Resumen Data Mining

Por: Pablo Federico Martín Luna

# Tema 1: Introducción al Data Mining

## ¿Qué es Data Mining?

La minería de datos es el proceso de extraer patrones y conocimiento de grandes conjuntos de datos, utilizando técnicas automatizadas y algoritmos sofisticados.

## Knowledge Discovery in Databases (KDD)

Proceso de descubrir patrones y relaciones interesantes en grandes conjuntos de datos almacenados en una base de datos. Se utilizan técnicas automatizadas y algoritmos para analizar los datos y extraer información útil e información relevante que puede ser utilizada para tomar decisiones informadas.

### Aspectos importantes de KDD

* Patrones validos: conocimiento contrastable con la realidad.
* Potencialmente útiles: relación con el objetivo que nos proponemos.
* Comprensibles: relacionado con el usuario que maneja el conocimiento extraído de los datos.

### Proceso y fases de KDD

1. Definir que se va a hacer
   1. Entender el dominio de aplicación.
   2. Entender el problema a resolver.
   3. Fijar los objetivos: asociación, clasificar, agrupar o predecir.
2. Selección de los datos
   1. Tipos de datos que se van a usar y sus fuentes.
   2. Se selección los datos y se extraen.
   3. Costoso en tiempo y esfuerzo.
3. Preprocesamiento

Preparación y limpieza de los datos distintos datos de manera que se puedan manejar en el resto de las fases. (70% del esfuerzo aprox.)

* + Datos necesarios.
  + Datos incompletos.
  + Datos redundantes.
  + Datos incorrectos.
  + Errores de transcripción.
  + Datos envejecidos.
  + Variaciones de datos.

1. Transformación

Tratamiento preliminar de datos, transformación y creación de nuevas variables.

* + Convertir datos categóricos a numéricos y viceversa.
  + Otras transformaciones: simplificar, agrupar, normalizar…
  + Reducción de la dimensionalidad.

1. Técnicas de Data Mining

Elección de las técnicas que se van a usar para encontrar patrones o información que estaba oculta extrayendo nuevo conocimiento que dé valor extra a los datos.

Tipos:

* + Predictivo.
  + Descriptivo.

Técnicas:

* 1. Regla de asociación

Dado un conjunto de registros, encontrar reglas de dependencia que ayuden a predecir la ocurrencia de un elemento basándonos en la presencia de otros.

Características:

* Cuando aparecen elemento/s por asociación aparecen otro/s.
* Se trabaja con atributos categóricos.
* Los atributos pueden expresarse en forma binaria.
* Los elementos de X no aparecen en Y.
  1. Patrones secuenciales

Es la extracción de patrones frecuentes relacionados con el tiempo u otro tipo de secuencia.

Características:

* Son similares a las reglas de asociación, pero los atributos tienen temporalidad o siguen una secuencia.
  1. Clasificación

Dado un conjunto de registros. Cada uno con un conjunto de atributos y donde uno de ellos se le denomina atributos de clase.

Encontrar un modelo para el atributo de clase en función de los valores del resto de atributos.

Hay que validar los resultados sobre un conjunto de datos de validación no usados hasta entonces.

Características:

* Trabajar con datos etiquetados, supervisado.
* Utilizar un conjunto de datos de entrenamiento y clasificar nuevos datos
* Encuentra los atributos que definen mejora una clase
  1. Regresión

Predicción del valor de una variable basándose en el valor que reflejan otras y siguiendo modelos de dependencia que pueden ser lineales o no lineales.

Cuidado con hacer hipótesis de causalidad.

Características:

* Predicen valores numéricos
* La clasificación se puede entender como la predicción de una clase
  1. Clustering

Dados unos datos y unas medidas de similitud entre ellos, encontrar agrupaciones similares.

Las medidas de similitud pueden ser la distancia euclídea para atributos continuos.

Características:

* Trabaja con datos no etiquetados, no supervisado
* Sugiere nuevos grupos con respecto a los atributos

1. Interpretación y evaluación

Los modelos o patrones obtenidos deberían aceptar o contradecir la hipótesis inicial. Los patrones deben presentarse de forma que sean entendibles. Es por ello por lo que las técnicas de visualización son muy útiles.

* Integrar los datos:
  + Usar el modelo obtenido en el sistema de información que se está trabajando.
  + Continuar con el proceso de KDD usando los nuevos datos que se obtienen.

## Cross-Industry Standard Process for Data Mining (CRISP-DM)

Diagrama

Descripción generada automáticamente

1. Comprensión del negocio:
   * Establecimiento de los objetivos del negocio.
   * Evaluación de la situación.
   * Establecimiento de los objetivos del Data Mining.
   * Generación del plan del proyecto.
2. Comprensión de los datos:
   * Recopilación inicial de los datos.
   * Descripción de los datos.
   * Exploración de los datos.
   * Verificación de la calidad de los datos.
3. Preparación de los datos:
   * Selección de los datos.
   * Limpieza de los datos.
   * Construcción de los datos.
   * Integración de los datos.
   * Formateo de datos.
4. Modelado:
   * Selección de la técnica de modelado.
   * Diseño de la evaluación.
   * Construcción del modelado.
   * Evaluación del modelado.
5. Evaluación:
   * Evaluación de resultados.
   * Revistar el proceso.
   * Establecimiento de los siguientes pasos o acciones.
6. Despliegue:
   * Planificación de despliegue.
   * Planificación de la monitorización y del mantenimiento.
   * Generación de informe final.
   * Revisión del proyecto.

## Casos de uso

* Asociación: Encontrar ítems o características que se dan en común.
* Clasificación: Crear un clasificador con una serie de individuos. Luego usarlo para asignar una clase a los individuos nuevos.
* Predicción: Dados unos datos históricos saber si un evento ocurrirá o no.
* Clustering: Agrupar individuos de manera que los individuos cercanos se les considere de un mismo grupo.

# Tema 2: Introducción al Data Mining