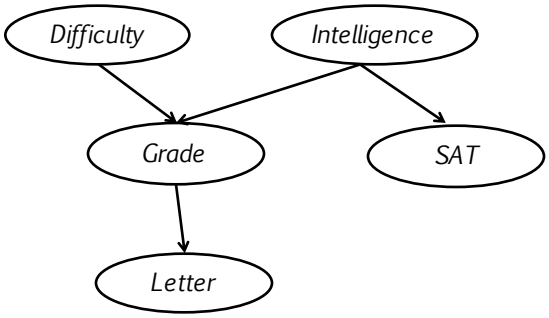
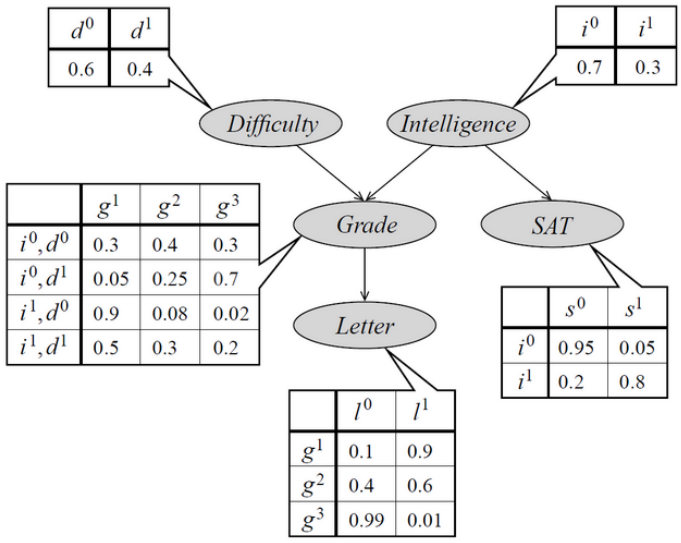
Ex Valor más probable: Bayesian Networks



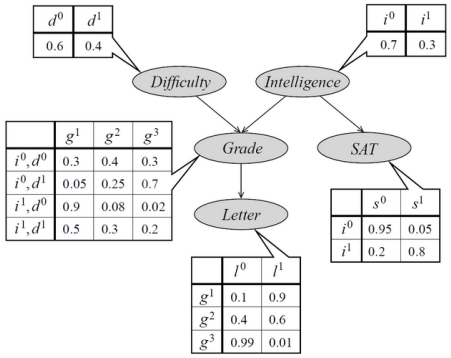
BN Factorization:

$$P(D,I,G,S,L) = P(D)P(I)P(G \mid D,I)P(S \mid I)P(L \mid G)$$

Ex: Valor más probable



Ex: Valor más probable



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

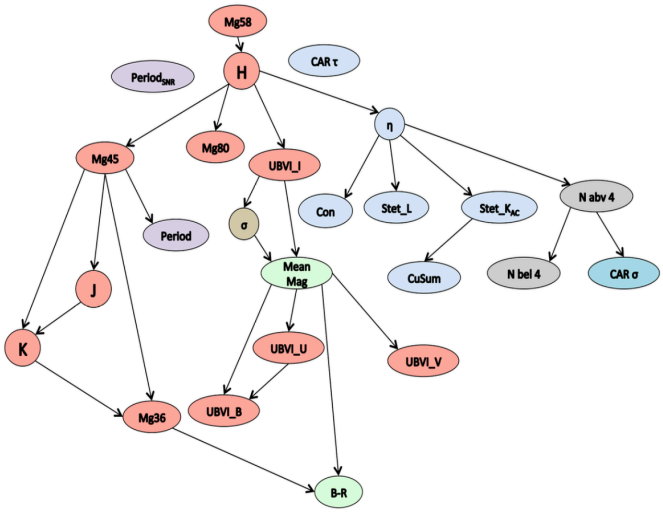
	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Ej: Qué hacemos si Grade es faltante, por ejemplo, tenemos una entrada:

x = [0 , 1 , ? , 0, 0] (D, I, G, S, L) respectivamente

Ex: Valor más probable

Ex: BN for astronomical time series with missing data



Datos Faltantes(Cont..)

- No siempre un dato faltante es un error. Ej, persona no tiene licencia de conducir, no usa tarjeta de crédito, etc.
- En esos casos es importante tener valores definidos como "no se aplica", etc.

Algunas técnicas de preprocesamiento

- **Binning:** Los datos se ordenan separándose en grupos (bins).
  - Smoothing by bin means: Cada valor en el bin es reemplazado por la media del bin.
  - Smoothing by bin boundaries: cada valor se reemplaza por el valor mínimo del bin o el máximo dependiendo de cuál sea el más cercano.

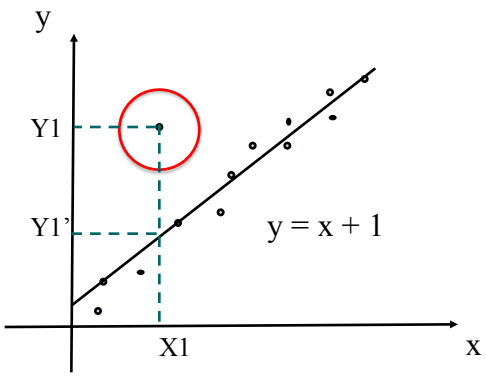
Este método se utiliza como herramienta de Discretización, smoothing, etc.

* Sorted data : 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Partition into (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Algunas técnicas de preprocesamiento

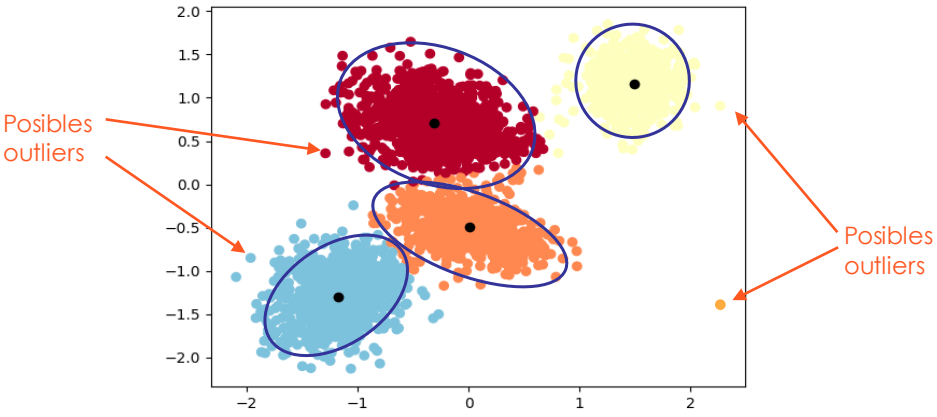
•Regresión para corrección: Algunos datos se "corrigen" en base a una función. Ej:

Regresión Lineal



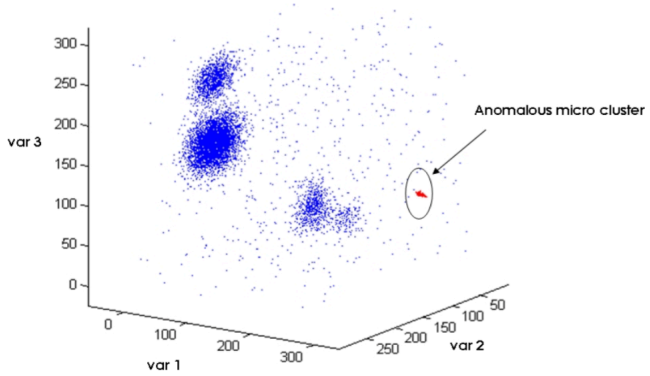
Algunas técnicas de preprocesamiento

- **Clustering** para la detección de outliers (candidatos a ser datos erróneos)



Algunas técnicas de preprocesamiento

- **Clustering** para la detección de outliers: Outliers are not always alone



Integración

- La información en la mayoría de los casos debe ser integrada desde múltiples fuentes de datos.

- Algunos problemas típicos:
  - *Identificación de la entidad*
  - *Redundancia*
  - *Detección y Resolución de conflictos entre valores*

Identificación de la entidad (entity identification problem)

- La misma entidad tiene distintos nombres en diferentes fuentes de datos, ej, customer_id, cust_number.

- Para esto se utilizad Metadata donde se almacena información sobre las entidades en cada fuente de datos, ej: nombre, significado, tipo de datos, rango, valores nulos, etc.

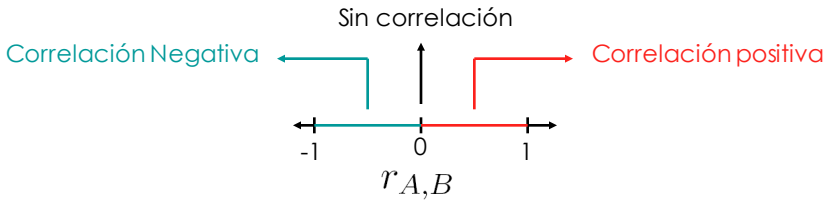
Redundancia

- Un atributo es redundante si puede ser derivado de otro. Errores en la identificación de la entidad suelen llevar a situaciones de redundancia
- Tablas no normalizadas también llevan a redundancias

- Puede ser detectada realizando un análisis de correlación:

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \overline{A})(b_i - \overline{B})}{N\sigma_A\sigma_B} \qquad -1 \leq r_{A,B} \leq 1$$

Redundancia (Cont.)



Detección y Resolución de conflictos entre valores

- Para la misma entidad, los valores del atributo proveniente de distintas fuentes de datos es diferente.
- Problema causado por diferencias de representación, escala, codificación, etc.

Detección y Resolución de conflictos entre valores (Cont..)

- **Diferencias en representación:** Por ejemplo, para una cadena de hoteles, el precio de una habitación en un hotel en distintas ciudades puede representar conceptos distintos, p.ej, uno incluye desayuno, impuestos u otro tipo de servicio que el precio en otra ciudad no contempla.
- **Diferencias en niveles de abstracción:** El total de ventas puede significar ventas totales en la cadena completa en una base de datos y en otra puede ser ventas totales en el hotel.

Detección y Resolución de conflictos entre valores (Cont..)

- **Diferencias de codificación:** Los valores para una entidad se nombran distintos en distintas fuentes de datos. Ej. Sexo ⟶ M,F ó 1,2 , etc.

- **Diferencias de escala:** Medidas en una base de datos pueden estar en centímetros, en otra en metros, etc.

- Al hacer matching entre atributos de diferentes fuentes de datos es necesario tener en cuenta la estructura de la información, ej. En una base de datos el descuento se aplicaba sobre un item y en otra el descuento se aplica sobre el total de la orden

Transformación

- Los datos se transforman y consolidan de tal forma de quedar listos para los procesos de minería de datos.
- La transformación puede envolver los siguientes procesos:
  - **Smoothing**: binning, regresión, clustering.
  - **Normalización:** Se modifica la escala de los atributos de tal forma que todos los valores queden dentro de un rango específico. Ej, 0 y 1, -1 y 1.
  - **Construcción de características** (feature construction): Nuevos atributos son construidos desde el mismo set de datos para mejorar el proceso de data mining.

Transformación(Cont..)

Agregación: Roll Up de algunos atributos (ventas mensuales, anuales, etc.)

Generalización: Datos son reemplazados por datos de niveles más altos (jerarquías de concepto). P. Ej, pasar de calles a comunas, ciudades a países, edad a un concepto más alto como "adulto-joven", "senior",etc.

Normalización

Útil en algoritmos que consideran distancias (KNN, clustering, etc.)

Normalización Min-Max: Realiza una transformación lineal en los datos originales. Si los rangos iniciales son *min_A y max_A,* y los rangos finales son *new_min_A* y *new_max_A,* la transformación de un valor v a un valor v' queda:

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

Normalización(Cont..)

Normalización z-score: Los valores para un atributo A son normalizados en base a la media y la desviación estándar:

$$v' = \frac{v - \overline{A}}{\sigma_A}$$

Normalización(Cont..)

Normalización decimal (by decimal scaling)**:** Los valores para un atributo A son normalizados moviendo los puntos decimales:

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

Donde *j* es el menor entero tal que $Max(|v'|) < 1$

Ej: rango inicial -971 a 990 ⟶ -0.971 a 0.990

Construcción de características

Se construyen nuevos atributos a partir de los existentes de tal forma de ayudar al proceso de data mining.
Ejemplo: Caso robo en cajeros automáticos

En algoritmos de clasificación se construyen inicialmente muchas características y luego en base a procesos de selección se dejan las mejores

Reducción de Datos

- A veces la cantidad de información hace que sea impracticable procesar la base de datos
- La idea es reducir la base de datos manteniendo la integridad en un alto porcentaje.
- Los algoritmos de minería de datos deben producir resultados muy similares en la base de datos reducida

Reducción de Datos

- Algunas estrategias de reducción son las siguientes:
  - Agregación del cubo de datos
  - Selección de un subconjunto de atributos
  - Reducción de dimensionalidad
  - Discretización (Bining, generación de rangos)
  - Jerarquías de concepto (Generalización).

Agregación del cubo de datos

- Algunos datos son agregados dependiendo de la información que se desea manejar.

- P. ej. Sólo se desean las ventas anuales, por lo tanto se deben sumar los datos mensuales.

- Se logra una importante reducción de la cantidad de información sin pérdida de datos necesarios para el análisis

Selección de un subconjunto de atributos

- Muchas veces existen atributos en las bases de datos que para cierto análisis son irrelevantes.
- Los patrones descubiertos sobre una cantidad reducida de atributos son más entendibles por el usuario.
- Se necesitan algoritmos eficientes, para **n** atributos existen 2^n posibles subconjuntos de atributos
- Algunos algoritmos conocidos son: Stepwise forward selection, Stepwise backward elimination, Plus R take away R, Exhaustive search, etc.