Sesión 2: Visualización, estructuración y preprocesamiento de datos

ML-IA: Módulo de Machine learning

Jacob Muñoz-Castro

Universidad de los Andes

5 de noviembre de 2020

Agenda

1. Tipos de datos
2. Tratamiento de datos
3. Limpieza e imputación
4. Tipos de gráficos

¿Qué son datos?

Colección de **objetos** definidos por **atributos**.

- Un atributo es una propiedad o característica de un objeto.
- Por ejemplo: color de ojos de la persona, temperatura, *intereses de la cuota esperada de microcrédito*, *calificación del marco integral de supervisión*.
- Recibe también el nombre de variable, característica, predictor, etc.
- Una colección de atributos describe un objeto.
- Recibe también el nombre de registro, punto, caso, muestra, entidad, entrada, instancia, etc.

Clasificación de atributos

Un mismo atributo se puede mapear a diferentes valores (altura en cm o ft, saldos en moneda local o extranjera).

Diferentes atributos pueden mapearse aun mismo conjunto de valores (identificaciones y edad son enteros).

- **Atributo discreto**: Tiene valores contables (naturales). Ej: códigos postales, conteo de clic, frecuencia de palabras en un texto.
- **Atributo continuo**: Tiene los reales como valor del atributo. Ej: temperatura, altura, peso, pagos, índices.

Tipos de datos

- Registrado: matriz de datos, documentos, transacciones.
- Ordenado: espacial, temporal, secuencial (genética).
- Gráfico: World Wide Web (www), estructuras moleculares.

Registro

Los datos registrados o de registro son los que convencionalmente pensamos cuando no mencionan datos.

Son una colección de registros con un conjunto fijo de atributos.

Matriz de datos

- Solo contienen datos numéricos que pueden ser representados por una matriz de m por n (m registros con n atributos).
- Cada registro puede pensarse como puntos en un espacio multidimensional, donde cada dimensión representa un atributo distinto.

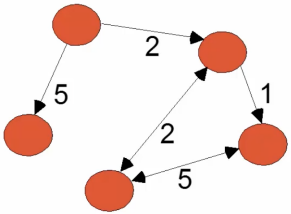
Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Datos de transacción/documental

Tipo especial de datos donde cada registro (transacción) es involucra un conjunto de elementos.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Datos gráficos




```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

Datos ordenados

Espacio $\times$ tiempo

Datos estructurados

Registros

Datos no estructurados

- Lo que no parezca registro: texto, presentaciones, imágenes, video, audio, PDF.
- Estrategia: transformarlo a terreno conocido (estructurar los datos).
- Pero puede llegar a ser costoso y demandante en tiempo y trabajo.

Calidad de datos

- Preguntas fundamentales sobre la calidad de nuestros datos:
  - ¿De qué problema nos debemos preocupar?
  - ¿Cómo podemos detectar problemas con los datos?
  - ¿Qué se puede hacer para responder estos problemas?
- Los problemas más comunes de calidad de datos:
  - Ruido y valores atípicos
  - Valores omitidos (Missing values)
  - Datos duplicados

Ruido

Señal inválida que se sobrepone sobre datos válidos. Ej: mala asignación de etiquetas por intervención humana.

Casas & González-Ramírez (2016). Productivity measures for the colombian manufacturing industry. Borradores de Economía; No. 947.

Valores atípicos (outliers)

Registros con características que son considerablemente diferentes al resto de los registros en nuestros datos.

A diferencia de ruido, los outliers son datos válidos.

Valores omitidos (NA)

- **Razones:** información que no se recolectó (NS/NR) o atributos que no aplican a todos los casos (ingreso anual de NUIP/TI).
- **Tratamiento:**
  - Eliminar registros
  - Estimar los valores
  - Ignorar los valores durante el análisis: no incluir un atributo plagado de valores omitidos es equivalente.
  - Reemplazar con todas los valores posibles (ponderados por sus probabilidades).

Valores duplicados

- Datos que aparecen más de una vez.
- De fácil tratamiento siempre y cuando sea de fácil identificación.
- Este problema es sustancialmente importante cuando se están uniendo bases de distintas fuentes Ej: dispersar mal un pago por identificaciones mal reportadas por entidades.

Limpieza

- Agregación
- Muestreo
- Reducción de dimensionalidad
- Selección de variables
- Discretización y binarización
- Transformación de atributos

Agregación

Combinar dos o más atributos o registros en un solo atributo o registro.
Propósito:

- Reducción de los datos: reducir el tamaño de de datos grandes puede representar un ahorro significativo en el procesamiento y la memoria. Ej: resumir cuotas de crédito a partir de capital e interés.
- Cambio de la escala de los datos: agregación vertical (agregar de clientes a entidades, de depositantes a departamentos, productos a persona).
- Estabilización de los datos: agregación de datos dinámicos puede volver menos variable y errático un panel de datos.

Muestreo

Esta es la técnica principal empleada para selección de datos.

- Estadísticamente es muy costoso en tiempo y recursos **obtener** los datos completos.
- En términos de minería de datos, es muy costoso en tiempo y recursos **procesar** los datos completos.

La clave de un buen muestreo es la **representatividad**.

- A veces es tan sencillo como tomar una muestra aleatoria (muestreo aleatorio simple).
- Otras veces, la pregunta a responder sugiere mantener proporciones constantes (muestreo estratificado).
- Muestras con y sin reemplazo.

Maldición de la dimensionalidad

Cuando la dimensionalidad (número de atributos) aumenta, nuestros datos se vuelven más dispersos en el espacio que estos ocupan.

Si añadimos suficientes variables, cada punto en el espacio se verá como un outlier.

SOLUCIÓN: reducción de dimensionalidad (sesión 6).

Selección de variables

- Variable redundante: contiene información casi duplicado.
- Variable irrelevante: contiene información poco útil.

SOLUCIÓN:

- Brute-force approach: intentar todas los subconjuntos de variables (p.e. stepwise).
- Embedded approach: la selección de variables hace parte del algoritmo (p.e. árboles de decisión)
- Filter approach: las variables se eligen antes de usar los algoritmos.
- Creación de variables (feature engineering): extracción, construcción y mapeo a nuevos espacios (exp, log, abs, estandarización, normalización...ej:detección de anomalías en series de tiempo con series de Fourier).

Exploración de datos

¿Qué es exploración de datos?

- Visualización y estimación para entender mejor las características de los datos.
- Ayuda a elegir las herramientas adecuadas para el procesamiento y el análisis.
- Hacer uso de habilidades humanas para reconocer patrones (no reconocibles por nuestras herramientas de análisis de datos).
- Ahora nos concentraremos en visualización.

Histogramas

- Muestra la distribución de valore de una sola variable.
- Dividir los valores en contenedores (intervalos) y muestra un gráfico de barras del número de registros para cada uno.
- La forma del histograma depende de la granularidad de los intervalos.

Diagrama de caja

Muestra la distribución de los datos.

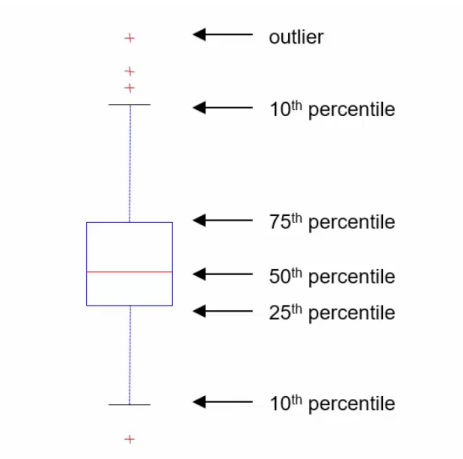


Gráfico de dispersión

- Los valores de los atributos determinan la posición del punto.
- Los gráficos de dispersión más comunes son de dos dimensiones.
- Usamos tamaño, forma y color para mostrar atributos complementarios.

Gráfico de contorno

- Usados para atributos continuos en una grilla espacial.
- Las lineas de contorno que forman los límites de las regiones conectan puntos con igual valor.