Universidad Autónoma de Baja California

Maestría y Doctorado en Ciencias e Ingeniería
Ingeniería en Computación

3. Pre-procesamiento de datos

Minería de Datos

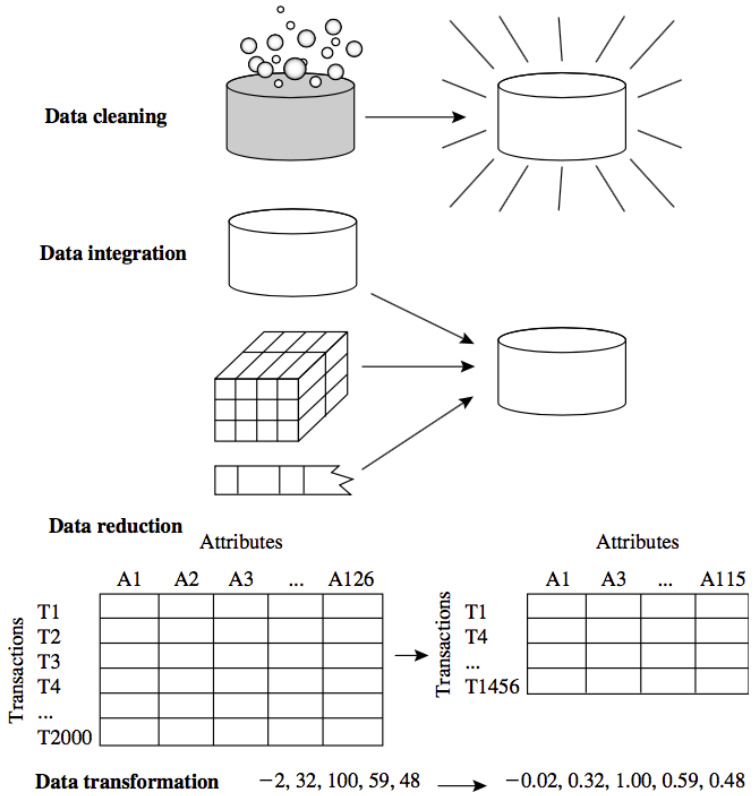
¿Para qué preprocesar los datos?

- ¿Cómo pueden ser preprocesados los datos para ayudar mejorar la calidad de los datos, y consecuentemente los resultados del minado?
- ¿Cómo pueden ser preprocesados los datos para mejorar la eficiencia y facilidad del proceso de minado?
- Para minar datos, se requiere calidad en los datos, esto incluyen
  - Exactitud. Que no contenga errores, ruido, ni se desvíe de valores esperados.
  - Integridad. Que todos los atributos contengan datos.
  - Consistencia. No contenga discrepancias entre los valores usados.
  - Puntualidad. Los datos son ingresados a tiempo en las bases de datos.
  - Credibilidad. Refleja la confianza que los usuarios le tienen a los datos.
  - Interpretabilidad. Todos los datos deberán ser fácilmente interpretados por los usuarios.

Técnicas de preprocesamiento

- **Limpieza de datos.** Pueden ser aplicados para eliminar el ruido y corregir inconsistencias en los datos.
- **Integración de datos.** Unifica datos de diversas fuentes hacia una sóla fuente de datos, tal como un almacén de datos.
- **Reducción de datos.** Puede reducir el tamaño de los datos mediante la agregación, eliminación de características redundantes, o agrupando.
- **Transformación de datos.** Puede ser aplicado para normalizar los datos escalando los valores dentro de un rango de [0,1].

Introducción. Tareas principales en el preprocesamiento de datos



Limpieza de datos

- Datos provenientes del mundo real tienden a estar incompletos, ruidosos, e inconsistentes.
- La limpieza de datos intenta rellenar valores faltantes, alisar ruido, identificar datos anómalos, y corregir inconsistencias en los datos.

Valores faltantes ¿qué se puede hacer?

1. Ignorar la tupla
2. Asignar el valor faltante manualmente
3. Usar una constante global para asignar valores faltantes
4. Usar una medida de tendencia central para el atributo (e.g. Promedio o Mediana) para asignar el valor faltante
5. Usar el Promedio o Mediana del atributo para todas las muestras pertenecientes a la misma clase
6. Usar el valor más probable para asignar el valor (e.g. Usando regresiones, árboles de decisión)

Datos ruidosos

- El ruido es un error aleatorio o varianza dentro de una medición de una variable.
- **Binning.** Estos métodos alisan los valores a través de la consulta de sus vecinos. Ver figura.
- **Regresión.** Con éste método se encuentra una función entre dos atributos para encontrar la mejor línea que prediga la una con la otra. También se puede realizar regresión multivariable.
- **Análisis de datos anómalos.** Usando análisis de grupos, cualquier dato que no entre a un grupo se puede eliminar.

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

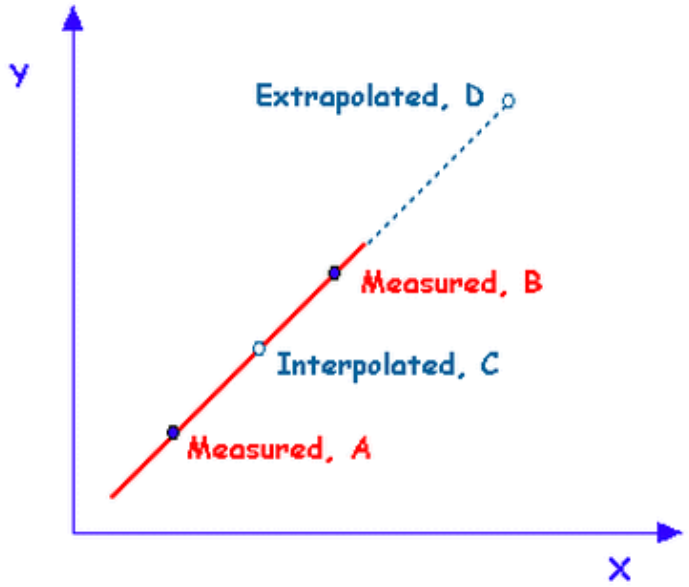
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34



Actividad

- Descargar los siguientes datasets con datos faltantes:
  - https://www.kaggle.com/datasets/aparnashastry/building-permit-applications-data
  - https://www.kaggle.com/datasets/jojoker/singapore-airbnb
  - https://archive.ics.uci.edu/ml/datasets/Heterogeneity+Activity+Recognition
- Utilizar un mínimo de 3 técnicas mostradas para el tratado de valores faltantes, aplicarlas por cada bases de datos

Reducción de datos

- La técnicas para la reducción de datos pueden ser aplicadas para obtener una representación reducida de los datos, pero en un volumen mucho menor, aún manteniendo la integridad de los datos originales. Esto es, minar en un conjunto de datos reducido debería ser más eficiente pero a su vez produce el mismo (o casi el mismo) resultado.

Transformación de datos y discretización de datos

- Estrategias para la transformación de datos:
  1. **Suavizado.** Sirve para remover ruido de los datos. Se pueden usar técnicas como, binning, regresiones, agrupación, etc.
  2. **Construcción de atributos.** Cuando nuevos atributos son agregados para facilitar el proceso de minado.
  3. **Agregación.** Donde se agregan resúmenes o agregan operaciones a los datos. Por ejemplo, las ventas totales diarias pudieran ser agregadas para posteriormente calcular las ventas mensuales y anuales.
  4. **Normalización.** Donde los datos de los atributos con escalados para caber dentro de un rango más pequeño, por ejemplo -1 a 1, o 0 a 1.
  5. **Discretización.** Donde valores crudos de atributos numéricos (e.g. edad) son reemplazados por etiquetas de intervalos (e.g. 0.10, 11-20, etc.) o etiquetas conceptuales (e.g. Joven, adulto, adulto mayor).
  6. **Generación conceptual de jerarquías para datos categóricos.** Donde atributos tal como calle pueden ser generalizados por conceptos de un nivel más alto, como ciudad o país.

Transformación de datos por normalización

- Las unidades de medida pueden afectar, por lo que se deben normalizar unidades, por ejemplo, pesos y dólares se manejarían como pesos.
- Para evitar dependencias de las unidades de medida, los datos deberían ser normalizados o estandarizados (sinónimo de normalización en términos de preprocesamiento de datos).
- Las normalizaciones más comunes transforman los rangos de los datos de un atributo a [-1,1] o [0,1].
- La normalización de datos intenta dar a todos los atributos un peso igual.

Técnicas comunes de normalización

- Normalización min-max:

$$v_i' = \frac{v_i - min_A}{max_A - min_A}$$

- Normalización z-score:

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A}$$

Actividad

- Investigar sobre la normalización
  - ¿Para qué sirve?
  - ¿Cuando usarse y no usarse?
  - ¿Qué tipos de normalización son más convenientes usar bajo qué circunstancias?
  - Cualquier otro detalle de interés que hayan encontrado respecto este tema.

Discretización de datos

- Discretización por binning.
  - Los valores de los atributos pueden ser discretizado mediante la aplicación de bins de igual longitud o frecuencia.
- Discretización por análisis de histograma.
  - Muy parecido al anterior, pero usando histogramas para definir los rangos.
- Otros métodos de discretización de datos.
  - Por análisis de agrupaciones
  - Por árboles de decisión
  - Por análisis de correlación

Generación conceptual de jerarquías para datos categóricos

- Los atributos categóricos tienen un número finito (pero posiblemente grande) de valores distintos, sin algún orden entre los valores. Por ejemplo, localidad geográfica, categorías de puesto laboral, tipo de artículo, etc.
- El concepto de jerarquías puede ser usado para transformar los datos hacia múltiples niveles de granularidad.
- Entre los métodos comunes de ésta tarea se encuentran:
  - …

Especificación de un orden parcial de los atributos explícitos a nivel esquema por usuarios expertos

- Si se tienen los siguiente atributos:
  - street
  - city
  - province_or_state
  - country
- Se puede definir una jerarquía de orden entre los esquemas de los atributos:
  - street < city < province_or_state < country

Especificación de una porción de una jerarquía por un grupo de datos explícitos

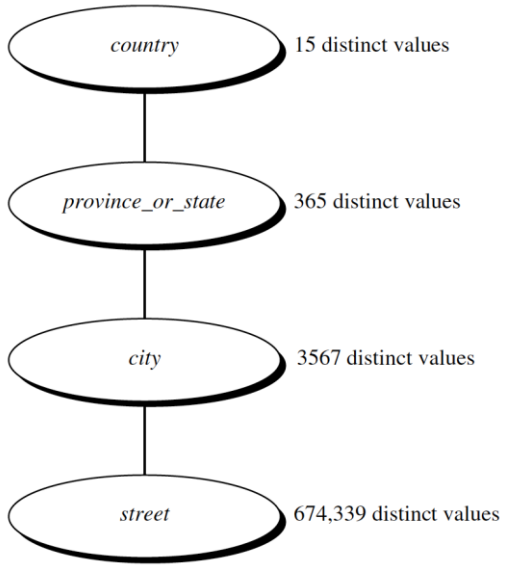
- Después de especificar que province y country forman una jerarquía a nivel esquema, un usuario pudiera definir niveles intermedios manualmente, tal como:

$$\{Alberta,\ Saskatchewan,\ Manitoba\} \subset prairies\_Canada$$

$$\{British\ Columbia,\ prairies\_Canada\} \subset Western\_Canada$$

Especificación de un conjunto de atributos, pero no de su orden parcial

- Los atributos con la mayor cantidad de valores distintos es puesto en el punto más bajo de la jerarquía. A como más bajo sea el número de valores distintos un atributo tenga, más alto será el concepto de jerarquía bajo el que será puesto.
- Por ejemplo, en una Base de Datos que contiene entradas de country, province_or_state, city, y street. Se tienen la siguiente cantidad de entradas de cada uno:
  - country (15)
  - province_or_state (365)
  - city (3567)
  - street (674,339).



Especificación de solamente un conjunto parcial de atributos

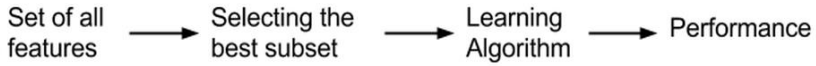
- Parecido al anterior pero con la diferencia que pudieran faltar atributos lo cual ayuden a definir las jerarquías, por lo que habría que manualmente determinar los atributos agrupables de manera jerárquica para iniciar el proceso de jerarquización.

Visión general de técnicas de reducción de datos

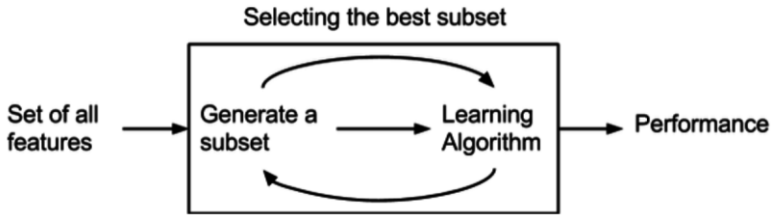
- **Reducción de dimensionalidad (feature selection).** Trata sobre el reducir la complejidad de los datos mediante la reducción de atributos.
  - Se eliminan atributos irrelevantes, atributos débilmente irrelevantes, y atributos duplicados. Por ejemplo, PCA, selección de subconjuntos de atributos.
  - Elegir solamente los atributos importantes ó reducer atributos mediante la creación de nuevos que resumen el comportamioento de uno o más atributos.
- **Reducción de numerosidad.** Estos reemplazan el volumen original de datos por formas de representación más pequeñas y alternas.
  - Técnicas paramétricas. Un modelo es usado para estimar los datos en lugar de los datos actuales, tal como regresiones.
  - Técnicas no paramétricas. Se almacenan representaciones de los datos, tal como histogramas, agrupaciones, muestreos, transformadas por wavelets, etc.
- **Compresión de datos.** Transformaciones son aplicadas para obtener representaciones de los datos originales de manera comprimida.

Reducción de dimensionalidad. Feature selection. Tipos de métodos

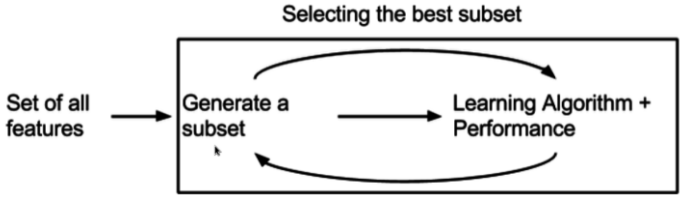
- **Filter.** Son utilizados durante la fase del preprocesamiento. Realizan la elección sin importer la técnica de ML que vaya a usarse.



- **Wrapper.** Son utilizados en conjunto con la técnica de ML de manera iterative (aunque no siempre) para poder elegir el major conjunto de atributos.



- **Embedded.** Son métodos que estan embebidos en la técnica de ML y no pueden separarse.



Reducción de dimensionalidad. Feature selection. Tipos de métodos

- **Filter.**
  - Information Gain
  - Chi-square test
  - Fisher's Score
  - Correlation Coefficient
  - Variance Threshold
  - Mean Absolute Difference (MAD)
  - Dispersion Ratio
  - Mutual Dependence
  - Relief
- **Wrapper.**
  - Forward selection
  - Backward elimination
  - Bi-directional elimination
  - Exhaustive selection
  - Recursive elimination
- **Embedded**
  - Regularization
  - Tree-based methods

Actividad

- Elegir dos opciones de métodos por filtro y aplicarlo con dos datasets a elegir por el estudiante
  - Presentar resultados. Atributos más importantes
  - Mostrar código
- Explicar cómo funcionan matemáticamte los dos métodos elegidos de feature selection
- No se tienen que limitar a los enlistados en esta presentación
- No repetir métodos entre los estudiantes. Usen el foro de Bb para ponerse de acuerdo