

CURSO MINERÍA DE DATOS

Introducción - Clase 2 - Bárbara Poblete

LIBRO DEL CURSO

- Introduction to Data Mining
- Autores: Pang-Ning Tan, Michael Steinbach, Vipin Kumar



MÉTODOS DE DM

- Predictivos
- Descriptivos

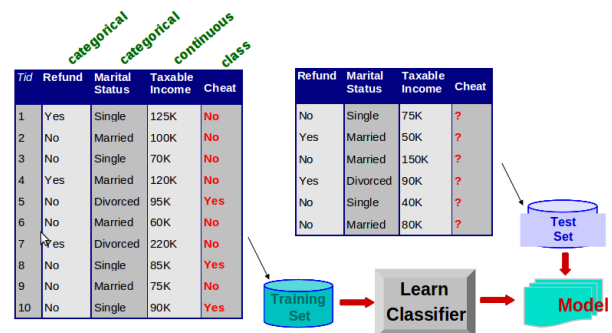


MÉTODOS DE DM

- Clasificación (Predictivo)
- Clustering (Descriptivo)
- Descubrimiento de Reglas de Asociación (Descriptivo)
- Descubrimiento de Patrones Secuenciales (Descriptivo)
- Regresión (Predictivo)
- Detección de Desviación (Predictivo)

CLASIFICACIÓN

- Set de Entrenamiento (atributos incluyendo clase)
- Busca modelar en atributo clase
- Objetivo: asignar la clase más correcta a records nuevos
- Set de Evaluación



CLASIFICACIÓN: APLICACIÓN 1

- Marketing directo
- Meta: Reducir costos de publicidad apuntando directamente a potenciales compradores.
- ¿Cómo?

CLASIFICACIÓN: APLICACIÓN 2

- Detección de Fraude
- Meta: Predecir transacciones fraudulentas en el uso de tarjetas de crédito
- ¿Cómo?

CLASIFICACIÓN: APLICACIÓN 3

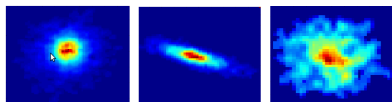
- Fidelidad de Clientes
- Meta: Predecir si es posible perder a un cliente a la competencia
- ¿Cómo?

CLASIFICACIÓN: APLICACIÓN 4

- Catalogación de exploración del espacio
- Meta: Predecir la clase (estrella o galaxia) de objetos en el espacio, en especial de objetos poco visibles, basándose en exploración de telescopios (3000 imágenes de 23.040x23.040 píxeles por imagen, del observatorio Palomar)
- ¿Cómo?

CLASIFICANDO GALAXIAS

- Tamaño de los datos:
 - 72M de estrellas, 20M de galaxias
 - Catalogo de objetos: 9GB
 - BD de imágenes: 150GB
- Atributos:
 - Características de las imágenes
 - Características de las ondas de luz, etc
- Clase:
 - Etapas de formación (temprana, intermedia, tardía)

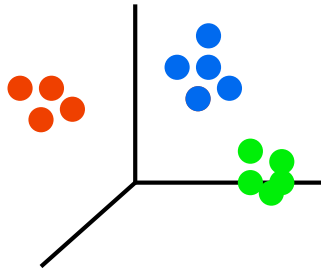


CLUSTERING

- Conjunto de puntos (datos), cada uno con un set de atributos y una medida de similitud
- Encontrar conjuntos tales que:
 - Puntos en un *cluster* sean más similares entre sí
 - Puntos en conjuntos diferentes sean menos similares entre sí

VISUALIZACIÓN DE CLUSTERING

- Clustering 3D basado en distancia Euclídeana
- Dist. intra-cluster es minimizada
- Dist. inter-cluster es maximizada



CLUSTERING APLICACIÓN I

- Segmentación de mercado
 - Meta: Subdividir un mercado en subconjuntos de clientes en donde cualquier conjunto es un potencial objetivo de marketing
 - ¿Cómo?

CLUSTERING APLICACIÓN 2

- Clustering de documentos
 - Meta: Encontrar grupos de documentos que son similares entre sí, basándose en las palabras más importantes que contienen.
 - ¿Cómo?

EJEMPLO

- Clustering de puntos: 3204 artículos del L.A. Times
- Medida de similitud: cuántas palabras tienen en común estos documentos (después de filtrar algunas palabras).

Category	Total Articles	Correctly Placed
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

REGLAS DE ASOCIACIÓN

- Dado un conjunto de records, cada uno contiene un número de elementos de una colección determinada
- Objetivo: Producir reglas de dependencia que predecirán la ocurrencia de un elemento (ítem) basándose en ocurrencias de otros ítems.

REGLAS DE ASOCIACIÓN

TID	Items
1	Pan, Coca-cola, Pañales, Leche
2	Cerveza, Pan
3	Cerveza, Coca-cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca-cola, Pañales, Leche

REGLAS DE ASOCIACIÓN APLICACIÓN 1

- Promoción de Marketing y Ventas
 - Sea la regla encontrada del tipo $\{\text{Queso}, \dots\} \rightarrow \{\text{PapasFritas}\}$

REGLAS DE ASOCIACIÓN APLICACIÓN 2

- Manejo de góndolas en los supermercados
- Meta: Identificar elementos que son comprados juntos por un número suficientemente grande de personas
- ¿Cómo?

REGLAS DE ASOCIACIÓN APLICACIÓN 3

- Manejo de inventario
- Meta: Una empresa de reparaciones a domicilio quiere anticipar la naturaleza de las reparaciones de sus consumidores, para mantener sus vehículos equipados y reducir nro. de viajes.
- ¿Cómo?

PATRONES SECUENCIALES

- Dado un set de objetos asociados a una línea de tiempo de eventos, encontrar los elementos que tengan fuertes dependencias secuenciales entre ellos
- Se forman reglas descubriendo patrones y luego se aplican restricciones de tiempo

REGRESIÓN

- Predecir el valor de una variable continua, en base a valores de otras variables, asumiendo modelo de dependencia lineal o no-lineal.
- Estadística y redes neuronales

DETECCIÓN DE DESVIACIÓN/ANOMALÍA

- Detectar desviaciones significativas de los valores normales

DESAFÍOS DE DM

- Escalabilidad
- Dimensionalidad
- Datos complejos y heterogéneos
- Calidad de los datos
- Distribución de los datos y propiedad
- Privacidad
- Streaming