Minería de datos aplicada

Pre procesamiento y Visualización*

Tareas principales en Pre-procedamiento de datos

- Limpieza de datos
- Integración de datos
- Reducción de datos
- Transformación de datos y discretización

Tareas principales en Pre-procedamiento de datos

- **Limpieza de datos**
- Integración de datos
- Reducción de datos
- Transformación de datos y discretización

- Rellenar los valores que faltan datos ruidosos, lisas, identificar o eliminar valores atípicos, y resolver las inconsistencias

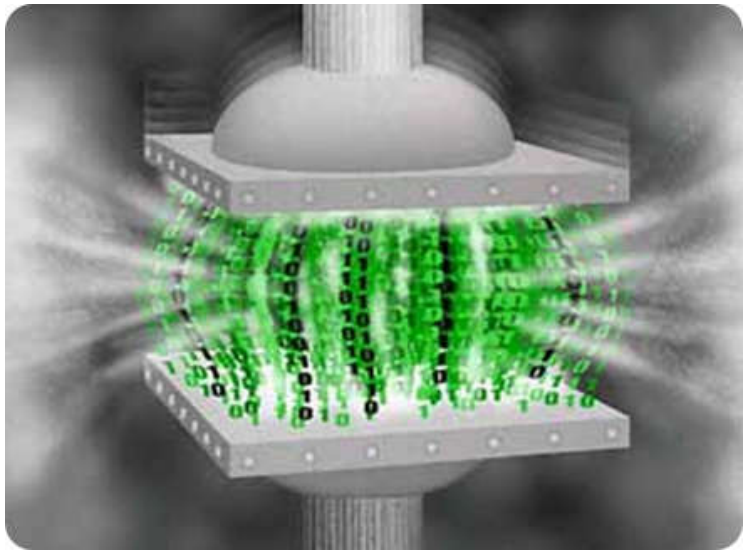
Tareas principales en Pre-procedamiento de datos

- Limpieza de datos
- **Integración de datos**
- Reducción de datos
- Transformación de datos y discretización

- integración de múltiples bases de datos, cubos de datos y/o archivos

Tareas principales en Pre-procedamiento de datos

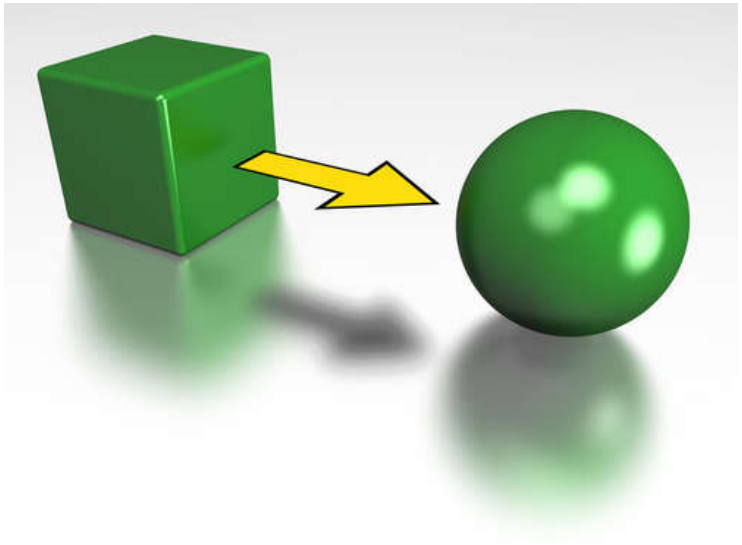
- Limpieza de datos
- Integración de datos
- **Reducción de datos**
- Transformación de datos y discretización



- Reducción de dimensionalidad
- Reducción de numerosidad
- Compresión de datos

Tareas principales en Pre-procedamiento de datos

- Limpieza de datos
- Integración de datos
- Reducción de datos
- **Transformación de datos y discretización**



- Normalización
- Concepto de generación de jerarquia

Limpieza de datos

!Los datos en el mundo real son sucios!

- **Incompletos**: carente de valores de atributos, que carece de ciertos atributos de interés, o que contiene sólo datos agregados
  - i.g. Ocupación = "" (datos no disponibles)
- **Ruidosos**: contienen ruido, errores, o valores atípicos
  - i.g. el sueldo = "- 10" (un error)
- **Inconsistentes**: contiene discrepancias en códigos o nombres
  - Edad = "42", Cumpleaños = "03/07/2010"
  - Fue calificación "1, 2, 3", ahora la calificación "A, B, C"
  - Discrepancia entre los registros duplicados
- **Intencional** (por ejemplo, disfrazado de datos que faltan)

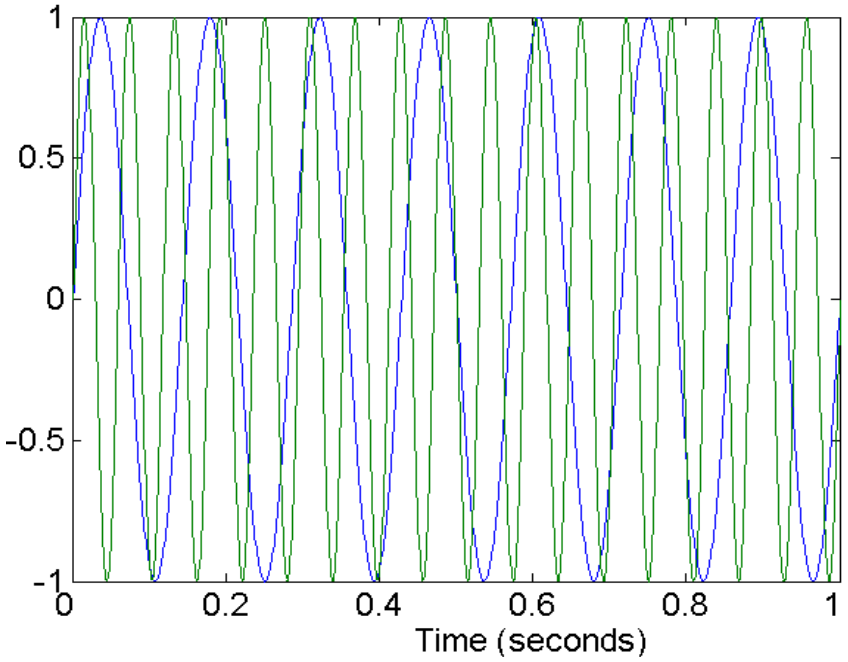
Datos (faltantes) incompletos

- Las razones de los valores que faltan
  - No se recoge la información (i.g. las personas se niegan a dar su edad y peso)
  - Los atributos pueden no ser aplicables a todos los casos (i.g. el ingreso anual no es aplicable a los niños)
- Tratar los valores faltantes o incompletos
  - Eliminar los objetos de datos
  - Estimar los valores perdidos
  - Ignorar el valor perdido durante el análisis
  - Reemplazar con todos los valores posibles (ponderados por sus probabilidades)

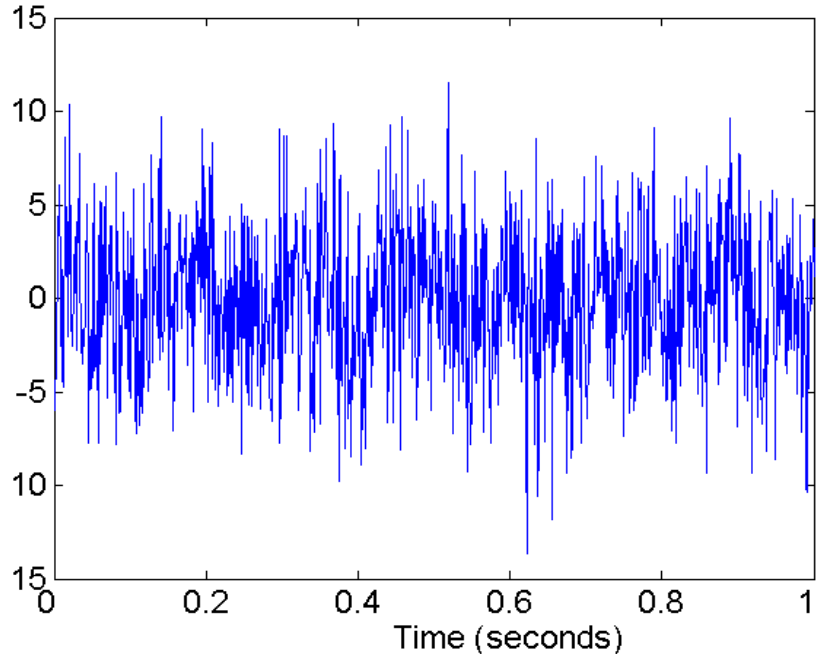
Datos ruidosos

El ruido se refiere a la modificación de los valores originales

- Ejemplos: la distorsión de la voz de una persona cuando se habla por un teléfono y "nieve" en la pantalla de la televisión



Dos ondas sinusoidales



Dos ondas sinusoidales + Ruido

Datos ruidosos

- Valores de atributos incorrectos pueden deberse a
  - Instrumentos de recolección de datos erróneos
  - Problemas de entrada de datos
  - Problemas de transmisión de datos
  - Limitación de la tecnología
  - Inconsistencia en la convención de nomenclatura

- Otros problemas con los datos que requieren la limpieza de datos
  - Registros duplicados
  - Datos inconsistentes

Manejo de datos ruidosos

- **Intervalos**
  - Primera clasificación de datos y la partición en los contenedores (igual frecuencia)
  - Entonces uno puede suavizar mediante la media sus “Cajones”, suavizar por promedio del “cajón”, suavizar por límites “Cajón”, etc.
- **Regresión o interpolación (investigue que es la imputación)**
  - Ajustes suaves a los datos (!Cuidado!)
- **La agrupación**
  - Detectar y eliminar los valores atípicos
- **Equipo multidisciplinar y la inspección humana**
  - Detectar valores sospechosos y comprobar por humanos (por ejemplo, hacer frente a los posibles valores atípicos)

Manejo de datos ruidosos : Detección de discrepancia en los datos

- El uso de metadatos (por ejemplo, dominio, rango, la dependencia, la distribución)
- Compruebe la sobrecarga de campo
- Compruebe regla de unicidad, la regla y la regla consecutiva nula
- Utilizar herramientas open source (e.g. PDI o OpenRefine) o comerciales (o algoritmos):
  - Depuración de datos: utilizar el conocimiento de dominio sencillo (por ejemplo, código postal, corrección ortográfica) para detectar errores y hacer las correcciones
  - Auditoría de datos: mediante el análisis de datos para descubrir las reglas y relaciones para detectar infractores (por ejemplo, la correlación y agrupación para encontrar los valores extremos)

Manejo de datos ruidosos : Migración e integración

- Herramientas de migración de datos: permiten transformaciones que se especificarán
- ETL (extracción / transformación / Carga) herramientas: permiten a los usuarios especificar transformaciones a través de una interfaz gráfica de usuario

Manejo de la redundancia en integración de datos

- los datos redundantes se producen a menudo cuando integración de múltiples bases de datos
  - Identificación de objetos: El mismo atributo o un objeto puede tener diferentes nombres en diferentes bases de datos
  - Derivable de datos: Un atributo puede ser un atributo de otra tabla, por ejemplo, los ingresos anuales "derivada"
- Atributos redundantes pueden ser capaces de ser detectados por análisis de correlación y el análisis de covarianza
- La integración cuidadosa de los datos de múltiples fuentes puede ayudar a reducir / evitar redundancias e inconsistencias y mejorar la velocidad de la minería y la calidad

Manejo de la redundancia en integración de datos

- los datos redundantes se producen a menudo cuando integración de múltiples bases de datos
  - Identificación de objetos: El mismo atributo o un objeto puede tener diferentes nombres en diferentes bases de datos
  - Derivable de datos: Un atributo puede ser un atributo de otra tabla, por ejemplo, los ingresos anuales "derivada"
- Atributos redundantes pueden ser capaces de ser detectados por análisis de correlación y el análisis de covarianza
- La integración cuidadosa de los datos de múltiples fuentes puede ayudar a reducir / evitar redundancias e inconsistencias y mejorar la velocidad de la minería y la calidad

Estrategias de reducción de datos

- La reducción de datos: obtener una representación reducida de la conjunto de datos que es mucho más pequeño en volumen, pero todavía produce los mismos (o casi el mismo) analítica resultados

- ¿Por qué la reducción de datos? – Una bodega de datos puede almacenar terabytes de datos. análisis de datos complejos puede llevar mucho tiempo para ejecutarse en el conjunto de datos completo.

- las estrategias de reducción de datos

Estrategias de reducción de datos

- reducción de dimensionalidad, por ejemplo, eliminar atributos sin importancia
  - Análisis de Componentes Principales (PCA)
  - selección de subconjuntos de características, la creación de operaciones
- reducción de numerosidad (algunos simplemente llamarlo: Reducción de Datos)
  - Regresión y Modelos log-lineal
  - Histogramas, el agrupamiento, el muestreo
  - la agregación cubo de datos
- Compresión de datos

Estrategias de reducción de datos

- Maldición de dimensionalidad
  - Cuando aumenta la dimensionalidad, los datos se vuelve cada vez más escasa
  - La densidad y la distancia entre los puntos, que es fundamental para la agrupación, el análisis de valores atípicos, se vuelve menos significativa
  - Las posibles combinaciones de subespacios crecerán exponencialmente
- Reducción de dimensionalidad
  - Evitar la maldición de la dimensionalidad
  - Ayudar a eliminar las características irrelevantes y reducir el ruido
  - Reducir el tiempo y el espacio requerido en la minería de datos
  - Permitir la visualización más fácil
- Las técnicas de reducción de dimensionalidad
  - la transformada
  - Análisis de componentes principales
  - técnicas supervisadas y no lineales (por ejemplo, la selección de características)

Análisis de componentes principales (PCA)

- Encuentra una proyección que captura la mayor cantidad de variación en los datos
- Los datos originales se proyectan en un espacio mucho más pequeño, lo que resulta en la reducción de dimensionalidad. Nos encontramos con los vectores propios de la matriz de covarianza, y estos vectores propios definimos el nuevo espacio

Análisis de componentes principales (PCA)

- N vectores de datos dados a partir de N-dimensiones, encontrar k ≤ n vectores ortogonales (componentes principales) que pueden ser mejor utilizados para representar datos
  - Normalizar los datos de entrada: Cada atributo cae dentro de la misma gama
  - Calcule k ortonormal (unidad) vectores, es decir, componentes principales
  - Cada dato de entrada (vector) es una combinación lineal de los vectores componentes principales k
  - Los componentes principales son ordenados en orden decreciente de "significado" o la fuerza
  - Puesto que los componentes están ordenados, el tamaño de los datos puede reducirse mediante la eliminación de los componentes débiles, es decir, los que tienen baja varianza (es decir, el uso de los componentes principales más fuertes, es posible reconstruir una buena aproximación de los datos originales)
- Que funciona para sólo datos numéricos

Selección de un subconjunto de atributos

- Otra forma de reducir la dimensionalidad de los datos
- atributos redundantes
  - Duplicar gran parte o todo de la información contenida en uno o más de otros atributos
  - Por ejemplo, el precio de compra de un producto y la cantidad de impuesto sobre las ventas pagado
- atributos irrelevantes
  - No contienen información que es útil para la tarea de minería de datos a la mano
  - Por ejemplo, los estudiantes de identificación es a menudo irrelevante para la tarea de predecir los estudiantes GPA

Generación de atributos

- Crear nuevos atributos (características) que puede capturar la información importante en un conjunto de datos de manera más eficaz que los originales
- Tres metodologías generales
  - extracción de atributos
    - -Dominio específico
  - mapeo de datos a un nuevo espacio (véase: la reducción de datos)
    - Por ejemplo, la transformación de Fourier, transformación wavelet, los enfoques múltiples (no cubierta)
  - atribuir la construcción
    - que combina las características (véase: patrones discriminativos frecuentes en el capítulo 7)
    - discretización de datos

THANK YOU!

ANY QUESTIONS?

Jun Akizaki - http://thepopp.com

Used Font: Roboto Light & Roboto Condensed Light

Icon: Font generated by flaticon.com under CC BY. The authors are: Stephen Hutchings.
Changed the color by Photoshop

World Map: http://www.tutsking.com/vectors/world-dots-map
Changed the color by Photoshop