



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Módulo **Minería de Datos** Diplomado

Por
Elizabeth León Guzmán, Ph.D.
Profesora
Ingeniería de Sistemas
Grupo de Investigación MIDAS

Preprocesamiento de Datos



Preparación de datos

“El propósito fundamental de la preparación de los datos es la manipulación y transformación de los datos sin refinar para que la información contenida en el conjunto de datos pueda ser descubierta o estar accesible de forma más fácil”

D. Pyle, 1999, pp. 90

Preprocesamiento de datos

- Datos malos -> extracción de patrones/reglas malas (poco útiles):
 - **Datos Incompletos**
 - **Datos con Ruido**
 - **Datos inconsistentes**
 - **Datos duplicados**

Preprocesamiento de datos

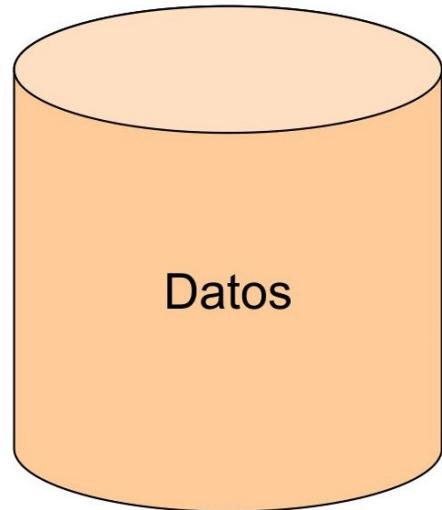
- Datos de **calidad**-> posible generación de patrones/reglas de calidad
 - **Recuperar información incompleta**
 - **Eliminar outliers**
 - **Resolver conflictos**
- Decisiones de calidad deben ser basadas en datos de buena calidad.

Preprocesamiento de datos

- **Reducción** del tamaño del conjunto de datos -> posible mejora de la eficiencia del proceso de Minería de Datos
- **Selección de datos relevantes:** **eliminando** registros duplicados, eliminando anomalías...
- **Reducción de Datos:** **Selección de** características, muestreo o selección de instancias, discretización.

Hecho: La preparación de datos (limpieza, transformación,... puede llevar la mayor parte del tiempo de trabajo (hasta un 90%).

Componentes de la Preparación de Datos



Limpieza

Integración

Transformación

Reducción

Limpieza de Datos

- Resuelve redundancias
- Chequea y resuelve problemas de ruido, valores perdidos, elimina outliers,...
- Resuelve inconsistencias/conflictos entre datos

Limpieza de Datos: Valores Perdidos

- Existen muchos datos que no contienen todos los valores para las variables.
 - Inferirlos
 - Ignorarlos
- Ignorarlos: No usar los registros con valores perdidos
 - Ventaja: Es una solución fácil.
 - **Desventajas:**
 - Perdida de mucha información disponible en esos registros.
 - No es efectiva cuando el porcentaje de valores perdidos por variable es grande.

Limpieza de Datos: Valores Perdidos

■ Remplazarlos:

- Constante global (altamente dependiente de la aplicación)
- Media del atributo
- Media del atributo for la clase dada (problemas de clasificación)

Possible interpretación:

Valores perdidos → “No importa”

Generar ejemplos u objetos artificiales con los valores del dominio del atributo faltante.

Ej: $X = \{ 1, ?, 3 \}$ generar ejemplos artificiales con el dominio del atributo $[0,1,2,3,4]$

$X_1 = \{ 1, 0, 3 \}, X_2 = \{ 1, 1, 3 \}, X_3 = \{ 1, 2, 3 \}, X_4 = \{ 1, 3, 3 \}, X_5 = \{ 1, 4, 3 \}$

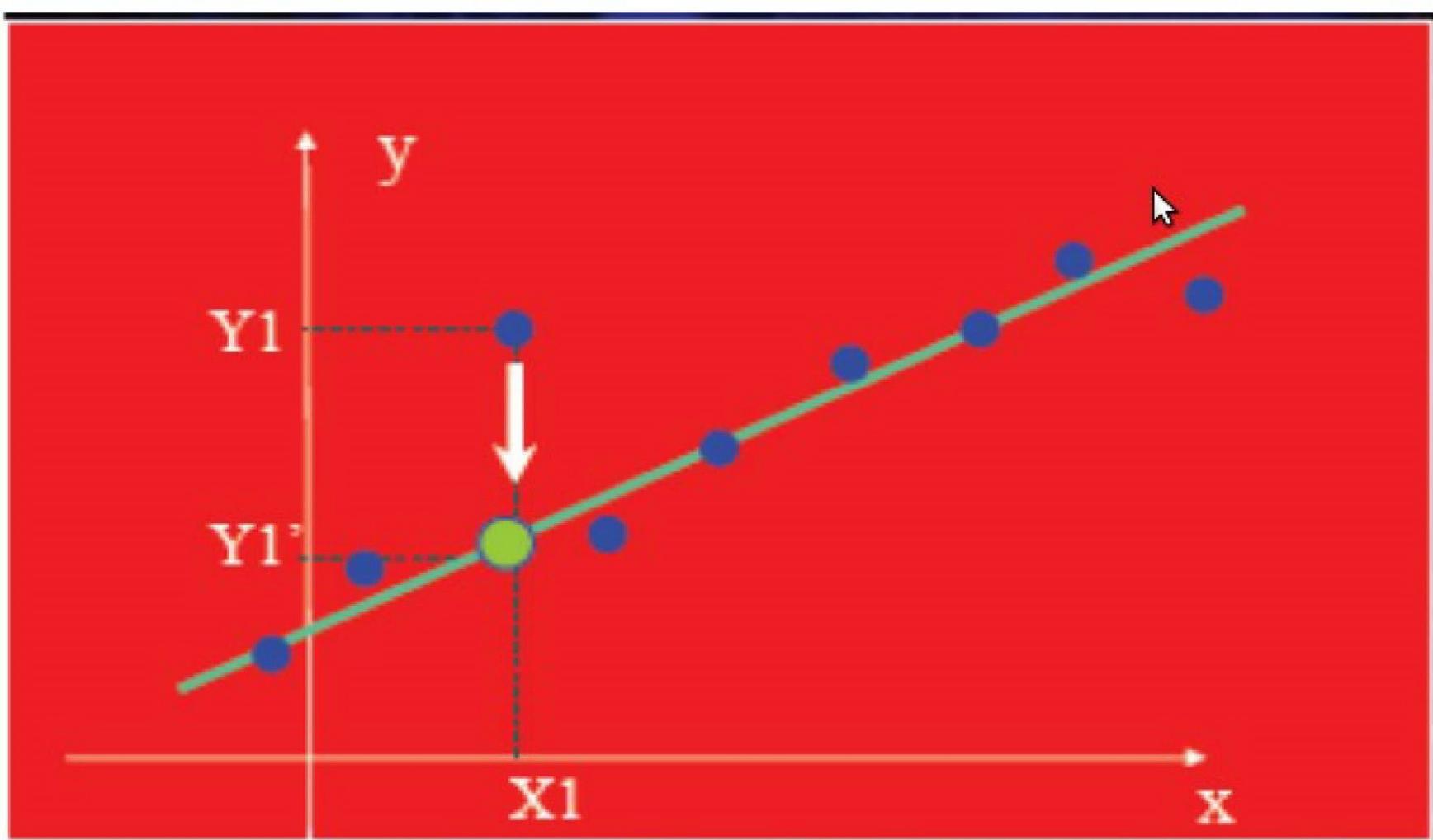
Limpieza de Datos: Valores Perdidos

Técnicas:

Regresion
Bayes,
Agrupación,
Aboles de decisión

Limpieza de Datos: Ruido

- Suavizamiento (Smoothing):



Limpieza de Datos: Detección “Outliers”

- Un atributo: encontrar mean y variance
Umbral = media +- 2 variance
- Basado en distancia: Multidimensional
 - Los ejemplos que no tienen vecinos son considerados “outliers”

Integración de Datos

- Obtiene los datos de diferentes fuentes de Información
- Resuelve problemas de representación y codificación
- Integra los datos desde diferentes tablas para crear información homogénea, ...

Integración de Datos

- Diferentes escalas:
 - Pesos vs Dolares
- Atributos derivados
 - Salario Mensual vs Salario Anual
- Solución
 - Procedimientos semiautomáticos
 - ETL
 - Minería

Transformación de Datos

- Los datos son transformados o consolidados de forma apropiada para la extracción de información. Diferentes vías:
 - Sumarización de datos
 - Operaciones de agregación, etc.

Bibliografía:

T. Y. Lin. Attribute Transformation for Data Mining I: Theoretical Explorations.
International Journal of Intelligent Systems 17, 213-222, 2002.

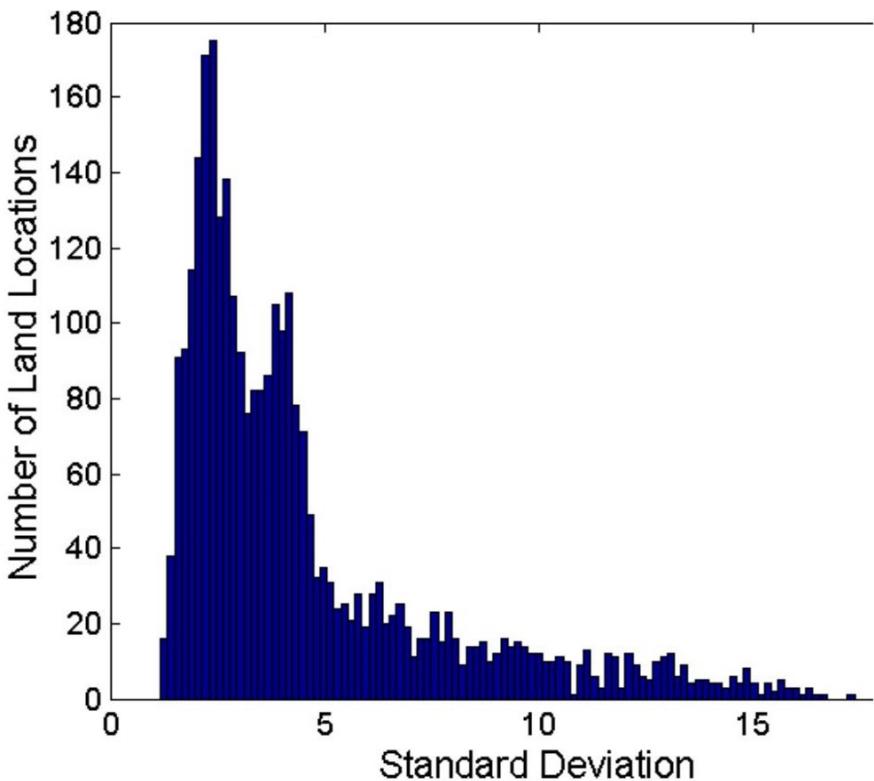
Agregación

- La combinación de dos o más atributos (u objetos) en un solo atributo (u objeto) propósito reducción de datos
- Reducir el número de atributos u objetos
 - Ciudades agregan en regiones, estados, países, etc
- Los datos agregados tiende a tener una menor variabilidad

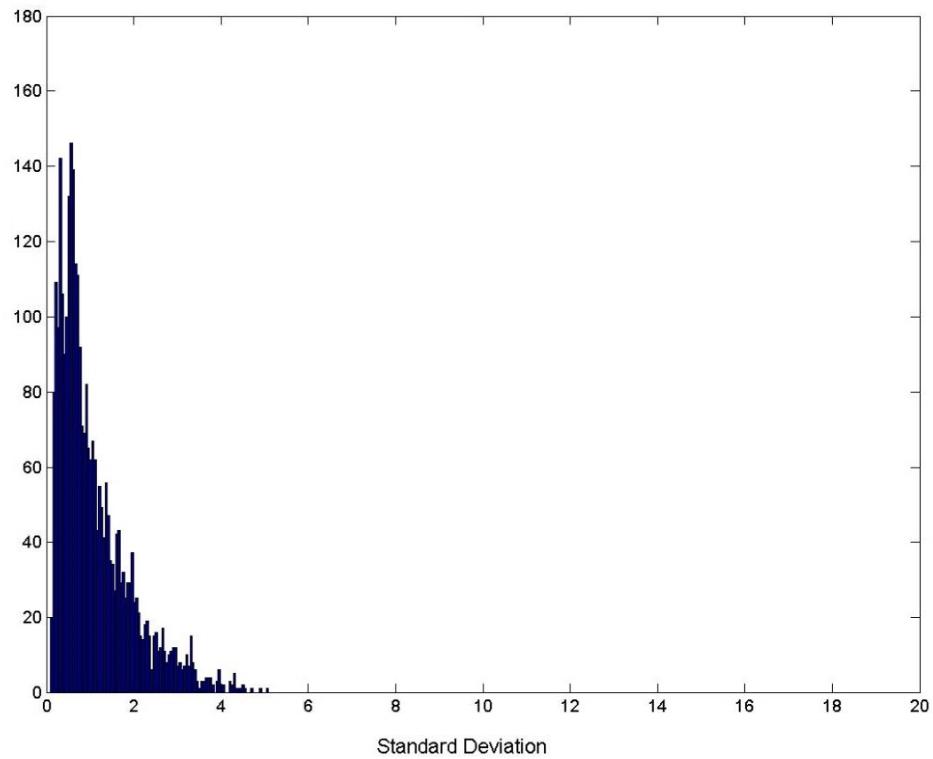
Ing. Elizabeth León
Guzmán PH.D

Agregación

Variación de la precipitación en Australia



Desviación estándar de la
precipitación mensual
promedio



Desviación estándar de la
precipitación media anual

Transformación de Datos: Normalización

- Normalización min-max

$$v' = \frac{v - min_A}{max_A - min_A} (new_max_A - new_min_A) + new_min_A$$

- Normalización z-score

$$v' = \frac{v - mean_A}{stand_dev_A}$$

Transformación de Datos: Normalización

- Normalización por escala decimal

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

- Donde j es el entero más pequeño tal que
 $\text{Max}(|v'|) < 1$

Reducción de Datos

- Discretización
- Selección de Instancias (objetos)
- Selección de características

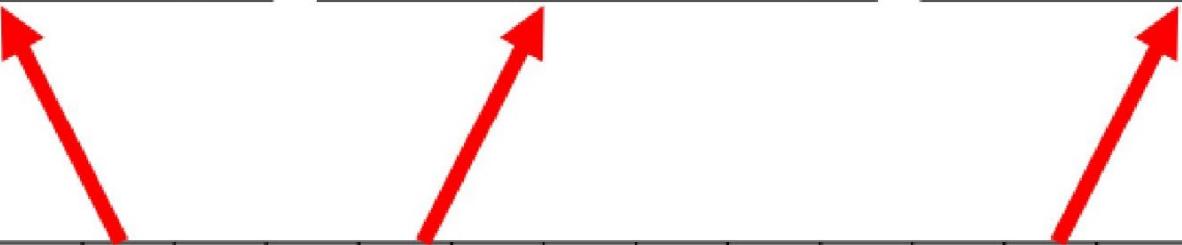
Reducción de Datos: Discretización

- Divide el rango de atributos continuos en Intervalos
- Almacena solo las etiquetas de los intervalos
- Importante para reglas de asociación y clasificación, algunos algoritmos solo aceptan datos discretos.

Reducción de Datos: Discretización

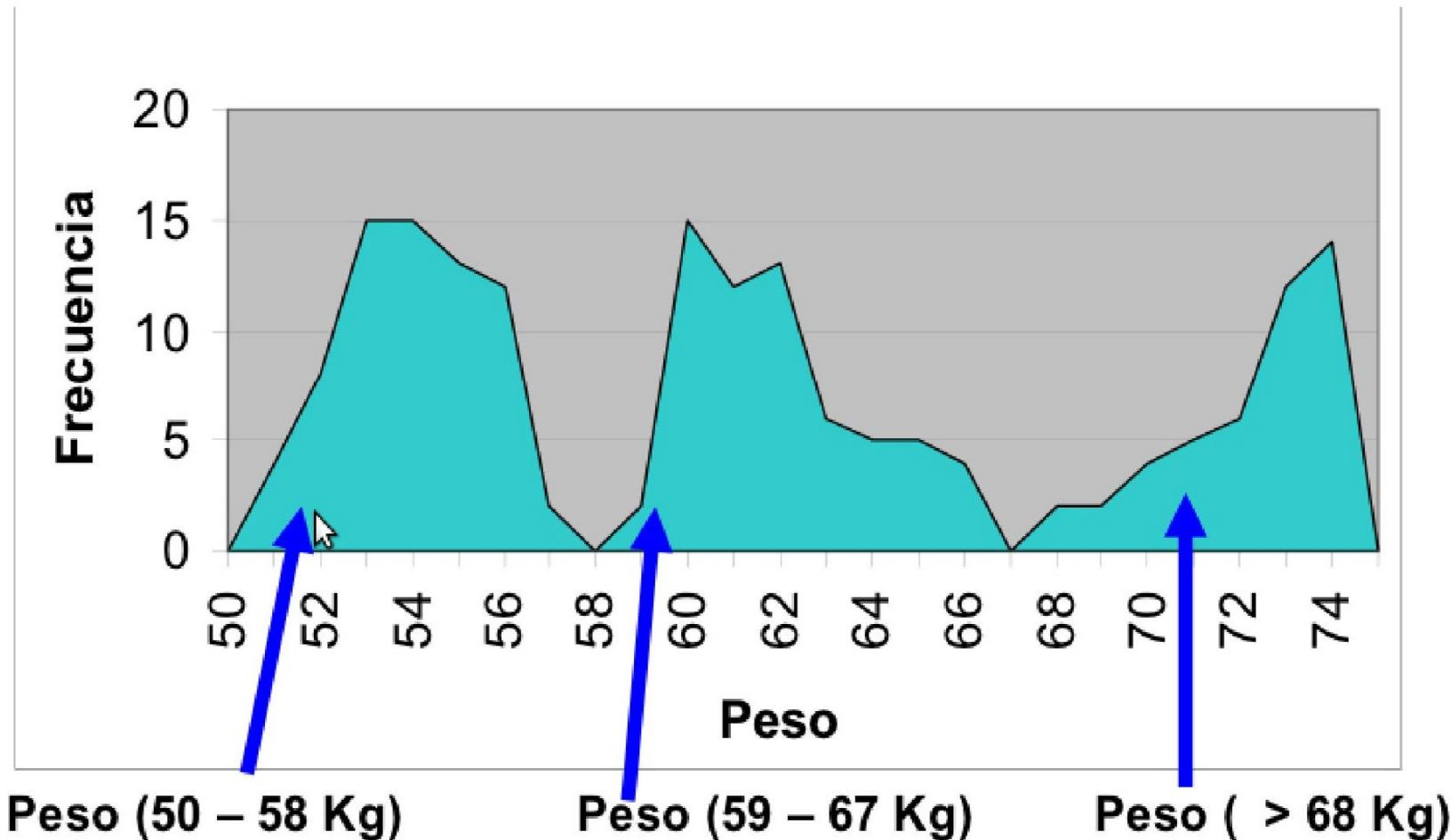
- Ejemplo:

	Edad (0-14 años)			Edad (16-24 años)			Edad (25 - 65 años)									
Edad	1	1	...	1	2	2	2	...	2	2	2	3	4	5	...	6
# de Autos	0	6	...	5	0	3	6	...	5	6	9	1	3	9	...	5



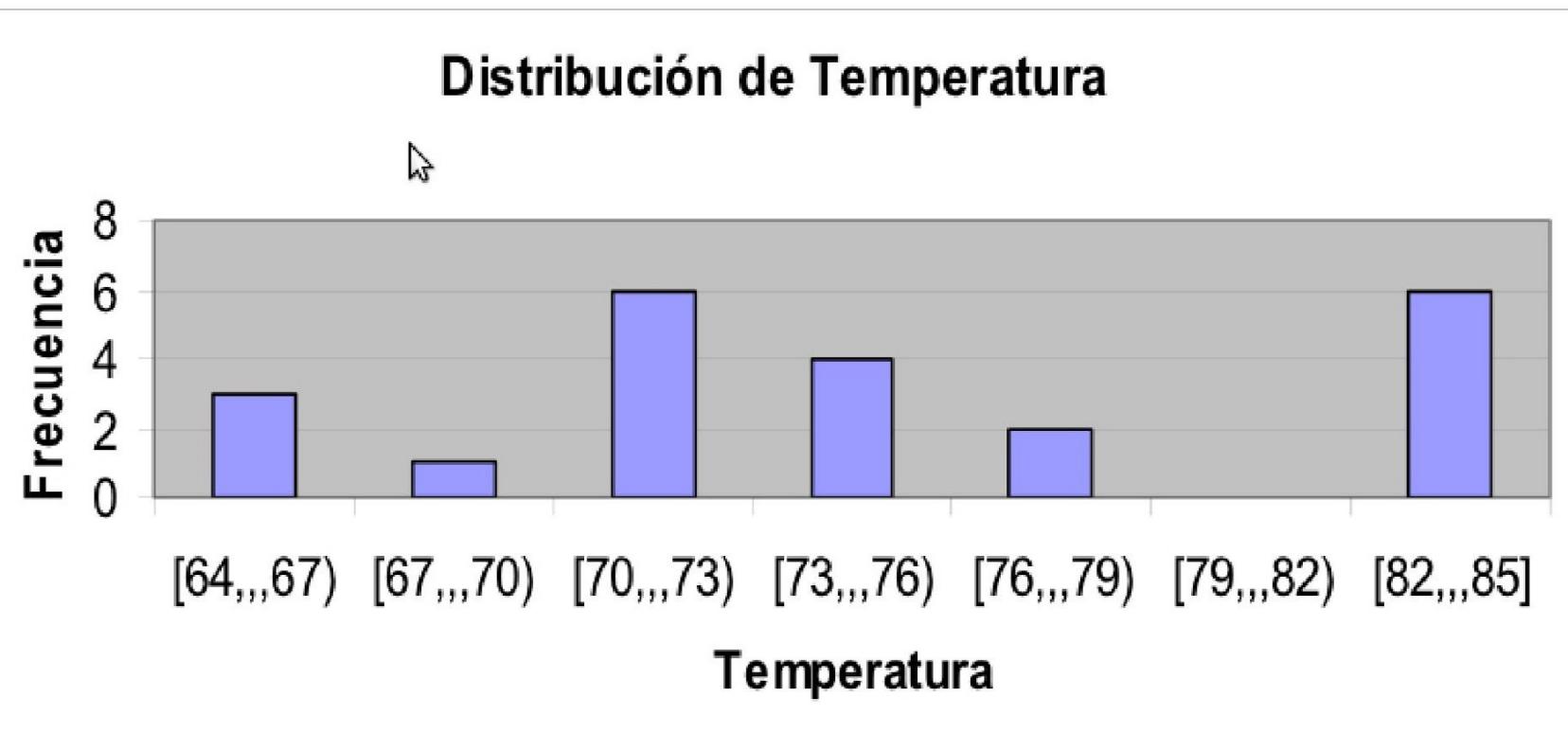
Reducción de Datos: Discretización

■ Distribución de Peso



Reducción de Datos: Discretización

- Igual amplitud

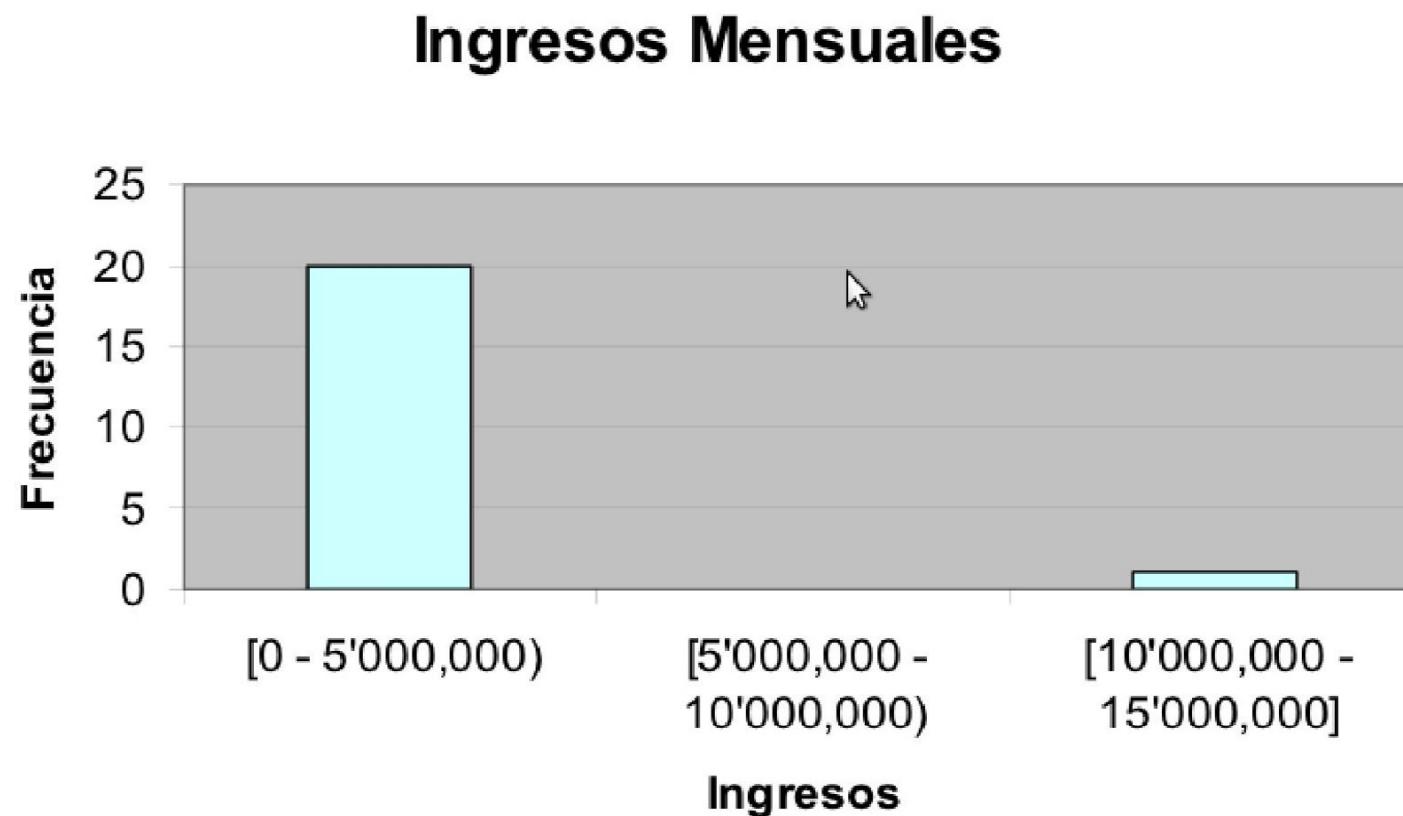


Valores de Temperatura:

63, 65, 66, 67, 70, 70, 71, 71, 72, 72, 73, 73, 74, 75, 76, 76, 82, 82, 83, 84, 85, 85

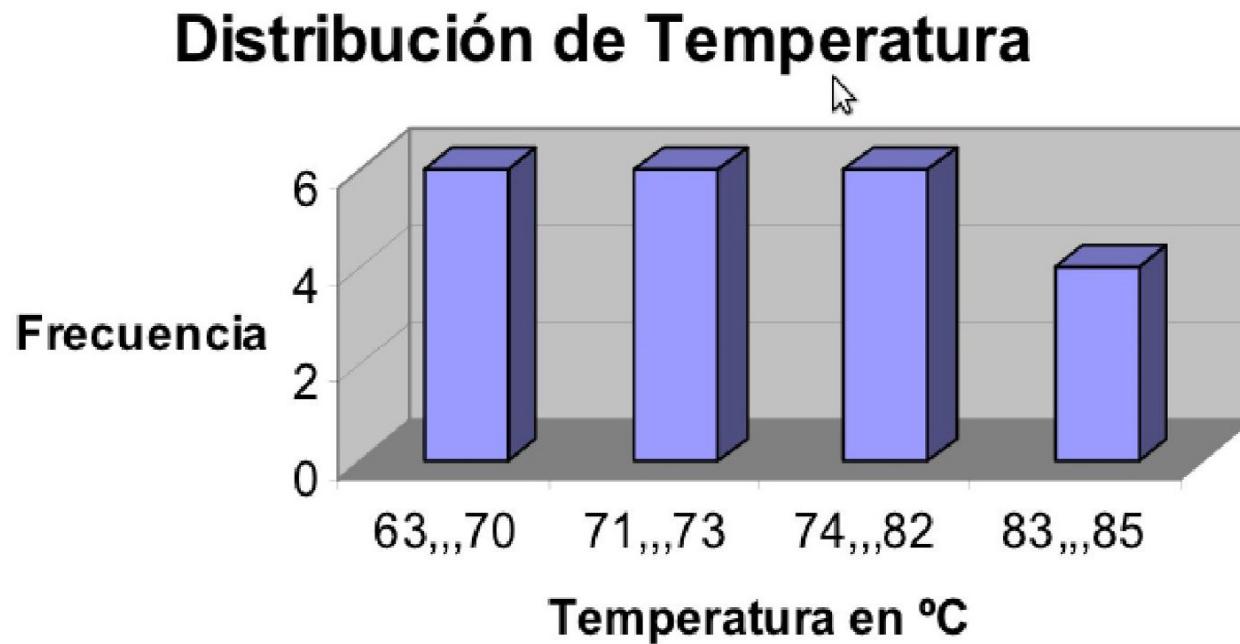
Reducción de Datos: Discretización

- Problemas Igual Amplitud



Reducción de Datos: Discretización

- Igual Frecuencia



Valores de Temperatura:

63, 65, 66, 67, 70, 70, 71, 71, 72, 72, 73, 73, 74, 75, 76, 76, 82, 82, 83, 84, 85, 85

Reducción de Datos: Discretización

- Ventajas de la igualdad en frecuencia
 - Evita desequilibrios en el balance o entre valores.
 - En la práctica permite obtener puntos de corte mas intuitivos.
- Consideraciones adicionales:
 - Se deben crear cajas para valores especiales
 - Se deben tener puntos de corte interpretables

Reducción de Datos: Discretización - BIN

- Valores numéricos que pueden ser ordenados de menor a mayor.
- Partitionar en grupos con valores cercanos
- Cada grupo es representado por un simple valor (media, la mediana o la moda).
- Cuando el numero de bins es pequeño, el limite mas cercano puede ser usado para representar el bin.

BIN

Ejemplo:

$$f = \{3, 2, 1, 5, 4, 3, 1, 7, 5, 3\}$$

ordenado:

$$F = \{1, 1, 2, 3, 3, 3, 4, 5, 5, 7\}$$

particionando en 3 **BINs**:

$$\{1, 1, 2, \quad 3, 3, 3, \quad 4, 5, 5, 7\}$$

representacion usando la moda:

$$\{1, 1, 1, \quad 3, 3, 3, \quad 5, 5, 5, 5\}$$

BIN

usando media:

$$\{1.33, 1.33, 1.33, \quad 3, 3, 3, \quad 5.25, 5.25, 5.25, 5.25\}$$

Remplazando por el limite mas cercano:

$$\{1, 1, 2, \quad 3, 3, 3, \quad 4, 4, 4, 7\}$$

Problema de **optimización** en la selección de **k bins**,
dado el numero de bins k: distribuir los valores en los
bins para **minimizar la distancia promedio** entre un
valor y la media o mediana del bin.

Reducción de Datos: Discretización - BIN

Algoritmo

1. Ordenar valores
2. Asignar aproximadamente igual numero de valores (v_i) a cada bin (el numero de bins es parámetro).
3. Mover al borde el elemento v_i de un bin al siguiente (o previo) si la distancia de error (ER) es reducida. (ER es la suma de todas las distancias de cada v_i a la media o moda asignada al bin).

Reducción de Datos: Discretización - BIN

- Ejemplo:

$$f = \{5, 1, 8, 2, 2, 9, 2, 1, 8, 6\}$$

- Partitionar en 3 bins. Los bins deben ser representados por sus modas

1. f ordenado = {1, 1, 2, 2, 2, 5, 6, 8, 8, 9}
2. Bins iniciales = {1, 1, 2, 2, 2, 5, 6, 8, 8, 9}
3. Modas = {1, 2, 8}
4. Total ER = $0+0+1+0+0+3+2+0+0+1 = 7$

- Después de mover dos elementos del bin2 al bin1, y un elemento del bin3 al bin2

$$\begin{aligned}f &= \{1, 1, 2, 2, 2, 5, 6, 8, 8, 9\} \\ \text{Modas} &= \{2, 5, 8\} \\ \text{ER} &= 4\end{aligned}$$

- Cualquier movimiento de elementos incrementa ER

Reducción de instancias: Muestreo

- El muestreo es la principal técnica empleada para la selección de datos.
A menudo se utiliza tanto para la investigación preliminar de los datos y el análisis de datos final.
- El muestreo se utiliza en la minería de datos, ya que el procesamiento de todo el conjunto de datos de interés es demasiado caro o consume tiempo.

Muestreo...

El principio fundamental para el muestreo efectivo es la siguiente:

- Utilizar una muestra de que funcionan tan bien como el uso de los conjuntos de datos completos -> **muestra es representativa**
- Una muestra es representativa si se tiene aproximadamente la **misma propiedad** (de interés) como el conjunto original de datos

Tipos de Muestreo

□ **Muestreo aleatorio simple**

Existe la misma probabilidad de seleccionar cualquier elemento en particular

□ **Muestreo sin reemplazo**

A medida que cada elemento está seleccionado, se elimina de la población

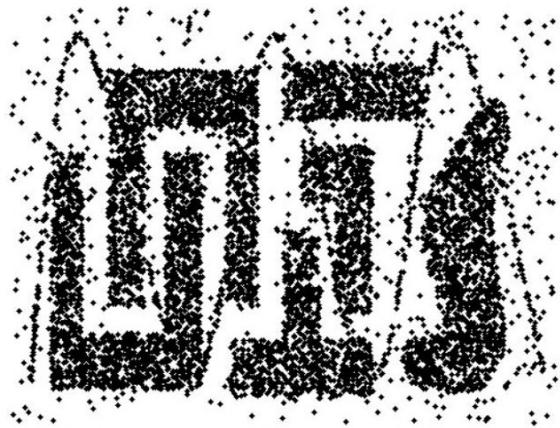
□ **El muestreo con reemplazo**

- Los objetos no se eliminan de la población, ya que son seleccionados para la muestra.
- En el muestreo con reemplazo, el mismo objeto puede ser recogido más de una vez

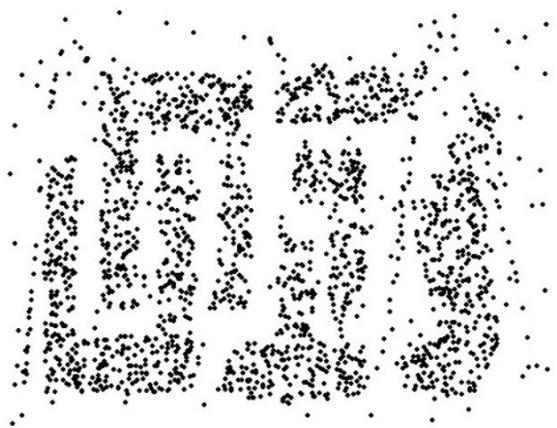
□ **El muestreo estratificado**

Dividir los datos en varias particiones, a continuación, tomar muestras al azar de cada partición

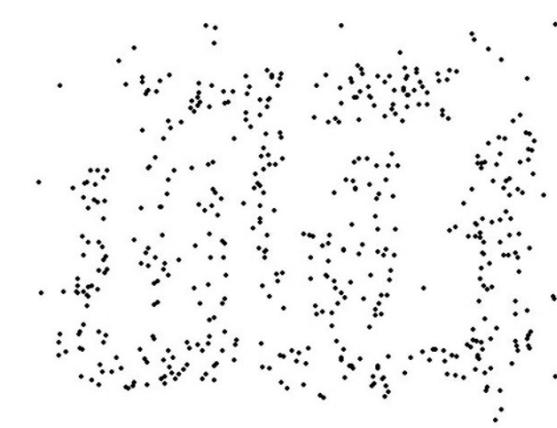
Tamaño de la muestra



8000 puntos



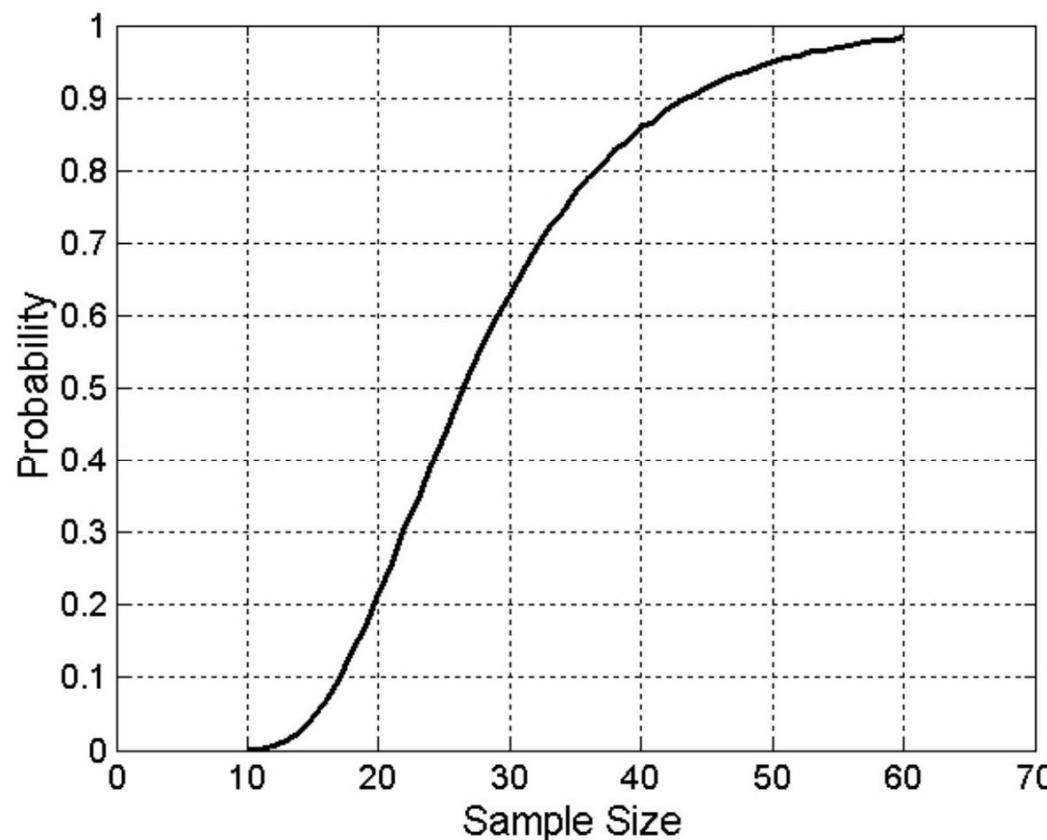
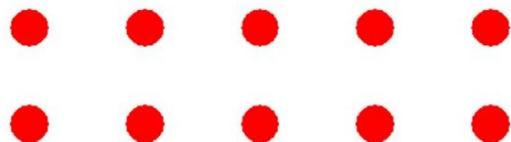
2000 Puntos



500 Puntos

Tamaño de la muestra

¿Qué tamaño de la muestra es necesario para conseguir al menos un objeto de cada uno de 10 grupos.



Selección de características subconjunto

Otra manera de reducir la dimensionalidad de los datos

- Características redundantes: Duplican gran parte o todo de la información contenida en uno o más otros atributos
Ejemplo: edad y fecha de nacimiento
- Características irrelevantes: No contienen información que es útil para la tarea de minería de datos
Ejemplo: Identificación de los alumnos suele ser irrelevante para la tarea de predecir perfil

Selección de características

1. Según la evaluación: filter wrapper	2. Disponibilidad de la clase: Supervisados No supervisado
3. Según la búsqueda: Completa $O(2^N)$ Heurística $O(N^2)$ Aleatoria ¿?	4. Según la salida del algoritmo: Ranking Subconjunto de atributos

Entropía

Distribución de las similaridades es una característica de la organización y orden de los datos en el espacio de n-dimensiones

Criterio para excluir dimensiones: cambios en el nivel del orden en los datos

Cambios medidos con **entropía**

Entropía es una medida global que es menor para configuraciones ordenadas y grande para configuraciones desordenadas

Entropía

Compara la entropía antes y después de remover una dimensión

Si las medidas son cercanas, el conjunto de datos reducido aproxima el original conjunto de datos

Entropía:

$$E = - \sum_{i=1}^{N-1} \sum_{j=i+1}^N \left([S_{ij}] \times \log(S_{ij})) + ((1-S_{ij}) \times \log(1-S_{ij})) \right)$$

Similaridad entre x_i y x_j

Entropía

El algoritmo esta basado en “sequential backward ranking”

La entropia es calculada en cada iteracion para decidir el “ranking” de las dimensiones.

Las dimensiones son gradualmente removidas

Entropía (Algoritmo)

- Comienza con todo el conjunto de datos F
- $E_F =$ entropia de F
- Por cada dimensión $f \in F$,
 - Remover una dimensión f de F y obtener el subconjunto F_f
 - $E_{Ff} =$ entropia de F_f
 - Si $(E_F - E_{Ff})$ es mínima
 - Actualizar el conjunto de datos $F = F - f$
 - f es colocada en la lista “rankeada”
- Repetir 2-3 hasta que solo haya una dimensión en F

Entropía (Algoritmo)

El proceso puede ser parado en cualquier iteración y las dimensiones son seleccionadas de la lista.

Desventaja: complejidad
Implementación paralela

Entropía

Para enumerar dimensiones (ranking)

Basado en la medida de similaridad (inversa a la distancia)

$$S_{ij} = e^{-\alpha D_{ij}} \quad \text{where } D_{ij} \quad \text{es la distancia} \quad \alpha = -(\ln 0.5)/D$$

$$S_{ij} = \left(\sum_{k=1}^n |x_{ik} = x_{jk}| \right) / n \quad \text{Hamming similarity (variables nominales)}$$

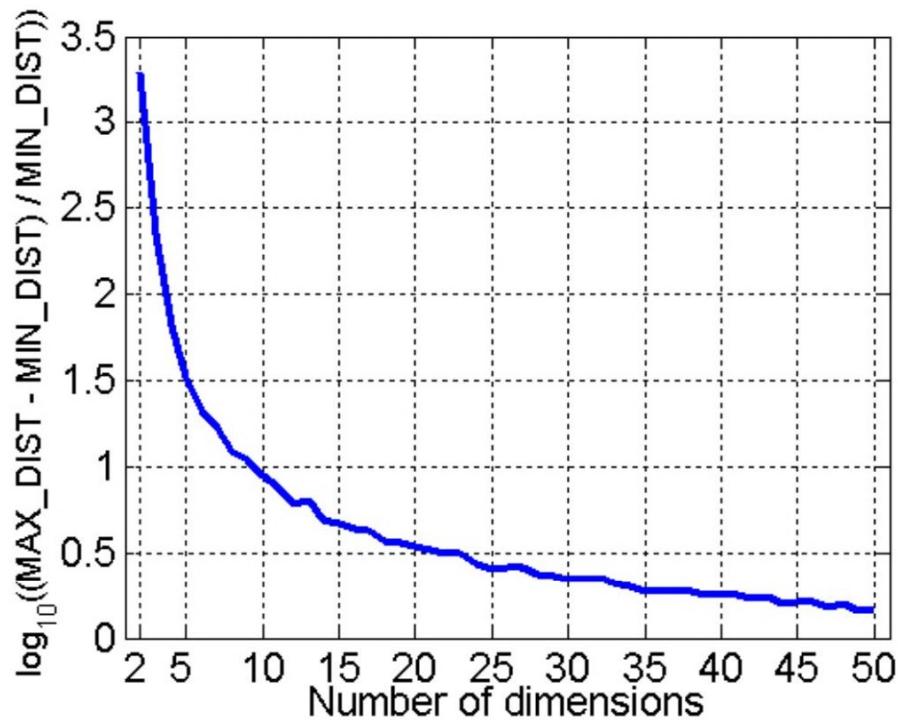
	F1	F2	F3		R1	R2	R3	R4	R5	
R1	A	X	1		R1		0/3	0/3	2/3	0/3
R2	B	Y	2		R2		2/3	2/3	0/3	
R3	C	Y	2		R3			0/3	1/3	
R4	B	X	1		R4				0/3	
R5	C	Z	3							

similaridades

La maldición de la dimensionalidad

Cuando aumenta la dimensionalidad, los datos se vuelven cada vez escasa en el espacio que ocupa

Las definiciones de la densidad y la distancia entre los puntos, lo cual es fundamental para el agrupamiento y la detección de las demás, pierden importancia



- Generar aleatoriamente 500 puntos
Calcular la diferencia entre la máxima y distancia mínima entre cualquier par de puntos

Reducción de dimensionalidad

Propósito:

- Evitar la maldición de la dimensionalidad
- Reduzca la cantidad de tiempo y memoria requeridos por los algoritmos de minería de datos
- Permitir datos a ser más fácil de visualizar
- Puede ayudar a eliminar las características irrelevantes o reducir el ruido

Técnicas

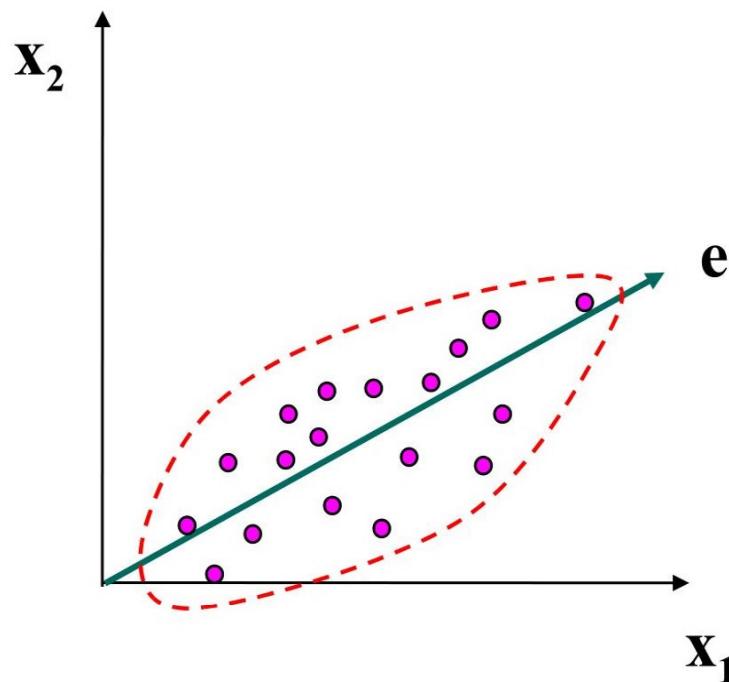
Principio de Análisis de Componentes

Descomposición de valor singular

Otros: técnicas supervisadas y no lineales

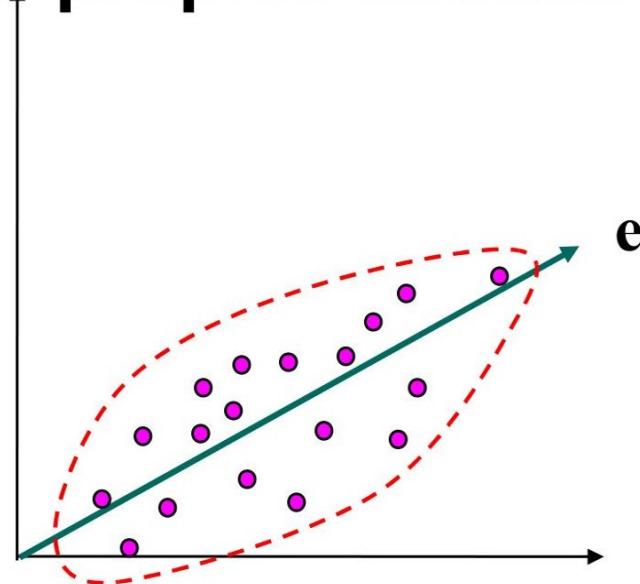
Reducción de dimensionalidad: PCA

- El objetivo es encontrar una proyección que captura la mayor cantidad de variación en los datos



Reducción de dimensionalidad: PCA

- Encontrar los vectores propios de la matriz de covarianza
- Los vectores propios definen el nuevo espacio



PCA

Principal Component Analysis

PCA

Identificar patrones en datos, y expresar los datos para realzar similaridades y diferencias

Encuentra un nuevo conjunto de dimensiones que captura la **variación** de los datos

PCA

Resultado:

- Conjunto de datos de menor dimensión
- Se reduce el ruido

“Benefico para algoritmos de minería”

PCA

Dimension 1: Captura la mayor variabilidad posible

Dimensión 2: Es ortogonal a la primera ,
captura mayor variabilidad del resto.
etc

Media, DS, Varianza

media

$$\bar{X} = \frac{\sum_{i=1}^N X_i}{N}$$

Desviación
estandar

$$s = \sqrt{\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N-1}}$$

Varianza

$$s^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N-1}$$

Solo en una dimensión

Matriz de Covarianzas

- Medida que permite encontrar que tanto las dimensiones varian de la media con respecto a cada una de las dimensiones.
- Medida entre 2 dimensiones

$$var(X) = \frac{\sum_{i=1}^N (X_i - \bar{X})(X_i - \bar{X})}{N-1}$$

$$cov(X, Y) = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{N-1}$$

• Ejemplo 2 dimensiones

	$Hours(H)$	$Mark(M)$	$(H_i - \bar{H})$	$(M_i - \bar{M})$	$(H_i - \bar{H})(M_i - \bar{M})$
Data	9	39	-4.92	-23.42	115.23
	15	56	1.08	-6.42	-6.93
	25	93	11.08	30.58	338.83
	14	61	0.08	-1.42	-0.11
	10	50	-3.92	-12.42	48.69
	18	75	4.08	12.58	51.33
	0	32	-13.92	-30.42	423.45
	16	85	2.08	22.58	46.97
	5	42	-8.92	-20.42	182.15
	19	70	5.08	7.58	38.51
	16	66	2.08	3.58	7.45
	20	80	6.08	17.58	106.89
Totals	167	749			1149.89
Averages	13.92	62.42			104.54

Matriz de Covarianzas

$$C = \begin{pmatrix} cov(x, x) & cov(x, y) & cov(x, z) \\ cov(y, x) & cov(y, y) & cov(y, z) \\ cov(z, x) & cov(z, y) & cov(z, z) \end{pmatrix}$$

Ejercicio

Calcular la matriz de covarianza de:

Item Number:	1	2	3
x	1	-1	4
y	2	1	3
z	1	3	-1

Ejercicio

x	y	z	$x - \bar{x}$	$y - \bar{y}$	$z - \bar{z}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})(z - \bar{z})$	$(y - \bar{y})(z - \bar{z})$
1	2	1	-0.33	0	0	0	0	0
-1	1	3	-2.33	-1	2	2.33	-4.66	-2
4	3	-1	2.67	1	-2	2.67	-5.34	-2
Total	4	6	3			5.00	-10.00	-4
media	1.33	2	1			2.5	-5	-2

$$Cov = \begin{pmatrix} 6.33 & 2.5 & -5 \\ 2.5 & 1 & -2 \\ -5 & -2 & 4 \end{pmatrix}$$

Eigenvectors

Eigenvectors son casos especiales de multiplicación de matrices:

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 11 \\ 5 \end{pmatrix}$$

Matriz de transformación

No es múltiplo del vector original

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = 4 \times \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

Eigenvalue

Cuatro veces el vector original

Vector transformado en su posición original de transformación

Propiedades de eigenvectors

Se encuentran para matrices cuadradas

No todas las matrices cuadradas tienen eigenvectors

Si una matriz $n \times n$ tiene eigenvectors, entonces tiene n eigenvectors

Los eigenvectors son perpendiculares (ortogonales)

Propiedades de eigenvectors

Ecalar el eigenvector a longitud 1 (estándar)

$$\begin{pmatrix} 3 \\ 2 \end{pmatrix} \text{ longitud es: } \sqrt{3^2 + 2^2} = \sqrt{13}$$

el vector con longitud 1 es:

$$\begin{pmatrix} 3\sqrt{13} \\ 2\sqrt{13} \end{pmatrix}$$

Eigenvalues

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = \boxed{4} \times \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

Eigenvalue asociado con el eigenvector

Ejercicio

$$\begin{pmatrix} 3 & 0 & 1 \\ -4 & 1 & 2 \\ -6 & 0 & -2 \end{pmatrix}$$

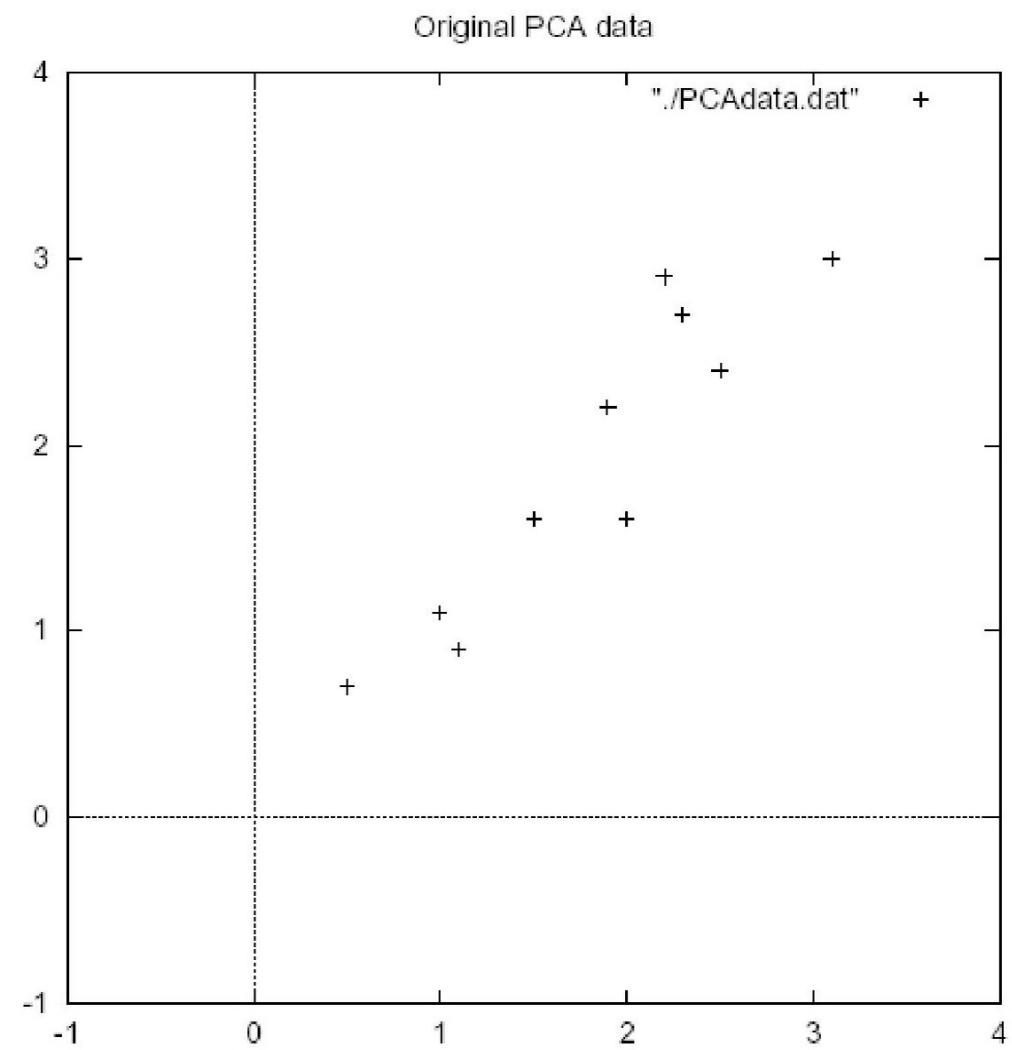
Cual de los siguientes vectores son eigenvectors de la matriz? Cual es su correspondiente eigenvalue?

$$\begin{pmatrix} 2 \\ 2 \\ -1 \end{pmatrix} \quad \begin{pmatrix} -1 \\ 0 \\ 2 \end{pmatrix} \quad \begin{pmatrix} -1 \\ 1 \\ 3 \end{pmatrix} \quad \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}$$

Método (Ejemplo)

Data =

	x	y
	2.5	2.4
	0.5	0.7
	2.2	2.9
	1.9	2.2
	3.1	3.0
	2.3	2.7
	2	1.6
	1	1.1
	1.5	1.6
	1.1	0.9



Método (Ejemplo)

- Restar la media

	x	y		x	y
	2.5	2.4		.69	.49
	0.5	0.7		-1.31	-1.21
	2.2	2.9		.39	.99
	1.9	2.2		.09	.29
Data =	3.1	3.0	DataAdjust =	1.29	1.09
	2.3	2.7		.49	.79
	2	1.6		.19	-.31
	1	1.1		-.81	-.81
	1.5	1.6		-.31	-.31
	1.1	0.9		-.71	-1.01

$$\bar{x} = 1.81 \quad \bar{y} = 1.91$$

Método (Ejemplo)

- Calcular la matriz de covarianzas

$$\text{cov} = \begin{pmatrix} 0.616555556 & 0.615444444 \\ 0.615444444 & 0.716555556 \end{pmatrix}$$

Método (Ejemplo)

- Calcular eigenvectors y eigenvalues de la matriz de covarianzas.

$$\text{eigenvalues} = \begin{pmatrix} 0.0490833989 \\ 1.28402771 \end{pmatrix}$$

$$\text{eigenvectors} = \begin{pmatrix} -0.735178656 & -0.677873399 \\ 0.677873399 & -0.735178656 \end{pmatrix}$$

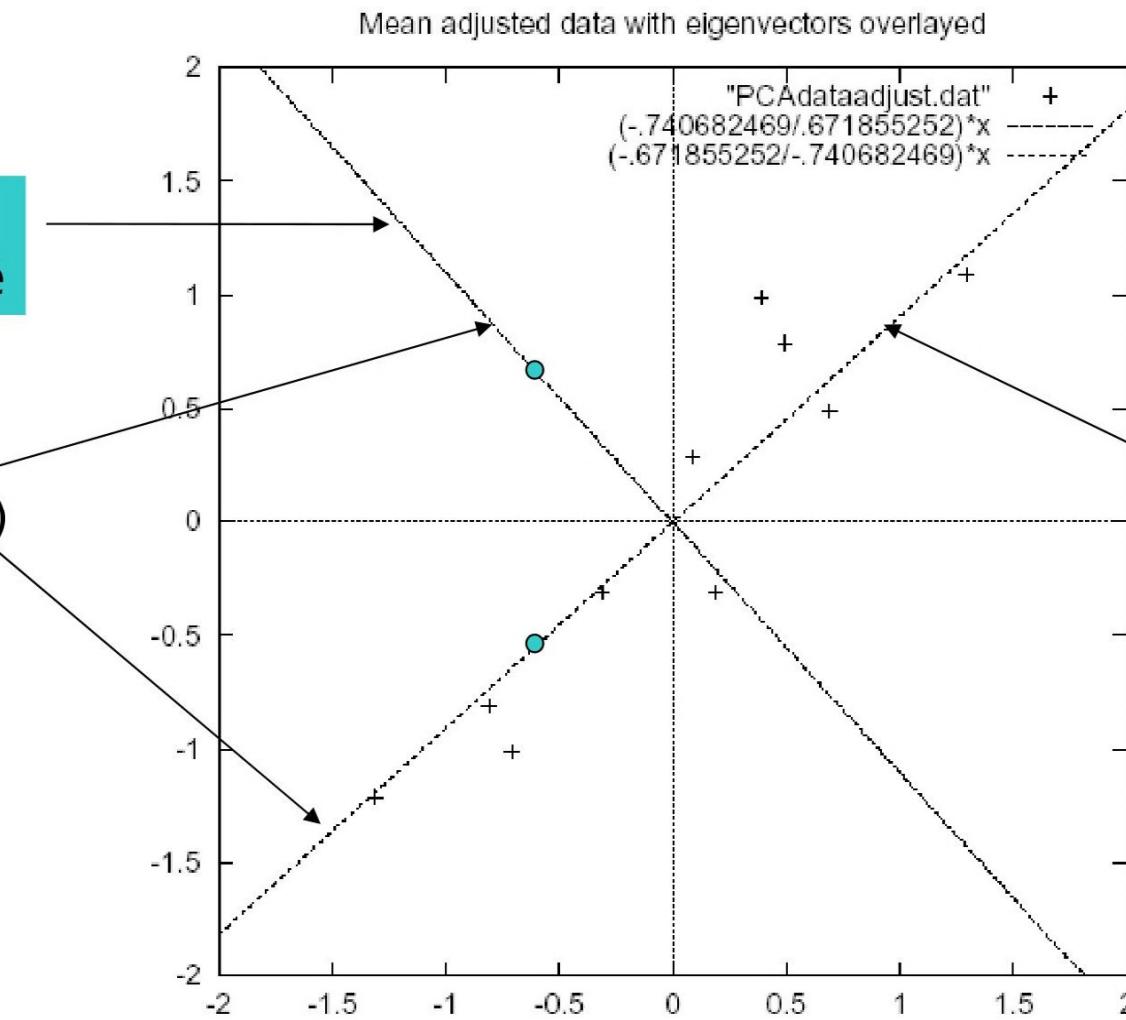
Método para calcular eigenvectors= Jacobi (Investigar)

Método (Ejemplo)

Menos importante

Eigenvectores
(perpendicular)

Entre los puntos.
Los puntos están
relacionados con
esta línea



Método (Ejemplo)

- Escoger componentes y formar vector (feature vector)

$$\text{FeatureVector} = (eig_1 \ eig_2 \ eig_3 \ \dots \ eig_n)$$

- Ordenar de eigenvalues de mayor a menor (orden de significancia)

Valores pequeños de eigenvalues indican que el eigenvector es menos importante.

$$\text{eigenvalues} = \begin{pmatrix} 0.0490833989 \\ 1.28402771 \end{pmatrix}$$
$$\text{eigenvectors} = \begin{pmatrix} -0.735178656 & -0.677873399 \\ 0.677873399 & -0.735178656 \end{pmatrix}$$

Componente Principal (mayor eigenvalue)

$$\begin{pmatrix} -.677873399 \\ -735178656 \end{pmatrix}$$

Método (Ejemplo)

Cuantos componentes principales son necesarios para tener una buena representación de los datos?

Analizar la proporción de la varianza (eigenvalues). Dividiendo la suma de los primeros m eigenvalues por la suma de todos los eigenvalues

$$R = \frac{\left(\sum_{i=1}^m \lambda_i \right)}{\left(\sum_{i=1}^n \lambda_i \right)}$$

90% es considerado bueno

Método (Ejemplo)

- Derivar nuevo conjunto de datos

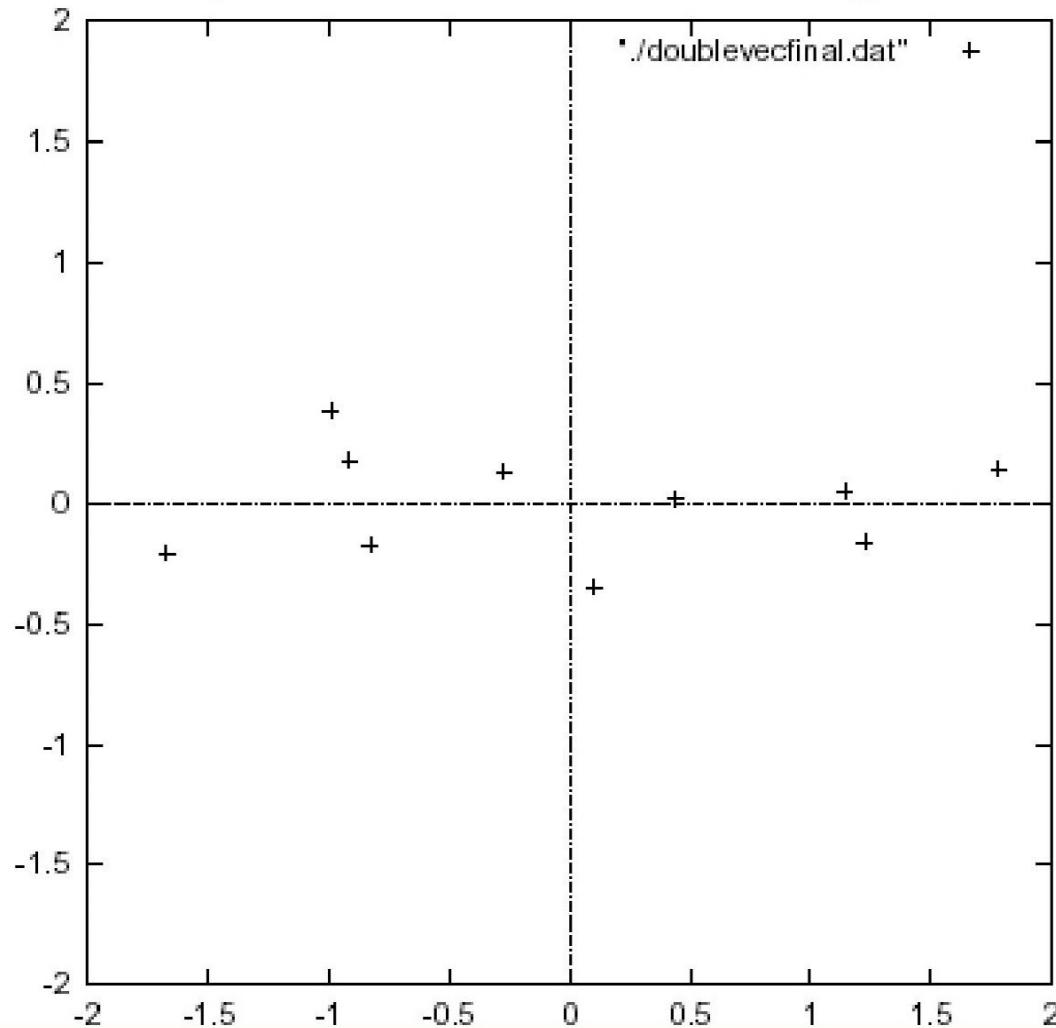
Datos ajustados
usando la media X Matriz de eigenvectors

	x	y
	-.827970186	-.175115307
	1.77758033	.142857227
	-.992197494	.384374989
	-.274210416	.130417207
Transformed Data=	-1.67580142	-.209498461
	-.912949103	.175282444
	.0991094375	-.349824698
	1.14457216	.0464172582
	.438046137	.0177646297
	1.22382056	-.162675287

Data transformed with 2 eigenvectors

Método (Ejemplo)

Nuevo conjunto usando los dos eigenvectores



Obtener el original conjunto de datos

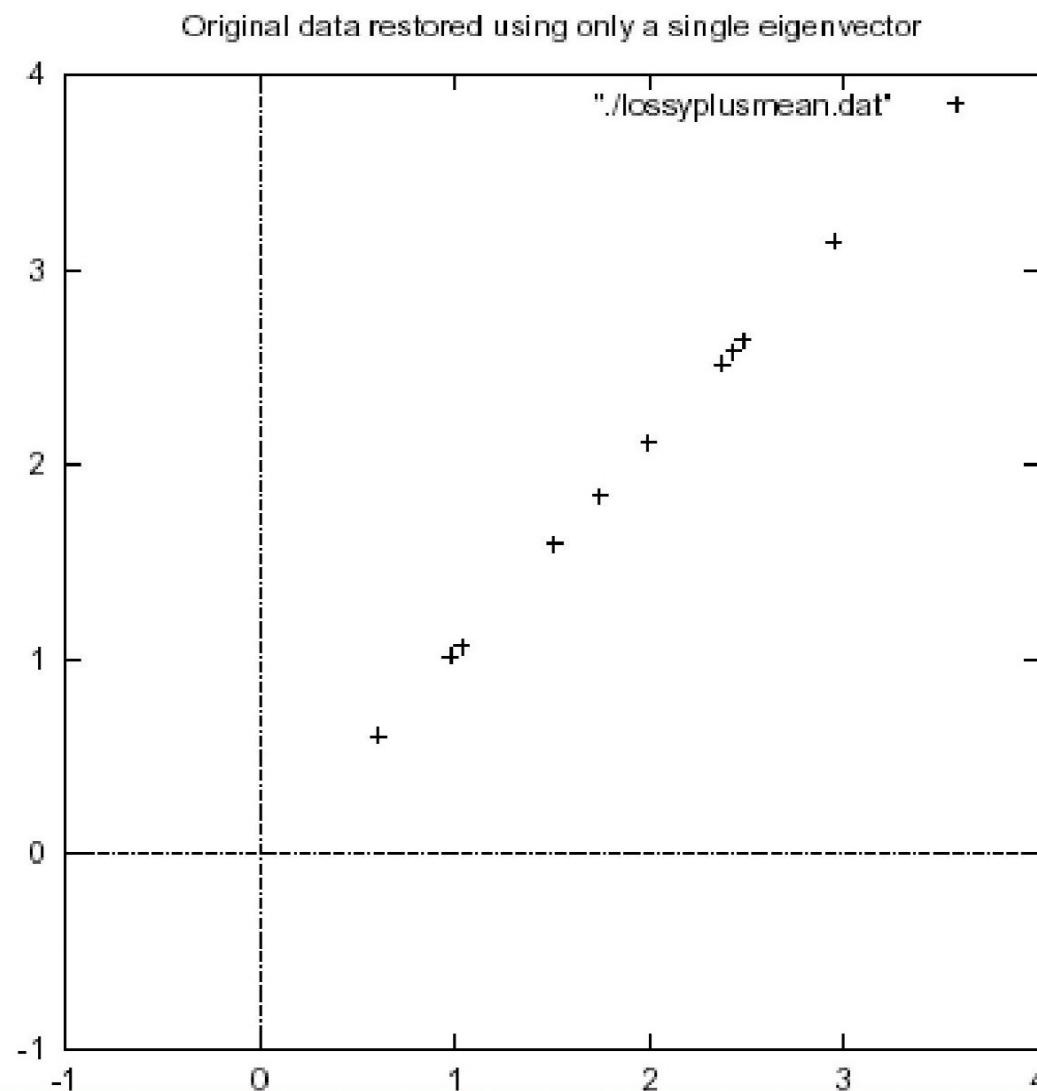
Nuevo conjunto X (eigenvector matrix) $^{-1}$

Nuevo conjunto X (eigenvector matrix) T

Transformed Data (Single eigenvector)

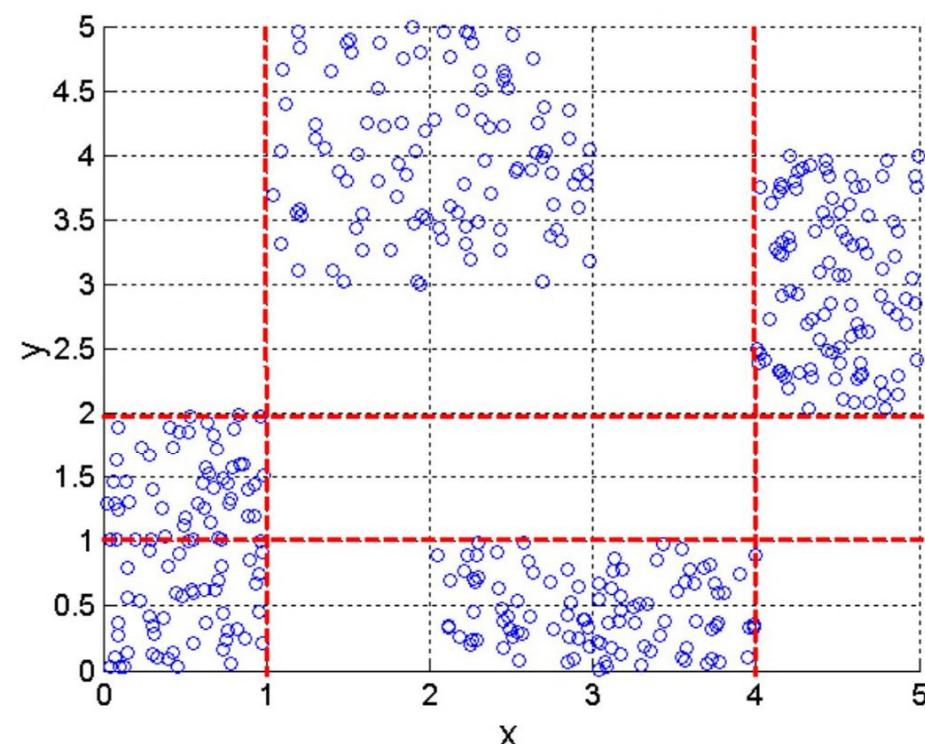
x
- .827970186
1.77758033
-.992197494
-.274210416
-1.67580142
-.912949103
.0991094375
1.14457216
.438046137
1.22382056

Obtener el original conjunto de datos

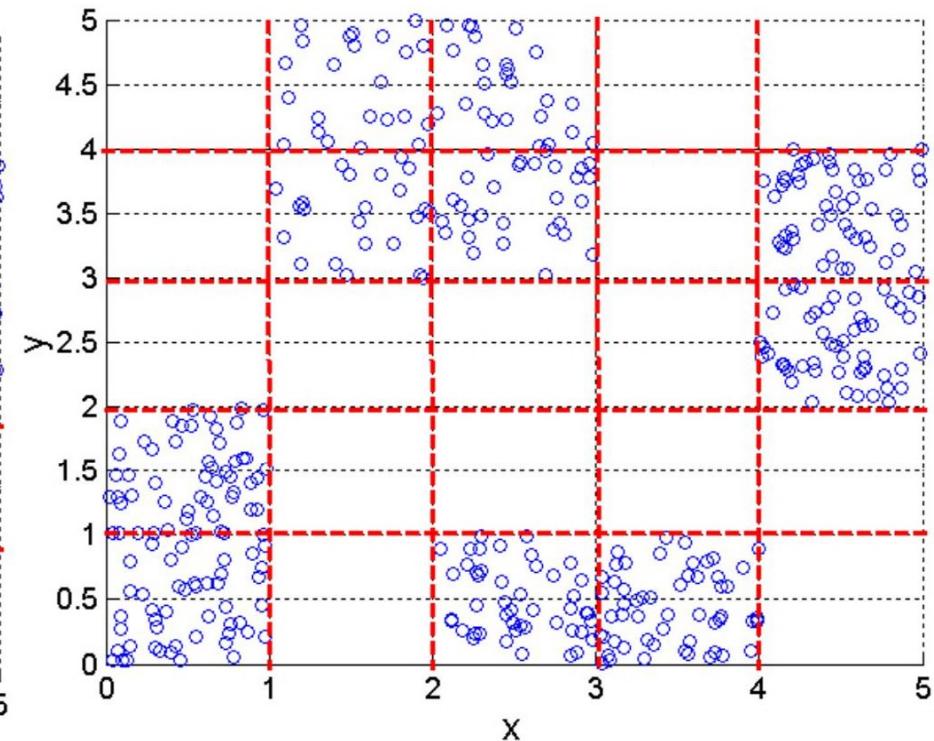


Discretización uso de las etiquetas de clase

■ Enfoque basado en la entropía

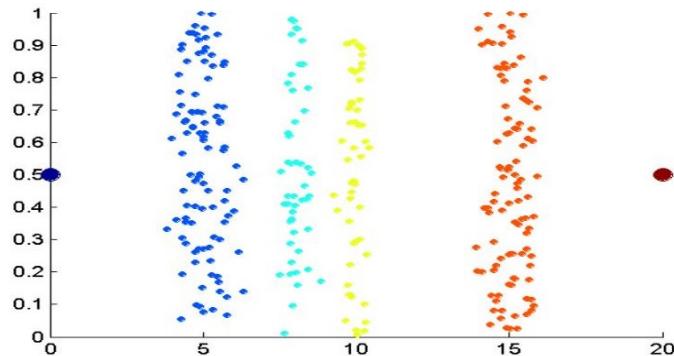


3 categorías, tanto para X e Y

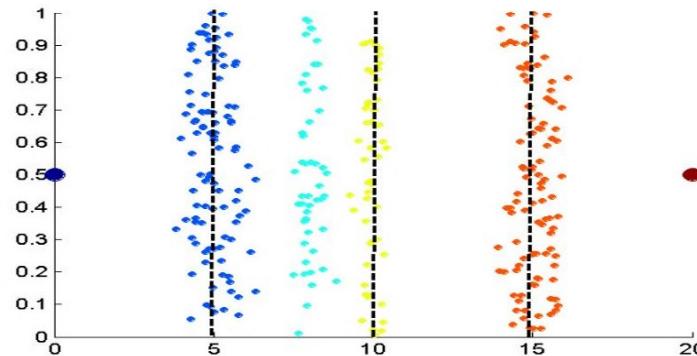


5 categorías tanto para X e Y

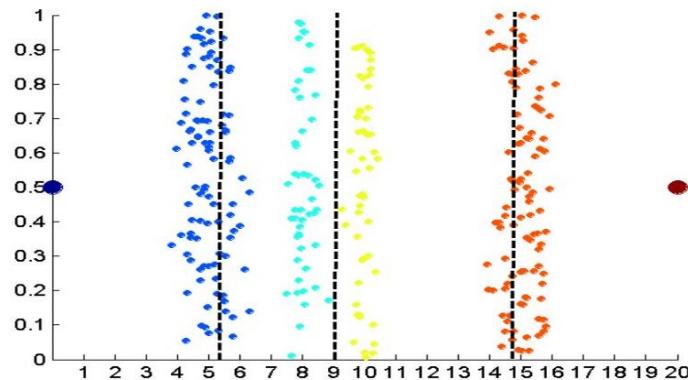
Discretización sin utilizar etiquetas de clase



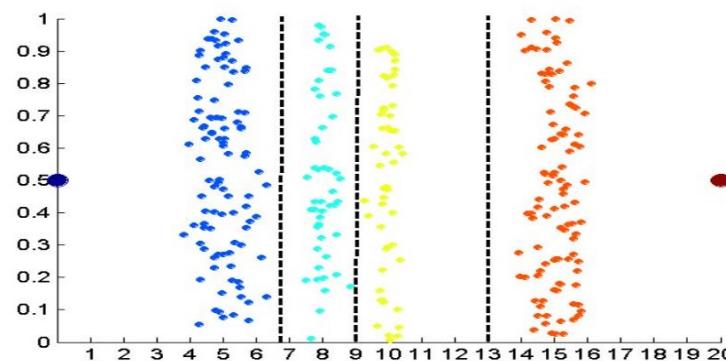
Datos



Amplitud del intervalo
de igualdad



la misma frecuencia



K-means

Bibliografía

- [1] Introduction to Data Mining. Tan, Steinbach, Kumar. 2006
- [2] Data Mining: Concepts, Models, Methods, and Algorithms. Mehmed Kantardzic. 2003
- [3] W. Kim, B. Choi, E-K. Hong, S-K. Kim. A Taxonomy of Dirty Data
- [4] Data Mining and Knowledge Discovery7, 81- 99, 2003