

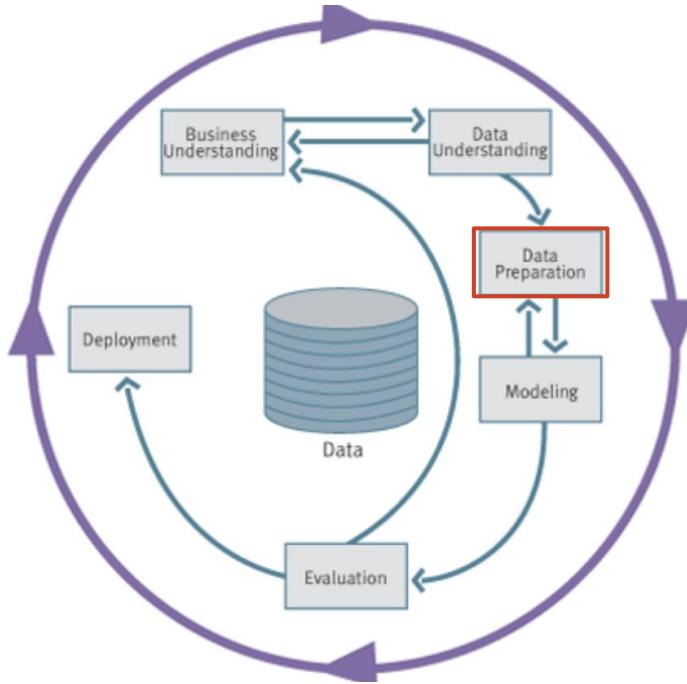
# Preparación/preprocesamiento de Datos

## Asignatura Minería de Datos

Por  
**Elizabeth León Guzmán, Ph.D.**  
Profesora  
Ingeniería de Sistemas y Computación

# CRISP-DM

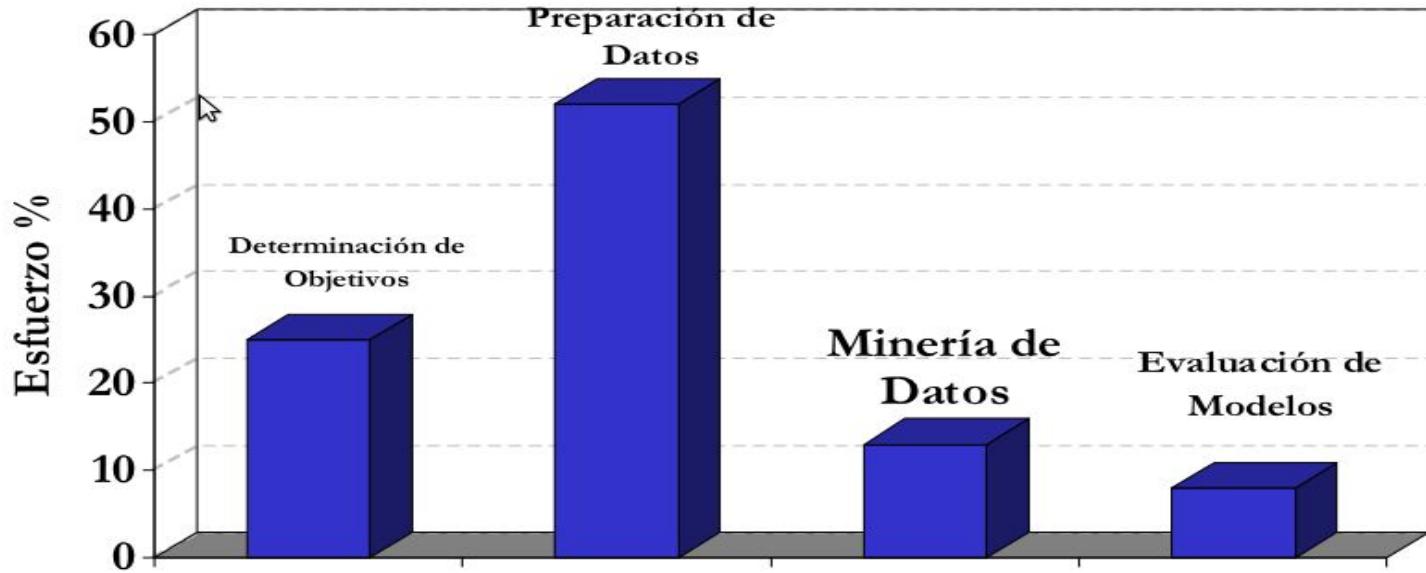
## CRoss Industry Standard Process for Data Mining Projects



# Preparación de datos

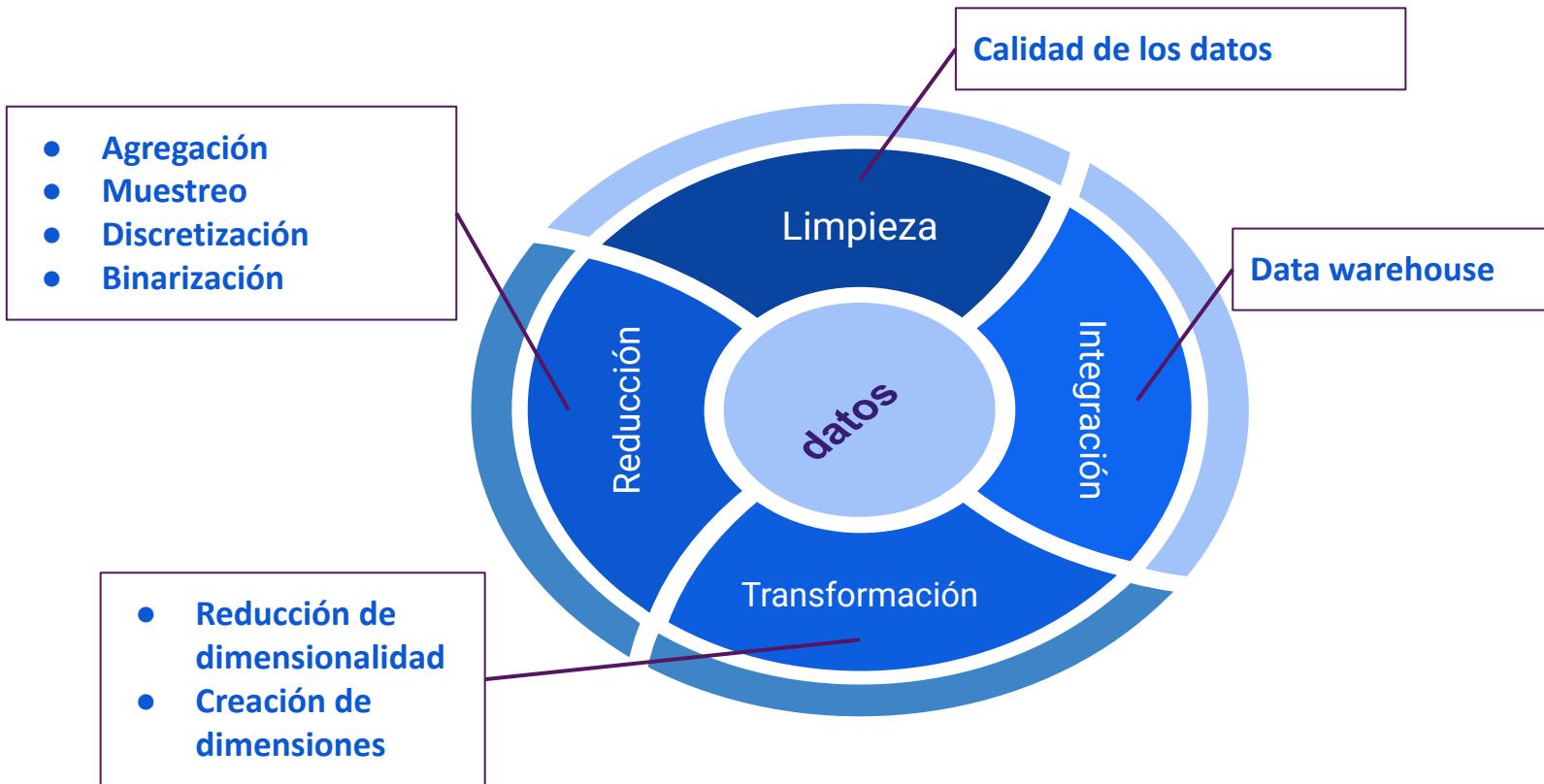
“El propósito fundamental de la preparación de los datos es la manipulación y transformación de los datos sin refinar para que la información contenida en el conjunto de datos pueda ser descubierta o estar accesible de forma más fácil”

D. Pyle, 1999, pp. 90



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

# Componentes en la Preparación de los Datos



# Calidad de los datos - problemas con datos

- Datos **incompletos**
- Datos con **ruido**
  - “Outliers”
- Datos **inconsistentes**
  - No son exactos
  - Sin actualizar
- Datos **Repetidos**



# Calidad de los datos

Decisiones de calidad deben ser basadas en datos de buena calidad.

## ¡Limpieza de los datos!

- Recuperar información incompleta
- Eliminar outliers, ruido
- Resolver conflictos y redundancias
- Eliminar duplicados

# Limpieza de Datos: Valores Perdidos

Existen muchos datos que no contienen todos los valores para las variables - Valores que faltan.

A	B	C	D
1	aa	20	Bog
1	de	21	Cart
2	?	23	Cart
3	ca	?	Med

Se puede:

- **Ignorarlos:** No usar los registros/variable con valores perdidos
- **Inferirlos o remplazarlos:** usar técnica de predicción

# Limpieza de Datos: Valores Perdidos

**Ignorarlos:** No usar los registros con valores perdidos

- **Ventaja:**
  - Es una solución fácil.
- **Desventajas:**
  - Pérdida 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

## Reemplazarlos:

- 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

## PREDICCIÓN

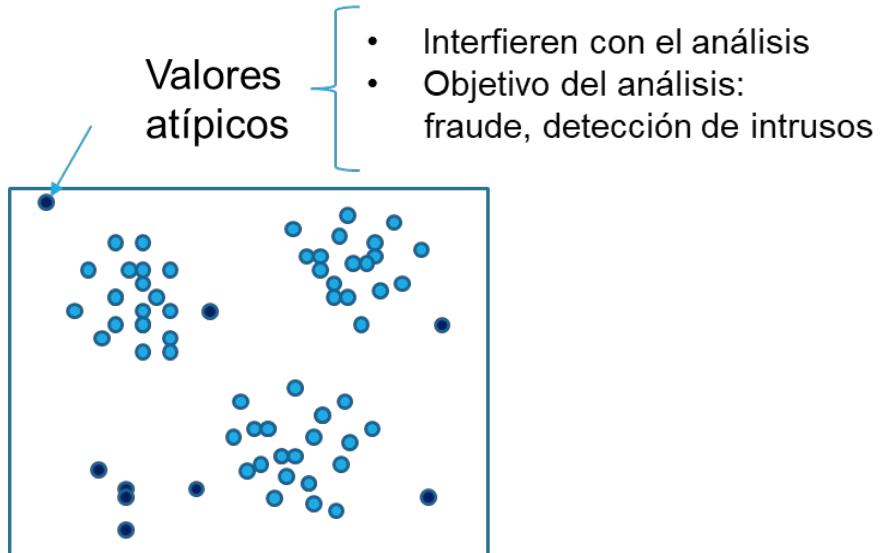
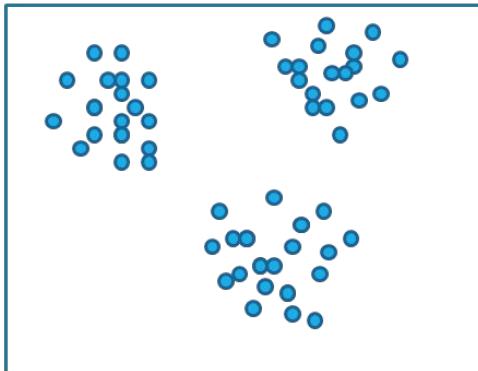
Técnicas:

A	B	C	D
1	aa	20	Bog
1	de	21	Cart
2	?	23	Cart
3	ca	?	Med

Regresion  
Bayes,  
Agrupación,  
Árboles de decisión

# Limpieza de Datos: Ruido

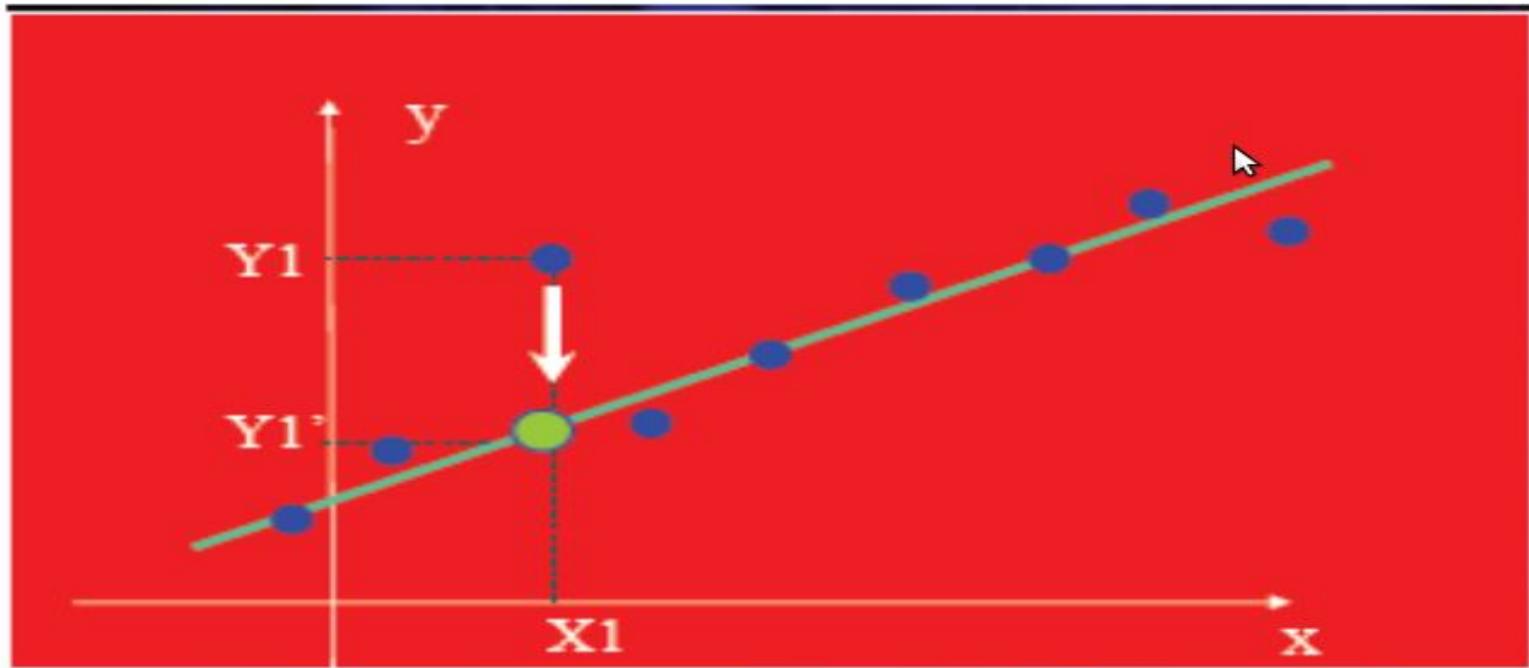
- Medida de error
- Distorsión de un valor
- Adición de un objeto extraño



- Espacio y tiempo: Procesamiento de señales e imágenes
- Minería: **Algoritmos robustos**

# Limpieza de Datos: Ruido

- Suavizamiento (Smoothing):



# Limpieza de Datos: Outliers

- Datos con características diferentes a los demás.
- Valores inusuales
- Objetos Anómalos
- Datos legítimos

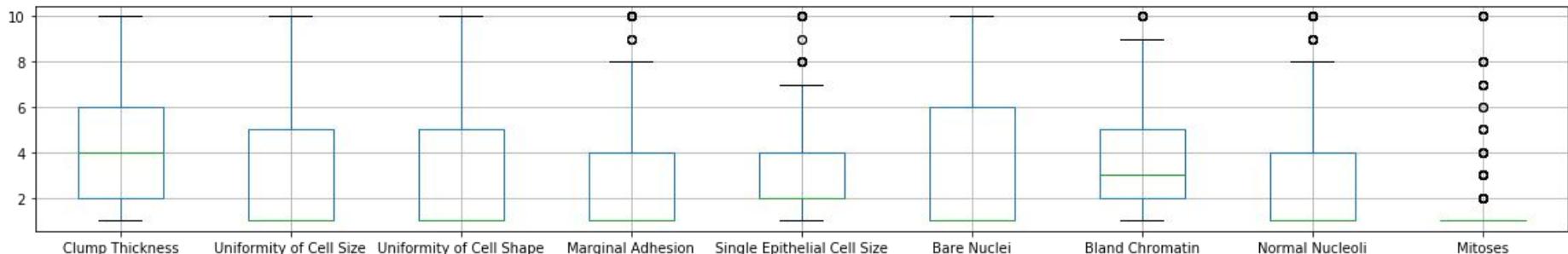
*Pueden ser de interés*

Detección de anomalías

# Limpieza de Datos: Outliers

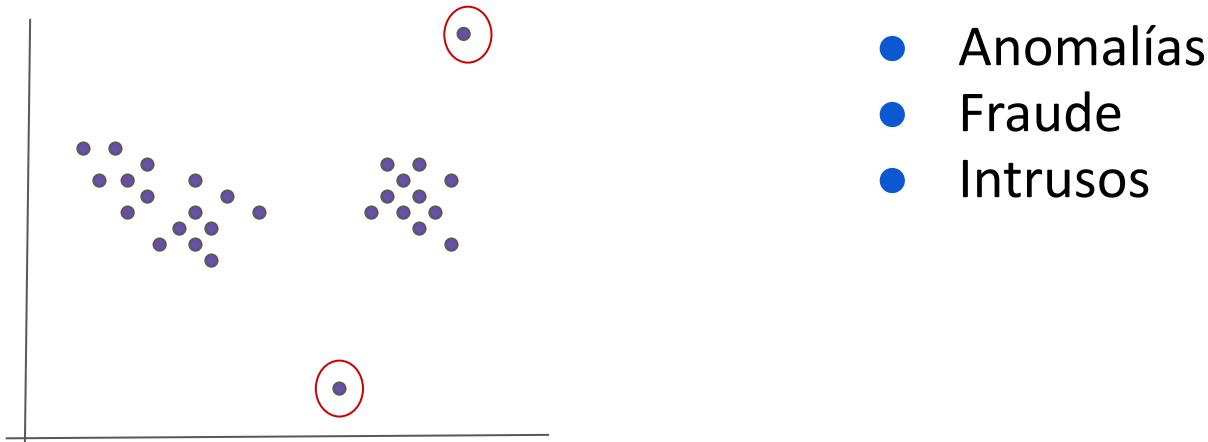
- Un atributo: encontrar *mean* y *variance*

**Umbral** = media +- 2 variance



# Limpieza de Datos: Outliers

- Basado en distancia: Multidimensional  
Los ejemplos que no tienen vecinos son considerados “outliers”

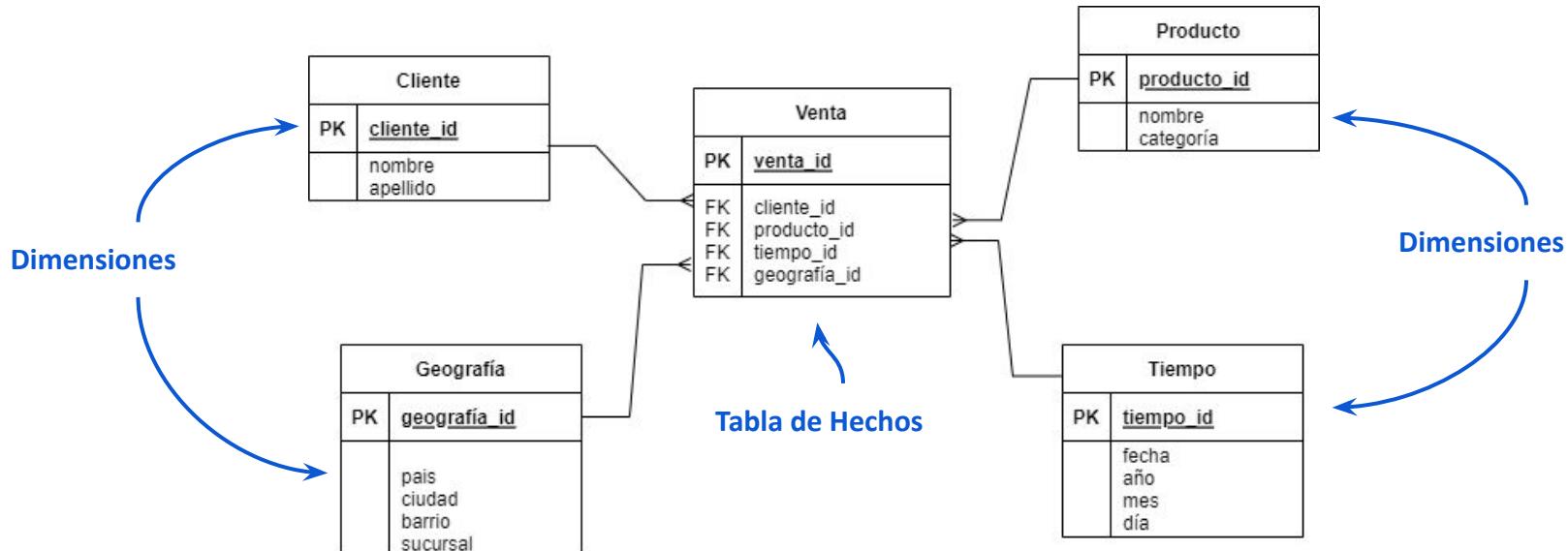


# Integración de Datos

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

*Ejemplo:* “Data warehouse” Modelamiento multidimensional

# Integración de Datos



# Integración de Datos

- Diferentes escalas:
  - Pesos vs Dólares
- Atributos derivados
  - Salario Mensual vs Salario Anual
- Soluciones:
  - Procedimientos semiautomáticos
  - ETL
  - Minería

# Preprocesamiento

- Transformación de Atributos /Variables
- Agregación
- Muestreo (Sampling)
- Discretización y Binarización
- Reducción de Dimensionalidad
- Selección de Subconjunto de Variables
- Creación de Variables

# Transformación de Variables

- Son transformaciones aplicadas a todos los valores de una variable.
- Para cada objeto la transformación es aplicada al valor de una variable.

## Ejemplos:

- Aplicar funciones
- Normalización

Cambiar la escala de valores y/o ajustar la distribución de datos sesgada a una distribución similar a Gauss a través de alguna "transformación monotónica" (función cuya derivada no cambia de signo)

# Transformación de Variables

## Aplicar funciones

Aplicación de una función a cada valor.

### Funciones

$$x^k, \log x, e^x, \sqrt{x}, 1/x, \sin(x), |x|$$

- Rangos muy grandes. Necesidad de comprimir (Ej: usando transformación log)
- Debe ser aplicado cuidadosamente ya que puede cambiar la naturaleza de los datos. Ej:  $1/x$  reduce la magnitud de valores mayores a 1, pero incrementa la magnitud de los valores entre 0 y 1.

1, 2, 3 es transformado a 1,  $1/2$ ,  $1/3$  y  
1,  $1/2$ ,  $1/3$  es transformado a 1, 2, 3

# Transformación de Variables

## Normalización o Estandarización

Objetivo: permitir que el conjunto de datos tenga una propiedad particular. Cuando se combinan variables, esta transformación es necesaria para evitar tener una variable con valores grandes dominantes en el resultado de los cálculos requeridos.

**Ejemplo:** edad y salario

- La diferencia entre edades es menor de 150
- Las diferencias en el salario son de miles de pesos

Las comparaciones serán dominadas por las diferencias del salario

# Transformación de Datos

## Normalización

Transformar la variable con distribución normal usando la media y la desviación estándar del atributo, la transformación usada es:

$$x' = \frac{x - \bar{x}}{S_x}$$

crear una nueva variable que tiene media cero y desviación estándar 1.

Sin embargo, dado los outliers esta transformación puede ser modificada. Reemplazar la media por la mediana, y la desviación estándar por la desviación estándar absoluta.

# 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

### Ejercicio:

20 23 25 26 29 34 35 37 39 40

Normalizar:

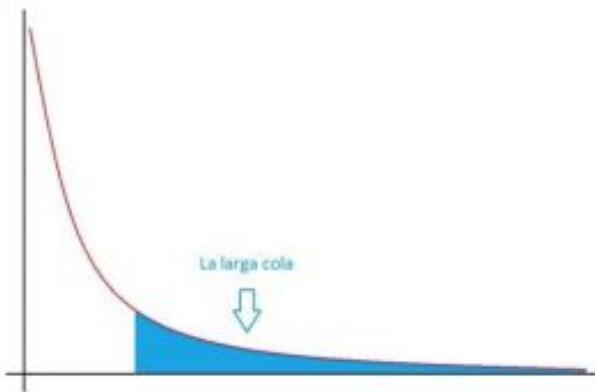
- entre 0 y 1
- z-score

# Transformación de Datos

## Normalización

### Logarítmica

Cuando la distribución tiene cola larga (Ej., Ingresos individuales, costos médicos individuales, etc.), la transformación logarítmica puede ajustar la distribución de datos a una distribución menos sesgada



$$x'_i = \log(x_i)$$

$$x'_i = \log(x_i + 1) : \text{En caso de que el valor sea cero}$$

$$x'_i = sgn(x_i)\log|x_i| = \frac{x_i}{|x_i|}\log|x_i| : \text{En caso de que el valor sea negativo}$$

$$x'^\lambda_i = \log(x_i + \sqrt{x_i^2 + \lambda}) : \text{Fórmula general}$$

# Transformación de Datos

## Normalización

- Normalización por escala decimal

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

Donde  $j$  es el entero más pequeño tal que  $\max(|v'|) < 1$

**Ejemplo:** la dimensión varía entre -9993 y 9923

max en valor absoluto es 9993

Entonces  $j$  es 4, dividir por 10000

-9993 normalizado es -0.9993

# Transformación de Datos

## Normalización

### Ejemplo:

Considere los siguientes datos:

-15 211 321 -451 555 601 781

valor máximo: 781

$$j=3; \max(|781|) < 1 \quad (781/10^3 = 0.781)$$

Los datos normalizados son

-0.015 0.211 0.321 -0.451 0.555 0.601 0.781

# Transformación de Datos

## Normalización

### Ejercicio

Normalizar el conjunto de datos “iris”

- Utilizando la función `preprocessing.normalize()` de Python
- Utilizar varias de las funciones de weka

# Transformación de Datos

## VARIABLES CATEGÓRICAS

Convertir una variable categórica en una variable numérica.

- Obligatoria para las técnicas que solo pueden manejar valores numéricos.
- *Codificación.* En minería de texto, *incrustación (embedding)* generalmente transforma a valores numéricos que contienen la misma semántica que los datos originales.
- Su selección es muy importante para el rendimiento del modelo.

# Transformación de Datos

## Variábles Categóricas

Algunas de las transformaciones de variables categóricas:

- Codificación de uno en caliente (one-hot encoding)
- Codificación de etiquetas (label encoding)
- Función hash (feature hashing)

# Transformación de Datos: Variables categóricas

## One-hot encoding

Convierte una columna categórica en múltiples columnas binarias (0 o 1), el número de columnas binarias corresponde con la cardinalidad de la columna.

**Ejemplo:** Si hay cuatro valores diferentes de la variable categórica, la codificación creará cuatro columnas nuevas, cada una de las cuales tiene 0 o 1

Original data
A
B
D
C
B
C
:



	Transformed data by one-hot encoding			
	A	B	C	D
1	1	0	0	0
0	0	1	0	0
0	0	0	0	1
0	0	0	1	0
0	0	1	0	0
0	0	0	1	0
:	:	:	:	:

# Transformación de Datos: Variables categóricas

## Codificación de Etiquetas

- Convierte los valores categóricos en enteros
- No es muy apropiado en la mayoría de los algoritmos de aprendizaje automático
- En caso de que la variable sea 'ordinal', la codificación de etiquetas puede funcionar

# Transformación de Datos: Variables categóricas

## Función Hash

Convertir una columna categórica en varias columnas mediante hashing. Se puede definir la cantidad de columnas nuevas, que puede ser menor que la cantidad de valores. En lugar de asignar 0 o 1 el hashing de características usa más de dos valores (-1, 0 o 1).



Original data	1	2	3
A	0	1	0
B	0	0	-1
D	0	0	1
C	-1	0	0
B	0	0	-1
C	-1	0	0
:	:	:	:

# Transformación de Datos: Binarización

## Proceso de binarizar un atributo categórico

Table 1: Conversión de atributo categórico a tres atributos binarios

Valor categórico	Valor entero	$X_1$	$X_2$	$X_3$
Superior	4	1	0	0
Alto	3	0	1	1
Básico	2	0	1	0
Bajo	1	0	0	1
Muy Bajo	0	0	0	0

Table 2: Conversión de atributo categórico a cinco atributos binarios asimétricos

Valor categórico	Valor entero	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$
Superior	4	1	0	0	0	0
Alto	3	0	1	0	0	0
Básico	2	0	0	1	0	0
Bajo	1	0	0	0	1	0
Muy Bajo	0	0	0	0	0	1

} Solo la presencia de un atributo

# Reducción de Datos

Reducción del tamaño del conjunto de datos



posible mejora de la eficiencia del proceso de  
Minería de Datos

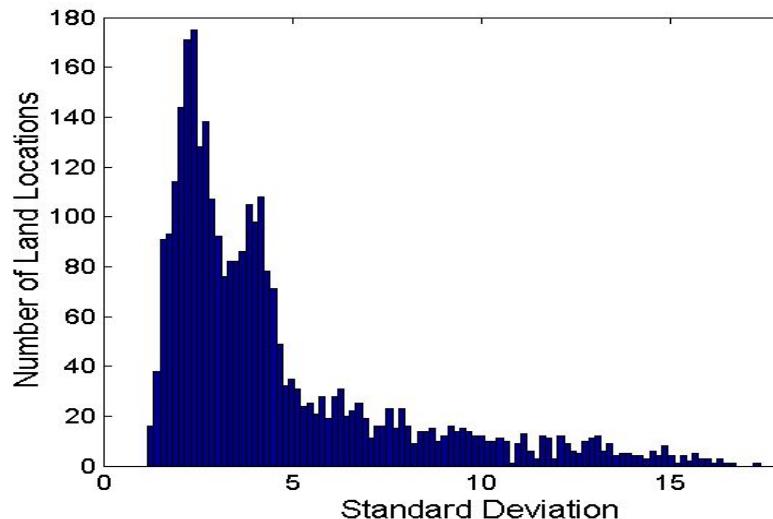
- Selección de instancias (muestreo)
- Selección de características (reducción de dimensionalidad)
- Reducción de valores (discretización)

# Reducción de datos: Agregación

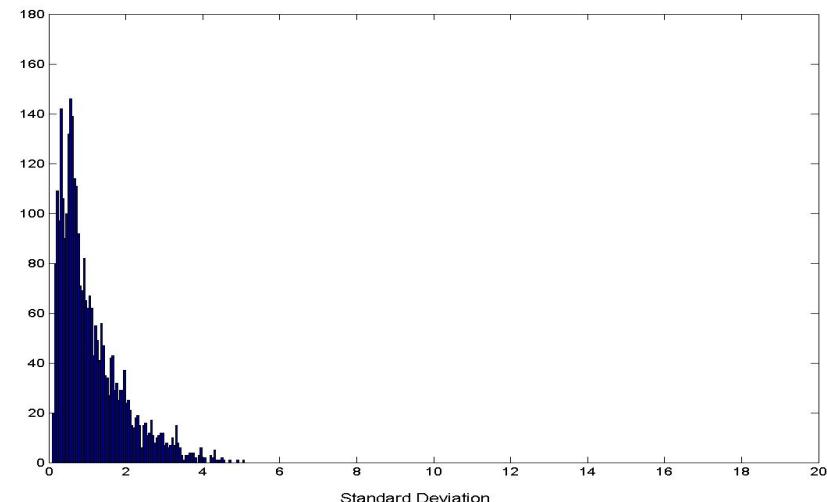
- La combinación de dos o más atributos (u objetos) en un solo atributo (u objeto) propósito reducción de datos
- Reducir el número de atributos u objetos  
Ciudades agregan en regiones, estados, países, etc
- Crear una transacción agregada con valores numéricos:  
promedio, suma.
- Los datos agregados tiende a tener una menor variabilidad

# Reducción de datos: Agregación

## Variación de la precipitación en Australia



Desviación estándar de la  
precipitación mensual promedio



Desviación estándar de la  
precipitación media anual

# Muestreo - *Sampling*

- Es la principal técnica empleada para seleccionar un **subconjunto de datos** del conjunto de datos a analizar. Usado tanto para la investigación preliminar de los datos como para el análisis final de los datos.
- En minería de datos, el muestreo se utiliza para **reducir los costos en recursos y tiempo** en el procesamiento del conjunto de datos.
- Usando muestreo se puede reducir el conjunto de datos a tal punto que un mejor algoritmo, pero más costoso, pueda ser utilizado.

## Principio fundamental del muestreo

Los resultados de usar una **muestra** del conjunto datos, obtenida de un *muestreo*, deben ser tan buenos como si se usará el conjunto completo de datos. Es decir, la muestra debe ser **representativa**.

## Muestra representativa

- Una muestra es representativa si se tiene aproximadamente la misma propiedad (de interés) que el conjunto original de datos.

**Ejemplo:** Si la media es de interés, entonces la muestra debe tener una media cercana a la del conjunto original.

El muestreo es un proceso estadístico por lo que la representación de una muestra pueda variar. La mejor forma de realizarlo es escogiendo un esquema de muestreo que garantice una probabilidad alta de seleccionar una muestra significativa (tamaño de la muestra y técnica de muestreo).

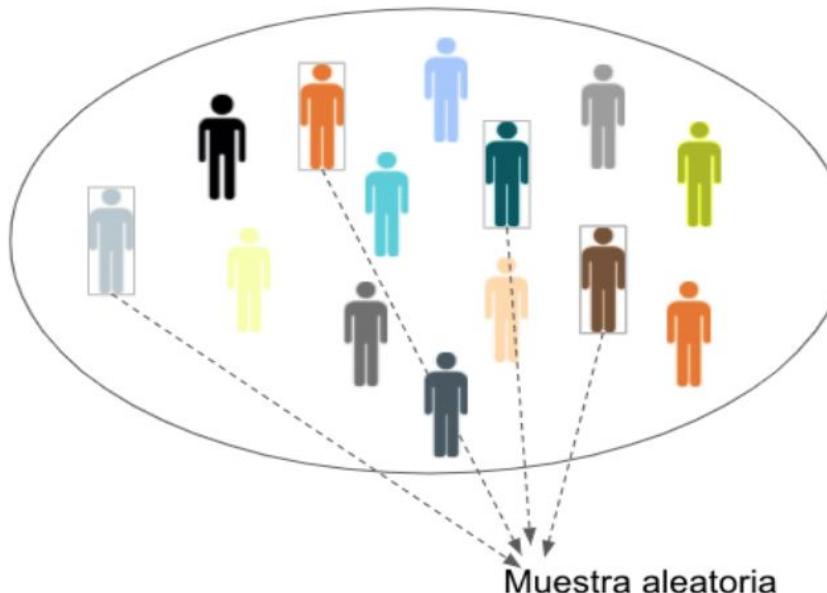
## Tipos de Muestreo

Los más básicos son:

- Muestreo aleatorio simple
- Muestreo sin reemplazo
- Muestreo con reemplazo
- Muestreo estratificado

## Muestreo Aleatorio Simple

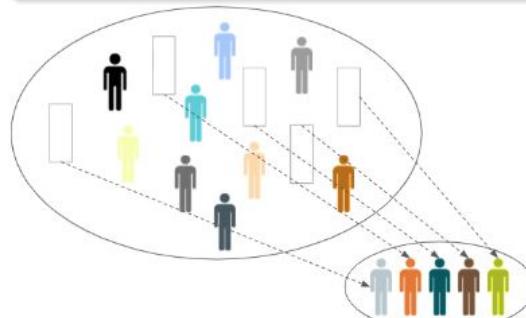
Es el muestreo más simple. La misma probabilidad de seleccionar cualquier elemento (objeto) en particular del conjunto de datos.



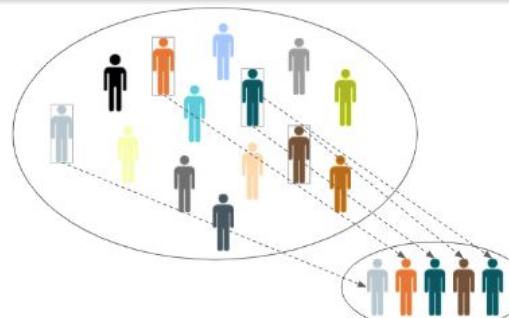
## Muestreo Aleatorio Simple

Tiene dos variaciones:

- **Sin Reemplazo** Cuando un objeto es seleccionado se remueve de la población del conjunto inicial, para que no sea nuevamente seleccionado.
- **Con Reemplazo** Los objetos no se eliminan de la población, ya que son seleccionados para la muestra. El mismo objeto puede ser seleccionado más de una vez, es decir en la muestra se puede repetir.



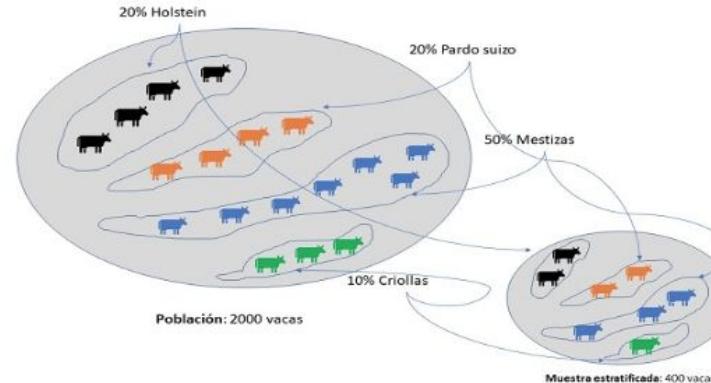
Muestra aleatoria sin reemplazo



Muestra aleatoria con reemplazo

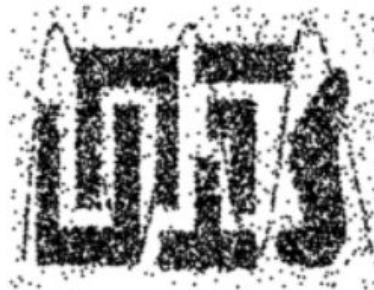
## Muestreo Estratificado

- Cuando la población tienen diferentes tipos de objetos (conjunto de datos con clase), la muestra puede fallar al no seleccionar objetos de las clases pequeñas.
- Para crear la muestra, se toman muestras al azar de cada partición. En un tamaño proporcional a los tamaños de las clases en el conjunto de datos inicial.



## Pérdida de Información

Entre más grande la muestra la probabilidad de ser representativa crece.



8000 puntos



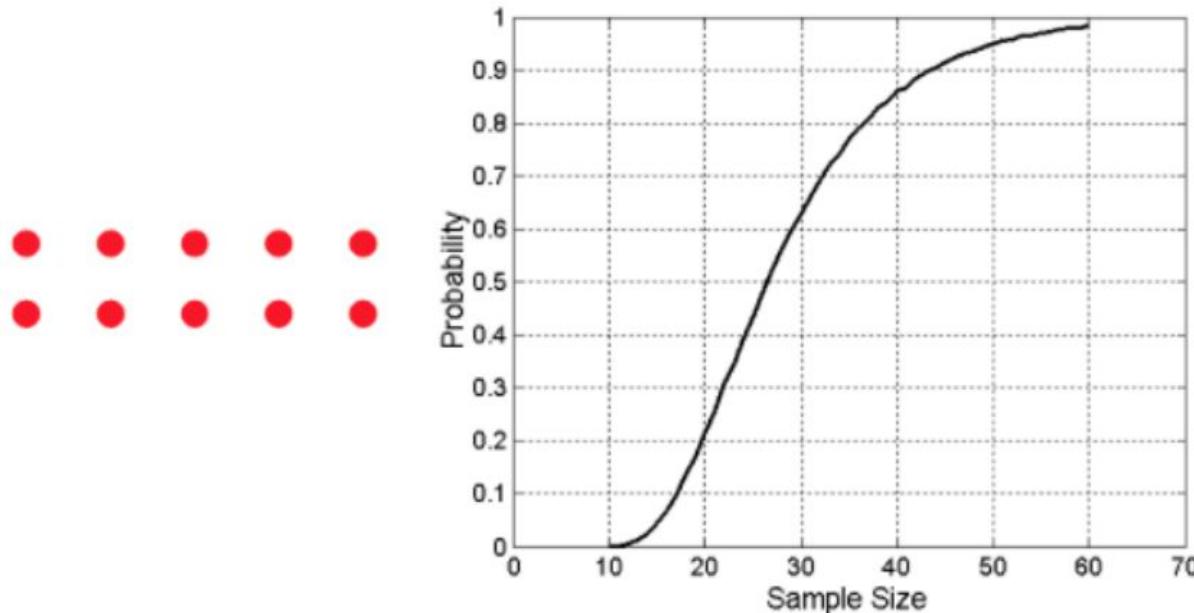
2000 Puntos



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?



## Muestreo Progresivo

- Muy difícil determinar el tamaño de la muestra, por lo que se pueden usar estrategias progresivas que incrementan el tamaño de la muestra.
- Inicia con un tamaño de muestra pequeño, y luego se incrementa hasta que el tamaño de la muestra sea suficiente.
- Requiere de un método para evaluar la muestra. Ej: La precisión de un modelo predictivo.

# Discretización

Proceso de convertir una variable continua en una variable ordinal

- Una variable con valores potencialmente infinitos, es mapeada a un conjunto pequeño de categorías  
**Ejemplo:** Temperatura (datos continuos: 30.5 , 28.4, 20.1, 42.1, ... )  
Temperatura (rangos como: **alta**, **media**, **baja**)
- Divide el rango de atributos continuos en *Intervalos*
- Almacena sólo las *etiquetas* de los intervalos
- Importante para algoritmos como reglas de asociación y clasificación, algunos algoritmos solo aceptan datos *discretos*.
- Pensado de tal forma que contribuya a un buen desempeño en la tarea de minería considerada

# Reducción de Datos: Discretización

El problema de discretización se vuelve en decidir cuántas categorías se requieren y en seleccionar apropiadamente los puntos de separación

1. Ordenar los valores de la variable
2. Dividir en n intervalos usando n-1 puntos de separación
3. Asignar los valores de un intervalo a la misma categoría.

# Reducción de Datos: Discretización

Ejemplo:

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

ordenado:

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

3 categorías

2 puntos de separación

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

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

# Reducción de Datos: Discretización

Se pueden usar métodos como el de igual amplitud, igual frecuencia y agrupación (clustering)

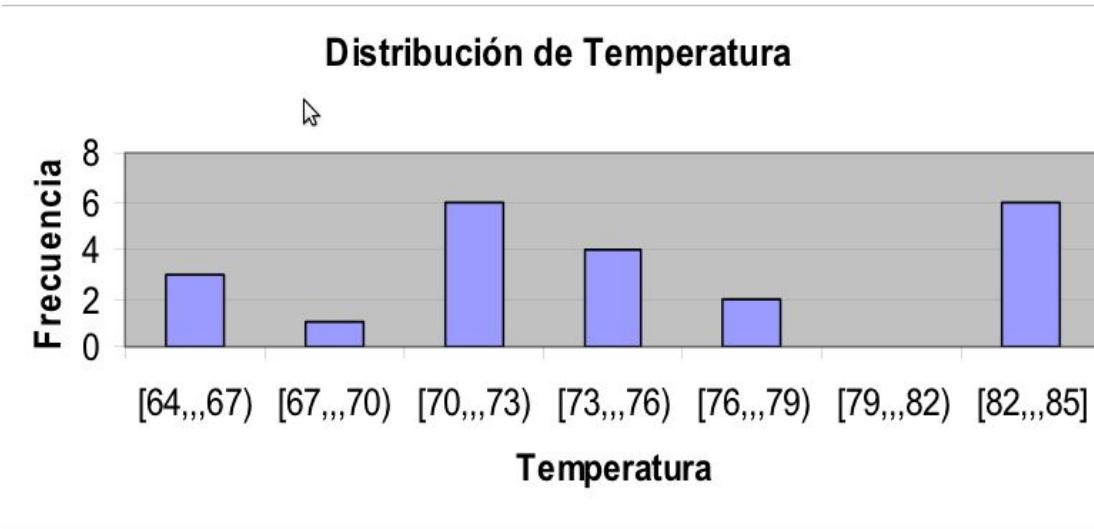
## Igual Amplitud

Se divide el rango de la variable en un número específico de intervalos de tal forma que cada intervalo tiene el mismo ancho.

Se puede ver afectada por “outliers”

# Reducción de Datos: Discretización

Igual amplitud



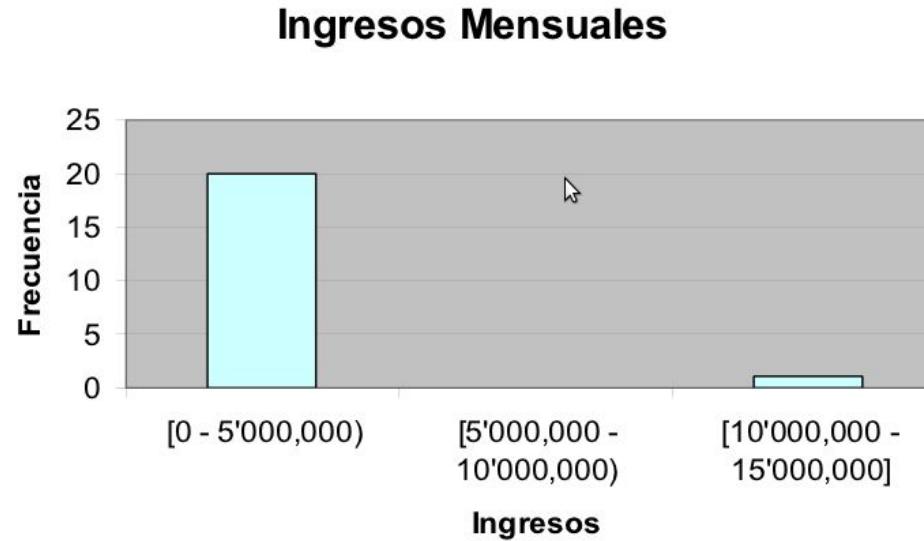
Intervalos de tres  
valores

**Valores de Temperatura:**

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

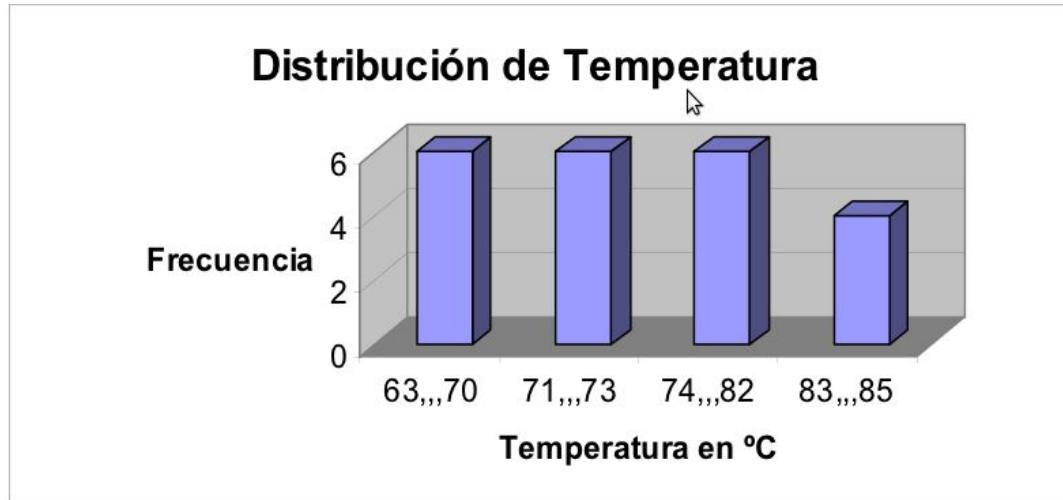
# Reducción de Datos: Discretización

Problemas Igual Amplitud



# Reducción de Datos: Discretización

## Igual Frecuencia



**Valores de Temperatura:**

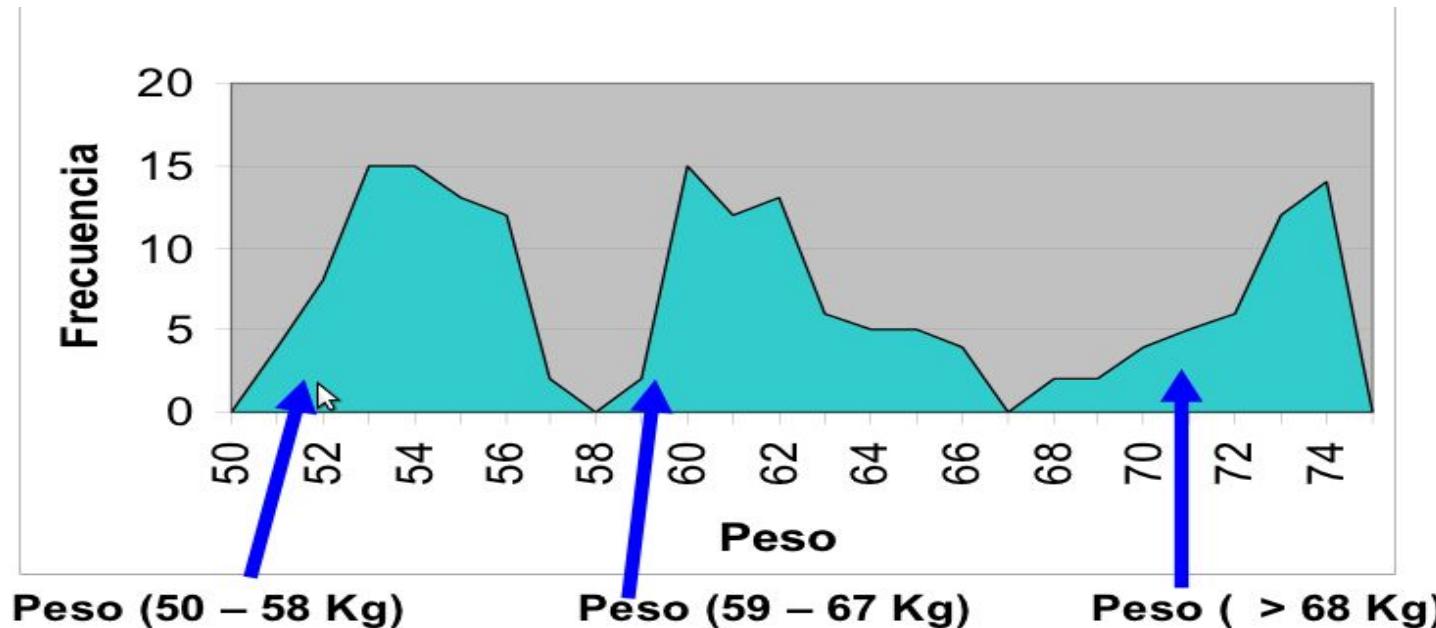
63, 65, 66, 67, 70, 70, 71, 71, 72, 72, 73, 73, 74, 75, 76, 76, 82, 82, 83, 84, 85, 85

# 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 más intuitivos.
- Consideraciones adicionales:
  - Se deben crear cajas para valores especiales
  - Se deben tener puntos de corte interpretables

# Reducción de Datos: Discretización

Distribución de Peso



# 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 número de bins es pequeño, el límite más cercano puede ser usado para representar el bin.

# Reducción de Datos: Discretización - 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**:

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

representación usando la moda:

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

# Reducción de Datos: Discretización - BIN

usando media:

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

Reemplazando por el límite más cercano:

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

Problema de **optimización** en la selección de **k bins**, dado el número 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 número 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

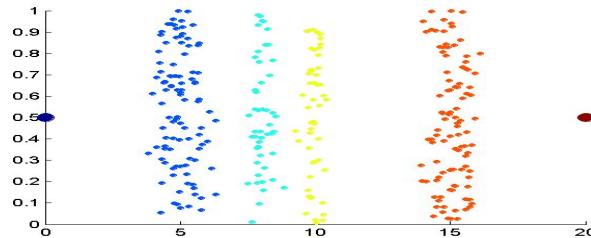
1.  $f \text{ ordenado } = \{1, 1, 2, 2, 2, 5, 6, 8, 8, 9\}$
2.  $\text{Bins iniciales} = \{1, 1, 2, 2, 2, 5, 6, 8, 8, 9\}$
3.  $\text{Modas} = \{1, 2, 8\}$
4.  $\text{Total ER} = 0+0+1+0+0+3+2+0+0+1 = 7$

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

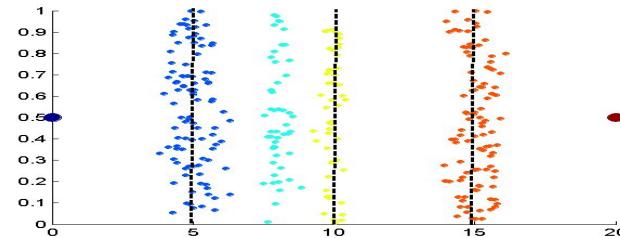
$$\begin{aligned}f &= \{1, 1, 2, 2, 2 \quad 5, 6, \quad 8, 8, 9\} \\ \text{Modas} &= \{2, 5, 8\} \\ \text{ER} &= 4\end{aligned}$$

- Cualquier movimiento de elementos incrementa ER

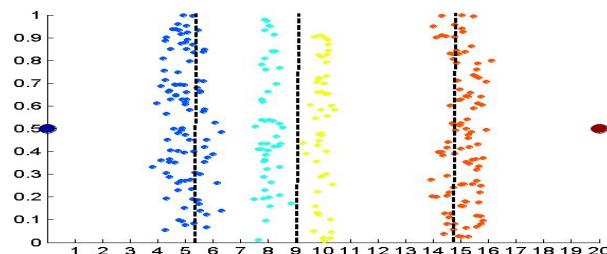
# Discretización sin utilizar etiquetas de clase



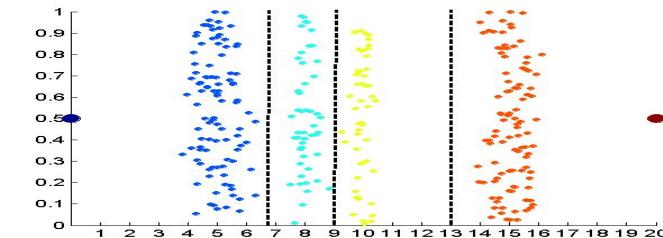
Datos



Amplitud del intervalo de  
Igualdad



la misma frecuencia



K-means

# Discretización Supervisada

- Se usa la información de la **clase** en la discretización

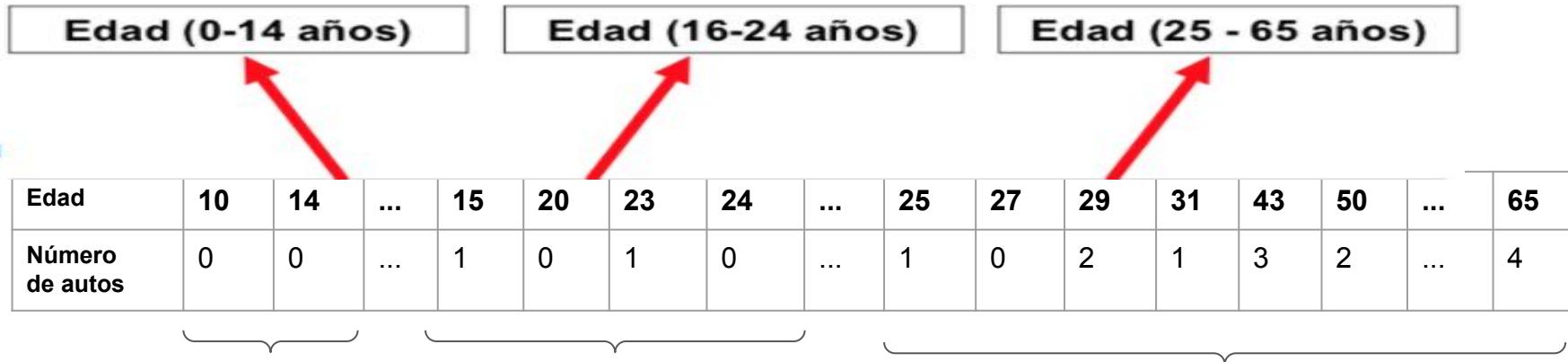
Ejemplo:

Edad	10	14	...	15	20	23	24	...	25	27	29	31	43	50	...	65
Número de autos	0	0	...	1	0	1	0	...	1	0	2	1	3	2	...	4

# Discretización Supervisada

- Se usa la información de la clase en la discretización

Ejemplo:



# Discretización Supervisada

## Basado en la entropía

- Colocar los puntos de separación de tal forma que maximicen la **pureza** de los intervalos (con respecto a la clase).
- Sin embargo, se requiere decisiones en cuanto a:
  - Nivel de pureza en el intervalo
  - Tamaño mínimo de los intervalos
- Los métodos basados en la **entropía** son las más usados.

# Discretización

## Entropía de un intervalo

Se tienen  $k$  categorías o clases,

$m_i$  es el número de valores en el  $i^{th}$  intervalo

$m_{ij}$  es el número de valores de la clase  $j$  en el intervalo  $i$

Entonces la entropía  $e_i$  del intervalo  $i^{th}$  es dado por:

$$e_i = - \sum_{j=1}^k p_{ij} \log_2 p_{ij}$$

donde  $p_{ij} = \frac{m_{ij}}{m_i}$  es la probabilidad de la clase  $j$  en el  $i^{th}$  intervalo

## Entropía total

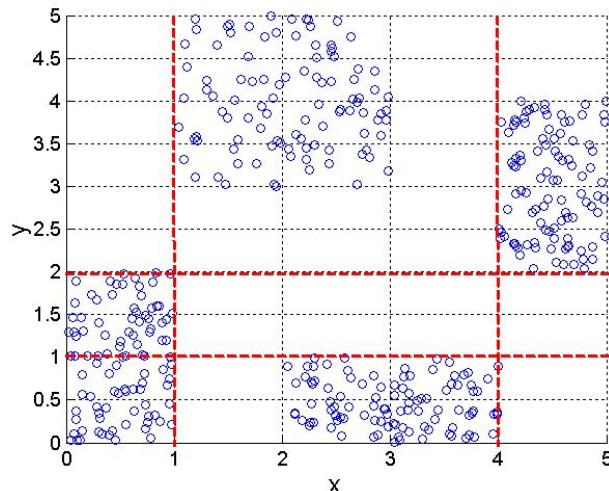
$$e = \sum_{i=1}^n w_i e_i$$

donde  $m$  es el número de valores,  $w_i = \frac{m_i}{m}$  es la fracción de valores en el intervalo  $i^{th}$ , y  $n$  es el número de intervalos

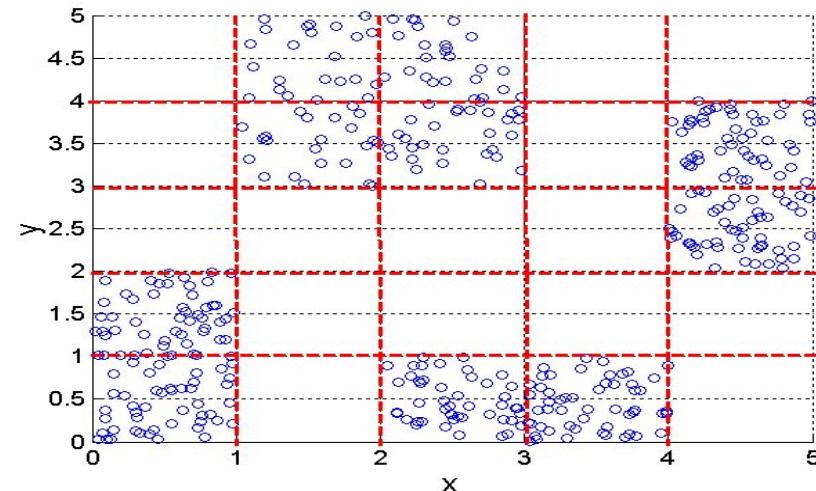
# Discretización

## Uso de etiquetas de clase

- Enfoque basado en la entropía



3 intervalos, tanto para X e Y



5 intervalos tanto para X e Y

Discretización de dos atributos X y Y. En dos dimensiones los puntos se pueden separar, en una dimensión no.

# Ejercicio

- *KBinsDiscretizer*

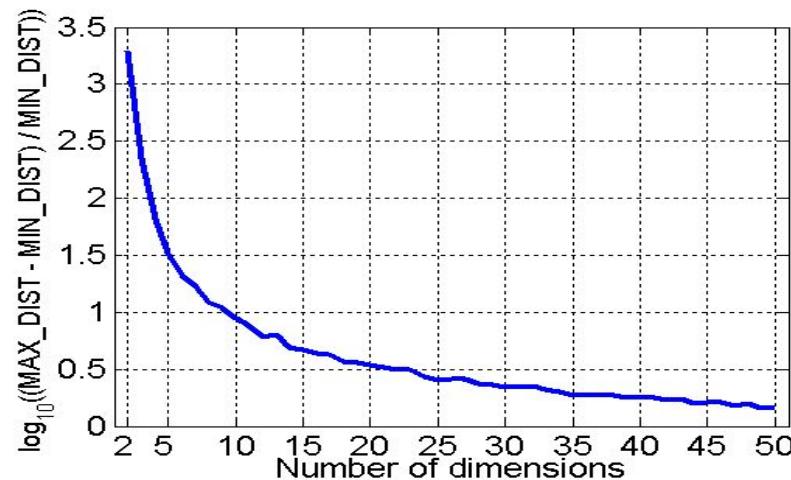
Biblioteca **Sklearn**

- WEKA
  - usando intervalos iguales (binning) y intervalos con igual frecuencia, para esto se sigue la secuencia  
`filter>attribute>unsupervised>discretize`
  - por entropía, para esto se sigue la secuencia:  
`filter>attribute>supervised>discretize.`

# La maldición de la dimensionalidad

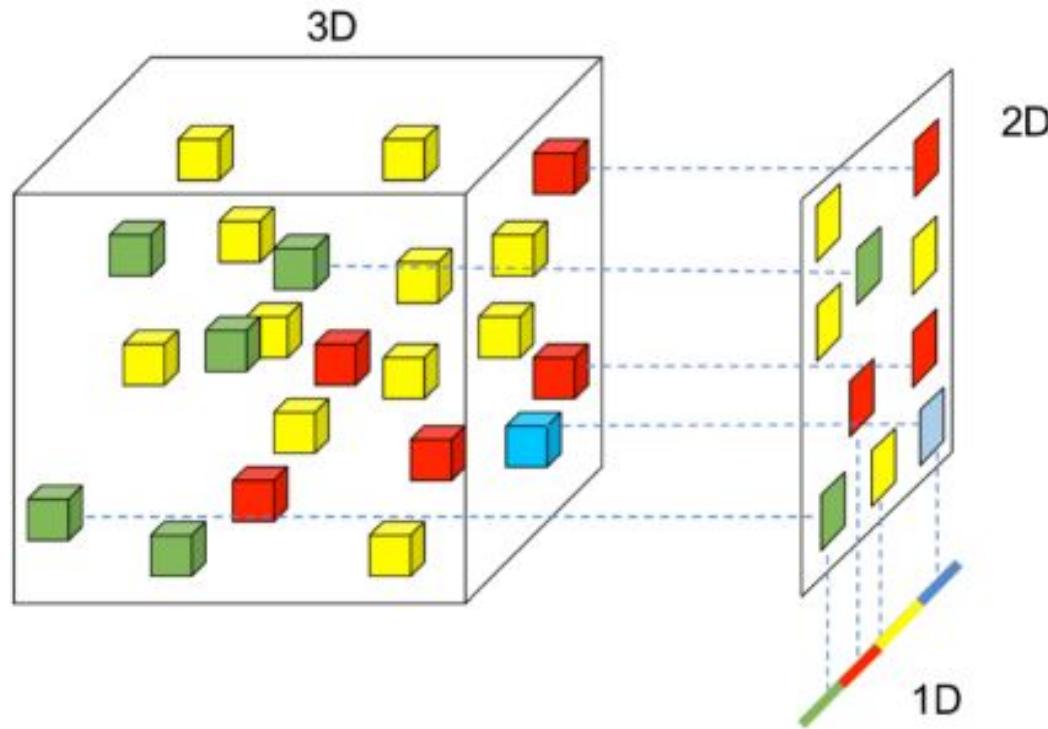
Cuando aumenta la dimensionalidad, los datos se vuelven cada vez más dispersos en el espacio que ocupan.

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.



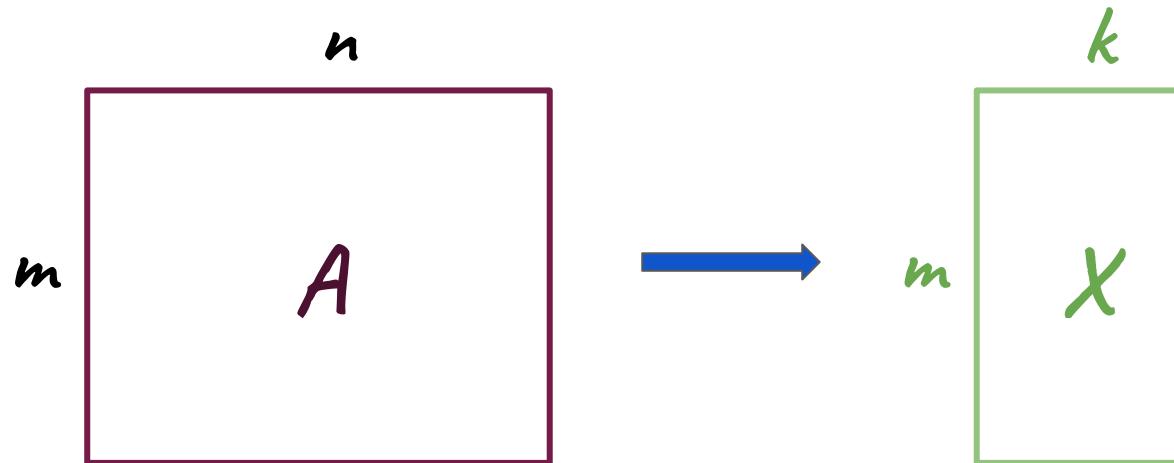
- Generar aleatoriamente 500 puntos  
Calcular la diferencia entre la máxima y  
distancia mínima entre cualquier par de puntos

# Reducción de dimensionalidad



# Reducción de dimensionalidad

Reducir  $n$  variables en  $k$  variables, donde  $k \ll n$



# Reducción de dimensionalidad

- Selección de características o dimensiones (**Feature selection**)
- Extracción de Características (**Feature extraction**)

# Reducción de dimensionalidad:

## Selección de características

- 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 estudiantes suele ser irrelevante para la tarea de predecir perfil de estudiantes .

# Reducción de dimensionalidad:

## Selección de características

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

# Selección de características

Estrategias o enfoques de Selección de características de acuerdo a la evaluación:

- **Embebido** (Embedded): La selección de características ocurre naturalmente como parte del algoritmo de minería. El algoritmo decide cuáles atributos usar y cuales no. Ejemplo: árboles de decisión.
- **Filtro** (Filter): Las características son seleccionadas antes de aplicar el algoritmo de minería con alguna técnica independiente de la tarea de minería a resolver. Ejemplo: seleccionar las características cuya correlación es baja.
- **Envoltura** (Wrapper): Usa el algoritmo de minería como una caja negra para encontrar los mejores atributos. Usa las medidas de validación del algoritmo.

## Medias y Varianzas

- Usado para problemas de clasificación
- Crear subconjuntos de acuerdo a la clase y *Normalizar* las medias por su varianza y compararlas
- Si las medias son muy distantes el interés por la dimensión incrementa dado que son potencialmente útiles para diferenciar las clases
- Dos conjuntos A y B (cada uno representando el subconjunto dividido por clases: A y B)  
 $n_1$  y  $n_2$  son los tamaños de los conjuntos de datos

$$SE(A - B) = \sqrt{var(A)/n_1 + var(B)/n_2}$$

### TEST:

$$|mean(A) - mean(B)|/SE(A - B) > \text{valor de umbral}$$

## Ejemplo: Medias y Varianzas

X	Y	C
0.3	0.7	A
0.2	0.9	B
0.6	0.6	A
0.5	0.5	A
0.7	0.7	B
0.4	0.9	B

- 1 Calcular medias y varianzas de cada una de las variables por cada clase:

$$X_A = \{0.3, 0.6, 0.5\}, X_B = \{0.2, 0.7, 0.4\},$$

$$Y_A = \{0.7, 0.6, 0.5\}, Y_B = \{0.9, 0.7, 0.9\}$$

- 2 Aplicar el test:

$$SE(X_A - X_B) = \sqrt{var(X_A)/n_1 + var(X_B)/n_2} = \\ \sqrt{0.0233/3 + 0.06333/3} = 0.1699$$

$$SE(Y_A - Y_B) = \sqrt{var(Y_A)/n_1 + var(Y_B)/n_2} = \\ \sqrt{0.01/3 + 0.0133/3} = 0.0875$$

# Selección de Características

## Ejemplo: Medias y Varianzas

Usando un umbral de 0.5

- $|mean(X_A) - mean(X_B)|/SE(X_A - X_B) = |0.4667 - 0.4333|/0.1699 = 0.1961 < 0.5$
- $|mean(Y_A) - mean(Y_B)|/SE(Y_A - Y_B) = |0.6 - 0.8333|/0.0875 = 2.6667 > 0.5$

X es candidata para reducción por que los valores de las medias son cercanos por lo que el test final esta por debajo del umbral. En cambio, la variable Y es significativamente mas alta que el umbral, por lo que no es candidata a reducción (es una variable potencialmente útil para distinguir las dos clases).

# Selección de Características:

## 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

# Selección de Características:

## 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

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

Similaridad entre  $x_i$  y  $x_j$

# Selección de Características:

## Entropía

- El algoritmo está basado en “sequential backward ranking”
- La entropía es calculada en cada iteración para decidir el “ranking” de las dimensiones.
- Las dimensiones son gradualmente removidas

# Selección de Características:

## Entropía

### Algoritmo

1. Comienza con todo el conjunto de datos  $F$
2.  $E_F$  = entropia de  $F$
3. 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 de “ranking”
4. Repetir 2-3 hasta que solo haya una dimensión en  $F$

# Selección de Características:

## Entropía

- El proceso puede ser parado en cualquier iteración y las dimensiones son seleccionadas de la lista.
- Desventaja: complejidad
- Implementación paralela

# Selección de Características:

## Entropía

Para enumerar dimensiones (ranking)

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

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

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

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

similaridades

# Extracción de Características

## Propósito:

- Evitar la maldición de la dimensionalidad
- Reducir tiempo y memoria requeridos por los algoritmos de minería de datos
- Permitir mejor visualización de los datos
- Ayudar a eliminar 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

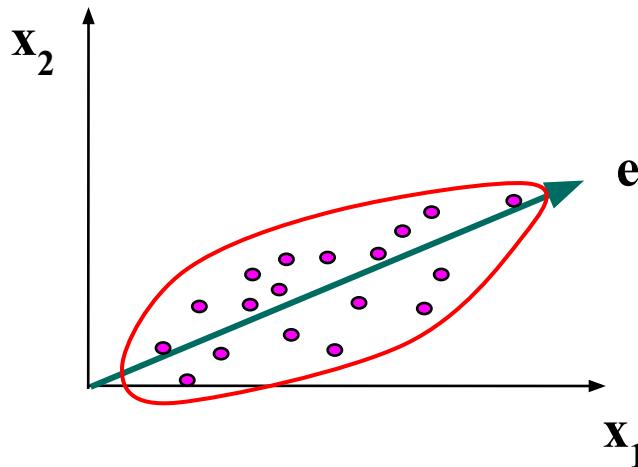
- Es el método de reducción de la dimensionalidad más ampliamente utilizado
- Inventado por Pearson (1901) y Hotelling (1933)
- Utilizado por primera vez en ecología (1954) bajo el nombre de “análisis de factores” (Análisis de factores principales)

# PCA

## Principal Component Analysis

# Reducción de dimensionalidad: PCA

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

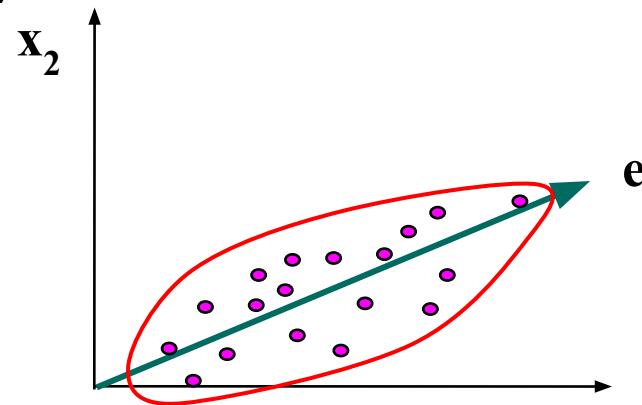


# Reducción de dimensionalidad: PCA

- Toma la matriz de  $m$  objetos por  $n$  variables, las cuales pueden estar correlacionadas, y las **resume** por ejes no correlacionados (principals components o ejes principales) que son **combinaciones lineales de las variables originales**  $n$
- Los primeros componentes mantienen la mayor variabilidad entre los objetos (varianza)

# Reducción de dimensionalidad: PCA

- Encontrar los vectores propios de la matriz de covarianza
- Los vectores propios definen el nuevo espacio (nuevos componentes)



# Reducción de dimensionalidad: PCA

*Dimension 1:* Captura la mayor variabilidad posible

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

etc

.

.

.

# Reducción de dimensionalidad: PCA

## Media, DS, Varianza

media

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

Desviación  
estándar

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

Varianza

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

Solo en una dimensión



# Reducción de dimensionalidad: PCA

## Matriz de Covarianzas

- Medida que permite encontrar que tanto las dimensiones varían 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
<b>Total</b>	<b>4</b>	<b>6</b>	<b>3</b>			5.00	-10.00	-4
<b>media</b>	<b>1.33</b>	<b>2</b>	<b>1</b>			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

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

Vector transformado en su posición original de transformación

Cuatro (4) veces el original

Eigenvector

# 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

Escalar 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

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

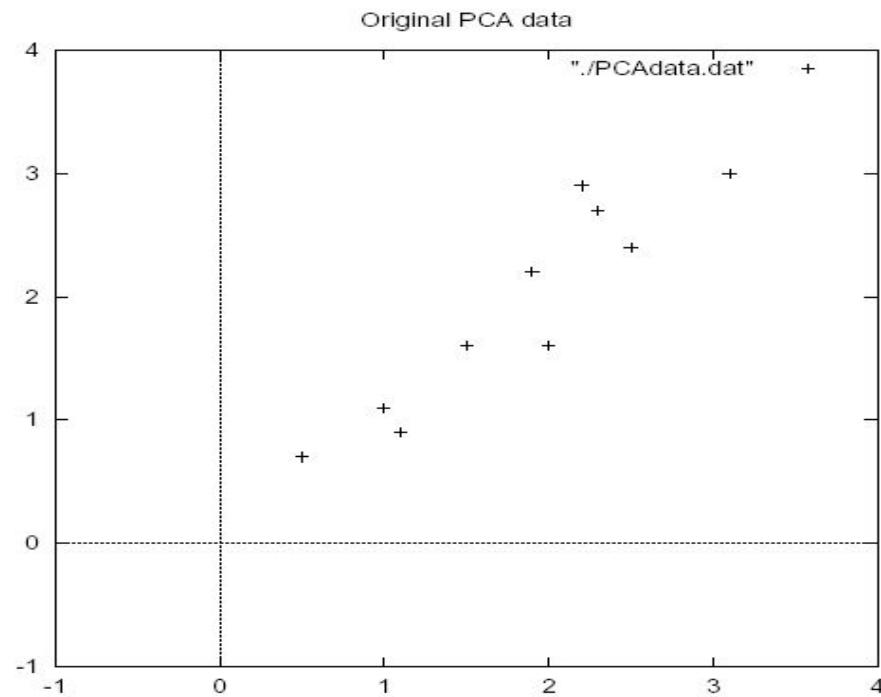
Cual de los siguientes vectores son eigenvectors de la matriz?  
Cual es su correspondiente eigenvalue?

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

# 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

	$x$	$y$		$x$	$y$
Data =	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
	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.

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

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

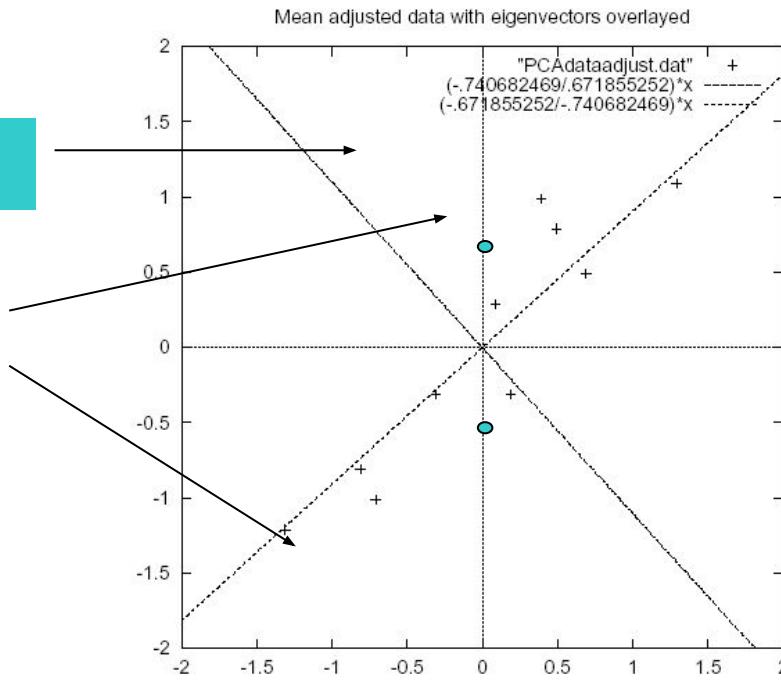
Método para calcular eigenvectors= Jacobi (Investigar)

# Método (Ejemplo)

Menos importante

Eigenvectors  
(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

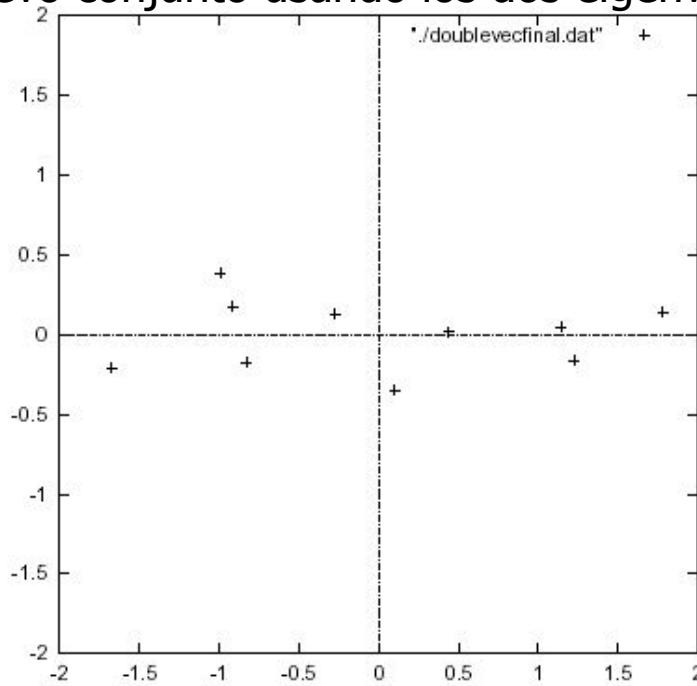
$$(\text{eigenvectors})^T X (\text{Datos ajustados usando la media})^T$$

	$x$	$y$
	-.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

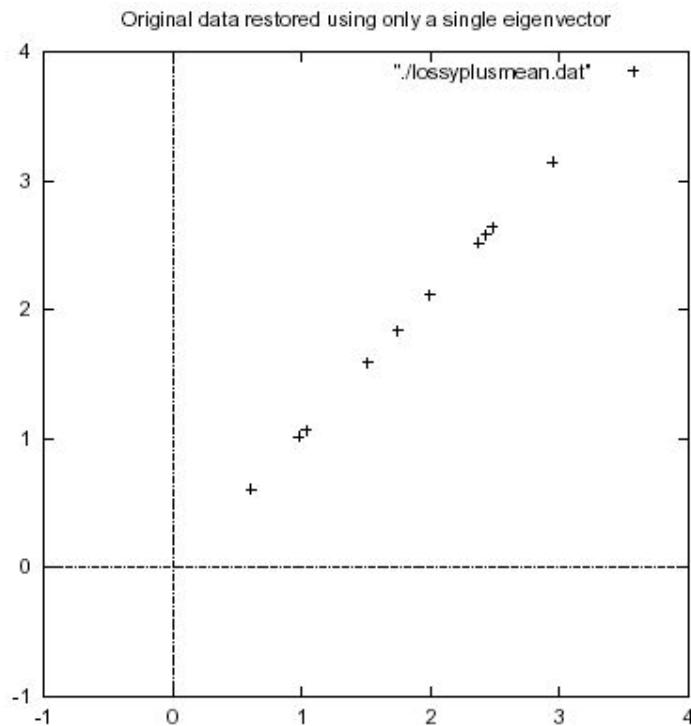
Nuevo conjunto X (eigenvector matrix) $^{-1}$

Nuevo conjunto X (eigenvector matrix) $^T$

Transformed Data (Single eigenvector)

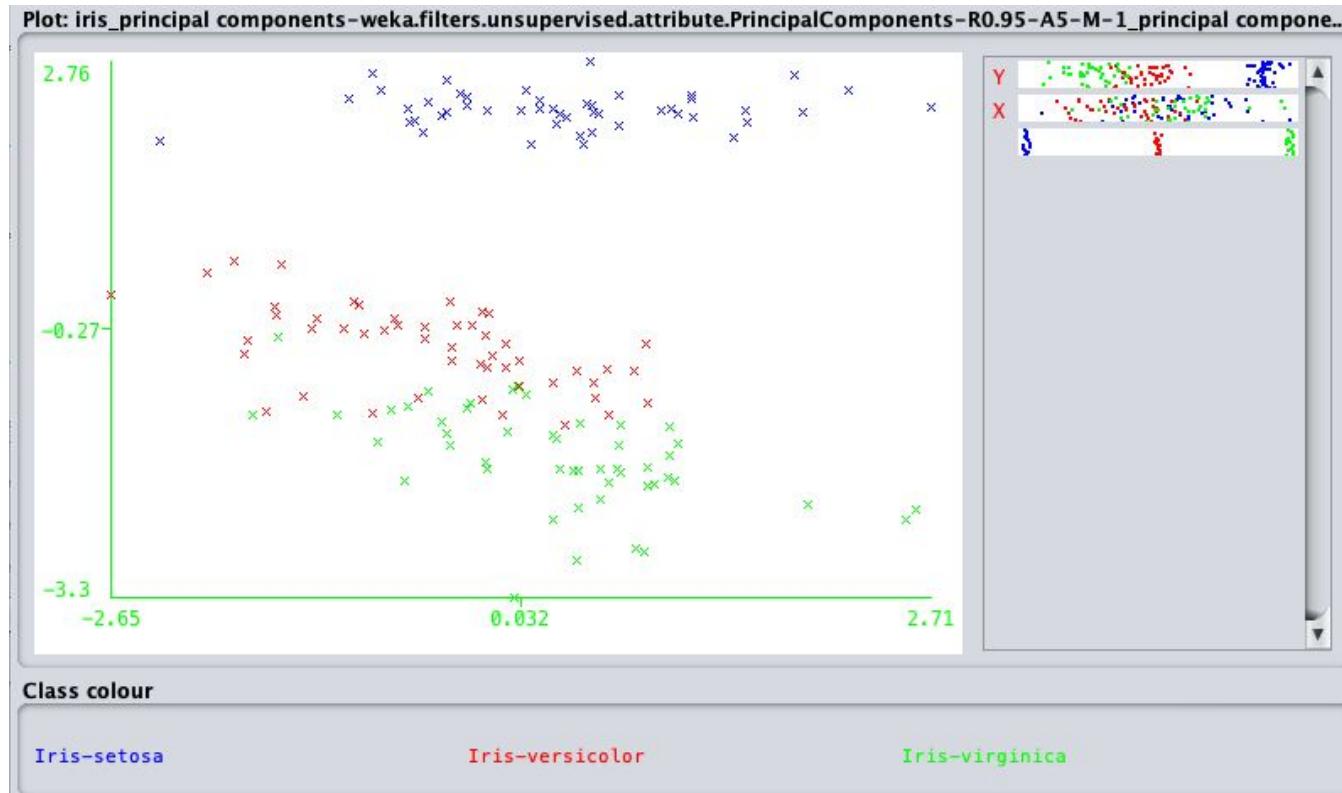
$x$
-.827970186
1.77758033
-.992197494
-.274210416
-1.67580142
-.912949103
.0991094375
1.14457216
.438046137
1.22382056

# Obtener el original conjunto de datos

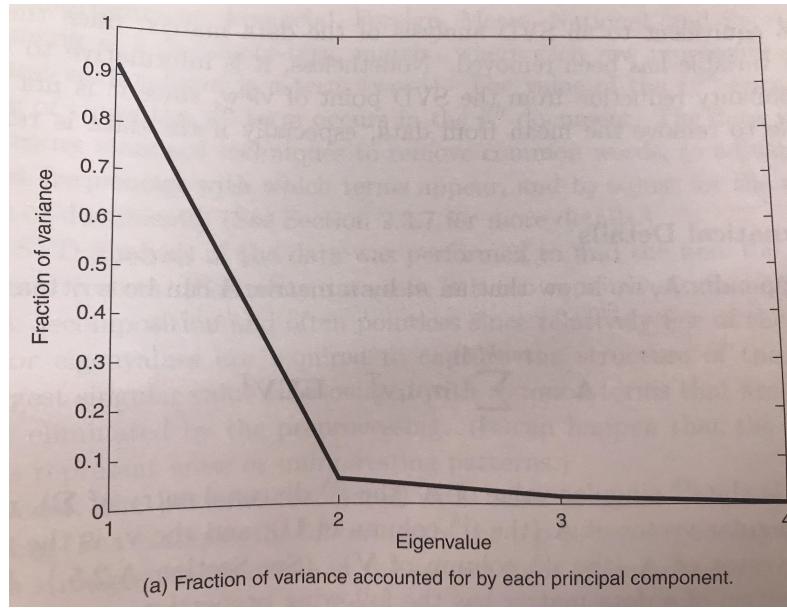


# PCA

## Conjunto de Datos - Iris



# PCA



(a) Fraction of variance accounted for by each principal component.

# PCA

## Python

```
eig_vals, eig_vecs = np.linalg.eig(cov_mat)
```

```
from sklearn.decomposition import PCA  
numComponents = 2  
pca = PCA(n_components=numComponents)
```

# Reducción de dimensionalidad:

## Otros métodos

- Non-negative Matrix Factorization (NMF)
- Multidimensional Scaling (MDS)
- Linear Discriminant Analysis (LDA)
- Generalized Discriminant Analysis (GDA): discriminante no lineal.
- Kernel PCA (matriz de kernel): función de kernel proyecta un conjunto de datos en un espacio de **mayor dimensión**, donde puede ser linealmente separable.

# Reducción de dimensionalidad: Manifold learning

- ISOMAP
- LLE
- t-SNE
- Hessian Eigenmap
- Spectral Embedding
- Modified LLE

# 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
- [5] Lindsay I Smith. “A tutorial on Principal Components Analysis”, 2002
- [6] Hinton, G.. “Visualizing High-Dimensional Data Using t-SNE” van der Maaten, L.J.P. Journal of Machine Learning Research (2008)