Jesús García Herrero

METODOLOGÍA DE ANÁLISIS DE DATOS

En esta clase concluimos el curso de Análisis de Datos con una visión de las metodologías del análisis de datos. Como se ha visto, este es un campo creciente, y por tanto hay muchas metodologías del descubrimiento del conocimiento en uso y bajo desarrollo. Algunas de estas técnicas son genéricas, mientras otros son de dominio específico.

Se destaca que, en general, un proyecto de KDD involucra un proceso iterativo e interactivo de búsqueda de modelos, patrones o parámetros. Los patrones descubiertos han de ser válidos, novedosos para el sistema (para el usuario siempre que sea posible) y potencialmente útiles. Por tanto es preciso evaluar la validez, utilidad y simplicidad de los patrones obtenidos mediante alguna de las técnicas de Minería de Datos. Debemos tener en cuenta que el objetivo final es incorporar el conocimiento obtenido en algún sistema real, tomar decisiones a partir de los resultados alcanzados o, simplemente, suministrar la información alcanzada a quien esté interesado.

El primer paso es la identificación de los datos. Para ello hay que imaginar qué datos se necesitan, dónde se pueden encontrar y cómo conseguirlos. Una vez que se dispone de datos, se deben seleccionar aquellos que sean útiles para los objetivos propuestos. Se preparan, poniéndolos en un formato adecuado. Una vez se tienen los datos adecuados se procede a la minería de datos, proceso en el que se seleccionarán las herramientas y técnicas adecuadas para lograr los objetivos pretendidos. Y tras este proceso llega el análisis de resultados, con lo que se obtiene el conocimiento pretendido. Se identifican las siguientes etapas:

- Comprensión del dominio de la aplicación, del conocimiento relevante y de los objetivos del usuario final.

- Creación del conjunto de datos: consiste en la selección del conjunto de datos, o del subconjunto de variables o muestra de datos, sobre los cuales se va a realizar el descubrimiento.

- Limpieza y preprocesamiento de los datos: Se compone de las operaciones, tales como: recolección de la información necesaria sobre la cual se va a realizar el proceso, decidir las estrategias sobre la forma en que se van a manejar los campos de los datos no disponibles, estimación del tiempo de la información y sus posibles cambios.

- Reducción de los datos y proyección: Encontrar las características más significativas para representar los datos, dependiendo del objetivo del proceso. En este paso se pueden utilizar métodos de transformación para reducir el número efectivo de variables a ser consideradas o para encontrar otras representaciones de los datos.

- Elegir la tarea de Minería de Datos: Decidir si el objetivo del proceso de KDD es: Regresión, Clasificación, Agrupamiento, etc.

- Elección del algoritmo(s) de Minería de Datos: Selección del método(s) a ser utilizado para buscar los patrones en los datos. Incluye además la decisión sobre que modelos y parámetros pueden ser los más apropiados.

- Minería de Datos: Consiste en la búsqueda de los patrones de interés en una determinada forma de representación o sobre un conjunto de representaciones, utilizando para ello métodos de clasificación, reglas o árboles, regresión, agrupación, etc.

- Interpretación de los patrones encontrados. Dependiendo de los resultados, a veces se hace necesario regresar a uno de los pasos anteriores.

- Consolidación del conocimiento descubierto: consiste en la incorporación de este conocimiento al funcionamiento del sistema, o simplemente documentación e información a las partes interesadas.

El proceso de KDD puede involucrar varias iteraciones y puede contener ciclos entre dos de cualquiera de los pasos. La mayoría de los trabajos que se han realizado sobre KDD se centran en la etapa de minería. Sin embargo, los otros pasos se consideran importantes para el éxito del KDD. Gran parte del esfuerzo del proceso de KDD recae sobre la fase de preparación de los datos, fase crucial para tener éxito, que se destaca en el proceso con las operaciones más habituales.

Finalmente se apuntan otras técnicas de interés no cubiertas en este curso, que permitirán complementar el proceso de búsqueda y generalización con paradigmas de la computación que comparten el interés en esta área: razonamiento basado en incertidumbre, técnicas meta-heurísticas de búsqueda y optimización, y algoritmos avanzados de aprendizaje (redes de neuronas, máquinas de vectores de soporte), disciplinas que entran dentro de la inteligencia computacional (soft computing).

Metodología de Análisis de Datos

Jesús García Herrero

Universidad Carlos III de Madrid

Universidad Carlos III de Madrid

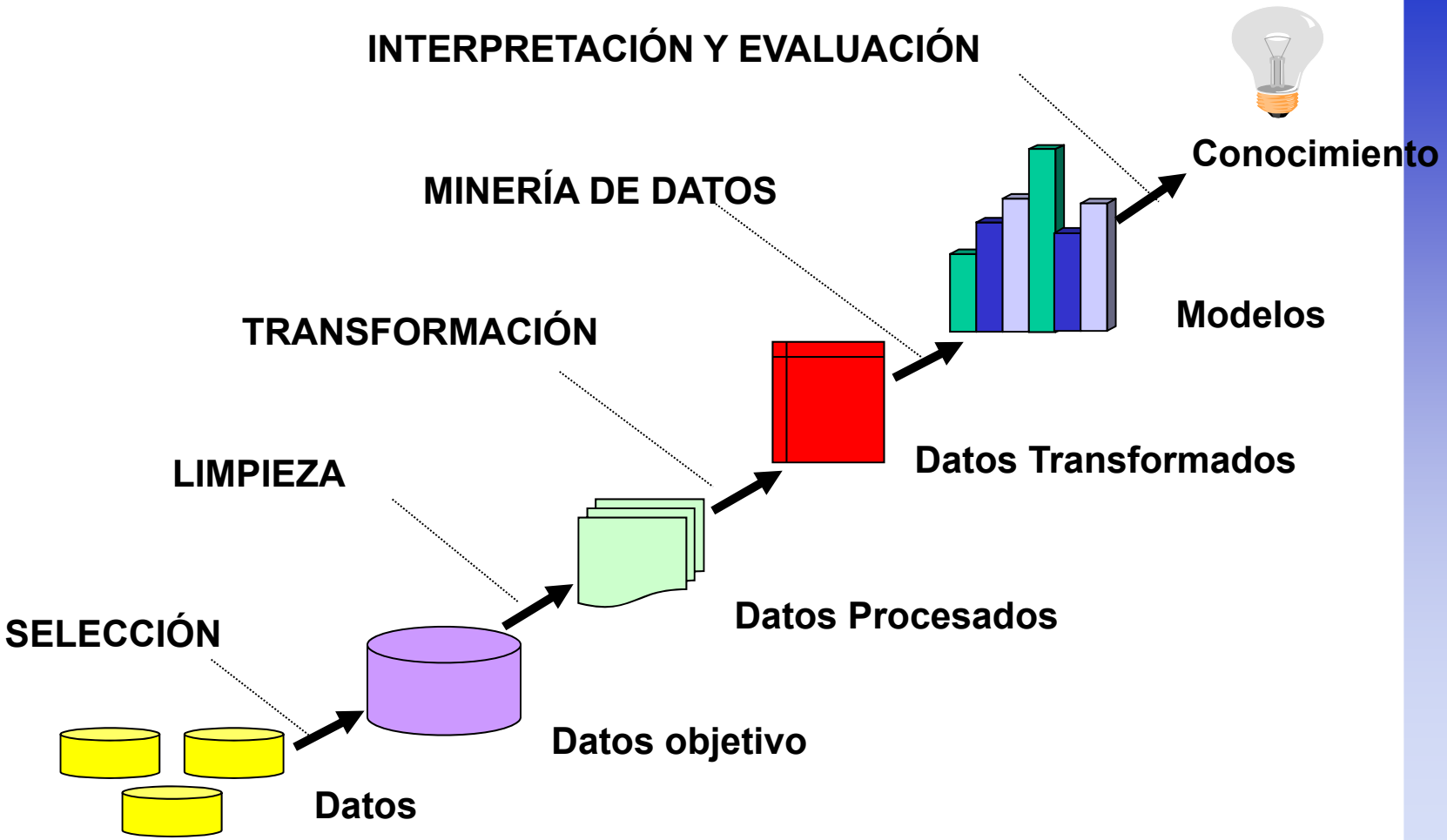
Septiembre 2009

ENFOQUES DE ANÁLISIS DE DATOS

- **Evaluación de Hipótesis ("*Top-Down*"). Técnicas estadísticas**
  - Propuesta de hipótesis
  - Determinar y recolectar datos necesarios para análisis
  - Evaluación de hipótesis para aceptar o rechazar, basadas en datos

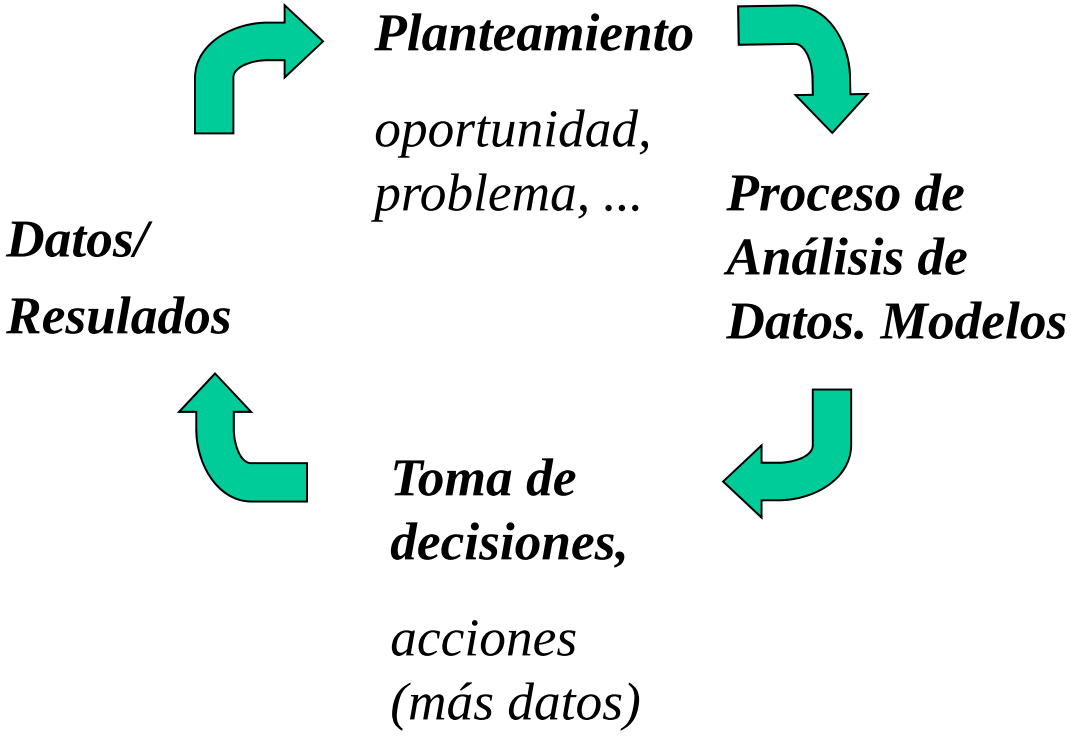
- **Descubrimento de conocimiento ("*Bottom-up*"). Técnicas de aprendizaje**
  - **Preparar datos disponibles para su exploración**
  - **Aprendizaje supervisado/dirigido**
    - Explicar un atributo particular (clasificación, predicción, ….)
  - **Aprendizaje no supervisado**
    - Buscar patrones significativos (agrupamiento, asociación)

El Proceso de KDD



El Proceso de KDD

- Contexto de un aplicación con Análisis de Datos.
  - Proceso interactivo e iterativo. Ensayo y error



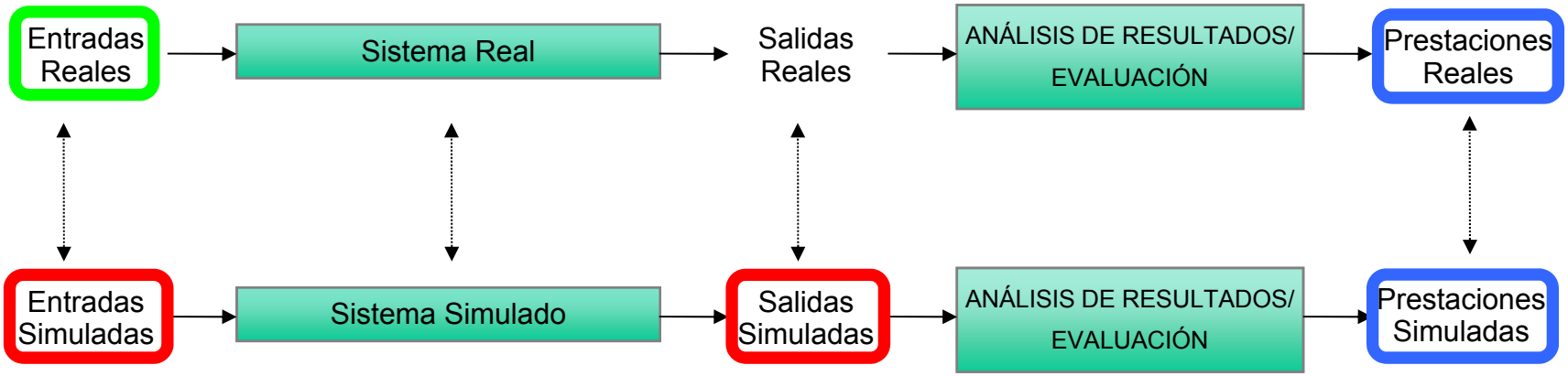
METODOLOGÍA

1. Formular el problema
2. Determinar la representación (atributos y clases)
   - directamente
   - hablando con expertos
   - a partir de otras técnicas (filtros)
3. Identificar y recolectar datos de entrenamiento (bases de datos, ficheros, …)
4. Preparar datos para análisis
5. Selección de modelo, construcción y entrenamiento
6. Evaluar lo aprendido
   - validación cruzada, expertos
7. Integrar la base de conocimiento a la espera de nuevos datos tras acciones

METODOLOGÍA NO SUPERVISADO

1. Formular el problema
   - búsqueda de afinidad, grupos, etc.
2. Representación
3. Identificar y recolectar datos de entrenamiento (bases de datos)
4. Preparar datos para análisis
5. Selección de modelo y construcción
   - agrupamiento de clientes (sin tener en cuenta la clase todavía)
6. Utilizar las estructuras encontradas para aplicar ap. Supervisado
   - predicción de abandono en cada grupo
7. Generar nuevas hipótesis a evaluar
   - características de grupos especiales. Búsqueda de más datos

MINERÍA DE DATOS EN SIMULACIÓN



Evaluación de relaciones entre diferentes tipos de variables

Búsqueda de patrones y relaciones significativas y útiles

- **Entradas reales** -> mejorar modelo
- **Entradas/salidas simulados** ->comprensión de resultados
- **Prestaciones** -> validación de resultados

SELECCIÓN DE DATOS DE ENTRADA

- Selección de datos
  - Aleatoriamente: conjuntos grandes. Verificación
  - Aquellos que se parecen más entre sí
  - Aquellos que se diferencian mas entre sí
  - Los datos que están en las fronteras entre las clases
  - Los datos que tienen mayores errores de clasificación se tratan (proporcionalmente) más veces
    - Boosting
  - Incremental: incorporar sucesivamente datos de un conjunto reserva
- Pre-procesamiento
  - Reducción del ruido (filtrado de datos)
  - Selección de atributos
  - Tratamiento de los valores desconocidos, discretización de valores numéricos

SELECCIÓN DEL MÉTODO

- Necesidades deligibilidad deseada
- Naturaleza de datos (estadísticos, linealmente separables, …)
- Volumen de datos
  - Cuando no caben en memoria principal se requiere esquema incremental
  - Cuando el tamaño es muy grande
    - Selección de datos
    - Mejor si los esquemas son escalables (crecen linealmente o casi).
    - Mejor si son paralelizables
- Incorporar explícitamente conocimiento del dominio
  - reglas ya conocidas a extender (FOIL)

FILTRADO DE ATRIBUTOS

- Los errores en los datos son muy comunes y pueden degradar fuertemente el análisis
- Se pueden aplicar técnicas que permitan identificar potenciales problemas, evitando o agilizando la supervisión manual.

  * **Mejora de árboles de decisión**
    - El ruido en los atributos debe incorporarse también en el entrenamiento para aprender a combatirlo
    - Descartar los ejemplos mal clasificados (y re-entrenar) frecuentemente reduce la complejidad de la estructura, con diferencias no significativas de prestaciones
      - Equivale a un proceso de poda global

  * **Regresión robusta**
    - Eliminar ejemplos separados más de 3σ
    - Estimadores de mínimo error absoluto o de mínima mediana de error cuadrático

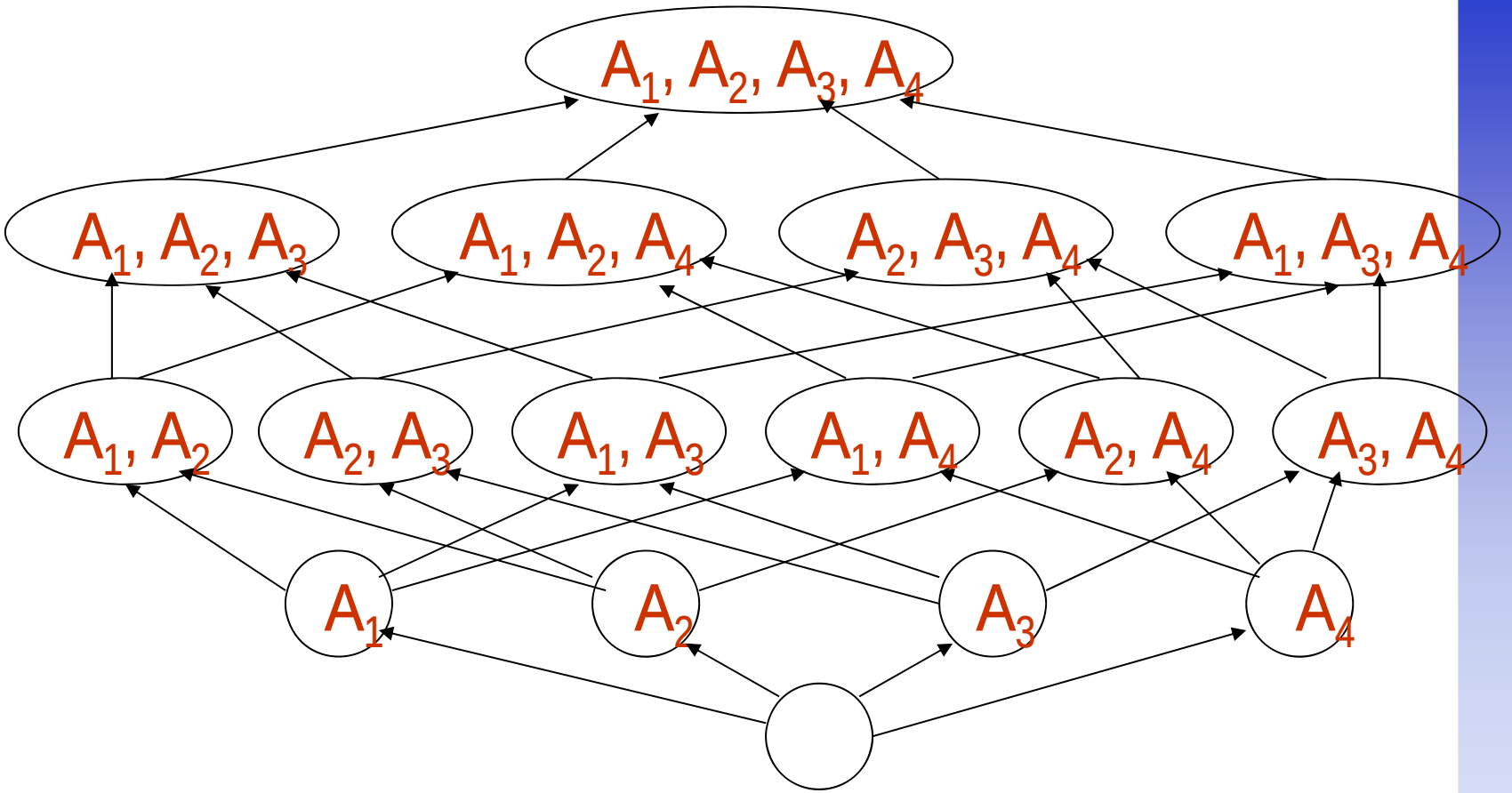
BUSQUEDA DE ATRIBUTOS

- Espacio de búsqueda: subconjuntos posibles de los atributos
  - Con F atributos hay 2^F grupos posibles
- Una exploración exhaustiva no es factible con atributos numerosos (>30)
- Se puede comenzar por
  - conjunto de atributos de entrada completo (*backward elimination*)
  - conjunto vaco de atributos (*forward selection*)
- Se puede realizar búsqueda
  - en escalada (*greedy*): mueve 1 atributo cada vez. encuentra óptimo local
  - mejor-primero: mantiene todas las ramas y puede hacer *backtracking*. Es exhaustivo si no se para
- La evaluación de cada nodo (subconjunto de atributos) se realiza llamando al algoritmo seleccionado (*wrapper*) o independientemente

EVALUAR CONJUNTOS DE ATRIBUTOS

- Técnicas independientes: filtro previo al aprendizaje
  - Técnicas estadísticas: máxima correlación con clase y mínima entre atributos
  - Máxima separación entre clases
    - Otros clasificadores: árboles, 1R
- Técnicas asociadas al propio proceso de aprendizaje: técnicas wrapper
  - Evaluación mediante validación cruzada, o con conjunto independiente

Ejemplo búsqueda



Ejemplo	Sitio de acceso: A_1	1ª cantidad gastada: A_2	Vivienda: A_3	Última compra: A_4	Clase
1	1	0	2	Libro	Bueno
2	1	0	1	Disco	Malo

Añadir nuevos atributos

Scheme: j48.J48 -C 0.25 -M 2

Attributes: 3

X, Y, CLASE

Number of Leaves : 34

Size of the tree : 67

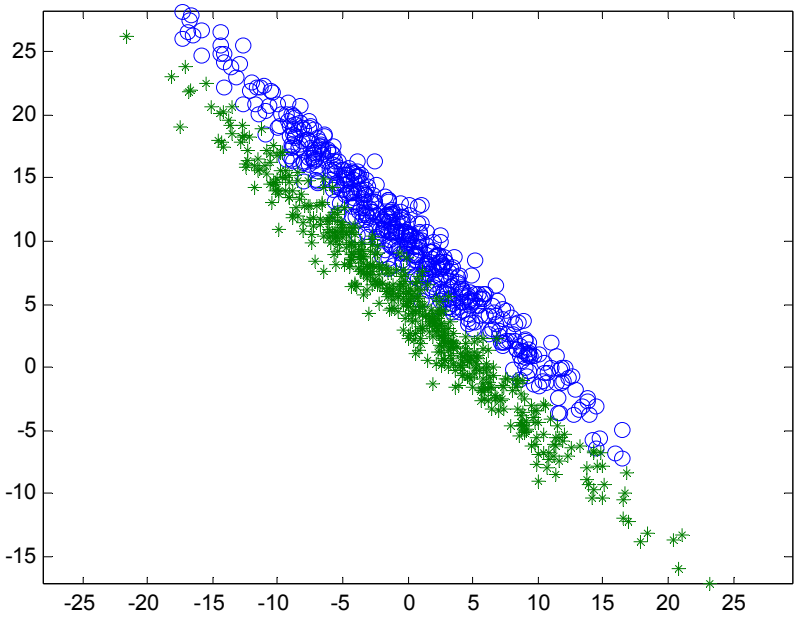
Scheme: j48.J48 -C 0.25 -M 2

Attributes: 4

X, Y, SUMA, CLASE

Number of Leaves : 2

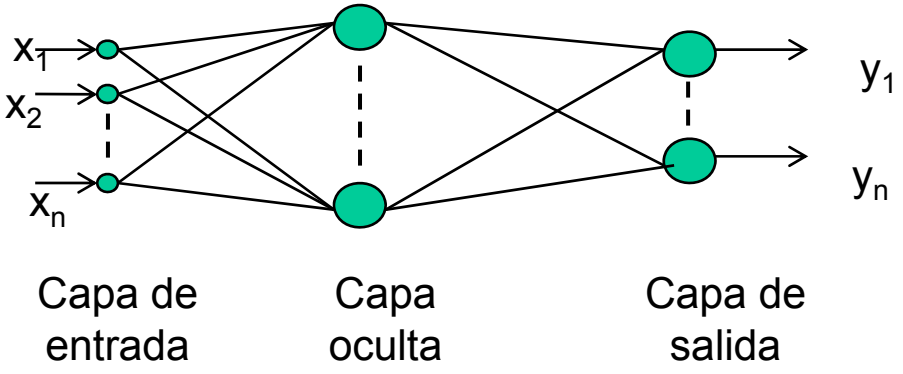
Size of the tree : 3

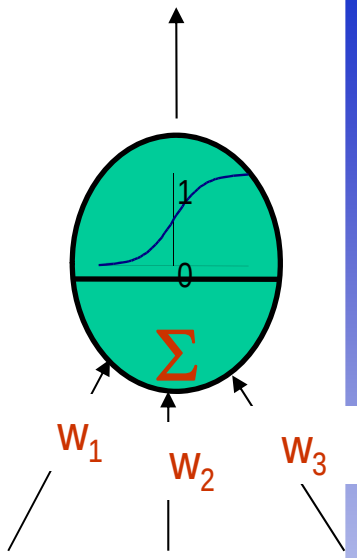


Dataset	NaiveBayes	j48.J48
XYsuma	96.7	96.72
XY	93.12	92.96

Otras técnicas aplicadas al análisis de datos

- Redes de neuronas artificiales:
  - Predicción/Clasificación, Clustering

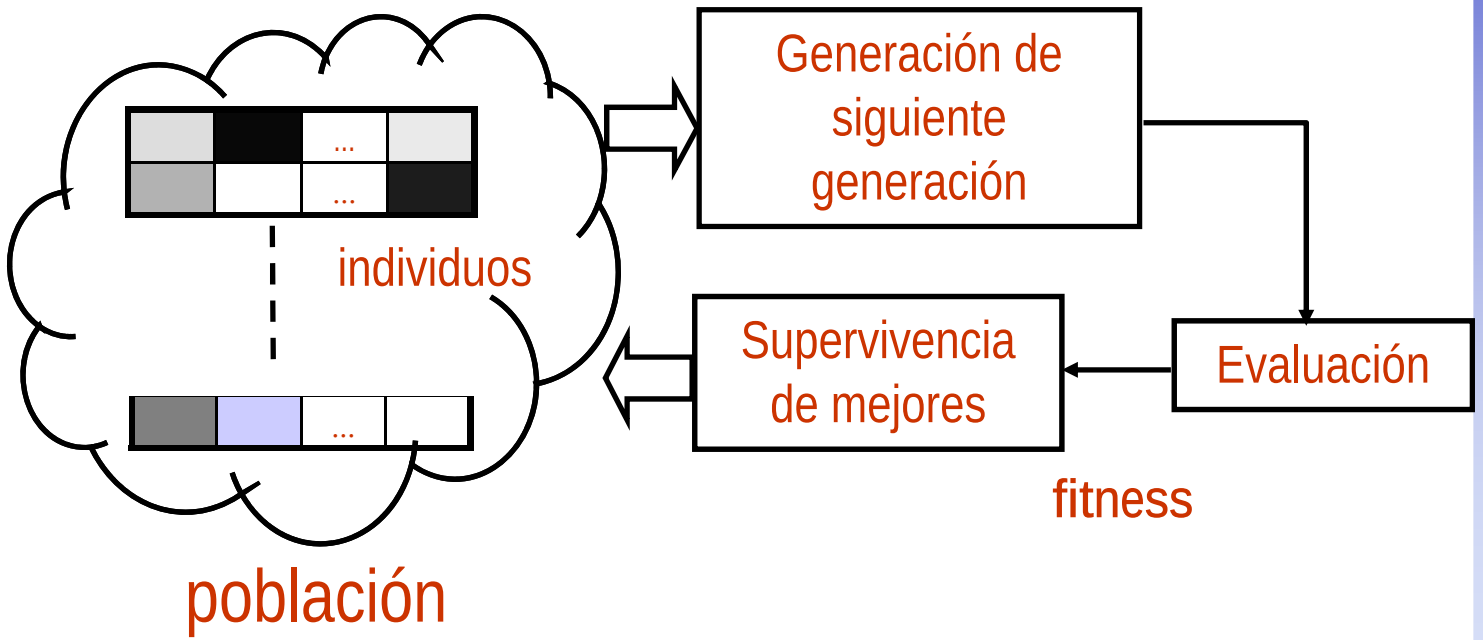




  - Interpolación de funciones no lineales con algoritmos de aprendizaje
  - Preperación de datos: entradas numéricas normalizadas
- SVM (*Máquinas de vectores de soporte)*
  - Predicción/Clasificación
  - Transformación del espacio de entrada en un espacio linealmente separable

Otras técnicas aplicadas al análisis de datos

- Algoritmos genéticos
  - Clasificación como búsqueda (optimización)
  - Selección de atributos, ajuste, ...
  - Meta-aprendizaje: *stacking*



Otras técnicas aplicadas al análisis de datos

- Lógica borrosa
  - Clasificación/predicción, agrupamiento