Diplomado *Business Intelligence*

Clase 3

Pre-Procesamiento de Datos

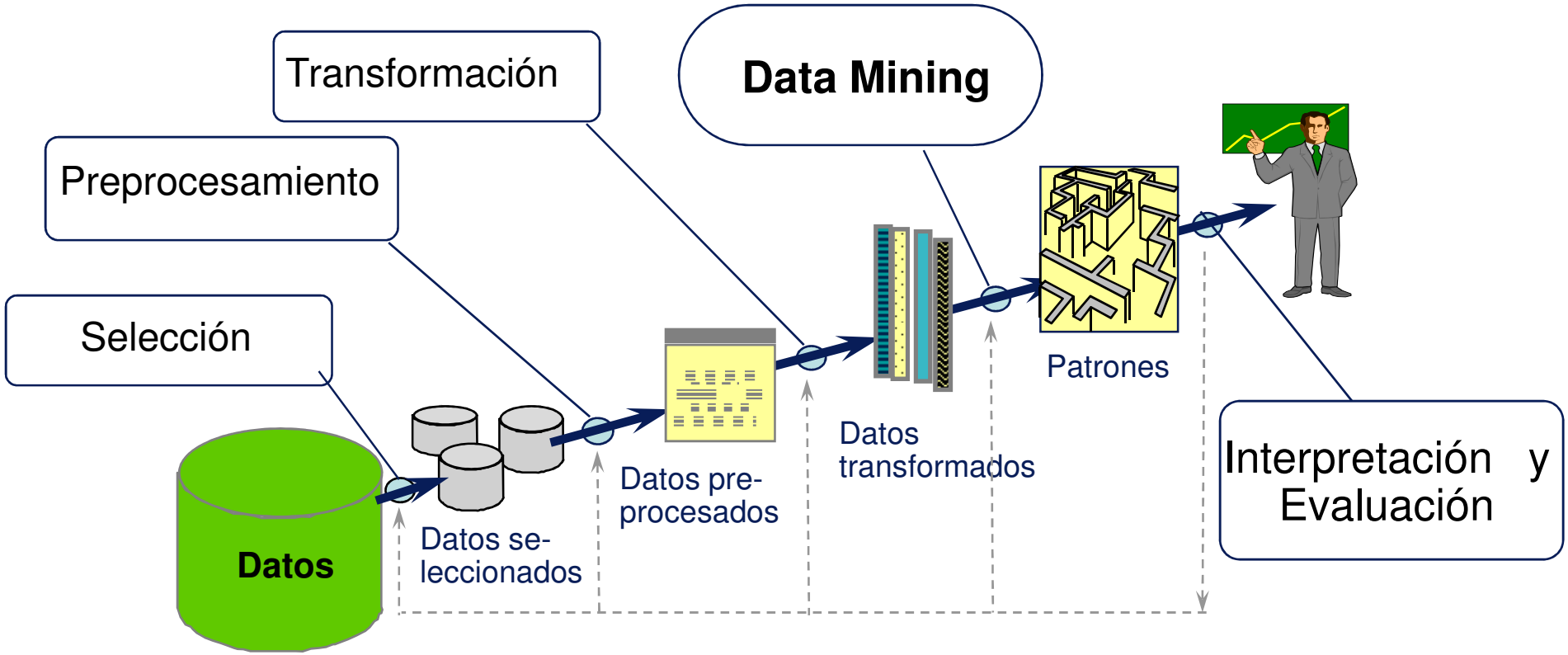
21 de Diciembre 2011

Sebastián Maldonado

Proceso de KDD

Knowledge Discovery in Databases

"KDD es el proceso no-trivial de identificar patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles dentro de los datos"



Nivel de datos

Nivel	Significado	Ejemplo	Operación permitida
Escala nominal	“Nombre” de objetos	número de telef.	comparación
Escala ordinal	“Orden” de objetos (sin distancia)	Notas (1, …, 7)	Transformación monótona
Escala de intervalo	Punto cero y unidad arbitrario	Temp. en grados Cel.	$f(x)=ax + b$ $(a>0)$
Escala de proporción	Dado el punto cero Unidad arbitraria	Peso en kg Ingreso en $	$f(x)=ax$
Escala absoluta	Dado el punto cero y la unidad	Contar objetos número de autos	$f(x)=x$

Limpieza de datos

- Tipos de Datos perdidos (Taxonomía Clásica) [Little and Rubin, 1987]:
  - **Missing Completely at Random (MCAR):**
    - Los valores perdidos no se relacionan con las variables en la base de datos
  - **Missing at Random (MAR):**
    - Los valores perdidos se relacionan con los valores de las otras variables dentro de la base de datos.
  - **Not Missing at Random or Nonignorable (NMAR):**
    - Los valores perdidos dependen del valor de la variable.

Valores Perdidos: la Historia

- La teoría y práctica con valores perdidos:
  - Antes de los 70's y los 70's: Procedimientos particulares para cada caso, no existe teoría. Ej. eliminación de casos (case deletion), single imputation (modelos ad-hoc).
  - Los 80's: Algoritmos basados en estimaciones de máxima verosimilitud, algoritmo EM.
  - Los 90's: Multiple Imputation, Cadenas de Markov (Markov Chain), Monte Carlo, Métodos Bayesianos.

Técnicas Populares de Tratamiento

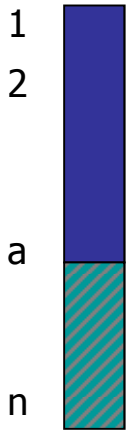
1. Eliminación de datos:
   - Eliminación de Casos (listwise or casewise deletion)
   - Eliminación de pares (o tuplas) de casos (pairwise data delection)

- Donde encontrarlo: La mayoría de paquetes estadísticos, SAS, SPSS, etc.

- Cuando Ocuparlo ➔ MCAR

Técnicas Populares de Imputation

2. Sustitución por la media (mediana y moda):



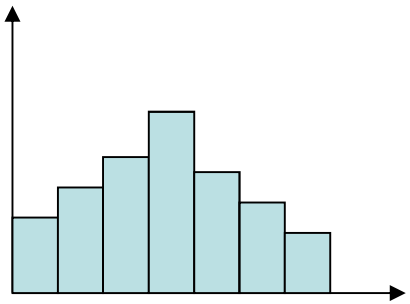
Observados

$y_1, y_2, \ldots, y_a$

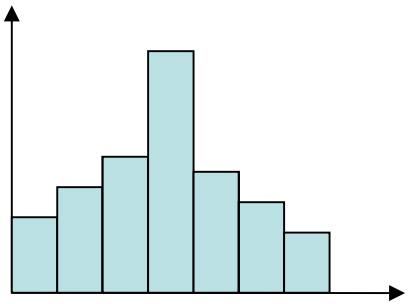
$y_{a+1}, \ldots, y_n$ Perdidos

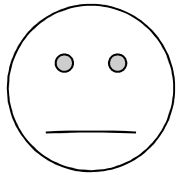
Reemplazar con $y_{obs} = \frac{1}{a}\sum_{i=1}^{a} y_i$

Antes



Después



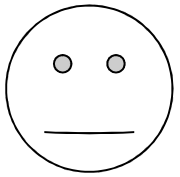


- Corrompe la distribución de Y

Técnicas Populares de Imputation

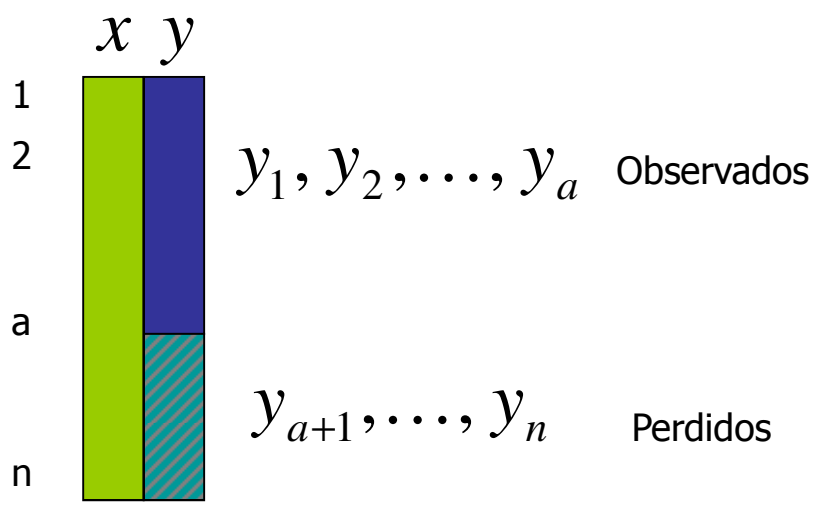
3. Simple Hot Deck:

- Reemplaza los valores perdidos con un valor aleatorio obtenido de la distribución de probabilidades de la variable.
- Preserva la distribución marginal de la variable.
- Distorsiona las correlaciones y covarianzas.

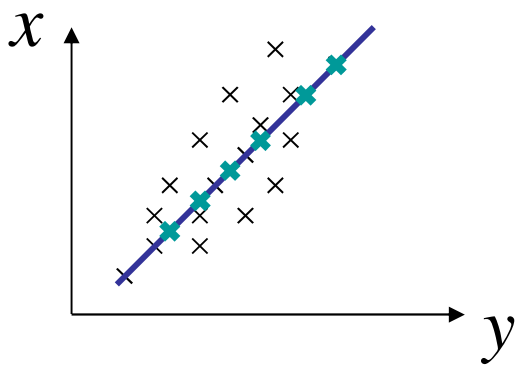


Técnicas Populares de Imputation

4. Métodos de Regresión:



- Reemplazar los valores perdidos con un valor obtenido a través de un modelo de regresión

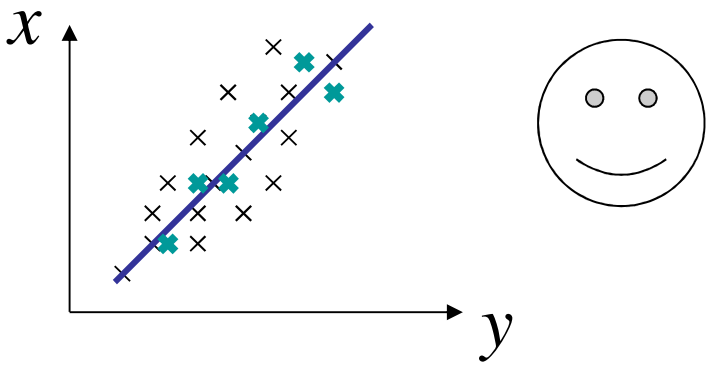


Problema: Esto aumenta las correlaciones

Técnicas Populares de Imputation

4. Métodos de Regresión:

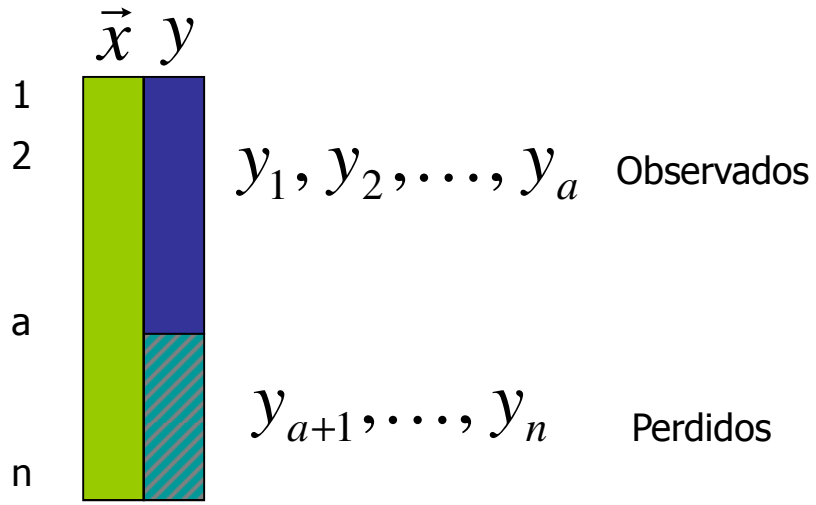
Mejor idea: Reemplazar los valores perdidos con un valor obtenido a través de un modelo de regresión más los residuos de éste



- Se requiere un modelo
- Se asume que los datos perdidos no dependen de los valores de y
- Es difícil de ocupar cuando se tiene que todos los campos presentan valores perdidos.

Técnicas Populares de Imputation

5. Métodos de Árboles de Decisión:



Reemplazar los valores perdidos con un valor obtenido a través de un modelo de Árboles de Decisión

- Se requiere un modelo
- Se asume que los datos perdidos no dependen de los valores de y
- Problemas con datos multivariados y categóricos con más de dos valores.

Técnicas Populares de Imputation

5. El Método EM:

Propósito del Método: Encontrar la distribución subyacente de los datos de muestreo.

Idea General:

- Si se tienen datos suficientes en un atributo, se pueden lograr estimaciones de máxima verosimilitud
- Si se tiene algo de conocimiento del problema entonces se pueden ajustar los parámetros para obtener valores de los datos perdidos ciertos.

Técnicas Populares de Imputation

5. El Método EM:

- Como Funciona (sin fórmulas):
  1. Darle valores a los parámetros del modelo.
  2. Repetir este paso hasta alcanzar el resultado deseado:
     a. Paso Expectation (E): Completar los datos dándole valores a los valores perdidos (dando por conocido el valor de los parámetros.
     b. Paso Maximitation (M): Calcular los mejores parámetros basados en los datos completos.
- Ejemplo, utilizar la distribución normal.
- Tipos de resolución:
  - Suave
  - Fuerte

Técnicas Populares de Imputation

6. Multiple Imputation:

- Está basado en técnicas de simulación (no estadística necesariamente)
- ¿Cómo funciona?:
  - Reemplazar cada una de los valores perdidos con m>1 valores simulados.
  - Se analizan cada uno de los m subconjuntos de la misma forma.
  - Combinar los resultados obtenidos.
- ¿Por qué usarla?
  - Es altamente eficiente con pocos datos y pocas muestras válidas.

Pro y Contra

- A favor:
  - Nos “olvidamos” del problema de los valores perdidos.
  - No descartamos información.

- En contra:
  - La técnicas Imputation alteran los resultados de los modelos.
  - El esfuerzo por encontrar una buena técnica de imputation puede no siempre valer la pena.

Lectura Recomendada:

www.eustat.es/document/datos/ct_04_c.pdf

Capítulo 2

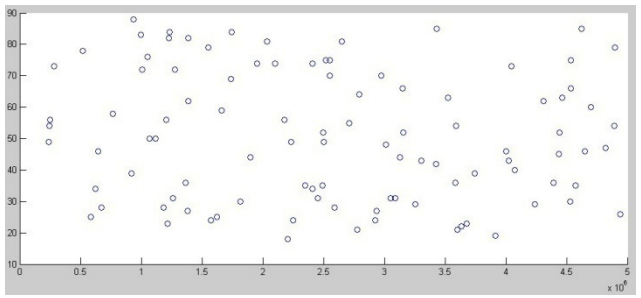
Trasnformación de Datos

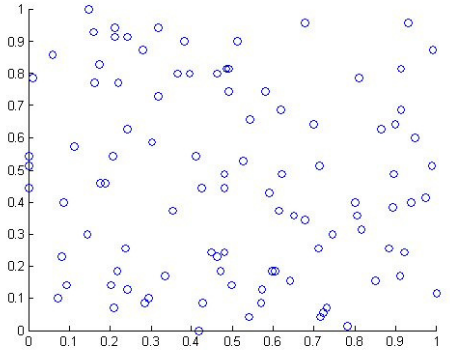
- ¿Qué significa “transformar” los datos?
  - Escalar con funciones matemáticas.
  - “Mapear”.
  - Discretizar.
  - Agregar Datos.

- ¿Para qué transformar los datos?
  - Mejorar la capacidad de discriminación de una variable.
  - Agrupar datos y reducir clases.
  - Dar significado “matemático” a variables.
  - Igualar “pesos relativos” de las variables.

Normalización

- "Aplicar una función matemática a una variable continua para cambiar el rango"
  - Iguala el tratamiento de cada variable.
  - Algunos métodos lo exigen.
- Ejemplo: 100 casos.
  - Ingreso: [100.000, 5.000.000]
  - Edad: [18, 80]

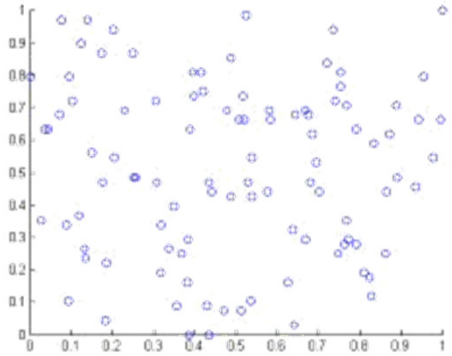




Normalización Escalamiento

- Escalamiento a [0, 1]

$$x_i^* = \frac{x_i - \min(X)}{\max(X) - \min(X)}$$



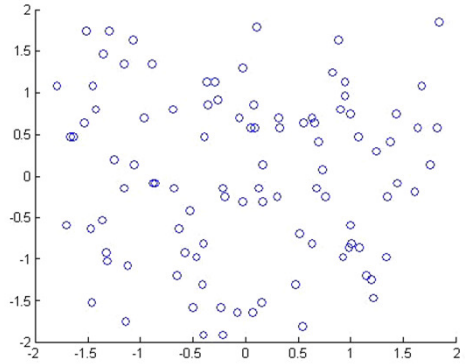
- Ventajas:
  - Sencillez de implementación.
  - Algunos métodos necesitan este tipo.
- Desventajas:
  - ¿Conozco siempre el rango?
  - No considera dispersión.
  - Cuidado con valores fuera de rango

Normalización

Puntaje Z (Variable Normal)

- Ajuste por media/varianza

$$z_i = \frac{x_i - \bar{X}}{\sigma(X)}$$



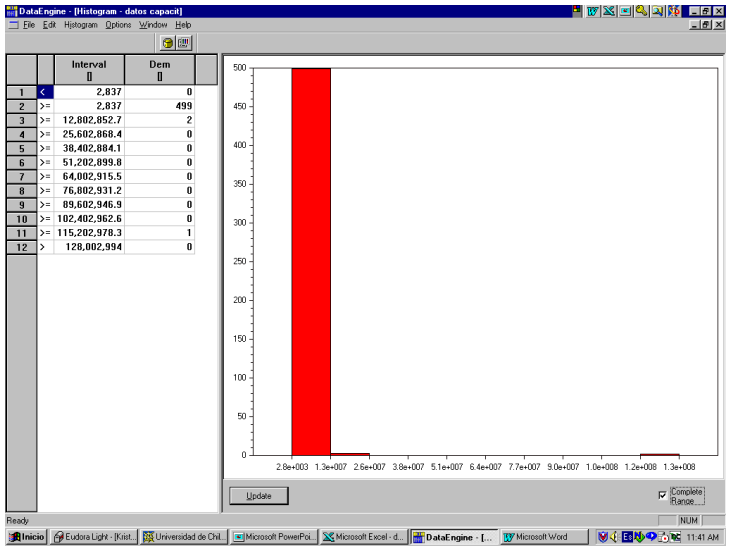
- Ventajas:
  - Considera propiedades estadísticas.
  - Se conoce media y desv. estándar.
- Desventajas:
  - $z \in (-\infty, +\infty)$
  - No todas las variables son normales.

"Mapeo"

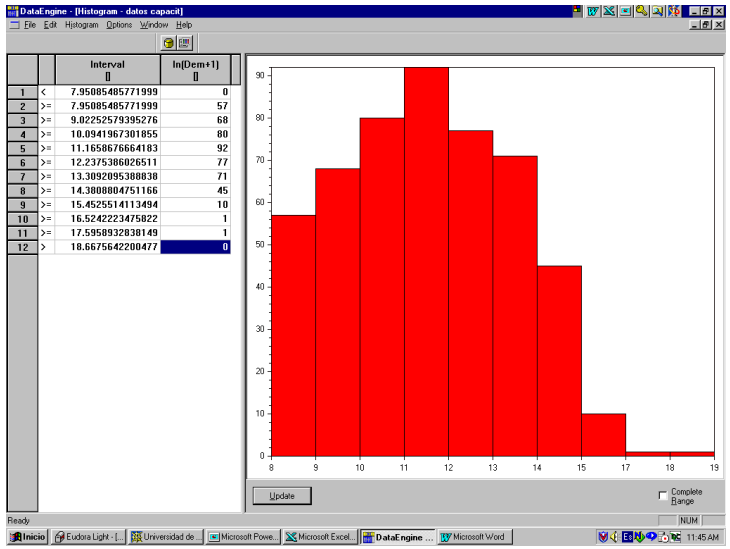
- "To Map": Crear una representación de una cosa en otra.
- Se refiere a transformar los datos a una nueva escala.
- Dos tipos clásicos:
  - A variables ordinales (ordenadas de alguna forma, pero sin distancia exacta).
  - A variables categóricas (sin orden alguno).

- Tipo Especial:
  - Logaritmo: Cuando variable continua concentrada en intervalo => $x' = \log(1 + x)$

Transformación de Atributos



F22, monto demanda
502 demandas, Valparaíso



F22, ln(monto demanda +1)
502 demandas , Valparaíso

"Mapeo"
Variables Ordinales

- Si la variable tiene un orden intrínseco (A – M – B):
  - Es posible asignar valor numérico a cada categoría:
    - A = 5; M = 2; B = 1.
    - "Distancia" entre A y B: d(A,B) = 4;
    - d(A,M)=3 y d(M,B) = 1.

- ¡¡Sólo utilizar si se está muy seguro de la distancia relativa!!
  - Puede "inutilizar" una variable.
  - Puede incorporar relaciones ficticias.

"Mapeo"
Variables Categóricas

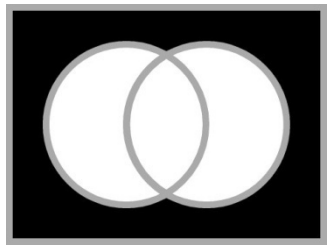
- Si no existe orden o distancia no es clara, utilizar variables "Dummy".
  - N Categorías => N-1 Variables binarias "Dummy".
- Método necesario para utilizar variables categóricas.
- Puede necesitar agregación previa.
- Aumenta el número de variables numéricas en el modelo.

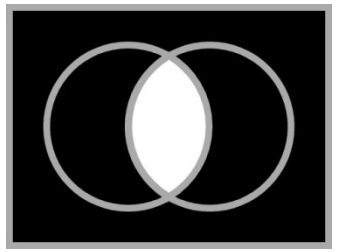
Agregación

- Corresponde a aplicar algún operador a dos o más variables.
- Permite reducir # de variables en la muestra.
- Útil sobre todo en variables categóricas:
  - Reduce variables dummy.
  - Explica mejor resultados (¿es la región importante o grupos de ellas?).
- En continuas, permite incorporar conocimiento del modelador al crear relaciones no triviales.

Agregación
Variables Binarias o Categóricas

- Se utilizan operadores lógicos.
  - O: Si alguna es verdadera.
    - Variable "SI_METROPOLITANA" a partir de "COMUNA".
  - Y: Si todas son verdaderas.
    - Variable "HOMBRE_JOVEN" a partir de "SEXO" y "RANGO_EDAD".
  - XOR: Si una y sólo una es verdadera.
    - Variable "SI_BICOLOR" a partir de "SI_BLANCO" y "SI_NEGRO".





- Regla: "O" a las más desagregadas luego "Y" según objetivo.

Agregación
Variables Continuas

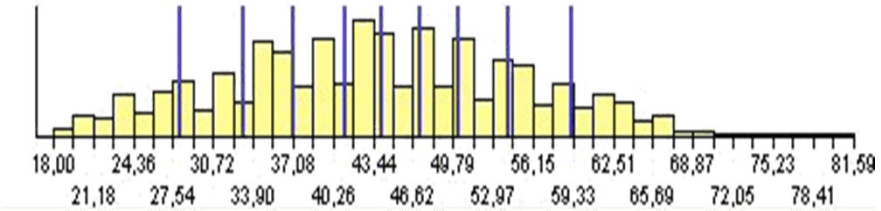
- Se utilizan funciones matemáticas.
  - Proporciones:
    - INGRESO/EDAD.
    - MARGEN = 1 - COSTO/VENTAS.
  - Suma y Resta:
    - COSTO = C_DIRECTO + C_INDIRECTO.
    - UTILIDAD = VENTAS-COSTOS.
  - Potencias (¡Raras!).
- ¡Incorporar el conocimiento!

Discretización

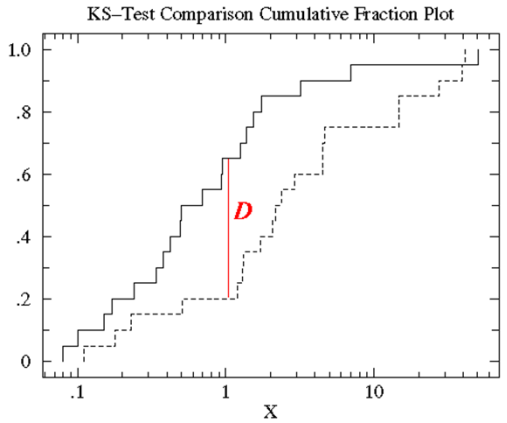
- Transformación de variables continuas a variables categóricas u ordinales.
- Deseable con algunos métodos (árboles de decisión).
- Aumenta número de variables, pero puede dar paso a mejor discriminación.
- ¿Cuándo usar?
  - Variable continua muy concentrada.
  - Variable continua con comportamiento no lineal.

Discretización Métodos

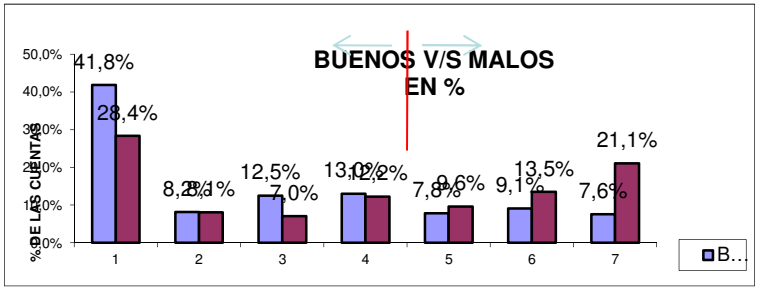
- Según percentiles (x% de la variable en cada grupo).



- Según K-S o Chi2: Elegir el punto de corte que maximiza diferencia. (¡Árboles!)



- Según grupo ad-hoc (no linealidad).



Selección de Atributos

- Primer paso: eliminación de variables altamente concentradas.
  - Porcentaje de valores concentrados en un único valor (95%, 99%)
  - Varianza menor a un cierto umbral (variables estandarizadas)
- Eliminación de variables con un porcentaje muy alto de valores perdidos (por ejemplo, >30%).
  - Dependerá del número de observaciones y de la naturaleza del problema.

Selección de Atributos

- Objetivo: elegir subconjunto de atributos relevantes, eliminando atributos que generan ruido y confunden al método.
- Mejor representación y comprensión del modelo. Entrenamiento más rápido, mejor clasificación.
- Métodos:
  - Filtros (Filters): Seleccionan los atributos en forma independiente del algoritmo de aprendizaje.
  - Envolvente (Wrappers): Evalúan atributos con el algoritmo de acuerdo a su poder predictivo.
  - Empotrados (Embedded): Realizan selección en el entrenamiento del algoritmo.

Filtros

- Correlación entre atributos y variable dependiente
- Relación entre atributo y variable dependiente
  - Test chi-cuadrado para atributos categóricos
  - ANOVA (Analysis of Variance), test KS para atributos numéricos
- Analisis de Componentes Principales

Test Chi-cuadrado: Independencia de dos variables

- Tenemos 2 variables categóricas
- Hipótesis: estas variables son independientes
- Independencia significa: Conocimiento de una de las dos variables no afecta la probabilidad de tomar ciertos valores de la otra variable

Test Chi-cuadrado: Tabla de contingencia

- Tabla de contingencia: matriz con r filas y k columnas, donde

  r=número de valores de variable 1

  k=número de valores de variable 2

Test Chi-cuadrado: Tabla de contingencia

- Ejemplo:
  Variable 1=Edad, variable 2=sexo
  Grado de libertad (degree of freedom):
  df=(r-1)(k-1)

Idea:
Comparar frecuencia esperada con frecuencia observada

Hipótesis nula:
variables son independientes

r=2

	Sexo		
Edad	masculino	femenino	Total
< 30	60	50	110
>= 30	80	10	90
Total	140	60	200

k=2

Test Chi-cuadrado: Test

Frecuencia esperada de una celda fe:

	Sexo		
Edad	masculino	femenino	Total
< 30	60	50	110
>= 30	80	10	90
Total	140	60	200

fe = (fr*fk)/n

con:

fr = frecuencia total en fila r

fk = frecuencia total en columna k

Ejemplo: r=k=1; fr=110; fk=140; n=200

fe = (110*140)/200=77

Test Chi-cuadrado: Frecuencia esperada

Frecuencia esperada vs. observada para todas las celdas:

	Sexo		
Edad	masculino	femenino	Total
< 30	77	33	110
>= 30	63	27	90
Total	140	60	200

	Sexo		
Edad	masculino	femenino	Total
< 30	60	50	110
>= 30	80	10	90
Total	140	60	200

Test Chi-cuadrado

H_0: Edad y sexo son independiente

H_1: Edad y sexo son dependiente (hay una relación entre edad y sexo)

df = 1 = (r-1)*(k-1)

Valor crítico de chi-cuadrado (df=1, α=0,01)=6,63 (ver tabla)

$$\text{Chi-cuadrado} = \sum \frac{(f_o - f_e)^2}{fe} = \frac{(60-77)^2}{77} + \frac{(50-33)^2}{33} + \frac{(80-63)^2}{63} + \frac{(10-27)^2}{27}$$

=27,8 > 6,63 => hay que rechazar H_0=>edad y sexo son dependientes

Correlación de Atributos

- Dos atributos están correlacionados si

$$A_1 \approx \alpha \cdot A_2 + \beta$$

- En modelos estadísticos los atributos deben ser independientes, por lo que no se deben agregar.
- Independientes: El valor de uno no tiene impacto en el valor del otro.
- Se debe eliminar uno sólo de los atributos, para mantener la información.

Filtro: Correlación de Pearson

- Se utiliza el siguiente coeficiente, con rango [-1,1]:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

- Si la correlación es ***crítica*** (|ρ|>0,8-0,9) es recomendable eliminar atributos altamente correlacionados para evitar problemas como multicolinealidad.

Ejemplo 1 - Limpieza

Id	ano_ing	mes_ing	nacion	fec_nac	sexo	region	e_civil	deuda	ingreso	var_obj
1	1999	3	N	23-01-1966	M	1	S	.	4800000	1
2	1999	3	N	16-08-1969	F	5	S	.	3600000	1
3	1999	4	N	23-08-1980	M	6	S	.	2400000	1
4	1999	4	N	05-03-1961	M	4	C	.	3000000	0
5	1999	5	N	27-08-1973	M	13	S	.	7200000	0
6	1999	6	N	01-01-1900	F	.	C	.	12000000	1
7	1999	7	N	11-01-1981	F	15	S	.	3600000	0
8	2000	3	N	25-05-1979	M	9	C	.	4800000	1
9	2000	6	N	02-08-1981	F	8	S	100000	7200000	1
10	2000	8	N	14-01-1981	F	13	S	.	9600000	0
11	2000	9	N	10-05-1976	M	12	C	.	.	0
12	2000	9	N	18-10-1972	M	13	D	.	7800000	0
13	2000	9	N	.	F	4	S	.	4080000	1
14	2000	11	N	29-03-1946	H	5	S	.	3000000	0
15	2001	1	N	18-02-1982	M	10	C	0	4140000	1
16	2001	3	N	01-01-1900	M	11	S	.	9360000	1
17	20001	6	N	08-08-1976	F	8	C	.	9600000	0
18	2002	2	N	.	M	3	V	.	12000000	1
19	2002	4	N	13-07-1981	F	3	S	0	24000000	0
20	2002	5	N	05-05-1973	M	13	C	200000	2760000	1
21	2002	.	N	29-05-1982	M	5	.	.	4800000	0
22	2003	1	N	05-12-1982	F	6	S	.	8136000	0
23	.	.	.	.	.	.	.	.	.	.
24	2003	6	N	03-05-1973	M	8	C	.	1479600	1

MCAR

MCAR

NMAR

MCAR

NMAR

NMAR

MAR

MCAR

Ejemplo 1 - Selección Atributos

Id	ano_ing	mes_ing	nacion	fec_nac	sexo	region	e_civil	ingreso	var_obj
1	1999	3	N	23-01-1966	M	1	S	4800000	1
2	1999	3	N	16-08-1969	F	5	S	3600000	1
3	1999	4	N	23-08-1980	M	6	S	2400000	1
4	1999	4	N	05-03-1961	M	4	C	3000000	0
5	1999	5	N	27-08-1973	M	13	S	7200000	0
6	1999	6	N	14-08-1981	F	13	C	12000000	1
7	1999	7	N	11-01-1981	F	15	S	3600000	0
8	2000	3	N	25-05-1979	M	9	C	4800000	1
9	2000	6	N	02-08-1981	F	8	S	7200000	1
10	2000	8	N	14-01-1981	F	13	S	9600000	0
11	2000	9	N	10-05-1976	M	12	C	6788891	0
12	2000	9	N	18-10-1972	M	13	D	7800000	0
13	2000	9	N	22-04-1977	F	4	S	4080000	1
14	2000	11	N	29-03-1946	M	5	S	3000000	0
15	2001	1	N	18-02-1982	M	10	C	4140000	1
16	2001	3	N	14-01-1975	M	11	S	9360000	1
17	2001	6	N	08-08-1976	F	8	C	9600000	0
18	2002	2	N	12-12-1979	M	3	V	12000000	1
19	2002	4	N	13-07-1981	F	3	S	24000000	0
20	2002	5	N	05-05-1973	M	13	C	2760000	1
21	2002	7	N	29-05-1982	M	5	S	4800000	0
22	2003	1	N	05-12-1982	F	6	S	8136000	0
24	2003	6	N	03-05-1973	M	8	C	1479600	1

ID!!!! **100% concentrada**

Ejemplo 1 - Transformación

ano_ing	mes_ing	fec_nac	sexo	region	e_civil	ingreso	var_obj
1999	3	23-01-1966	M	1	S	4800000	1
1999	3	16-08-1969	F	5	S	3600000	1
1999	4	23-08-1980	M	6	S	2400000	1
1999	4	05-03-1961	M	4	C	3000000	0
1999	5	27-08-1973	M	13	S	7200000	0
1999	6	14-08-1981	F	13	C	12000000	1
1999	7	11-01-1981	F	15	S	3600000	0
2000	3	25-05-1979	M	9	C	4800000	1
2000	6	02-08-1981	F	8	S	7200000	1
2000	8	14-01-1981	F	13	S	9600000	0
2000	9	10-05-1976	M	12	C	6788891	0
2000	9	18-10-1972	M	13	D	7800000	0
2000	9	22-04-1977	F	4	S	4080000	1
2000	11	29-03-1946	M	5	S	3000000	0
2001	1	18-02-1982	M	10	C	4140000	1
2001	3	14-01-1975	M	11	S	9360000	1
2001	6	08-08-1976	F	8	C	9600000	0
2002	2	12-12-1979	M	3	V	12000000	1
2002	4	13-07-1981	F	3	S	24000000	0
2002	5	05-05-1973	M	13	C	2760000	1
2002	7	29-05-1982	M	5	S	4800000	0
2003	1	05-12-1982	F	6	S	8136000	0
2003	6	03-05-1973	M	8	C	1479600	1

Antigüedad **Edad** **0 y 1** **agrupar binarizar** **agrupar binarizar** **ln(X)**