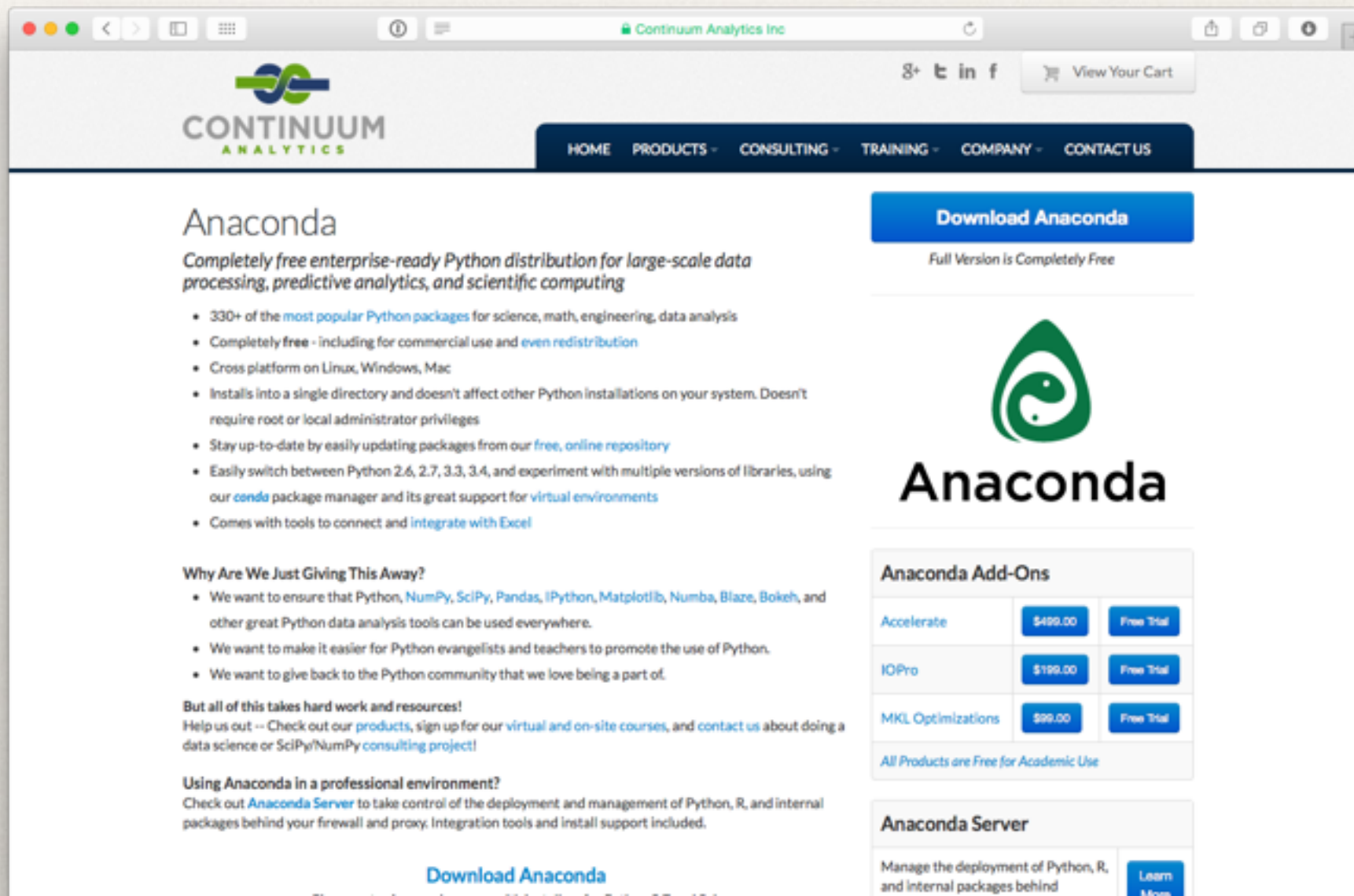


scikit-learn

scikit-learn

Anaconda es una distribución libre del lenguaje de programación Python para procesamiento a gran escala de datos, análisis predictivo y la computación científica, que tiene como objetivo simplificar la gestión y distribución de paquetes.

Utilizada el sistema de gestión de paquetes es Conda.



The screenshot shows the Anaconda website homepage. The header includes the Continuum Analytics logo, navigation links (HOME, PRODUCTS, CONSULTING, TRAINING, COMPANY, CONTACT US), and social media icons. The main content area features the Anaconda logo, a description of the distribution, a list of features, and a 'Download Anaconda' button. The right sidebar contains 'Anaconda Add-Ons' and 'Anaconda Server' sections.

Anaconda
Completely free enterprise-ready Python distribution for large-scale data processing, predictive analytics, and scientific computing

- 330+ of the most popular Python packages for science, math, engineering, data analysis
- Completely free - including for commercial use and even redistribution
- Cross platform on Linux, Windows, Mac
- Installs into a single directory and doesn't affect other Python installations on your system. Doesn't require root or local administrator privileges
- Stay up-to-date by easily updating packages from our free, online repository
- Easily switch between Python 2.6, 2.7, 3.3, 3.4, and experiment with multiple versions of libraries, using our conda package manager and its great support for virtual environments
- Comes with tools to connect and integrate with Excel

Why Are We Just Giving This Away?

- We want to ensure that Python, NumPy, SciPy, Pandas, IPython, Matplotlib, Numba, Blaze, Bokeh, and other great Python data analysis tools can be used everywhere.
- We want to make it easier for Python evangelists and teachers to promote the use of Python.
- We want to give back to the Python community that we love being a part of.

But all of this takes hard work and resources!
Help us out -- Check out our products, sign up for our virtual and on-site courses, and contact us about doing a data science or SciPy/NumPy consulting project!

Using Anaconda in a professional environment?
Check out Anaconda Server to take control of the deployment and management of Python, R, and internal packages behind your firewall and proxy. Integration tools and install support included.

Anaconda Add-Ons

Add-On	Price	Free Trial
Accelerate	\$499.00	Free Trial
IOPro	\$199.00	Free Trial
MKL Optimizations	\$99.00	Free Trial

All Products are Free for Academic Use

Anaconda Server

Manage the deployment of Python, R, and internal packages behind

[Learn More](#)

Principales interfaces

- ❖ Creación del objeto con el modelo:

```
m = Model()
```

- ❖ Estimación del modelo:

```
m.fit(data, [target])
```

- ❖ Realización de estimaciones:

```
m.predict(data)
```

- ❖ Transformación de los datos:

```
datos = obj.transform(data)
```


Ejemplo de sesión con scikit-learn

```
# Importación de las librerías
```

```
from sklearn.neighbors import RadiusNeighborsClassifier
```

```
# Creamos el modelo sin entrenar:
```

```
model = RadiusNeighborsClassifier(radius = 1)
```

```
# Entrenamos el modelo:
```

```
model.fit(X,y)
```

```
# Predecimos la clase para tres puntos diferentes:
```

```
print model.predict([[0.8, 1.5], [2.3, 2.8], [2, 2]])
```

Tabla de contenido

- ❖ Aprendizaje automático y minería de datos
- ❖ scikit-learn
- ❖ Reglas de asociación (APriori)
- ❖ Regresión lineal
- ❖ Overfitting y validación cruzada

Reglas de asociación

Las **reglas de asociación** se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos.

Las reglas de asociación se suelen expresar de la forma

$$X \Rightarrow Y$$

donde X e Y son items o conjuntos de items (ítemsets). Por ejemplo:

$$\{cerveza\} \Rightarrow \{pañales\}$$

$$\{pan, leche\} \Rightarrow \{huevos\}$$

$$\{cebollas, vegetales\} \Rightarrow \{carne\}$$

Conceptos básicos

- ❖ Soporte (Support): es el porcentaje de transacciones en las que aparece X. El soporte de una regla de asociación es el porcentaje de transacciones que contiene X e Y

$$supp(X \Rightarrow Y) = supp(X \cup Y)$$

- ❖ Confianza (Confidence): es la fracción de las transacciones en las que aparece X y también aparece Y

$$conf(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(Y)}$$

- ❖ Mejora de la confianza (Lift): es la fracción del soporte observado para una regla respecto al teórico suponiendo independencia.

$$lift(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X) \times supp(Y)}$$

- ❖ Convicción (Conviction): es el ratio de la frecuencia esperada para X que sucede sin Y, es decir, la frecuencia en la que la regla realiza una predicción errónea

$$conv(X \Rightarrow Y) = \frac{1 - supp(Y)}{1 - conf(X \cup Y)}$$

Associative Learning Algorithms

Dentro del aprendizaje no supervisado hay una familia de algoritmos que se suele denominar Associative Learning Algorithms, e incluyen a los siguientes:

- ❖ Apriori Algorithm
- ❖ Equivalence Classification Algorithm (Eclat)
- ❖ PrefixSpan
- ❖ FP-Growth

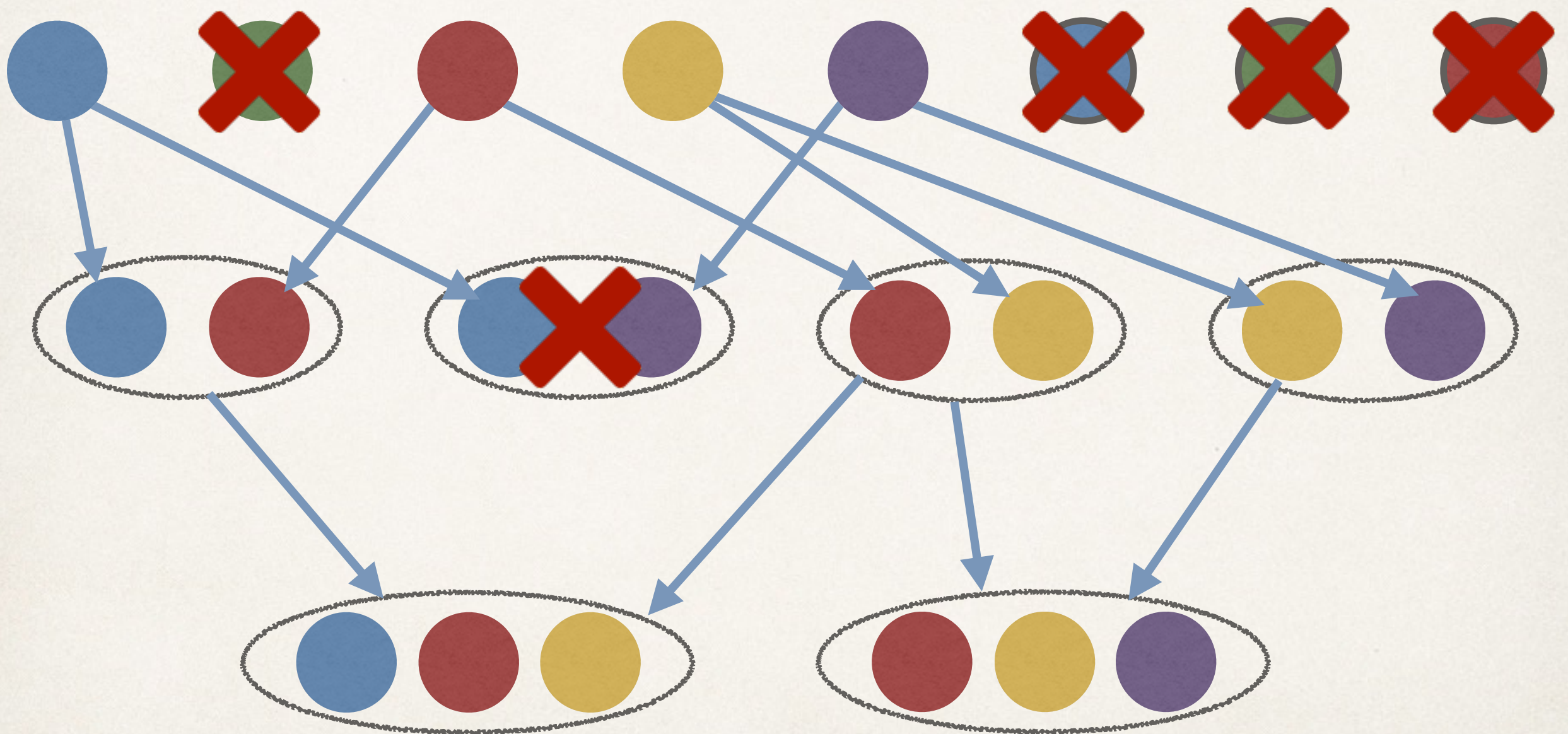
APriori

El algoritmo **APriori** es probablemente el más conocido para la búsquedas de reglas de asociación.

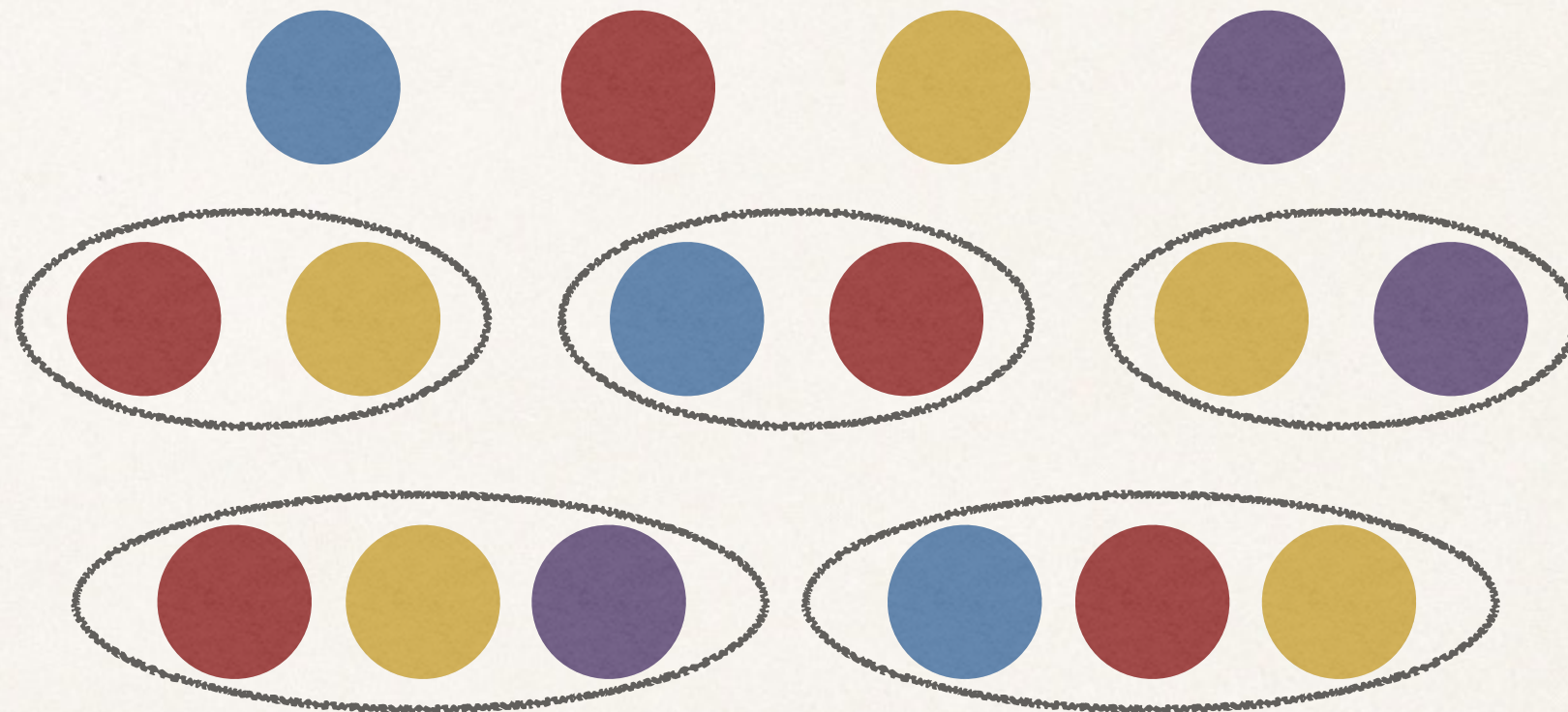
Originariamente se aplicó a buscar asociaciones entre productos comprados en supermercados, por lo que a algoritmos similares se les llama a veces también **market basket analysis**.

APriori

Selección de los ítemsets con un mínimo de soporte



APriori



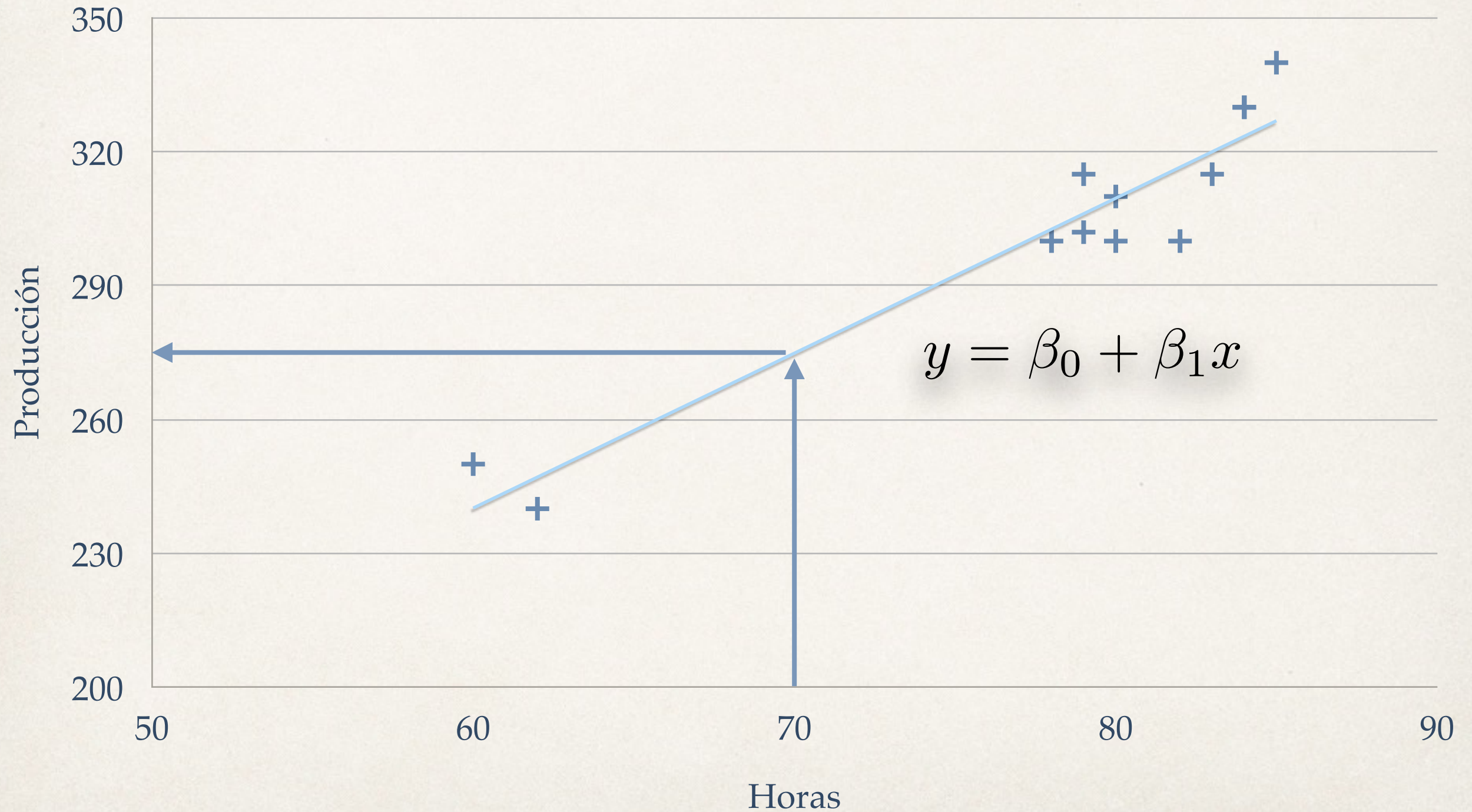
A partir de los ítemsets se crean reglas de asociación
con un mínimo de confianza



Tabla de contenido

- ❖ Aprendizaje automático y minería de datos
- ❖ scikit-learn
- ❖ Reglas de asociación (APriori)
- ❖ **Regresión lineal**
- ❖ Overfitting y validación cruzada

Regresión Lineal



Notación y datos de entrenamiento

$$y = \beta_0 + \beta_1 x$$

- ❖ x : variable explicativa
- ❖ y : variable dependiente u objetivo
- ❖ β_i : parámetros
- ❖ m : número muestras de entrenamiento

Horas (x)	Producción (y)
80	300
79	302
83	315
84	330
78	300
60	250
82	300
85	340
79	315
84	330
80	310
62	240

Función de esfuerzo

Hipotesis:

$$f(x_i; \beta_0, \beta_1) = \beta_0 + \beta_1 x_i$$

Parámetros: β_0, β_1



Función de esfuerzo (mínimos cuadrados):

$$J(\beta_0, \beta_1) = \frac{1}{2m} \sum_{i=1}^m (f(x_i; \beta_0, \beta_1) - y_i)^2$$

Objetivo: minimizar $J(\beta_0, \beta_1)$

Formulación matricial

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \dots \\ \beta_0 + \beta_1 x_n \end{bmatrix} = \begin{bmatrix} 1 + x_1 \\ 1 + x_2 \\ \dots \\ 1 + x_n \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = X\beta$$

Los valores de la regresión se pueden obtener mediante la expresión

$$\beta = (X^T X)^{-1} X^T Y$$

Función de esfuerzo

Hipotesis:

$$f(x_i; \beta_0, \beta_1) = \beta_0 + \beta_1 x_i$$

Parámetros: β_0, β_1

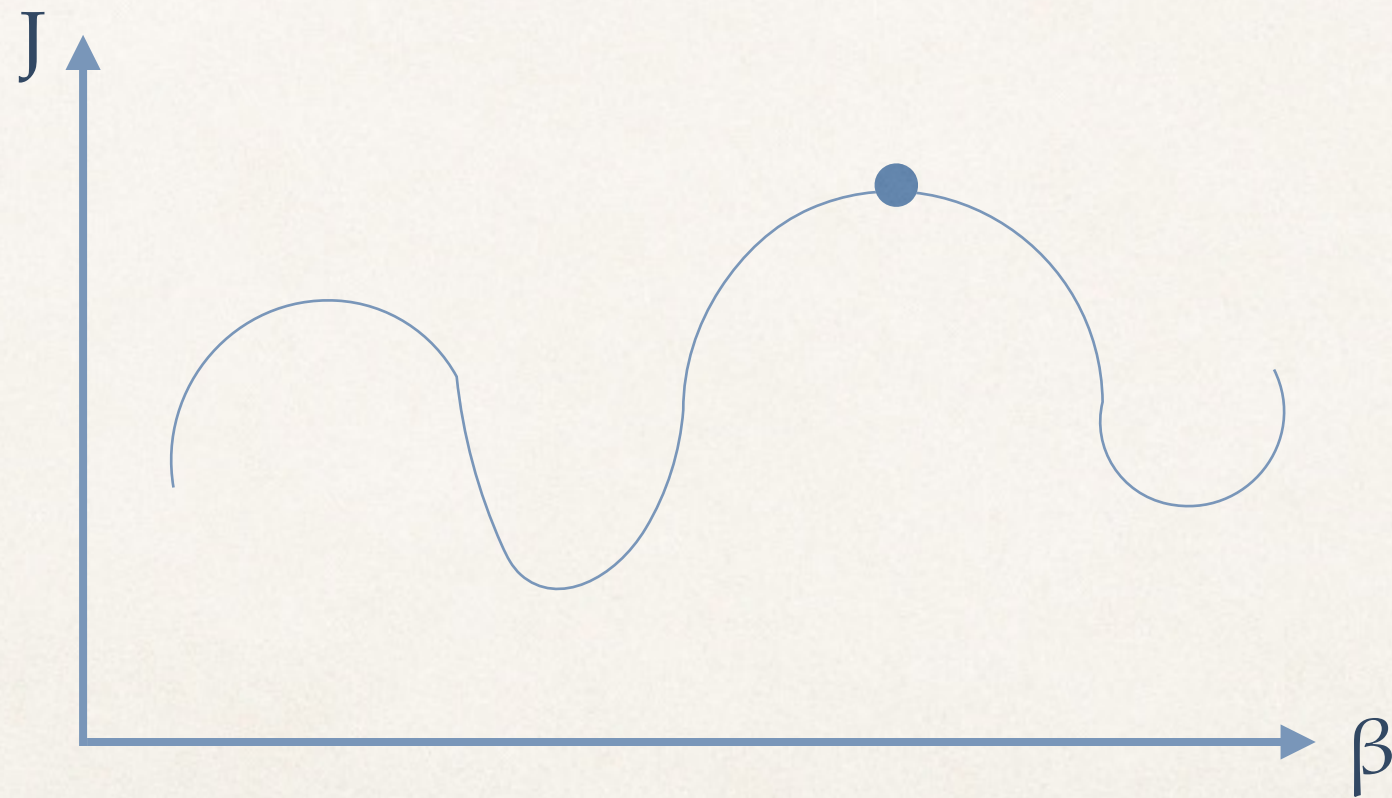
Función de esfuerzo (mínimos cuadrados):

$$J(\beta_0, \beta_1) = \frac{1}{2m} \sum_{i=1}^m (f(x_i; \beta_0, \beta_1) - y_i)^2$$

Objetivo: minimizar $J(\beta_0, \beta_1)$

Algoritmo de descenso de gradiente

$$\beta_i := \beta_i - \alpha \frac{\partial}{\partial \beta_i} J(\beta_0, \beta_1) \quad \alpha: \text{ration de aprendizaje}$$



Evaluación del modelo (R^2)

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

donde

$$SS_{tot} = \sum_{i=0}^m (y_i - \bar{y})^2$$

El modelo es mejor cuando R^2 se acerca a 1, es decir $SS_{res} = 0$.

$$SS_{res} = \sum_{i=0}^m (y_i - f(x_i; \beta_0, \beta_1))^2$$

El R^2 determina la calidad del modelo para replicar los resultados, indicando la proporción de variación de los resultados que se puede explicarse mediante el modelo.

Regresión lineal múltiple

Los modelos se pueden crear con varias variables explicativas

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n = \sum_{j=0}^n \beta_j x_j$$

Un caso particular es la regresión de polinomios

$$y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \dots + \beta_n x^n$$

Variables Categóricas

Las variables categóricas no pueden ser introducidas tal cual en un modelo lineal han de ser transformadas.

Para esto se suele realizar una nuevas variables lógicas que son ciertas cuando aparece la categoría en la variable original.

Ciudad	Madrid	Barcelona	Palma
Madrid	1	0	0
Barcelona	0	1	0
Palma	0	0	1
Madrid	1	0	0
Barcelona	0	1	0

Regularización

La regularización permiten evitar “overfitting” añadiendo penalizaciones a la función de esfuerzo utilizada en el ajuste.

La dos más utilizadas son Ridge

$$J(\beta) = \frac{1}{2m} \sum_{i=1}^m (f(x_i) - y_i)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

y LASSO (Least Absolute Shrinkage and Selection Operator)

$$J(\beta) = \frac{1}{2m} \sum_{i=1}^m (f(x_i) - y_i)^2 + \lambda \sum_{j=1}^p \beta_j$$

Selección de variables

- ❖ Eliminar variables con baja varianza
- ❖ Análisis univariante
- ❖ Eliminación recursiva de variables

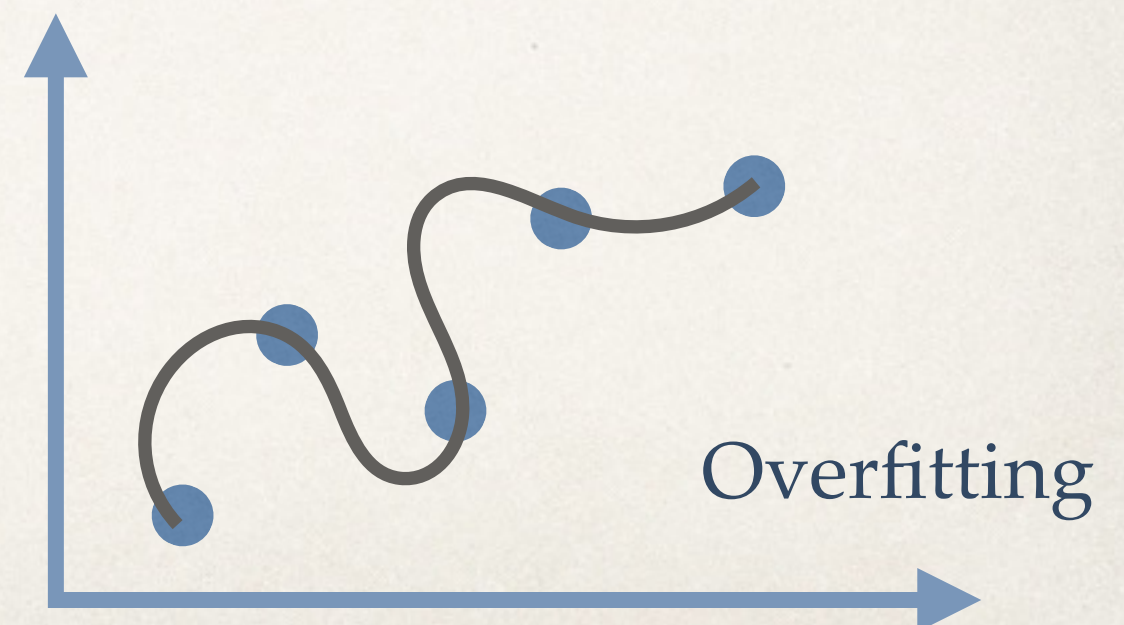
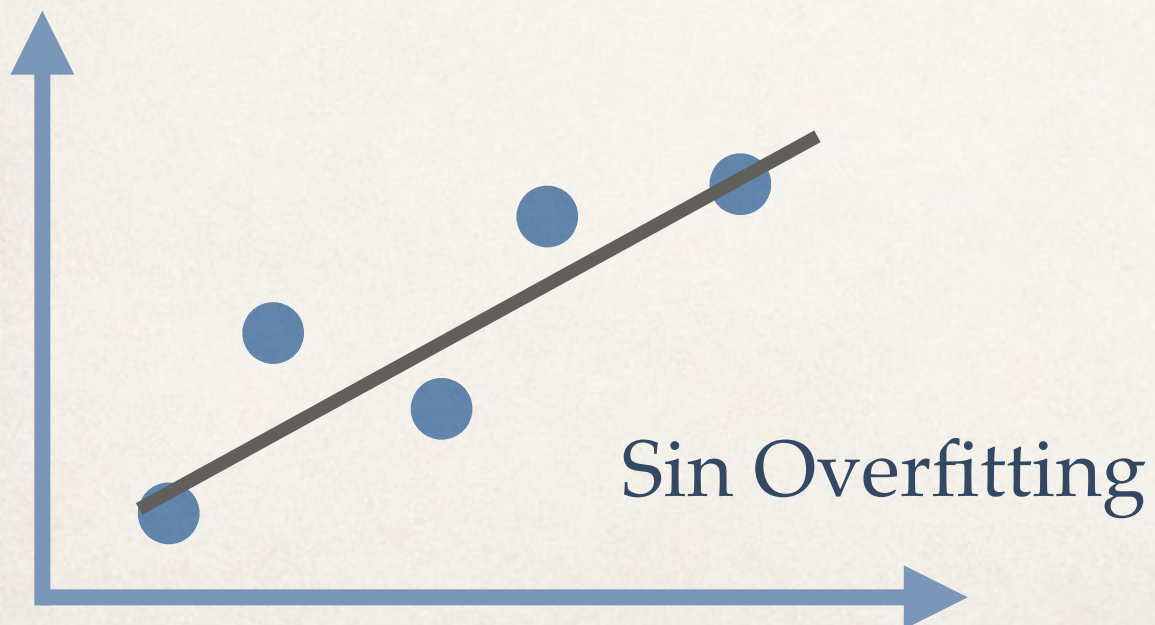
Tabla de contenido

- ❖ Aprendizaje automático y minería de datos
- ❖ scikit-learn
- ❖ Reglas de asociación (APriori)
- ❖ Regresión lineal
- ❖ Overfitting y validación cruzada

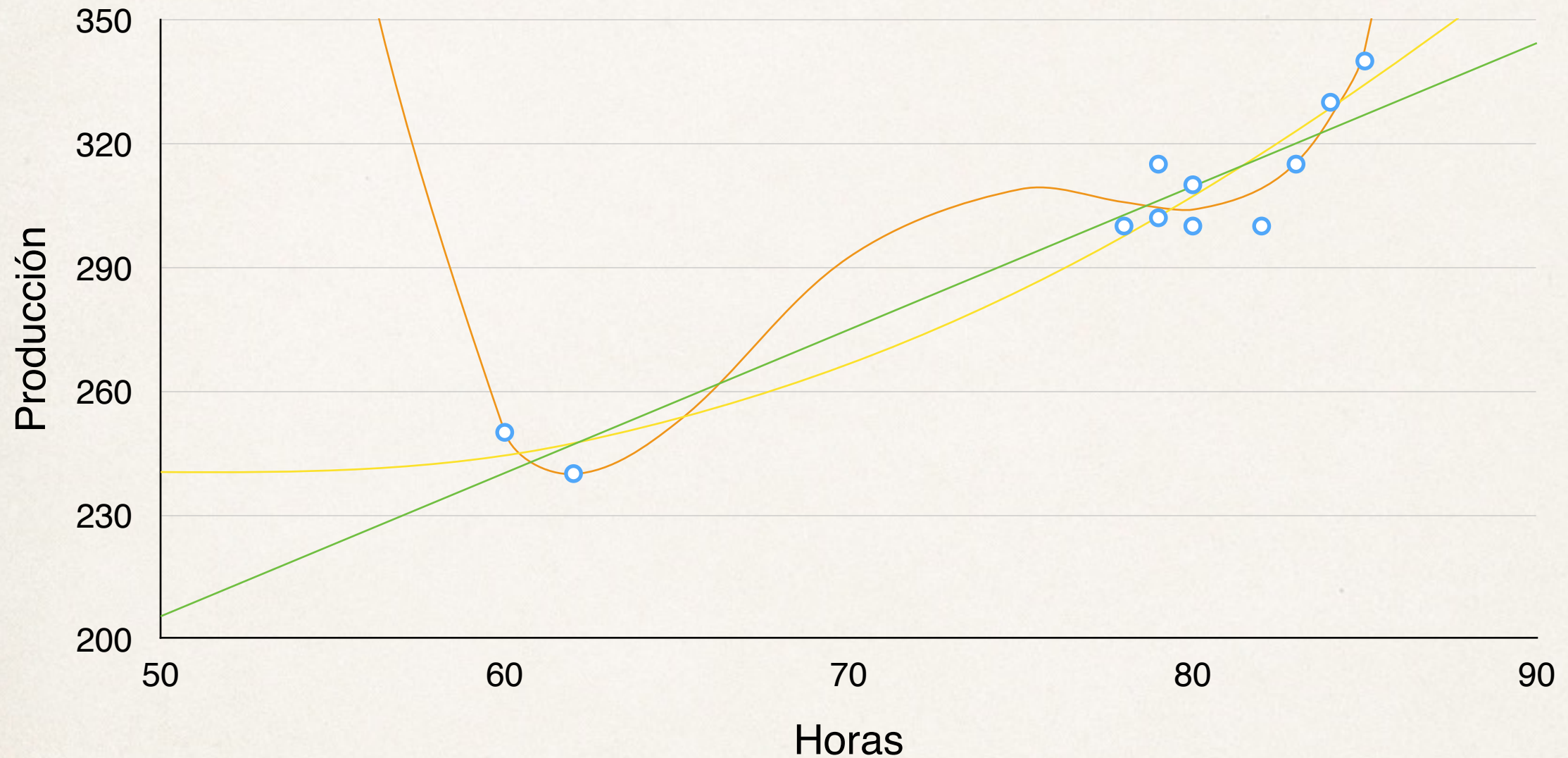
El problema del sobreajuste

El **sobreaprendizaje** o **sobreajuste** (overfitting) aparece cuando el algoritmo de aprendizaje memoriza el ruido existente en los datos con los que se está creando el modelo.

Esto no es deseable ya que las predicciones que se obtengan con estos modelos no serán precisas.



Ejemplo de sobreajuste



Al aumentar la complejidad del modelo puede aparecer overfitting. En esta situación el modelo aprende el ruido existente en los datos.

Datos de entrenamiento y datos de prueba

- ❖ Uno de los procedimientos utilizados para evitar los problemas de sobre ajuste es la utilización de conjuntos de datos diferentes para el entrenamiento, validación y test de los modelos:
 - Entrenamiento: los datos utilizados para el entrenamiento de los modelos
 - Validación: los datos utilizados para la selección del modelo.
 - Test: los datos utilizados para estimar el error del modelo seleccionado.
- ❖ Los datos se suelen repartir en proporciones 50 / 25 / 25 o 70 / 15 / 15.

Entrenamiento

Validación

Test

Validación cruzada

- ❖ La **validación cruzada** se puede utilizar para validar los modelos cuando el conjunto de datos es reducido.
- ❖ En la **validación cruzada** los datos se dividen en N grupos. El algoritmo se utiliza en todos menos uno de los grupos y se prueba en el restante. Posteriormente los grupos son rotados y para que el algoritmo aprenda con todos los datos.

	Datos A	Datos B	Datos C	Datos D
Iteración 1	Test	Entrenamiento	Entrenamiento	Entrenamiento
Iteración 2	Entrenamiento	Test	Entrenamiento	Entrenamiento
Iteración 3	Entrenamiento	Entrenamiento	Test	Entrenamiento
Iteración 4	Entrenamiento	Entrenamiento	Entrenamiento	Test