VNIVERSITAT DE VALÈNCIA OpenCourseWare

Procesado y Análisis de Datos Ambientales.
Curso 2009-2010.
José D. Martín, Emilio Soria, Antonio J. Serrano

TEMA 1: INTRODUCCIÓN AL PROCESADO Y ANÁLISIS DE DATOS

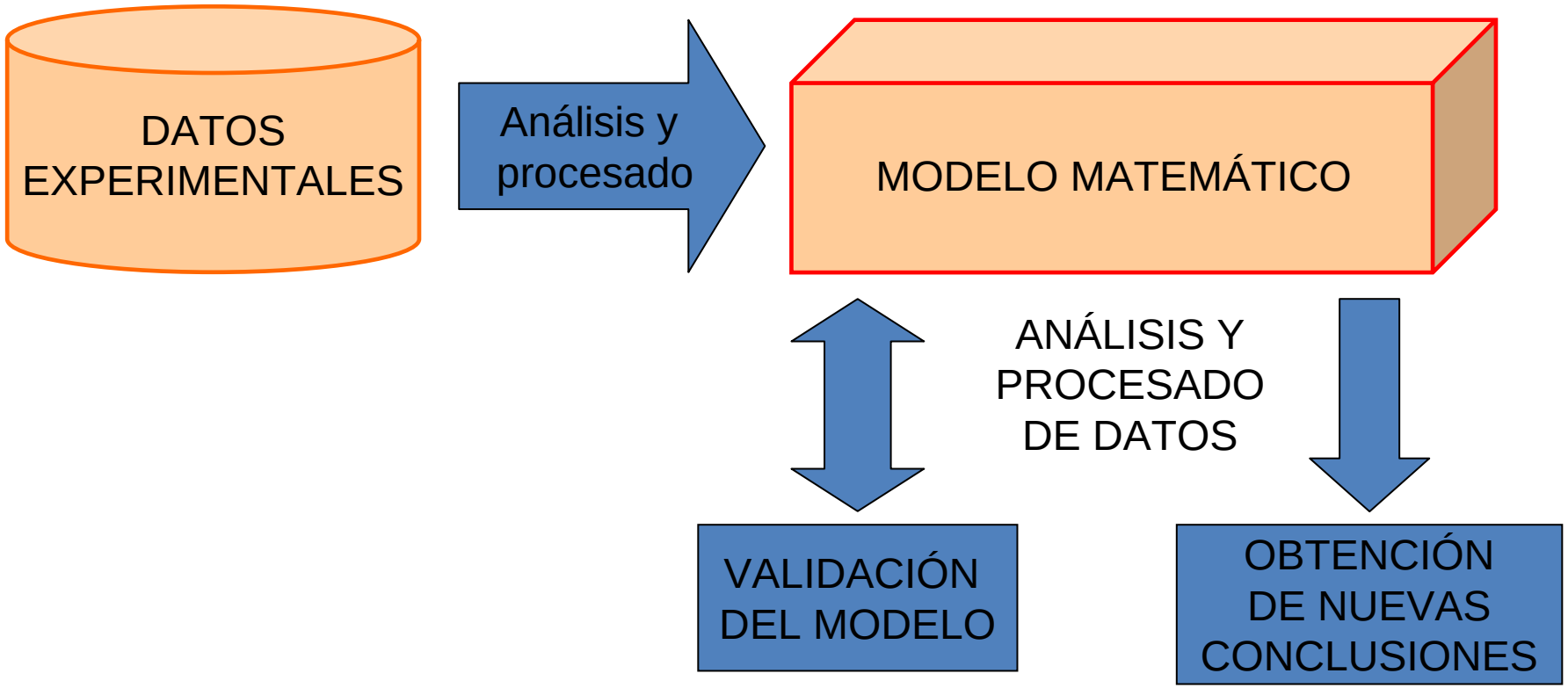
ÍNDICE

- Introducción.
- Selección de variables.
- Preprocesado.
- Clases de modelos
- Generalización y sobreajuste.
- Extracción de conocimiento.

ÍNDICE

INTRODUCCIÓN

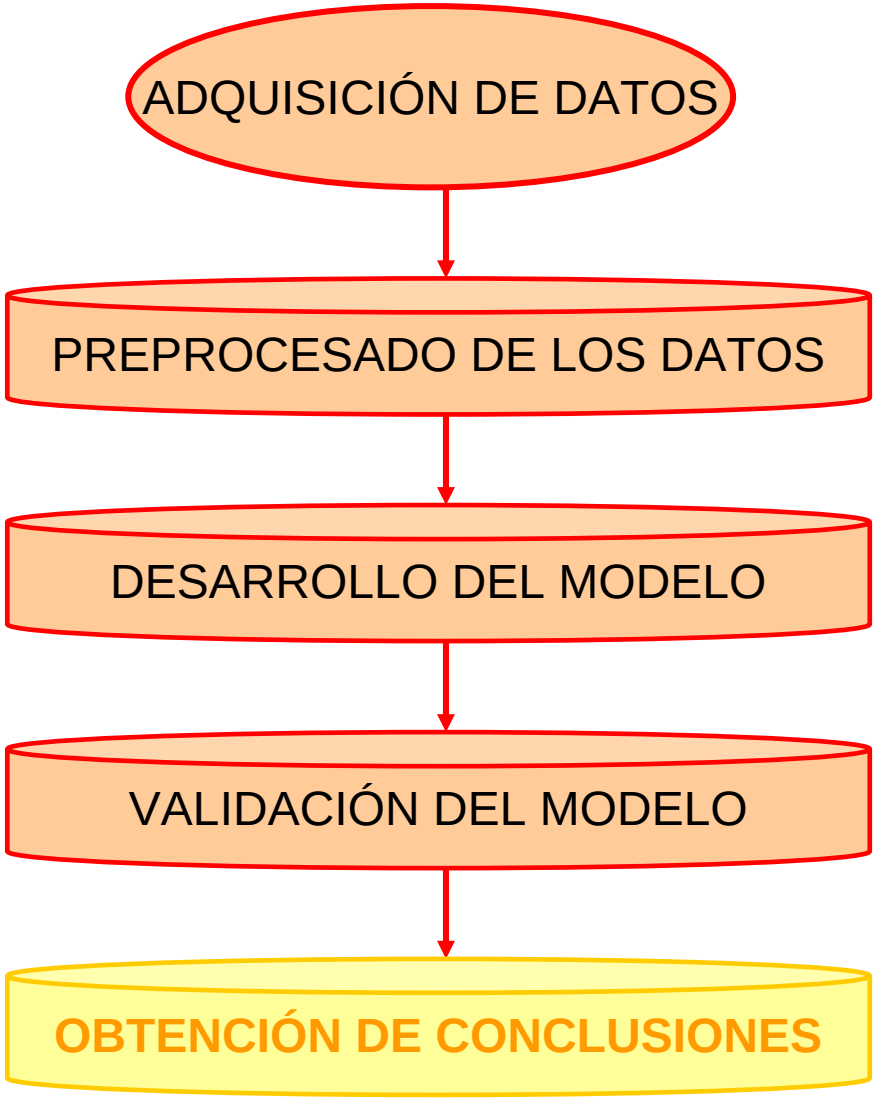
Gran cantidad de datos: Generación de conocimiento.



Tipos de problemas a resolver

- Clasificación.
- Modelado.
- Predicción.
- Agrupamiento.
- Estimación de densidades de probabilidad.

Pasos a seguir



ÍNDICE

- Introducción.
- Selección de variables.
- Preprocesado.
- Clases de modelos
- Generalización y sobreajuste.
- Extracción de conocimiento.

SELECCIÓN DE VARIABLES

- Relación entre el número de parámetros y el de patrones.
- El número de entradas al modelo afecta a la complejidad de los modelos.
- Las entradas no necesarias acaban siendo “ruido”.
- Extracción de conocimiento.

ÍNDICE

- Introducción.
- Selección de variables.
- **Preprocesado.**
- Clases de modelos
- Generalización y sobreajuste.
- Extracción de conocimiento.

PREPROCESADO DE LOS DATOS

- Preparación de los datos.
- Análisis exploratorio de los datos.
- Reducción de la dimensionalidad.
- Filtrado de los datos.

Preparación de los datos

1. Eliminación / Interpolación de datos incompletos.

2. Codificación de los datos.

3. Normalización:

$$y_k = \frac{x_k - \overline{x_k}}{\sigma_k}$$

Media cero y desviación estándar unidad

$$y_k = a \cdot \frac{1 - e^{-\beta \cdot x_k}}{1 + e^{-\beta \cdot x_k}}$$

Reducción de rango

$$y_k = \left(\frac{x_k - m_x}{M_x - m_x} \right) \cdot \left(M_y - m_y \right) + m_y$$

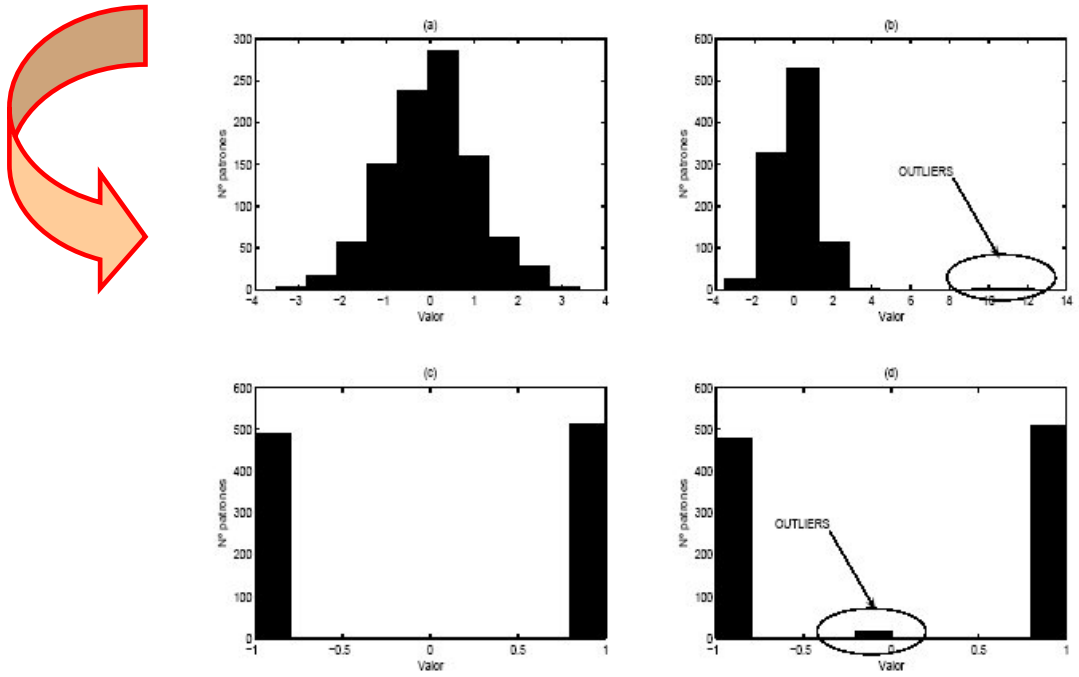
Transformación de rango

Análisis exploratorio de los datos

Distribución de probabilidad conocida ⇒ **Tests estadísticos**

Distribución de probabilidad desconocida:

- **Parámetros estadísticos.**
- **Detección de *outliers*.**



REPRESENTACIONES ÚTILES

Histogramas.

Diagramas de dispersión.

Agrupamiento (*clustering*).

Gráfico de probabilidad normal.

Autocorrelación.

Correlación cruzada.

Reducción de la dimensionalidad

1. Selección de características.
2. Extracción de características.

Problemas de clasificación

Mín. distancia intraclase y máx. interclase

MATRICES DE COVARIANZA

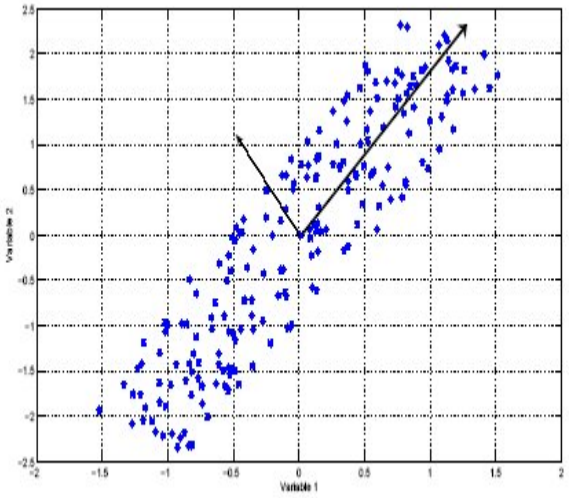
ANÁLISIS DE COMPONENTES PRINCIPALES (PCA)

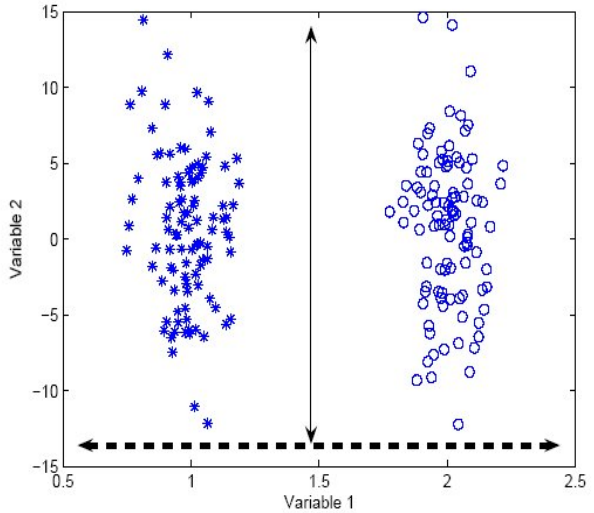
ANÁLISIS DISCRIMINANTE LINEAL (LDA)

TÉCNICAS GEOMÉTRICAS

ANÁLISIS DE FOURIER

…





Filtrado de los datos

- **Eliminar interferencias del proceso de medida. Ej: ruido de 50 Hz en señales eléctricas.**

- **Continua realimentación de todo el preprocesado de datos (por ej., nuevos *outliers* debido a la reducción de la dimensionalidad) hasta llegar a una convergencia del proceso.**

ÍNDICE

- Introducción.
- Selección de variables.
- Preprocesado.
- Clases de modelos
- Generalización y sobreajuste.
- Extracción de conocimiento.

CLASES DE MODELOS

Modelos lineales y no lineales:

- Complejidad / Interpretabilidad.
- Plasticidad / Estabilidad.
- Precisión / Generalización.
- Complejidad en la obtención de los parámetros.

$$m_1 \Rightarrow y = w_1 \cdot x_1 + ... + w_N \cdot x_N = \sum_{k=1}^{N} w_k \cdot x_k$$ → Lineal en parámetros y variables de entrada

$$m_2 \Rightarrow y = w_1 \cdot \varphi_1(x_1) + .. + w_N \cdot \varphi_N(x_N) = \sum_{k=1}^{N} w_k \cdot \varphi_k(x_k)$$ → Lineal solamente en parámetros

$$m_3 \Rightarrow y = \varphi[w_1 \cdot \varphi_1(x_1) + .. + w_N \cdot \varphi_N(x_N)] = \varphi\left[\sum_{k=1}^{N} w_k \cdot x_k\right]$$ → No lineal

- Capacidad de modelado.

Modelos paramétricos y no paramétricos

Modelos paramétricos:

- Modelo conocido.
- Ajuste a un polinomio de un cierto grado.

Modelos no paramétricos:

Los datos definen el modelo: árboles de decisión, histogramas, etc. Se utilizan ante un total desconocimiento del problema abordado, con muchos datos o con pocas variables de entrada.

Modelos semiparamétricos:

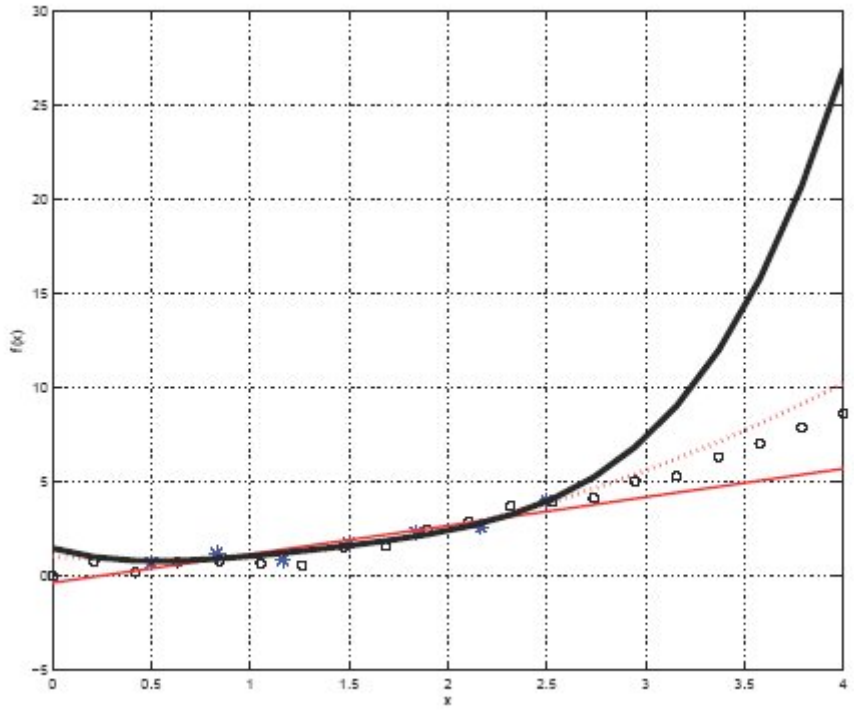
El modelo tiene una forma funcional que no es necesario definir de manera estricta.

ÍNDICE

- Introducción.
- Selección de variables.
- Preprocesado.
- Clases de modelos
- Generalización y sobreajuste.
- Extracción de conocimiento.

GENERALIZACIÓN Y SOBREAJUSTE (I)

SOBREAJUSTE (*OVERFITTING*). CONTROL DEL NÚMERO DE PARÁMETROS



Ajuste polinómico de mayor grado se ajusta mejor a los datos

¡¡ Un polinomio de grado 7 generalizaría mal!!

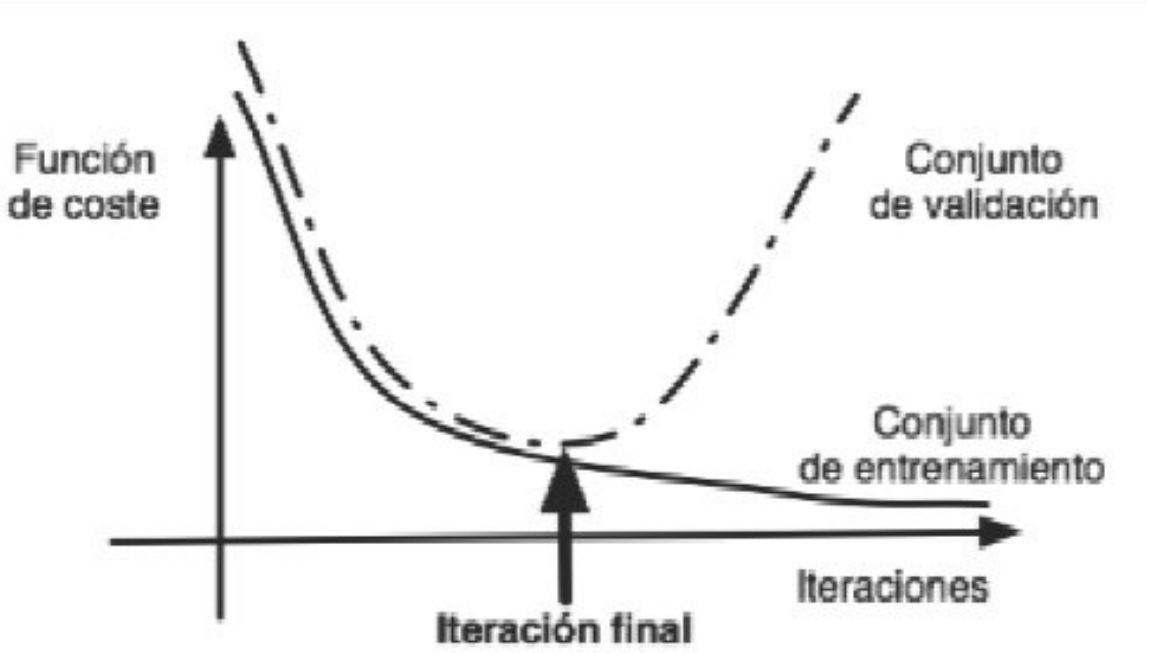
¡¡ Como hay 7 puntos un polinomio de grado 7 se ajustaría perfectamente!!

¡¡ Mejor solución el polinomio de grado 2!!

GENERALIZACIÓN Y SOBREAJUSTE (II)

SOBREENTRENAMIENTO (*OVERTRAINING*). CONTROL DEL NÚMERO DE ITERACIONES

Utilización de un conjunto de datos de generalización

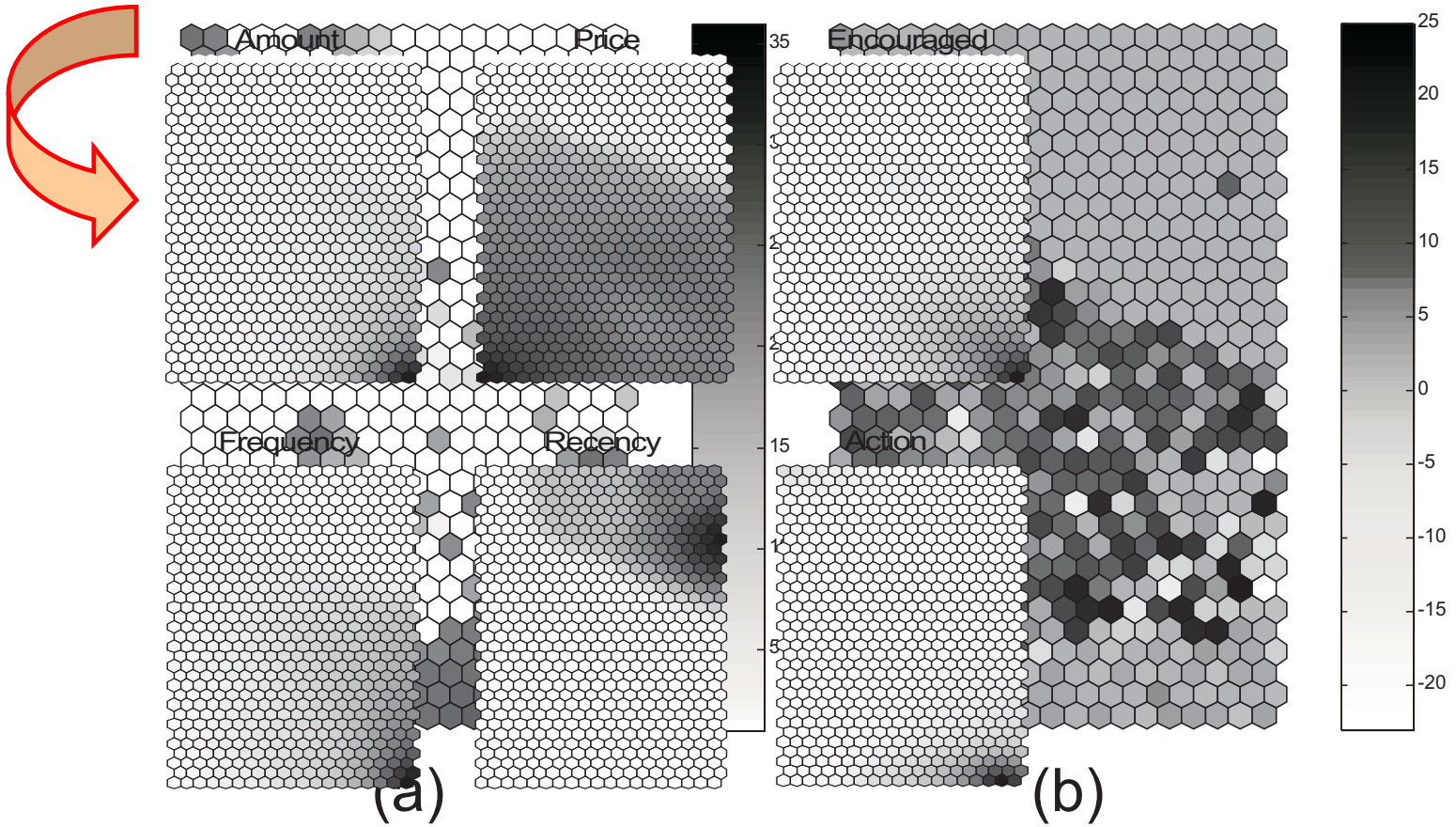


ÍNDICE

- Introducción.
- Selección de variables.
- Preprocesado.
- Clases de modelos
- Generalización y sobreajuste.
- Extracción de conocimiento.

EXTRACCIÓN DE CONOCIMIENTO

- Análisis de relevancia de las entradas.
- Obtención de reglas.
- Representación del mapeo entrada-salida.



RESUMEN

- Proceso realimentado.
- Número de entradas al modelo y EDA juegan un papel muy importante.
- Es necesario comprobar capacidad de generalización.
- Buen ajuste no es sinónimo de buen modelo.
- La elección del modelo debe ser adecuada a la complejidad del problema.