Aprendizaje automático

Departamento de Ingeniería del Software
e Inteligencia Artificial

Aprendizaje automático

- Introducción
- Aprendizaje no supervisado
  - Algoritmos de agrupamiento jerárquico
  - Algoritmos de agrupamiento basado en particiones
- Aprendizaje supervisado
  - k vecinos más cercanos
  - Árboles de decisión
  - Perceptrón multicapa

INTRODUCCIÓN

¿Qué entendemos por aprendizaje?

- No se puede hablar de inteligencia sin aprendizaje

- La **capacidad de aprendizaje** permite realizar nuevas tareas que previamente no podían realizarse, o bien realizar mejor (más rápidamente, con mayor exactitud, etc.) las que ya se realizaban, como resultado de los cambios producidos en el sistema al resolver problemas anteriores.

La capacidad de aprendizaje no puede añadirse a posteriori.

- Aprendizaje
  - "El acto, proceso o experiencia de adquirir conocimiento o aptitudes"

Aprendizaje Automático *(machine learning)*

- La **IA simbólica** trata con representaciones simbólicas de alto nivel (cercanas al entendimiento humano). Adecuada para:
  - Representar conocimiento humano y razonar con él
  - Resolver problemas bien-definidos o de índole lógica
- La **IA subsimbólica** es capaz de tratar con representaciones cercanas al problema y extraer conocimiento de ellas
- Existen aproximaciones de aprendizaje de conceptos simbólicas
  - Ej. Algoritmos de Winston y de Mitchell
- Sin embargo, el aprendizaje automático ha experimentado una explosión en su desarrollo desde la década de los 90 gracias a aproximaciones subsimbólicas que se benefician de
  - La abundancia de datos almacenados
  - El aumento en la capacidad de cálculo de los ordenadores

Aprendizaje Automático *(machine learning)* e IA subsimbólica

- En el aprendizaje automático, mediante algoritmos de aprendizaje se extraen “reglas” o “patrones” de los datos.
- Las aproximaciones de aprendizaje subsimbólico proporcionan
  - Mayor robustez frente al ruido
  - Mayor capacidad de escalamiento (nuevos ejemplos o nuevas variables)
  - Mayor capacidad para trabajar con cantidades ingentes de datos
  - Menor dependencia del experto
    - Éste puede ayudar a la selección de variables o formas de representación relevantes y a la interpretación de los resultados

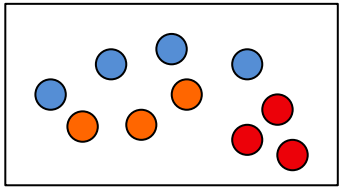
Tipos de Aprendizaje

- La IA subsimbólica está relacionada con la estadística
  - Sin embargo, la estadística requiere que se cumplan hipótesis y sigue una aproximación más “formal”
  - Mientras que la IA subsimbólica sigue una aproximación más “ingenieril”

- Aunque existen diferentes formas de clasificar el aprendizaje nos centraremos en el aprendizaje según el grado de realimentación:
  - Supervisado,
  - No supervisado
  - Por refuerzo

Tipos de aprendizaje: según el grado de realimentación

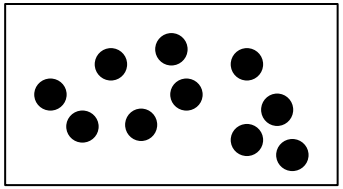
Aprendizaje supervisado



- Hay que suministrar al sistema ejemplos clasificados
- El objetivo es descubrir “reglas” de clasificación

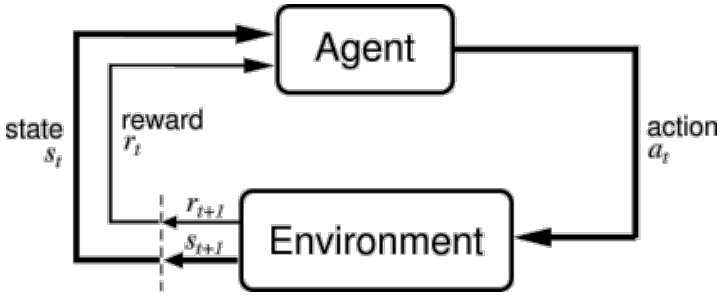
Aprendizaje no supervisado o por descubrimiento

- No hay información sobre la clase a la que pertenecen los ejemplos
- Objetivo: descubrir patrones en el conjunto de entrenamiento que permitan agrupar y diferenciar unos ejemplos de otros



Aprendizaje por refuerzo

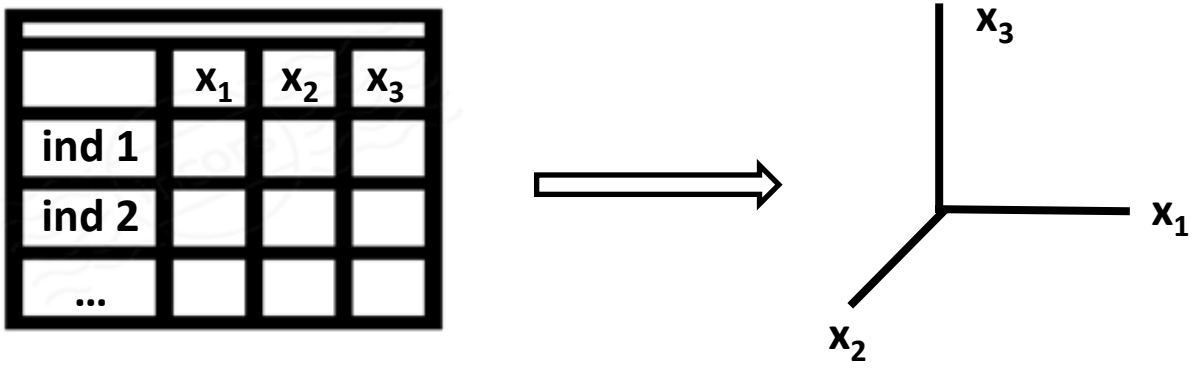
- El sistema recibe algún tipo de recompensa (positiva o negativa) cada vez que produce una respuesta, ajustando su comportamiento en función de dicha recompensa
- Típico sistema de aprendizaje robótico
- No lo veremos en esta asignatura



REPRESENTACIÓN DE LOS INDIVIDUOS

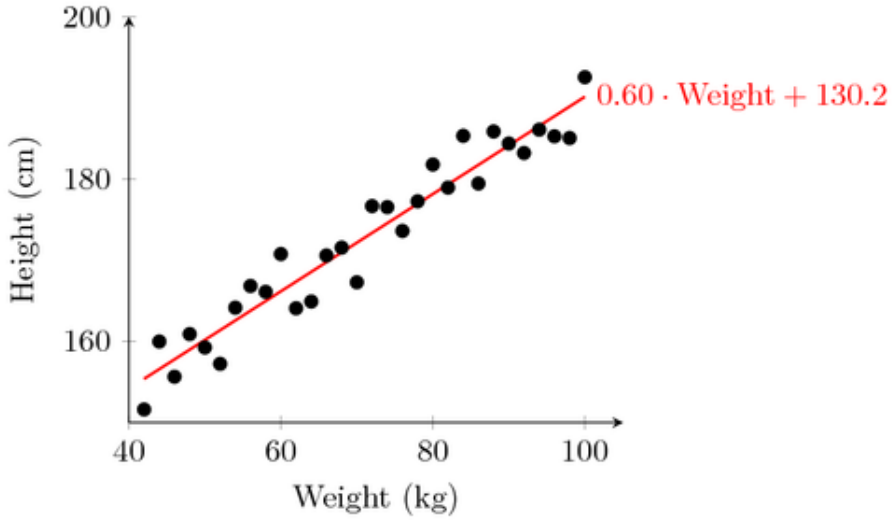
Representación de los individuos

- En el aprendizaje supervisado y no supervisado partimos de datos que son un conjunto de descripciones de objetos
- Tendremos un número de individuos o elementos (n) descrito por un conjunto de variables (m)
  - Nuestros datos se representan en forma de **matriz $n \times m$** donde las filas son los individuos y las columnas las variables
  - Las variables pueden ser de diferente naturaleza
    - Cuantitativa: variables reales o enteras
    - Cualitativa: variables categóricas u ordinales (p.ej. grupos de edad)
  - Las variables son las dimensiones en las que representamos a los individuos
    - P.ej. Si tenemos 3 variables cuantitativas nuestros individuos pueden considerarse como puntos en un espacio tridimensional



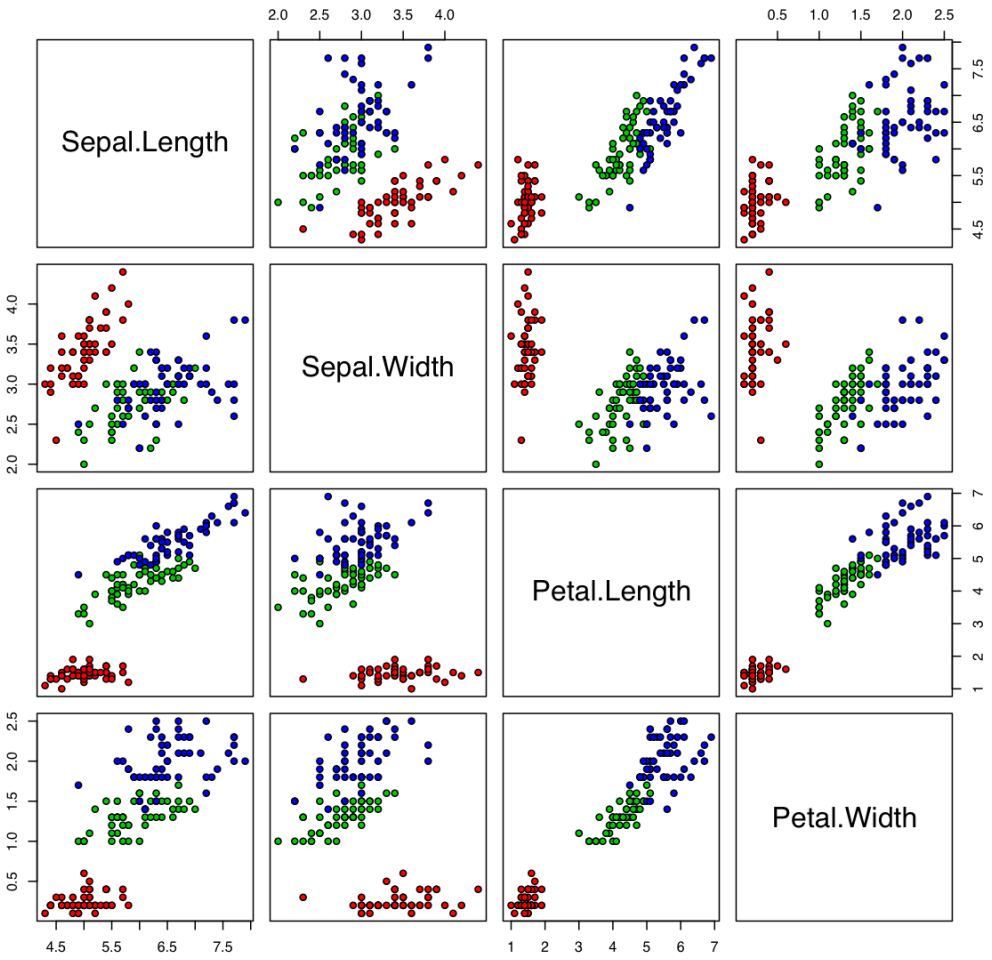
Representación de los individuos en el espacio

- Nuestros individuos son puntos representados en un espacio m-dimensional
  - De hecho, así los consideran muchos algoritmos de aprendizaje automático
- Suele ser de gran ayuda poder inspeccionar los datos visualmente, para ello se suele utilizar el diagrama de dispersión (*scatter plot* en inglés)
  - Se usa habitualmente para ajustar una regresión lineal



- Nosotros no podemos ver los datos en más de dos (o tres) dimensiones
  - Sin embargo, podemos mostrar las variables 2 a 2, si no son demasiadas

Ejemplo de diagramas de dispersión



Conjunto de datos de la flor del Iris

Es un conjunto clásico de aprendizaje estadístico

Variables de representación:

- Longitud del sépalo (cuantitativa)
- Ancho del sépalo (cuantitativa)
- Longitud del pétalo (cuantitativa)
- Ancho del pétalo (cuantitativa)
- Especie de Iris (categórica: setosa, virginica, versicolor)

Cada gráfico muestra las variables cuantitativas dos a dos y la especie usando el color.

El color es la “tercera” dimension en cada gráfico (**rojo** setosa, **azul** virginica y **verde** versicolor)

Importancia de la adecuada representación de los individuos

- Los algoritmos de Aprendizaje Automático son sensibles a la forma en que los individuos de nuestro problema están representados. Por ejemplo:
  - La representación puede contar con variables redundantes o no relevantes para el problema
    - Sin embargo, no siempre se puede determinar a priori cuáles son
  - Los datos se verán afectados por ruido, errores de medida, valores perdidos
- Los algoritmos de aprendizaje automático trabajan bien cuando el conjunto de datos tiene numerosos individuos que cubren bien la casuística que se puede dar en la vida real
  - Un algoritmo puede aprender sobre individuos “parecidos” a los que ya conoce, pero si hay “regiones” del espacio de representación sin individuos, de ellas es complicado aprender nada
    - Si cuando el algoritmo está en producción se le presenta un individuo que no se parece a los que empleó en su aprendizaje, su comportamiento será imprevisible!!!
  - Esto es más relevante cuanto mayor es el número de variables (como veremos más adelante).

Preparando los datos para el análisis

- Es una tarea crucial que afecta a los resultados, aunque no ahondaremos en ello. Incluye operaciones como:
  - Selección de variables relevantes para el problema, puede incluir
    - Transformación de variables iniciales
      - Cambiar la escala de representación de las variables (normalizar)
        - » Típicamente, las variables se escalan (se ponen en escala entre 0 y 1) o se estandarizan (se hace que tengan media 0 y varianza 1)

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad \textit{escalar}$$

$$x' = \frac{x - \bar{x}}{\sigma} \quad \textit{estandarizar}$$

¿Cuál uso? Depende del problema pero en general:
- valores pequeños y rango pequeño -> no hace falta
- distribución original normal -> estandarizar
- en otro caso -> escalar

        - » También se puede usar la transformada logarítmica de una variable para "concentrar" su rango,
        - » etc.
      - Discretizar una variable numérica, es decir, dividir su rango de valores posibles en categorías ordenadas
        - » Edad: [0 – 100] ➔ Bebé, Niño, Adolescente, Joven, Adulto, Anciano
      - Transformación de variables categóricas en variables binarias (una variable por categoría)
        - » Cuando las categorías no tienen orden

	Nacionalidad
id1	Español
id2	Frances

	Español	Francés
id1	1	0
id2	0	1

Preparando los datos para el análisis

- Combinación de variables iniciales
  - P. ej. Agrupando aquellas variables que están muy correlacionadas
- Eliminación de variables no relevantes o redundantes

– Tratamiento de valores perdidos. Es posible que algún valor de alguna variable para algún individuo no esté disponible en nuestra matriz
  - Si faltan pocos datos quizás podemos descartar esas filas
  - Otras opciones: asignar el valor medio en variables cuantitativas,0 o el valor modal (el más repetido) en variables cualitativas

	Edad	**Altura**	**Peso**	**Educación**
id1	20	175	80	Grado
id2	?	158	?	Secundaria
id3	28	166	65	Grado
id4	24	190	83	?
…	…	…	…	…

Entendiendo los datos

- Antes de acometer una tarea de aprendizaje automático conviene entender los datos lo mejor posible, para ello suele ser recomendable "trabajarlos" previamente
  - Visualización de datos:
    - Permite determinar qué variables están más relacionadas entre sí y cuál es la naturaleza de la relación (lineal, exponencial, etc)
    - En los problemas de clasificación permite ver cómo de bien se separan las clases y qué variables separan mejor
  - Calcular estadísticos descriptivos:
    - Tendencia central: media, mediana, moda
    - Dispersión: desviación típica, valores mínimos y máximos, percentiles
    - Bivariantes: coeficiente de correlación o tablas de contingencia
  - Representar la distribución de las variables
    - Observar valores más frecuentes y valores extremos (¿tienen sentido o son aberrantes?)
  - Usar técnicas de **aprendizaje no supervisado** para encontrar *estructura* en los datos