TEMA 1: INTRODUCCIÓN AL PRE-PROCESAMIENTO DE DATOS

Motivación

- El aprendizaje automático extrae conocimiento a partir de bases de datos
  - Gran potencial de aplicación
- Desafortunadamente
  - Las bases de datos reales están influenciadas por factores negativos como
    - Inconsistencias
    - Ruido
    - Valores perdidos
    - Outliers
    - Tamaños muy grandes

Bases de datos de baja calidad implican generar conocimiento de baja calidad

Definición

- "The fundamental purpose of data preparation is to manipulate and transforrm raw data so that the information content enfolded in the data-set can be exposed or made more easily accesible."

Dorian Pyle: Data Preparation for Data Mining, Morgan Kaufmann Publishers, 1999, pp 90

- El objetivo fundamental de la preparación de datos es manipular y transformar los datos originales de tal forma que la información contenida en ellos pueda ser expuesta o facilitar el acceso a ella

Importancia de la Preparación de Datos

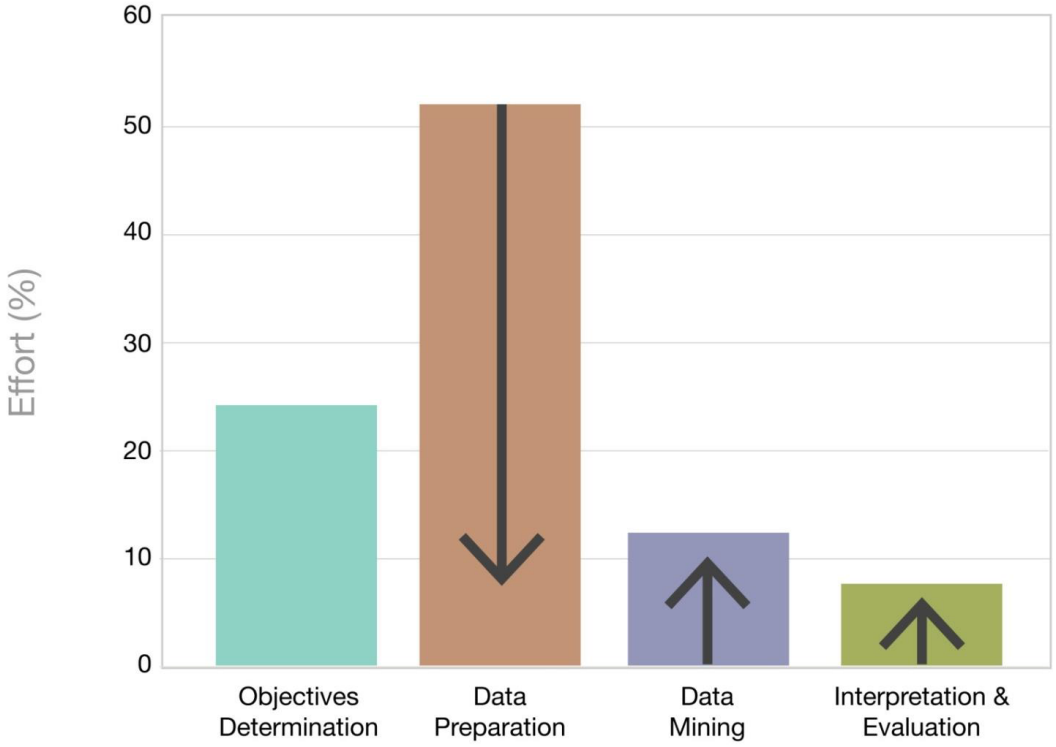
- 40% de los datos impuros
  - Sin limpiar (pre-procesar) los datos la calidad del conocimiento (patrones/reglas) obtenido por las técnicas de data mining se reduce en gran medida (poco útil)
- Limpiarlos por humanos
  - Muchas personas
  - Tarea laboriosa y complicada
  - Mucho tiempo
  - Suele llevar a errores

Importancia de la Preparación de Datos

- Tipos de impurezas (errores) de los datos
  - Valores inconsistentes (inexactos)
    - Valores para los que no se ha comprobado su validez
  - Ejemplos y variables redundantes
    - Ejemplos duplicados
    - Variables que pueden ser derivadas a partir de otras
  - Valores incompletos (valores perdidos):
    - Desconocidos
    - No almacenados
    - Irrelevantes
  - Valores "Outliers"
    - Valores fuera del rango habitual de la variable en estudio
  - Valores con ruido
  - Gran dimensionalidad de los datos

Importancia de la Preparación de Datos

- El pre-procesamiento de datos consume una parte muy importante del tiempo total de un proceso de minería de datos



Importancia de la Preparación de Datos

- ¿Qué incluye la Preparación de Datos?
  - "El Pre-procesamiento de Datos" / "La Preparación de Datos" engloba a todas aquellas técnicas de análisis de datos que permite mejorar la calidad de un conjunto de datos de modo que las técnicas de extracción de conocimiento/minería de datos puedan obtener mayor y mejor información (mejor porcentaje de clasificación, reglas con más completitud, etc.)

Importancia de la Preparación de Datos

- ¿Qué incluye la Preparación de Datos?
  - **Colección e integración de datos**
    - Proceso de integrar los datos provenientes de diferentes fuentes
  - Transformación de datos
    - Proceso de transformación de datos de tal forma que la técnica de aprendizaje pueda ser aplicada o ser más eficiente
  - Limpieza de datos
    - Proceso de corrección de errores e inconsistencias
    - Filtrado de ejemplos con errores
    - Reducción de variables con demasiado detalle
  - **Detección de outliers**
    - Proceso de detección de valores que presenten un comportamiento muy diferente del resto de valores de una variable

Importancia de la Preparación de Datos

- ¿Qué incluye la Preparación de Datos?
  - **Normalización de datos**
    - Proceso en el que los datos se expresan en la misma unidad de medida, escala o rango
  - Imputación de valores perdidos
    - Proceso de rellenado de las variables con valores perdidos asignando valores intuitivos (apropiados)
  - Identificación de ruido
    - Proceso de detección de errores aleatorios o variaciones en los datos
  - Reducción de datos
    - Proceso de obtención de una representación reducida en volumen
      - Produciendo resultados analíticos iguales o similares

Importancia de la Preparación de Datos

- La preparación de datos genera "datos de calidad", los cuales pueden conducir a patrones/reglas de calidad
  - Recuperar información incompleta
  - Eliminar outliers, ruido
  - Resolver conflictos
  - …
- La preparación de datos puede generar un conjunto de datos más pequeño que el original, lo cual puede mejorar la eficiencia del proceso de Minería de Datos
  - Eliminar registros duplicados
  - Eliminar anomalías
  - Selección de variables
  - Selección de instancias (muestreo)

Tareas de la Preparación de Datos

- Colección e integración de datos
  - Objetivo: Integrar los datos provenientes de diferentes fuentes de información en un data set único
    - Se utilizan funciones que establecen como se integran los ejemplos en la estructura común
    - Los datos de bases de datos relacionales se unifican en un registro único

Tareas de la Preparación de Datos

- Colección e integración de datos
  - Resolución de duplicidades e inconsistencias
    - Valores mal escritos (Pespi-cola)
      - Análisis de similitud entre palabras (no es trivial)
      - Ejemplo: distancia de edición (edit distance)
        - La distancia de edición entre dos strings a y b d(a,b) es el número mínimo de operaciones que transforman a en b.
        - Operaciones:
          - Insertar
          - Borrar
          - Sustituir
        - Distancia de edición entre pespi y pepsi: 2
          - Sustituir la s por la p: pespi → peppi
          - Sustituir la p por la s: pespi → pepsi
      - Analizar las distancia de los valores de una variable categórica y tomar medidas en consecuencia

Tareas de la Preparación de Datos

- Colección e integración de datos
  - Resolución de duplicidades e inconsistencias
    - Varios valores para el mismo concepto
      - Ejemplo: Pepsi, Pepsi-cola
      - Implican un nuevo valor discreto, revisar cuidadosamente la lista de valores discretos para cada atributo y unificarlos
    - Edad: 27 años - Fecha nacimiento: 16/03/1954

Creación de filtros específicos del problema para tratar estos problemas

Tareas de la Preparación de Datos

- Colección e integración de datos
  - Resolución de problemas de representación, escala, o forma de codificar
    - Sexo: V/M – M/F
    - Dinero: Euros – Dólares
    - Peso: Kg - Libras
    - Sueldo: Anual – Mensual
    - Precio: Con / Sin impuestos
- Algunos de ellos se pueden detectar con técnicas de EDA
  - Ejemplo: número de valores para la variable sexo (histograma)

Creación de funciones para solucionar problemas de escala, representación, etc…

Tareas de la Preparación de Datos

- Colección e integración de datos
  - Detección de variables redundantes: una variable es redundante si puede obtenerse a partir de otras
    - Atributos redundantes
      - x, x^2
    - Atributos iguales nombrados de forma diferente en diferentes tablas
      - Id-cliente vs. num-cliente
  - Análisis de correlaciones: dejar solamente una de las correlacionadas

Tareas de la Preparación de Datos

- Transformación (discretización) de datos
  - Transformar los valores numéricos en discretos o viceversa
  - Es una tarea esencial al trabajar con atributos discretos si aplicamos técnicas de minería de datos que solo acepten atributos numéricos (o viceversa)

Tareas de la Preparación de Datos

- Algunos algoritmos tienen métodos propios para tratar con datos incompletos o con ruido
  - En general no son muy robustos, lo normal es realizar previamente la limpieza de los datos

W. Kim, B. Choi, E.-D. Hong, S.-K. Kim

A taxonomy of dirty data.

Data Mining and Knowledge Discovery 7, 81-99, 2003

- Limpieza de datos: incluye los siguientes tratamientos
  - Completar / Imputar valores perdidos
  - Tratar valores con ruido
  - Identificar "outliers"

Tareas de la Preparación de Datos

- Valores perdidos: Los datos no siempre están disponibles
  - Muchos ejemplos pueden no tener valor asociado para ciertas variables
- Los valores perdidos (datos faltantes) pueden deberse a:
  - Errores técnicos (de equipamiento)
  - Inconsistencia con otros datos almacenados (y por tanto borrados)
  - Datos no ingresados
  - Considerados irrelevantes al momento de ser almacenados

Tareas de la Preparación de Datos

- Datos con ruido: error aleatorio o varianza en una variable medida
- Valores de atributos incorrectos debido a:
  - Instrumentos de medición erróneos
  - Problemas en la entrada de datos
  - Problemas en la transmisión
  - Limitaciones tecnológicas

Tareas de la Preparación de Datos

- Normalización de datos
  - Transformar los valores de tal forma que todos los atributos estén en el mismo rango (mejor)
  - Normalización min-max

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ejemplo: queremos normalizar el atributo cuyo rango de entrada es [12.000, 98.000] al rango [0.0, 1.0]. El valor 73.600 es transformado

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$$

Tareas de la Preparación de Datos

- Normalización de datos
  - Normalización por escala decimal

$$v' = \frac{v}{10^j}$$

    - donde j es el entero más pequeño tal que Max(|v' |)<=1

  - Ejemplo: si el valor del atributo varía entre -986 y 917, el valor máximo del atributo en valor absoluto es 986. Para normalizar se divide entonces por 1000 ($j = 3$):
    - -986 --normalizado--> -0.986

Tareas de la Preparación de Datos

- Normalización de datos
  - Normalización z-score

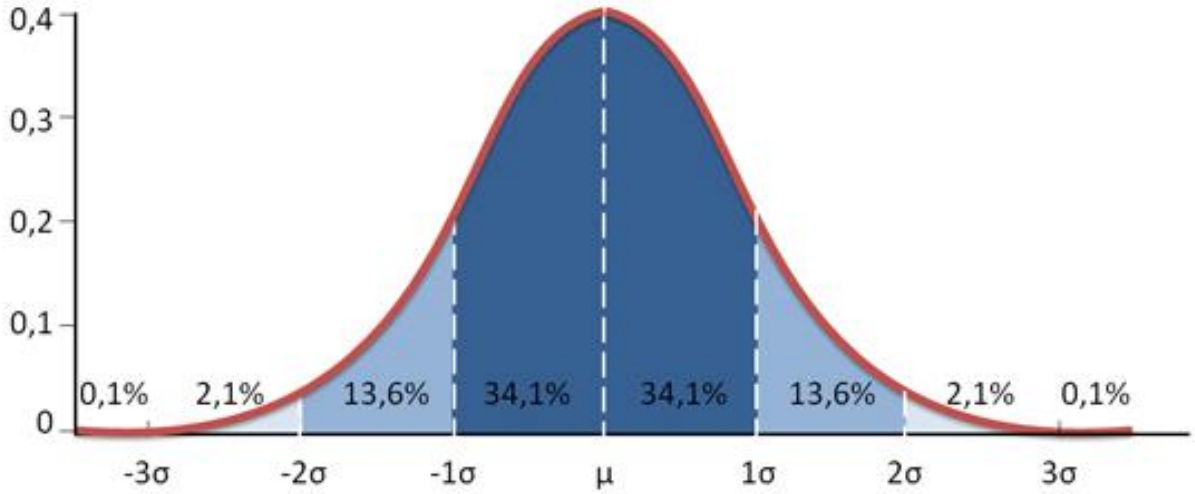
$$v' = \frac{v - \pi_A}{\sigma_A}$$

- "Resuelve" el problema de los outliers
- Ejemplo: sea π = 54,000 y σ = 16,000. El valor 73.600 es transformado

$$\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$

Tareas de la Preparación de Datos

- Normalización z-score



Tareas de la Preparación de Datos

- Detección de outliers: datos con características considerablemente diferentes a la mayoría del resto de datos
  - Métodos basados en estadística
    - Utilizan la media, desviación estándar
    - El valor v_i es un outlier si se cumple una de las dos condiciones siguientes

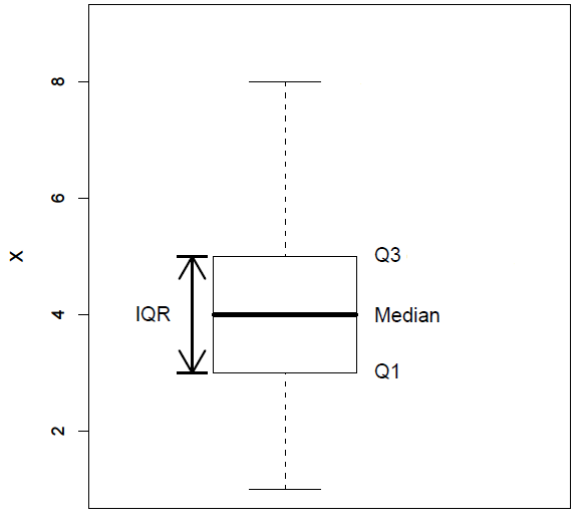
$$v_i > \pi_i + k * \sigma_i$$
$$v_i < \pi_i - k * \sigma_i$$

donde v_i es el valor a comprobar, π_i es la media de atributo i, σ_i es la desviación estándar del atributo i y k es un entero positivo

    - Pueden generar muchos falsos positivos

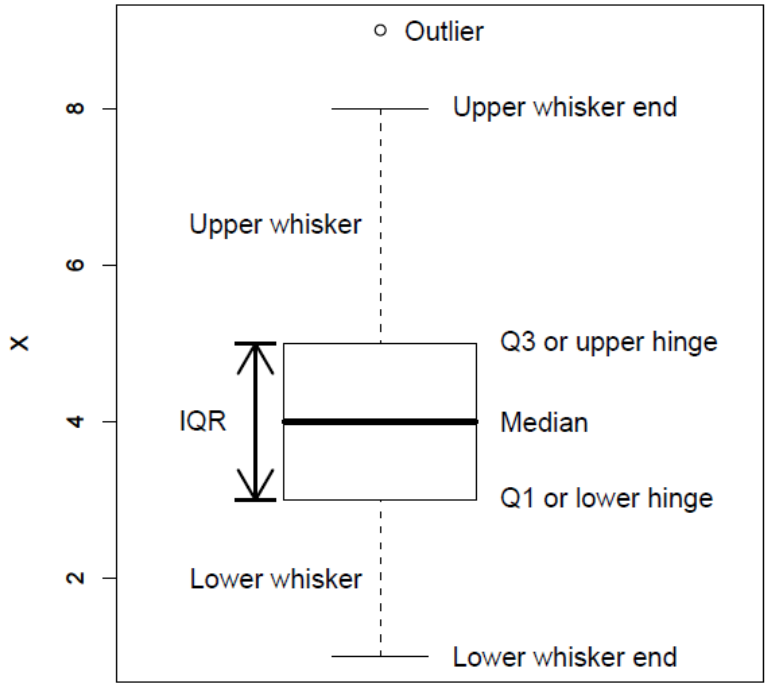
Tareas de la Preparación de Datos

- Rango inter cuartil: Q3-Q1
- Los valores de los cuartiles están definidos por:
  - El primer valor (Q1) es aquel para el que un cuarto de valores de la variable son menores que él
  - El segundo (Q2) es aquel para el que la mitad de los valores de la variable son menores que él
  - El tercero (Q3) es aquel para el que tres cuartas partes de los valores de la variable son menores que él
- Boxplot



Tareas de la Preparación de Datos

- Rango inter-cuartil
  - IQR = Q3 – Q1
  - Outlier si
    - dato > Q3 +1.5*IQR
    - dato < Q1 - 1.5*IQR
  - Outlier extremo si
    - dato > Q3 + 3*IQR
    - dato < Q1 - 3*IQR



Limpieza de datos

- La no detección de un outlier puede ser un problema importante si el atributo se normaliza posteriormente
  - Mayoría de datos estarán en un rango pequeño
    - Puede ocasionar poca precisión o sensibilidad para algunos métodos de minería de datos
- Tratamiento de outliers
  - Ignorarlo: si el método es robusto ante estos datos
  - Eliminar la variable (solución extrema): si hay otra variable correlacionada con datos mejores
  - Eliminar el ejemplo: puede producir un sesgo si son casos especiales
  - Reemplazar el valor: nulo, mínimo, máximo, media, moda, etc…
  - Discretizar (variable continuas): asignar a la categoría más baja o alta