M2.2.2 Modelos Supervisados y No Supervisados

Programa Big Data y Business Intelligence

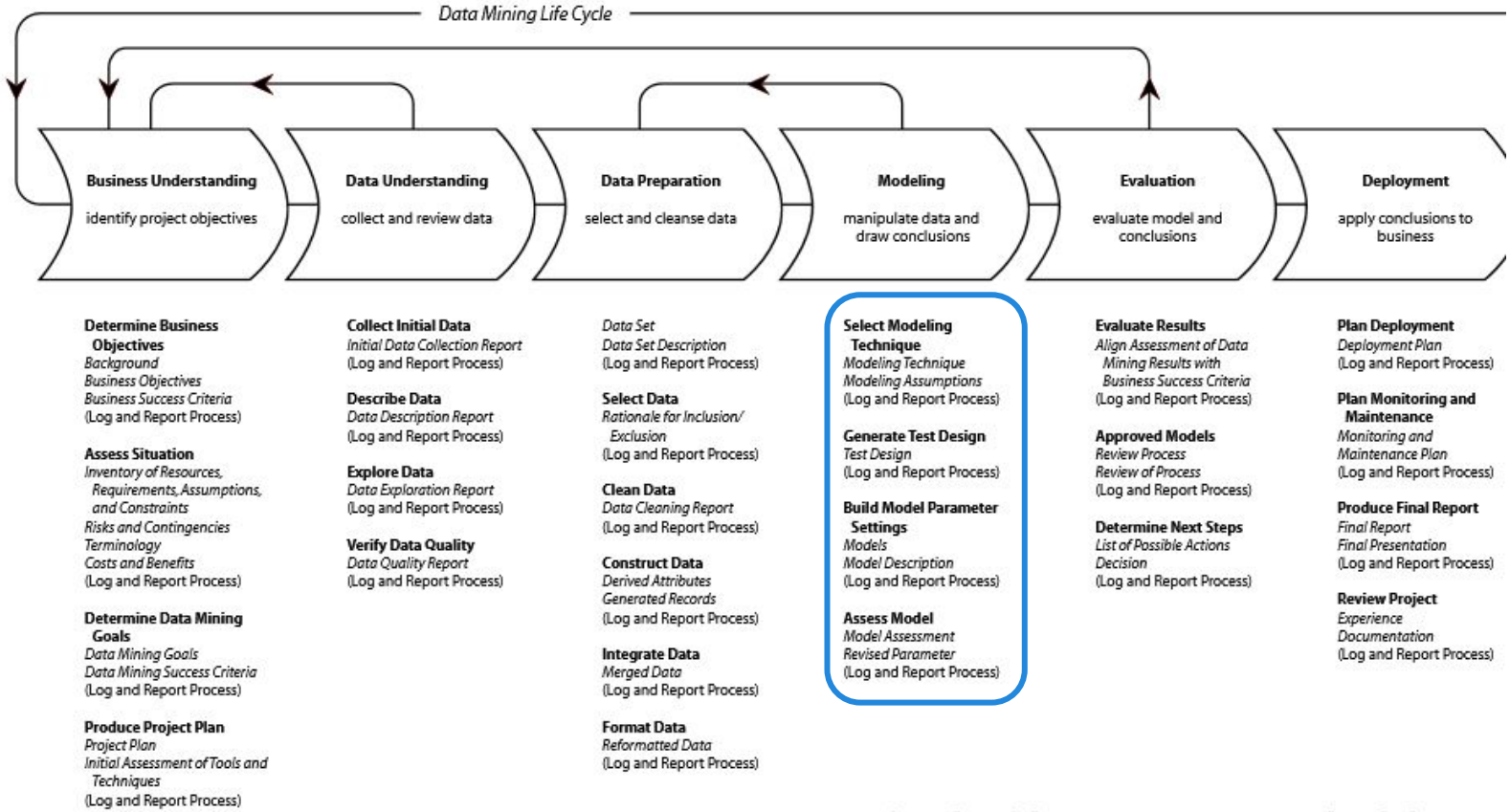
Enrique Onieva
enrique.onieva@deusto.es
https://twitter.com/EnriqueOnieva
https://www.linkedin.com/in/enriqueonieva/

Deusto
Facultad de Ingeniería
Ingeniaritza Fakultatea

(Antes de nada)

Descarga el directorio "src" que encontrarás en la carpeta de materiales en tu PC

En estas sesiones



Introducción a Machine Learning

- Terminología
- La importancia de la preparación de los datos
- Modelos y métodos
- El teorema del No Free Lunch
- Tipos de Aprendizaje-Modelos
  - Supervisados
  - No supervisados
  - Semi-supervisados

Terminología

- Pensemos en una tabla, como Excel, donde hay columnas, filas y celdas…
  - Columna → una columna incluye los datos de un determinado tipo. Todos los datos en ella deben de tener la misma escala y tener un significado relativo.
    - (Calidad de datos)
  - Fila → cada fila representa una entidad u observación
    - Las columnas describen las propiedades de esta
  - Celda → valor en una determinada fila y columna
    - Booleano, categoría, entero, real, texto...

	A	B	C	D
1		Column 1	Column 2	Column 3
2	Row 1	2.2	2.3	1
3	Row 2	2.3	2.6	0
4	Row 3	2.1	2	1

Terminología

- Perspectiva Estadística
  - Variables dependientes
  - Variable independientes

	A	B	C
1	X1	X2	Y
2	2.2	2.3	1
3	2.3	2.6	0
4	2.1	2	1
5			

- Perspectiva Ciencias de la Computación
  - Fila → Entidad, instancia, ejemplo...
  - Columna → Atributo, característica…
  - Podemos hablar de atributos de entrada y salida

	A	B	C	D
1		Attribute 1	Attribute 2	Output Attribute
2	Instance 1	2.2	2.3	1
3	Instance 2	2.3	2.6	0
4	Instance 3	2.1	2	1
5				

Terminología

- Perspectiva Estadística
  - Variable de Salida = f(Variables de Entrada)
  - Variable de Salida = f(Vector de Entrada)
  - Variable dependiente = f(Variables Independientes)
  - Y=f(X)
- Perspectiva Ciencias de la Computación
  - Salida = programa(atributos de entrada)
  - Predicción = programa(instancia)

Preparación de Datos

- Las técnicas de minería de datos (generalmente)
  - Trabajan con una única tabla de datos
  - Trabajan únicamente con los datos que hay en esa tabla
    - No consideran información (por evidente que pueda ser) más allá de los valores almacenados en la tabla
  - Si tenemos en un campo fecha “01/01/2016”
    - No saben que ese día fue viernes, y que el viernes es fin de semana (o no)
    - No saben que ese día es Año Nuevo, y que es festivo
    - No saben que el segundo “‘01” significa “enero”
- Por eso es importante la preparación de datos para su procesamiento por estas técnicas

Preparación de Datos

- Por eso hablaremos de datasets, en lugar de tablas, bases de datos…
- Por eso hablaremos de atributos en lugar de columnas, campos...
  - Estos atributos pueden extraerse de la base de datos
  - En (muchas) ocasiones deben de calcularse a partir de la información almacenada

Preparación de Datos

- Son dependientes del problema a abordar. Deben ser relevantes para la pregunta:
  - Quiero estudiar las ventas por mes
    - "01/01/2016" → Atributo MES = Enero
  - Quiero ver si recibo más compras a primeros de mes
    - "01/01/2016" → Atributo DIA_DEL_MES = 1
  - Quiero ver si hay más ventas días festivos
    - "01/01/2016" → Atributo FESTIVO = TRUE
  - Quiero ver si las ventas se elevan en fines de semana
    - "01/01/2016" → Atributo FIN_DE_SEMANA = FALSE

Preparación de Datos

- Son dependientes del problema a abordar. Deben ser relevantes para la pregunta:
  - Si tengo datos de ventas, y quiero estudiar clientes, para cada cliente, puedo extraer (entre otros)
    - Número de compras, Antiguedad,...
    - Frecuencia de compra (compras por mes)
    - Compra promedio (gasto promedio)
    - Categoría de productos adquiridos más frecuentemente
    - Preferencias de envío (ordinario o urgente / nacional o internacional)
    - Tipo de cliente (Normal o premium)

Preparación de Datos

- Son dependientes del problema a abordar. Deben ser relevantes para la pregunta:
  - Si tengo documentos, puedo extraer (entre otros)
    - Número de palabras, Idioma, Contiene Imagenes?, Temática
    - Cuántas veces aparece el término “Big Data” (o cualquier otro). Ojo:
      - “Big Data” ≠ “big data” ≠ “big-data” ≠ “Big data”
      - Requerirá de una preparación especial (dejar caracteres, minúscula, sin acentos...)
      - “Candidato” ≠ “Candidata”
      - Se puede extraer la raíz “Candidat” → “Candidato” = “Candidata” = “Candidatura”
    - Procedencia del autor
    - Fecha de creación/modificación → antigüedad

Preparación de Datos

- Son dependientes del problema a abordar. Deben ser relevantes para la pregunta:
  - Si tengo enlaces, puedo extraer (entre otros)
    - Dominio, Activo o Caído, Sitio web, Contiene enlaces...
    - [Todo lo comentado para documentos, sobre el contenido]
  - Si tengo imágenes, puedo extraer (entre otros)
    - Procesando con software especializado → Qué objetos aparecen
    - Sin software especializado
      - Color promedio, mediano, modal… Color promedio, mediano, modal… en determinada zona
      - Porcentaje de rojo, verde, azul…

Preparación de Datos

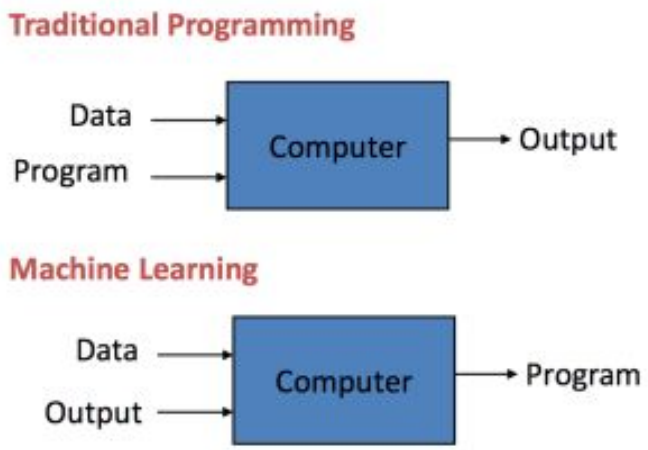
- Son dependientes del problema a abordar. Deben ser relevantes para la pregunta:
  - Para cualquier atributo es importante tener siempre presente la información que nos interesa y la que no
    - ¿Nos interesa la fecha de nacimiento del cliente?
      - ¿Sólo nos interesa si es mayor de 65 o no?
      - ¿Sólo nos interesa si su edad está en los rangos 20-30, 30-40, 40-50, 50-60...?
      - ¿Sólo nos interesa la cercanía a su cumpleaños?
    - ¿Nos interesa su localidad?
      - ¿Sólo nos interesa si es nacional o internacional?
      - ¿Sólo nos interesa el país?
      - ¿Sólo nos interesa la distancia desde nuestro almacén?
      - ¿Sólo nos interesa el idioma oficial de su localidad?

Modelos y Métodos

- Modelo: representación específica hecha a partir de los datos
  - Árbol de decisión
  - Red Neuronal
  - Conjunto de coeficientes
- Método (Algoritmo): proceso seguido para obtener un modelo
  - C4.5
  - Regresión lineal por mínimos cuadrados
  - “Probar coeficientes a lo loco hasta que esto funcione”

Modelos y Métodos

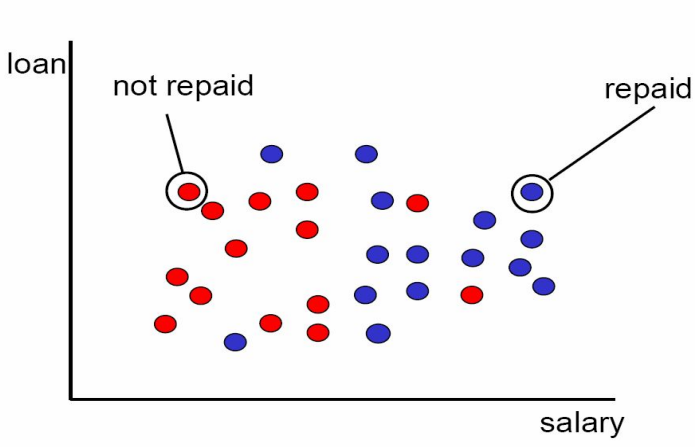
- Modelo (entrenado): resultado de la aplicación de un método sobre unos datos específicos
  - En la mayoría de los casos, el método (de aprendizaje) se encargará de dar valores a las variables que forman el modelo
    - Árbol de decisión
    - Red Neuronal
    - Conjunto de coeficientes

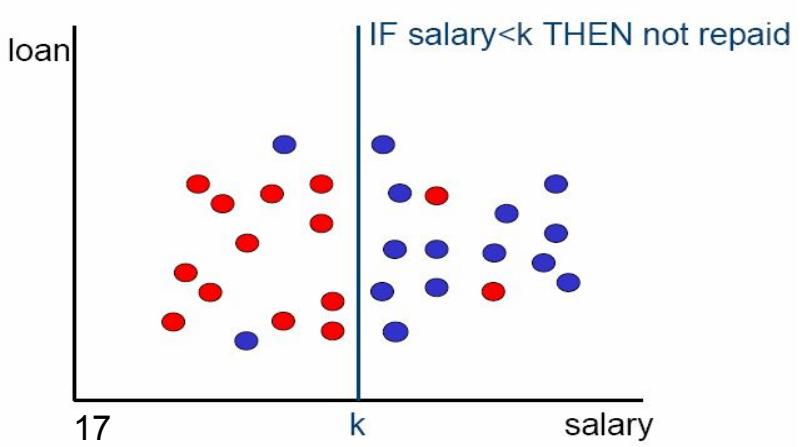


- http://machinelearningmastery.com/basic-concepts-in-machine-learning/

Modelos y Métodos

- Modelo = Método(Datos)
- Riesgo de un crédito
  - Modelo → Una regla: "If salary<k THEN not repaid"
  - Método → "Probar 1000 valores aleatorios de k y asignarlo al que mejor resultado dé"
  - Modelo (entrenado) → "If salary<k THEN not repaid"
    - Siendo k el valor obtenido por el método





Modelos y Métodos

- Modelos: existen multitud de ellos
  - Redes neuronales
  - Árboles de decisión
  - Áreas sobre los datos
  - Modelos probabilísticos
  - … cualquier cosa que, a partir de datos de entrada, nos pueda generar una salida
    - If salary<k THEN not repaid
    - If salary<k AND loan<m THEN not repaid
    - If $x_1 \cdot salary + x_2 \cdot loan < x_3$ THEN not repaid
    - If $x_1 \cdot salary^2 + x_2 \cdot salary + x_3 < x_4$ THEN not repaid

Modelos y Métodos

- Métodos: también existen multitud de ellos
  - Ajuste por mínimos cuadrados
  - Entrenamiento de redes neuronales
  - Métodos iterativos
  - … cualquier cosa que nos pueda concretar los valores de los parámetros del modelo
    - Probar 100 valores aleatorios y coger el mejor
    - Probar todos los posibles valores y coger la mejor
    - Configuración
    - Asignar valores a los parámetros en función de una
    - Fórmula, proceso, algoritmo, “magia”...

Modelos y Métodos

- Determinísticos
  - Para los mismos datos de entrada, y la misma configuración inicial, SIEMPRE obtendrán el mismo resultado
    - Probar k={5,10,15,20,25} miles de euros y devolver el mejor resultado
- Estocásticos
  - Para los mismos datos de entrada, y la misma configuración inicial, NO TIENEN POR QUÉ obtener el mismo resultado
    - Probar k={5 valores aleatorios entre 5 y 25} miles de euros y devolver el mejor resultado

El teorema del “No Free Lunch”

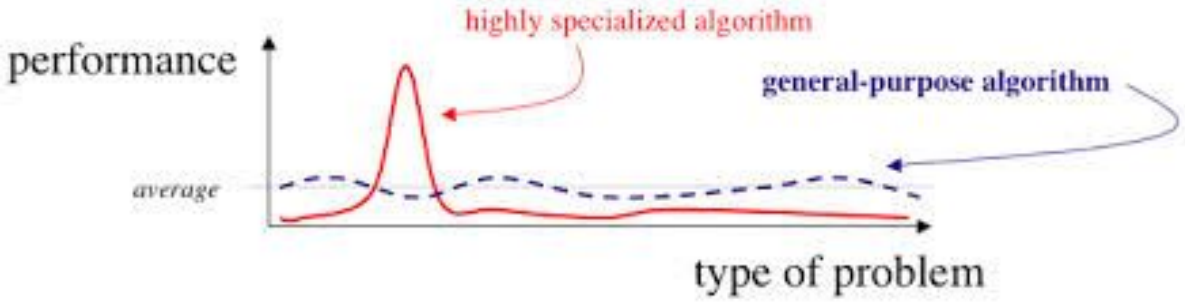
- Imagina que juntas en un estadio a los campeones del mundo de:
  - Tenis
  - Fútbol
  - Baloncesto
  - Esgrima
  - Ping-pong
  - …
- Compiten uno a uno en todos los deportes
  - Cada vitoria suma un punto
  - ¿Quién ganará?

El teorema del "No Free Lunch"

- Imagina que juntas en un ordenador a los mejores métodos-modelos para:
  - Predecir fugas de clientes del Banco 1,
  - Idem para Banco 2, Banco 3,...
  - Idem para Tienda de Barrio 1, Tienda de Barrio 2,...
  - Predecir el clima, terremotos, resultados deportivos,...
  - Segmentar clientes de Negocio 1, Negocio 2, Negocio 3,...
  - …
- Compiten uno a uno en todos los ámbitos
  - Cada vitoria suma un punto
  - ¿Quién ganará?

El teorema del "No Free Lunch"

- "No learning algorithm has an inherent superiority over other learning algorithms for all the problems"
  - "Any two algorithms are equivalent when their performance is averaged across all possible problems"
    - Interesante de leer la sección de "Interpretations of NFL results"
    - https://en.wikipedia.org/wiki/No_free_lunch_in_search_and_optimization



Tipos de Aprendizaje-Modelos

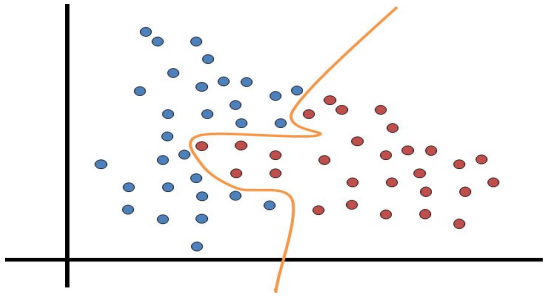
- Supervisados (Predictivos)
  - Cuando disponemos del valor que quisiéramos que nuestro modelo diera ante una determinada entrada
  - Tenemos datos "etiquetados"
    - Con la clase deseada o Con el valor esperado
  - Realizan predicciones del valor de salida a partir de datos

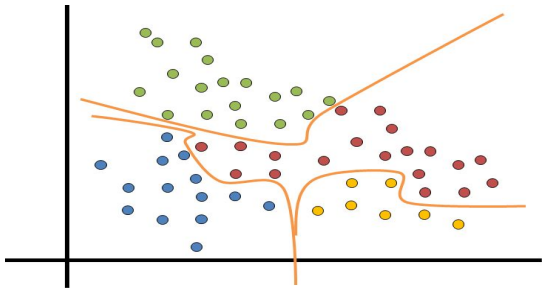
Deuda	Salario	Moroso
100.000	10.000	SI
110.000	30.000	NO
80.000	50.000	NO
90.000	45.000	NO

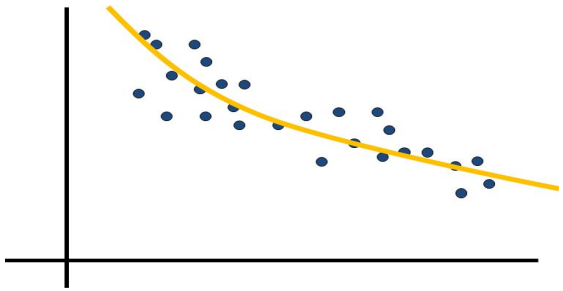
Salario	Edad	Préstamo
10.000	25	100.000
30.000	50	110.000
50.000	45	20.000
45.000	27	250.000

Tipos de Aprendizaje-Modelos

- Supervisados (Predictivos)
  - Clasificación: Cuando la variable a predecir es una categoría.
    - Binaria: {Sí, No}, {Azul, Rojo}, {Fuga, No Fuga}...
    - Múltiple: {Comprará Producto1, Producto2...}...
    - Ordenada: {Riesgo Bajo, Medio, Alto}...
  - Regresión: Cuando la variable a predecir es una cantidad
    - Precio, cantidad, tiempo,...

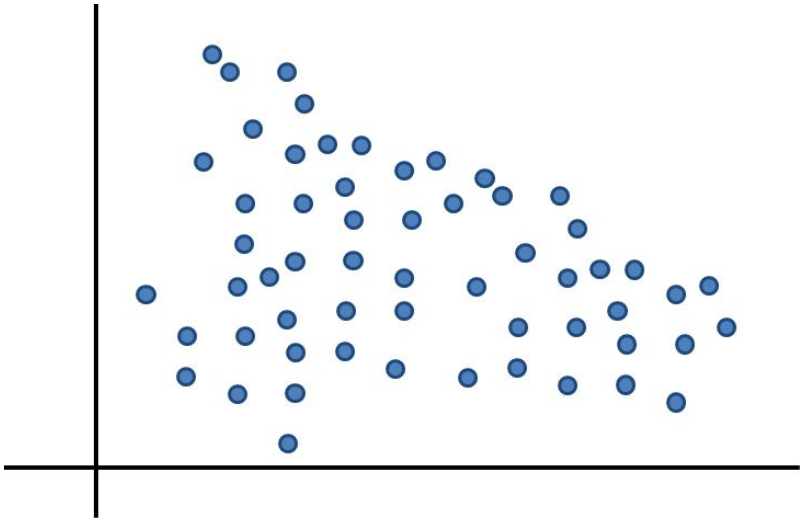






Tipos de Aprendizaje-Modelos

- No Supervisados (Descriptivos)
  - Cuando NO disponemos del valor que quisiéramos que nuestro modelo diera ante una determinada entrada
  - Su objetivo es modelar y describir la estructura o distribución interna de los datos
  - Muchas aplicaciones reales
  - hacen uso de estos datos
    - Hay más datos, y son baratos
    - Etiquetarlos puede ser costoso
    - Fáciles de obtener

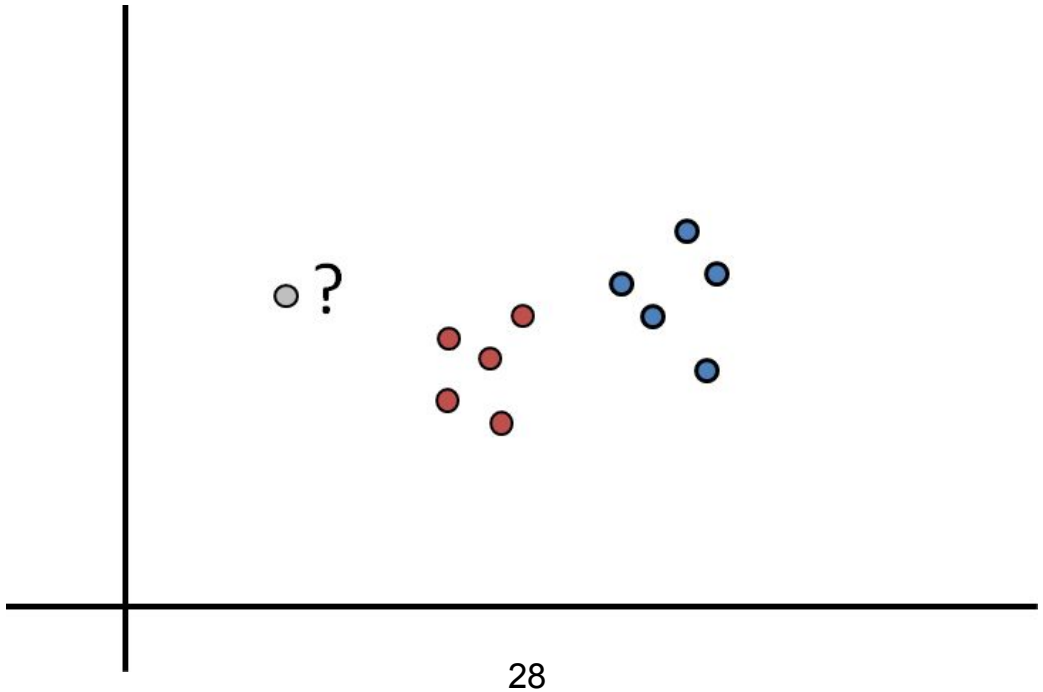


Tipos de Aprendizaje-Modelos

- No Supervisados (Descriptivos)
  - Agrupamiento - Clustering: Buscan encontrar grupos dentro de los datos de elementos similares
    - Clientes con hábitos de compra similares
    - Productos vendidos en fechas similares
  - Asociación: Buscan reglas que describen la mayor parte posible de los datos de los que se disponen
    - Productos que se compran juntos

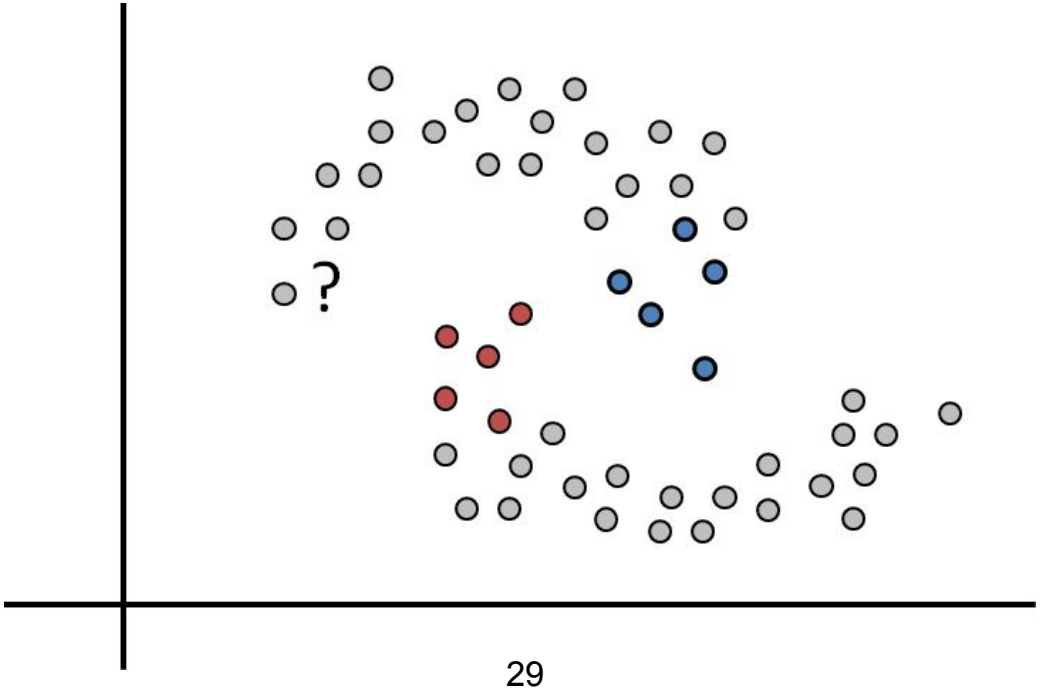
Tipos de Aprendizaje-Modelos

- Semi-Supervisados
  - Se dispone únicamente de datos etiquetados
  - ¿De qué color dirías que debería ser el punto gris?



Tipos de Aprendizaje-Modelos

- Semi-Supervisados
  - Se disponen de datos etiquetados y no etiquetados
  - ¿De qué color sería ahora el punto indicado?



Enrique Onieva
enrique.onieva@deusto.es
https://twitter.com/EnriqueOnieva
https://www.linkedin.com/in/enriqueonieva/