Los datos en Tratamiento Inteligente de Datos

Maria-Amparo Vila
vila@decsai.ugr.es

Grupo de Investigación en Bases de Datos y Sistemas de Información Inteligentes https://idbis.ugr.es/
Departamento de Ciencias de la Computación e Inteligencia Artificial
Universidad de Granada

Introducción al tema

Estructura de la presentación

1. Introduccion ideas básicas acerca de los datos
2. Tipos de Datos
3. Problemas de calidad
4. Exploración de los datos
   - 4.1 Exploracion estadística
   - 4.2 Visualización de los datos
5. Transformaciones de los datos
6. Problemas de reducción de variables.
   - 6.1 Selección de variables
   - 6.2 Cambio de coordenadas: componentes principales
7. Problemas de cambio de escala.

Introducción a los datos

Datos de partida

La estructura de datos más habitual para trabajar con DM es el **Dataset**

items\variables	V_1	V_2	$\cdots\cdots$	V_N
o_1	d_{11}	d_{12}	$\cdots\cdots$	d_{1N}
$\vdots$	$\vdots$	$\vdots$	$\cdots\cdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\cdots\cdots$	$\vdots$
o_M	d_{M1}	d_{M2}	$\cdots\cdots$	d_{MN}

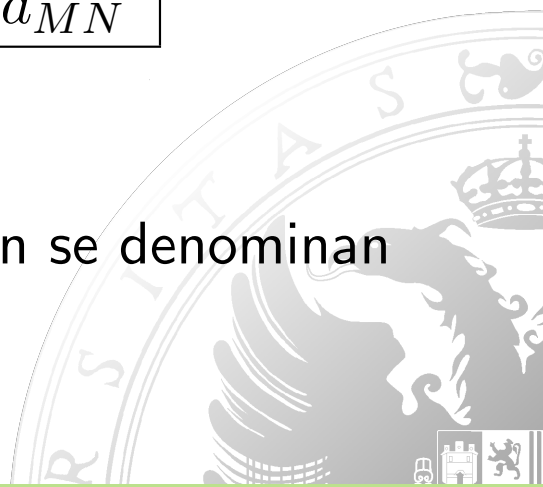
Introducción a los datos

Datos de partida

La estructura de datos más habitual para trabajar con DM es el **Dataset**

items \ variables	V_1	V_2	$\cdots\cdots$	V_N
o_1	d_{11}	d_{12}	$\cdots\cdots$	d_{1N}
$\vdots$	$\vdots$	$\vdots$	$\cdots\cdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\cdots\cdots$	$\vdots$
o_M	d_{M1}	d_{M2}	$\cdots\cdots$	d_{MN}

- los items representan los casos, objetos etc.
- Las variables pueden ser de muchos tipos. Tambien se denominan factores
- Puede haber datos perdidos



Introducción a los datos

Datos de partida

Los data set pueden obtenerse a partir de datos previos, mediante transformaciones, resúmenes etc. En algunos casos este es un punto clave (selección de factores, text mining etc.)

Introducción a los datos

Datos de partida

Los data set pueden obtenerse a partir de datos previos, mediante transformaciones, resúmenes etc. En algunos casos este es un punto clave (selección de factores, text mining etc.)

Existen problemas en los que la estructura de data set no es adecuada:

- Estructuras transaccionales
- Minería de grafos (se buscan patrones de estructuras)
- Minería de secuencias (Biocomputación)

Introducción a los datos

Datos de partida

Los data set pueden obtenerse a partir de datos previos, mediante transformaciones, resúmenes etc. En algunos casos este es un punto clave (selección de factores, text mining etc.)

Existen problemas en los que la estructura de data set no es adecuada:

- Estructuras transaccionales
- Minería de grafos (se buscan patrones de estructuras)
- Minería de secuencias (Biocomputación)

En la mayoría de los casos se pueden transformar una representaciones en otra con objeto de aplicar la técnica adecuada

A partir de ahora,salvo indicación en contra, nos centraremos en la estructura de data set

Tipos de datos

Atributos numéricos

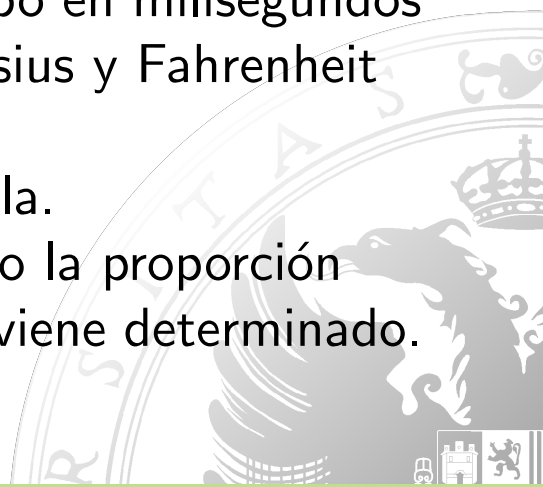
Tiene un dominio numérico lo que permite realizar operaciones aritméticas.
Se pueden clasificar en:

- **Discretos**:
  Son numeros enteros o naturales.
  Habitualmente resultados de conteo.
  Permiten cálculos numéricos, según su dominio.
  No confundir con atributos categoricos transformados.
  Nivel de juego 1 2 o 3 **no es** un atributo numérico

Tipos de datos

Atributos numéricos

- **Continuos**:
  Corresponden a números reales.
  Admiten cálculos y métodos estadísticos más avanzados
  A veces son demasiado detallados y pueden tener problemas de redondeo.
  Tienen siempre un punto de partida (valor cero) y un factor de escala. Según esto se clasfican en:
  - **"Interval"** El cero y la escala son arbitrarios. (tiempo en milisegundos y punto de partida arbitrario, la temperatura en Celsius y Fahrenheit etc.)
  - **"Ratio"** El cero no se elige pero si el factor de escala. (Distancias,altura, peso, volumen etc.) Tiene sentido la proporción
  - **"Absolute"** Tanto el cero como el factor de escala viene determinado. (Cualquier forma de porcentaje, frecuencia etc.)

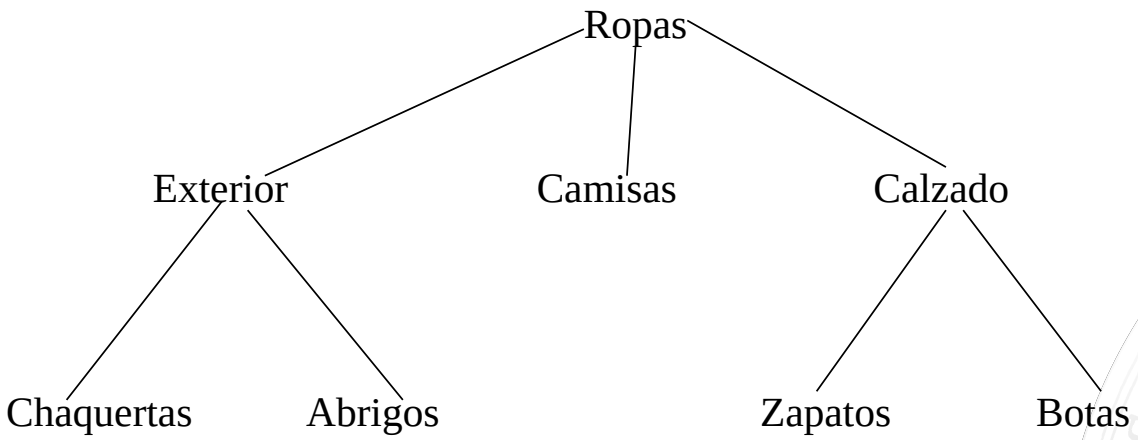


Tipos de datos

Atributos simbólicos, categóricos o nominales

Tienen un dominio discreto de valores no numéricos.
No permiten operaciones aritméticas. En principio solo la igualdad
Pueden admitir una estructura jeráquica. con distintos **niveles de granularidad**

Ejemplo de entidades simbólicas



Tipos de datos

Atributos simbólicos, categóricos o nominales

Algunos dominios de atributos simbólicos están ordenados (cursos académicos, pronóstico de una enfermedad etc.). Se denominan **atributos ordinales**, y admiten operadores de comparación.
Los atributos binarios (presencia/ausencia) son una forma de atributo ordinal
El ejemplo más claro de atributo ordinal es la fecha. Que admite granularidad
Ejemplo

$$FECHA \longrightarrow MES \longrightarrow TRIMESTRE \longrightarrow AÑO$$

Calidad de los datos

Accuracy (corrección y precisión)

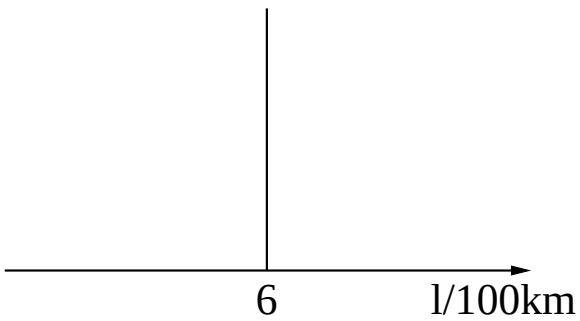
Accuracy *Parecido entre el valor del dato y el verdadero valor del atributo.*

- Para el caso de atributos numéricos:
  - Pueden exisitir errores de redondeo y hay que unificar precisión.
  - Pueden existir valoraciones imprecisas: intervalares o difusas. Habrá que tratarlas con herramientas adecuadas: semejanzas, fuzzy clustering, reglas de asociación difusas etc.
- Para el caso de atributos simbólicos
  .Se detectan errores en el dato
  - Hay detección sintáctica
  - Hay detección semántica

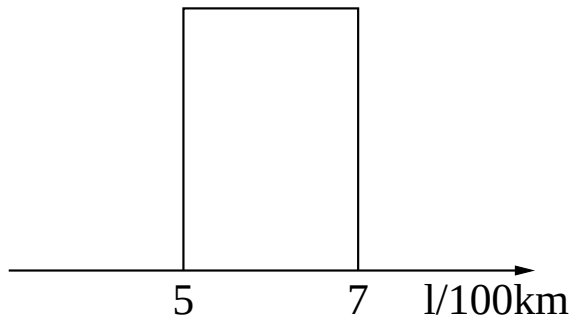
Calidad de los datos

Accuracy (corrección y precisión)

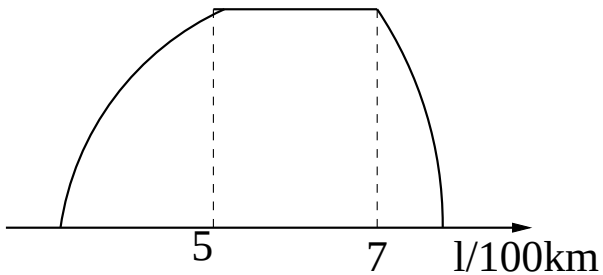
Ejemplo de valores imprecisos



Consumo 6



Consumo entre 5 y 7



Consumo moderado

Calidad de los datos

Completeness (Datos Completos)

Completeness Seguridad de que hay suficientes datos y de que no falta ningun valor de un atributo

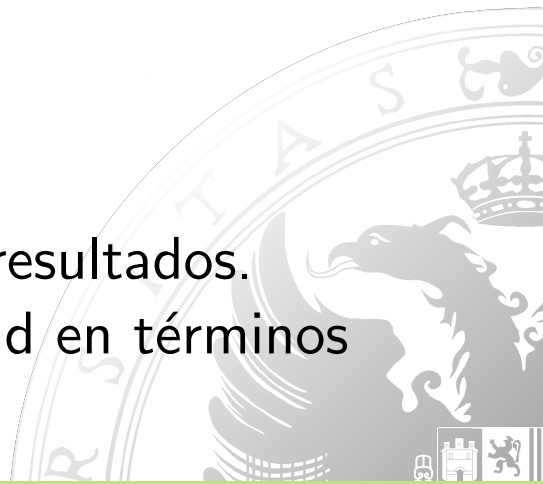
- Valores perdidos en los atributos:

  . Para el caso de atributos numéricos hay técnicas estadísticas que veremos al final

  . Para el caso de atributos simbólicos puede haber conocimiento previo que se pueda usar. (dependencias funcionales etc.)
- Falta de items:
  1. Registros o tuplas perdidas
  2. Información sesgada
  3. Datos dispersos

  . En este caso es dificil asegurar la calidad de los resultados.

  . Una gran cantidad de datos no asegura su calidad en términos del dominio de los items (casos 2 y 3)



Calidad de los datos

Otros problemas con la calidad de los datos

Anomalias (Outliers) Se trata de items que realmente no pertenecen al colectivo que se quiere estudiar y que distorsionan la regularidad que se busca

Desfase temporal (timeliness) Se refiere al hecho de que los datos o parte de ellos no tengan la misma "actualidad" que otros.

Datos desequilibrados Puede darse a dos niveles:

- Existen muchos items de un tipo y muy pocos de otro (problemas de clasificación/predicción)
- Los valores de un atributo numérico son mucho mayores que los de otros con lo que tiene mucho más peso

Datos Duplicados Aparecen, cuando se fusionan bases de datos. Habrá que limpiar los datos previamente

Exploración de los datos

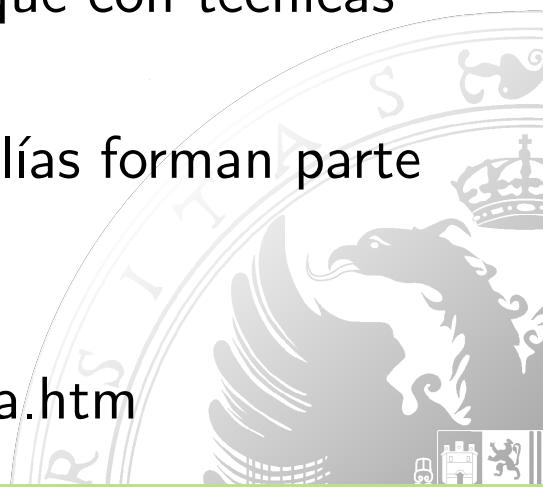
Por qué y para qué

Motivaciones para explorar los datos

- Ayuda a elegir las mejores herramientas para preprocesar y analizar
- Permite formular hipótesis iniciales sobre patrones a extraer ya que se explota la habilidad del ser humana para reconocer patrones

El análisis exploratorio de Datos(EDA) 1977

- Es debido a Tuckey
- El EDA está enfocado a la visualización. Supone que con técnicas adecuadas se puede extraer conocimiento directo.
- Para Tuckey el Clustering y la detección de anomalías forman parte del EDA
- Más información http://www.itl.nist.gov/div898/handbook/eda/eda.htm



Exploración de los datos

Exploración basada en estadística descriptiva

Distribución de frecuencias Se pueden considerar:

- Frecuencias absolutas: $F(d_i)$ número de veces que aparece un determinado valor d_i

  Cálculo de las frecuencias absolutas
  - Atributos discretos: Categóricos y numéricos enteros y finitos (con no muchos datos en el dominio). Simple conteo sobre los valores del dominio D
  - Atributos continuos. Se impone la dicretización, el dominio se transforma en discreto mediante intervalos.
    Normalmente. Si $D = [A, B]$ y queremos m intervalos se eligen de igual amplitud

    $$[a_1, a_2], ...[a_{m-1}, a_m] \ , \ a_1 = A, \ a_i = a_{i-1} + (B - A)/m \ \forall i = 2, ..m$$

    El problema de la discretización puede ser complejo para temas de visualización y asociación. En algunos casos es un tema de preprocesamiento.

Exploración de los datos

Exploración basada en estadística descriptiva

Distribución de frecuencias Se pueden considerar:

- Frecuencias relativas: $f(d_i)$ razón entre la frecuencia absoluta y el número total de items $f(d_i) = F(d_i)/M$, M es el número total de items. Se puede dar también en porcentajes $(f(x_i) * 100)$
- Frecuencias acumulativas Se definen sólo para datos ordenados. Supongamos el conjunto de valores $D = \{d_1, ..d_m\}$ y $d_1 \leq d_2 \leq ... \leq d_n$, se define:

$$fa(d_i) = \sum_{j=1}^{j=i} f(d_j)$$

Cuando se utilizan porcentajes $fa(d_i)$ nos indica el porcentaje de la población $\leq d_i$ y conduce a concepto de:

Exploración de los datos

Exploración basada en estadística descriptiva

Distribución de frecuencias

- Percentil Sea s un valor entre 0 y 100. Definimos:

$$p_s = max\{d \in D / fa(d) \leq s\}$$

Es decir el mayor valor del dominio que tiene por debajo al s por ciento de la población. Cuando $s = 0, 25, 50, 75, 100$ se denominan cuartiles

Medidas de centralización

- Para datos numéricos
  Media: $\overline{d} = \frac{\sum_{d \in D} d}{M}$
- Para todo tipo de datos:
  Moda: el valor más frecuente
  Mediana: El percentil 50.

Exploración de los datos

Exploración basada en estadística descriptiva

Medidas de dispersión

- Para datos numéricos
  Varianza $s^2 = \frac{\sum_{d \in D}(d-\bar{d})^2}{M-1}$, s es la Desviación típica

  Media de la desviación absoluta $AAD = \frac{\sum_{d \in D}|d-\bar{d}|}{M}$
  Mediana de la desviación absoluta
  $MAD = mediana\{|d - \bar{d}|; d \in D\}$
- Para todos los datos:
  Rango intercuartiles $r = p_{75} - p_{25}$

Exploración de los datos

Exploración basada en estadística descriptiva

Exploración de la relación entre atributos numéricos

Matriz de covarianzas Sean las variables V_j, V_k

$$cov(V_j, V_k) = \frac{\sum_{i=1}^{M}(d_{ij} - \overline{d_j})(d_{ik} - \overline{d_k})}{M-1}$$

Matriz de correlación Sean las variables V_j, V_k

$$corr(V_j, V_k) = \frac{cov(V-j, V_k}{s_j s_k}$$

- Cuando $corr(V_j, V_k) \approx 1$ o $corr(V_j, V_k) \approx -1$ existe una relación lineal entre ambos atributos y uno de ellos puede expresarse en función de otro. Esto puede servir para reducir variables.

Exploración de los datos

Visualización

Ideas básicas Se pueden visualizar:

Objetos Se representa un objeto en un gráfico, como un valor de uno, dos o tres atributos. Aparecen, nubes de puntos, colores distintos etc.

Atributos Según sean:

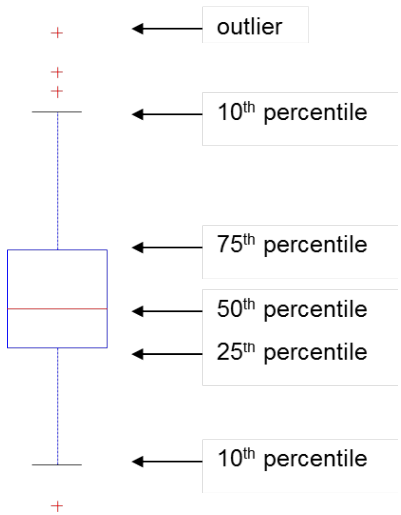
- Atributos categóricos: Diagramas de barras, diagramas de sectores. Uso de colores en la representaciones de objetos etc.
- Atributos numericos y ordinales: Histograma de frecuencias absolutas, relativas, acumulativas, **diagrama de cajas (box plot)**

Relaciones Representaciones conjuntas de atributos **Nube de puntos (scatter plot)** Histogramas bidimensionales. Representacion conjunta de diagramas de cajes etc.

Exploración de los datos

El diagrama de cajas (Tukey)

Es otra forma de ver la distribución de los datos



Para Atributos numéricos y continuos se suele hacer también usando, media en lugar de mediana y $\pm 1s, 2s, 3s$.. en lugar de percentiles. Los outliers se consideran a partir de $\pm 3s$

Exploración de los datos

Ejemplo

El dataset del IRIS

Puede obtenerse en:

http://www.ics.uci.edu/ mlearn/MLRepository.html

- Cuatro atributos numericos: longitud de pétalos y sépalos, ancho de pétalos y sépalos
- Un atributo de categórico:*setosa*, *virginica y versicolour*
- 150 items, 50 de cada tipo

Exploración de los datos

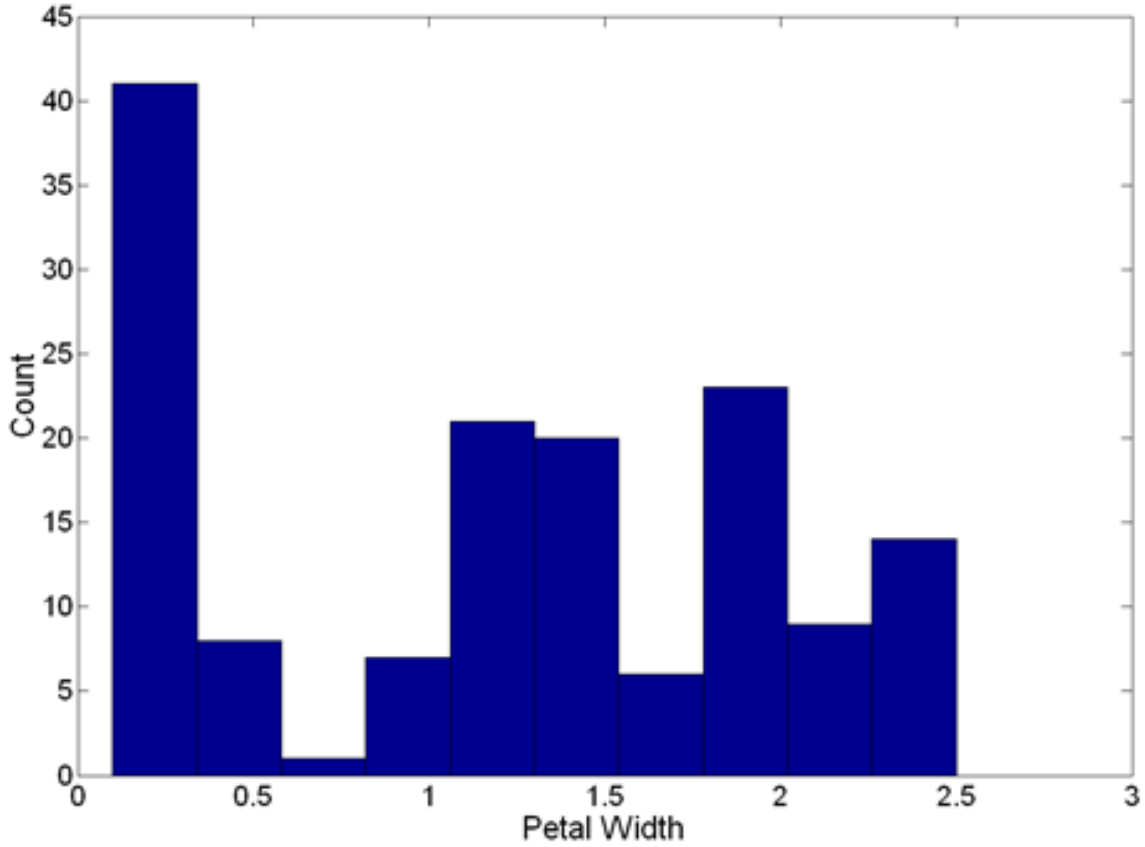
Ejemplo: medidas estadisticas

Numeric columns | Nominal columns

Row ID	D sepal l...	D sepal ...	D petal l...	D peta
Minimum	4.3	2	1	0.1
Maximum	7.9	4.4	6.9	2.5
Mean	5.843	3.057	3.758	1.199
Std. deviation	0.828	0.436	1.765	0.762
Variance	0.686	0.19	3.116	0.581
Overall sum	876.5	458.6	563.7	179.9
No. missings	0	0	0	0

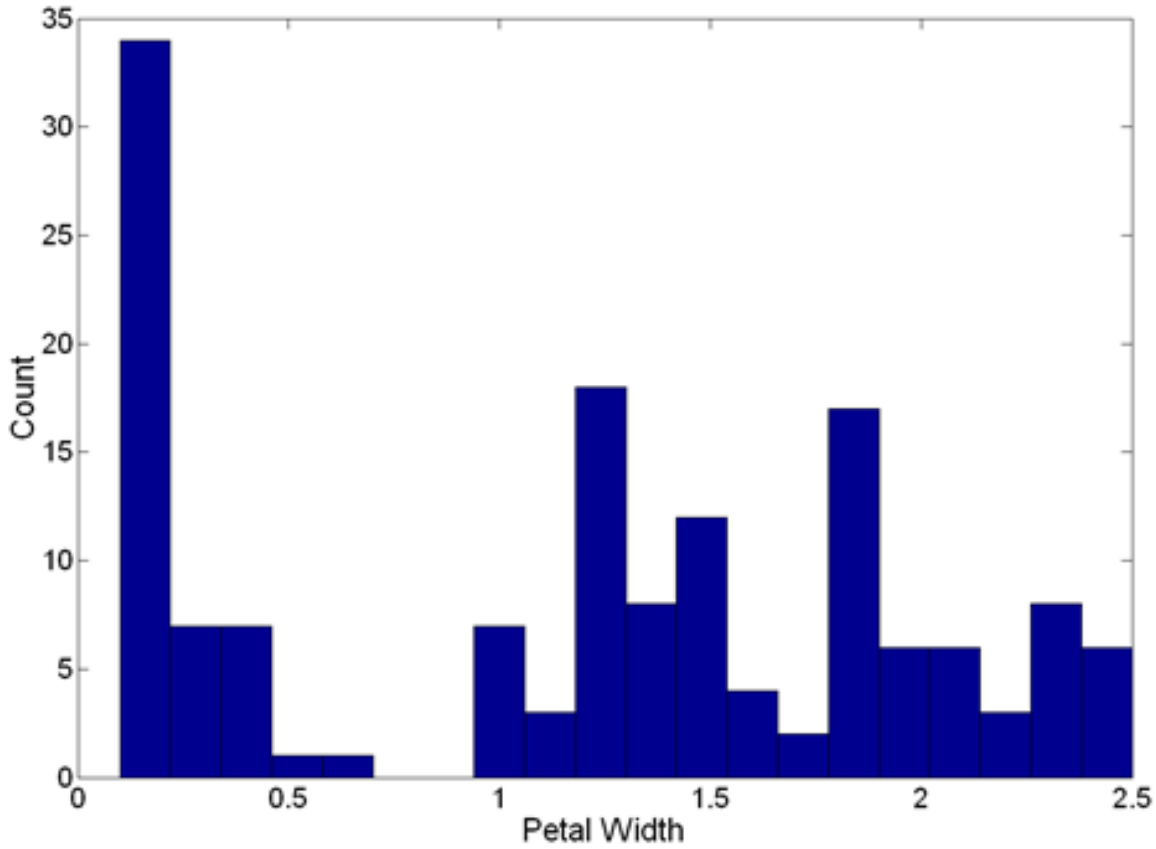
Exploración de los datos

Ejemplo: histograma con 10 intervalos



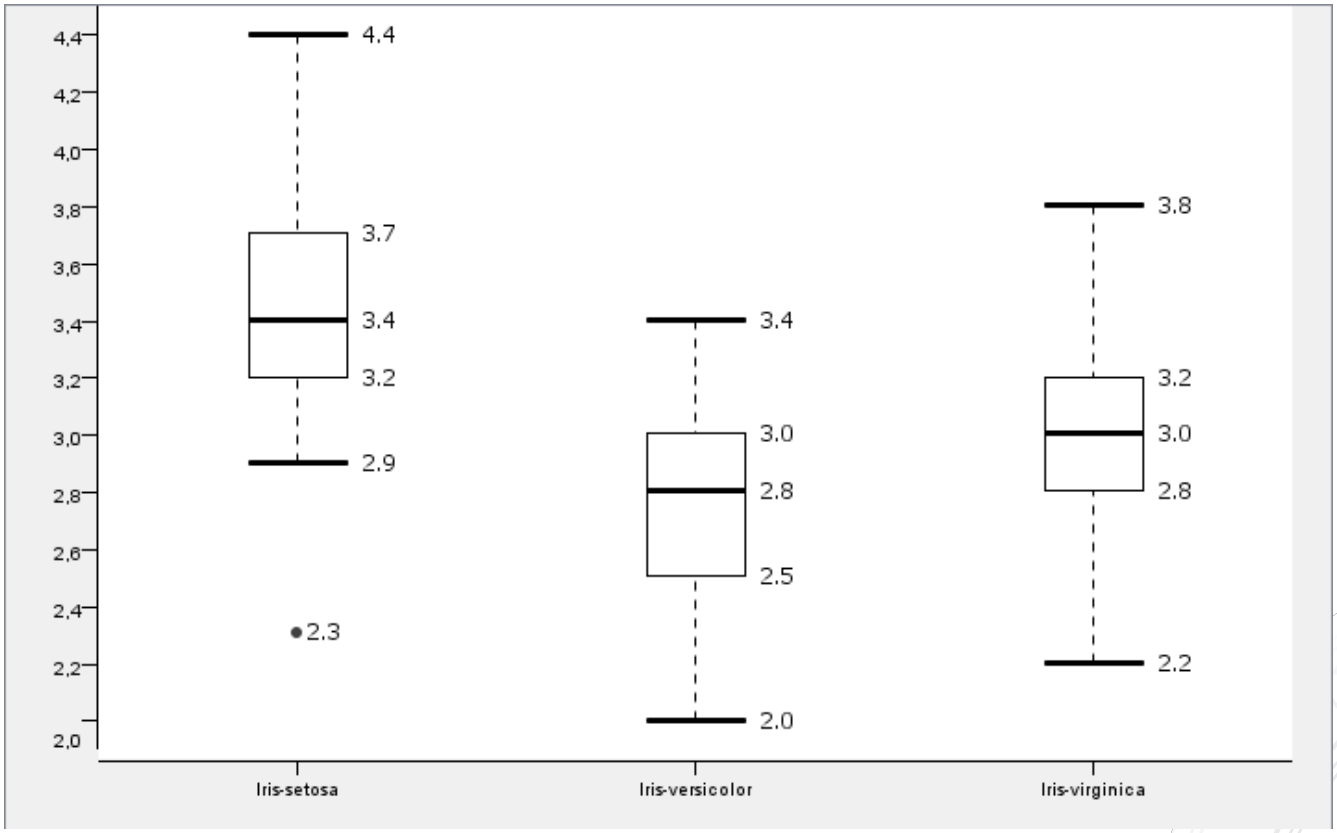
Exploración de los datos

Ejemplo: histogramas con 20 intervalos



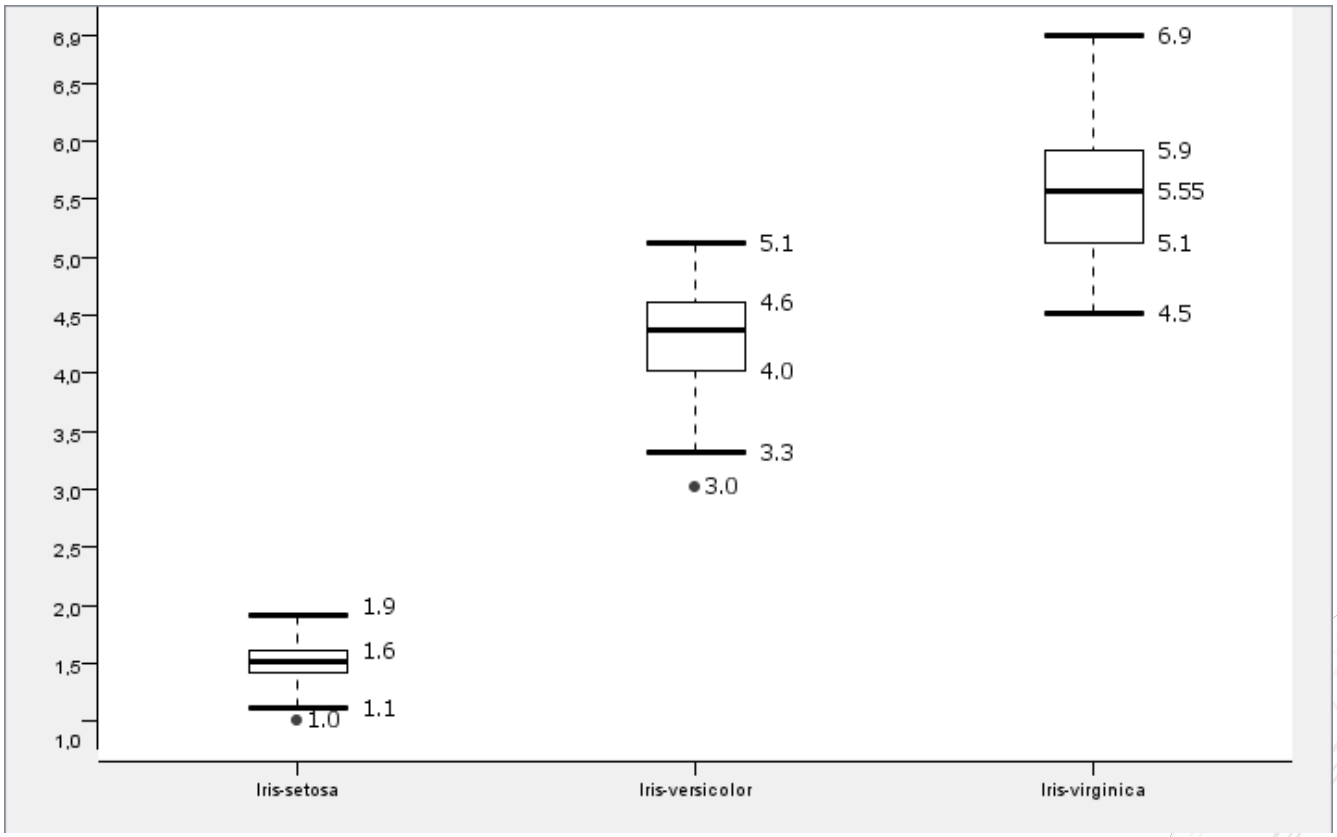
Exploración de los datos

Ejemplo: diagrama de cajas(petal length)



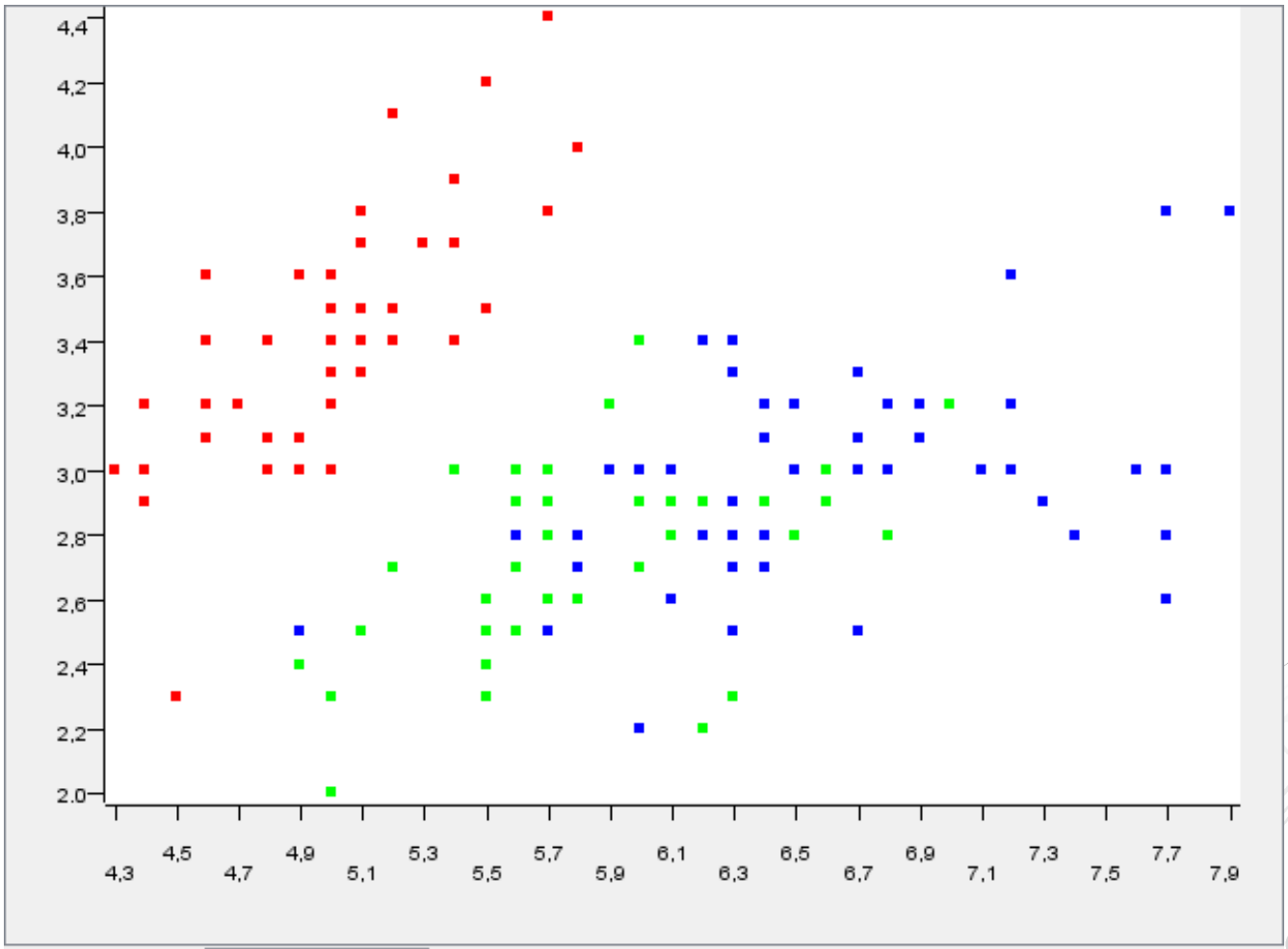
Exploración de los datos

Ejemplo: diagrama de cajas (petal width



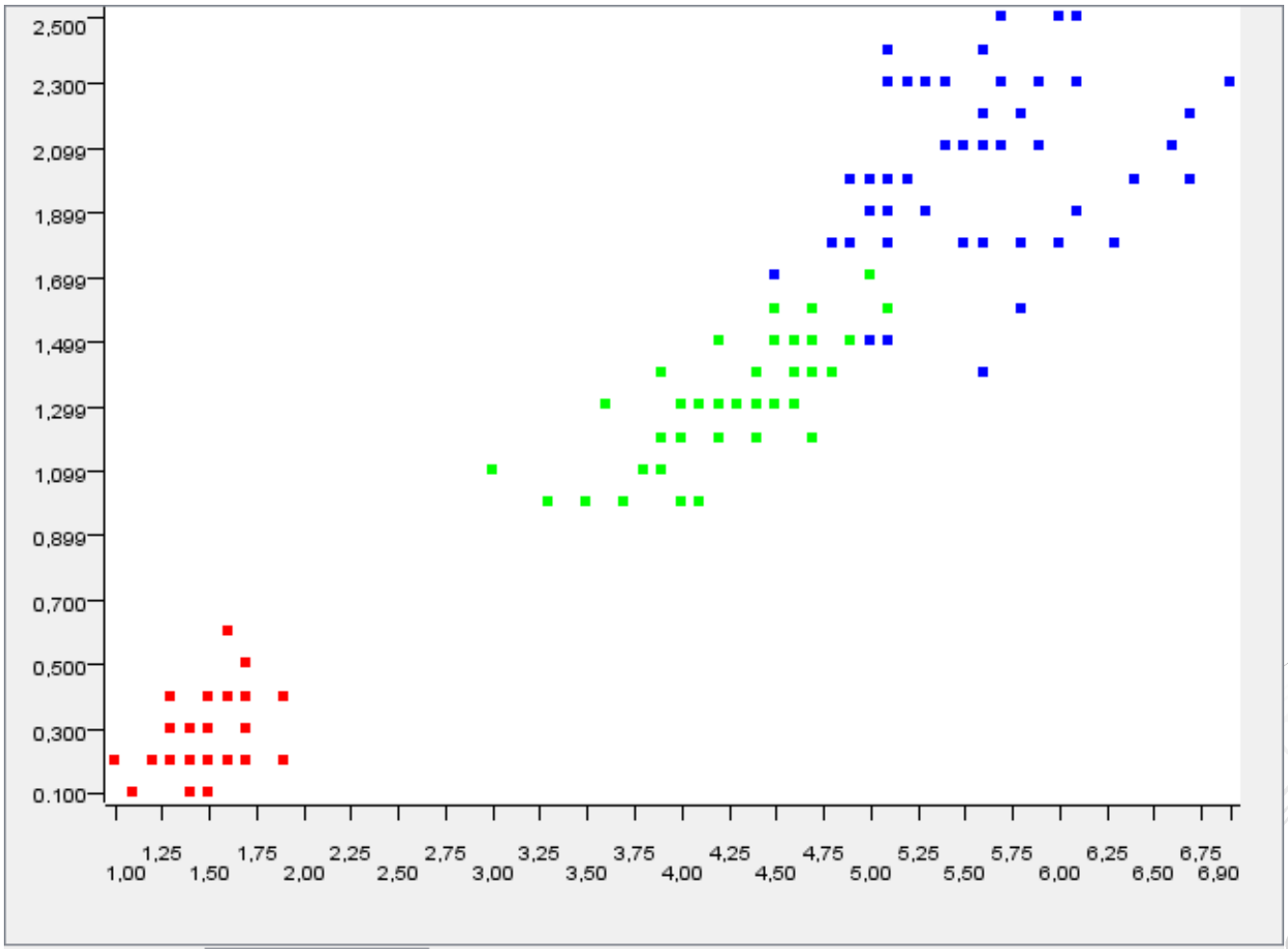
Exploración de los datos

Ejemplo: gráfica de puntos (sepal length/sepal width)



Exploración de los datos

Ejemplo: gráfica de puntos (petal length/petal width)



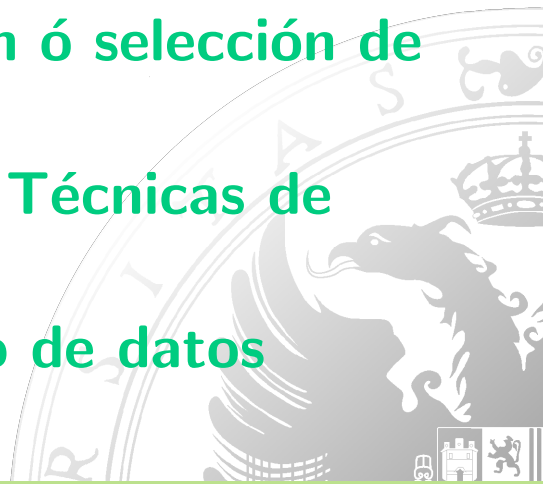
Transformaciones en los datos

Ideas básicas

Antes de aplicar técnicas de DM, en la mayoría los casos es necesario preprocesar (transformar) los datos

Por qué transformar los datos?

- Los datos necesitan ser transformados porque no pueden ser tratados directamente. (Cambio de fecha de nacimiento a edad)
- Los datos son demasiado detallados:**agregar, resumir, discretizar, transformar**
- Hay demasiadas variables: **Técnicas de reducción ó selección de factores**
- Hay mucha diferencia entre los rangos de valores. **Técnicas de cambio de escala**
- Hay mucho datos perdidos **Técnicas tratamiento de datos perdidos**



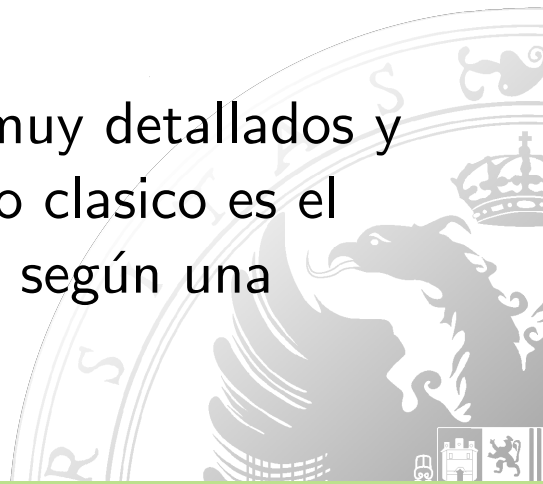
Transformaciones en los datos

Agregación, resumen, discretización

Agregación

Agregar datos consiste en combinar varios objetos para obtener uno nuevo. En general es necesario agregar cuando la información es demasiado detallada. Tenemos:

- Agregación horizontal (resumen) Los datos son muy detallados a nivel de objeto y hay que resumir los atributos. Son las situaciones clásicas de OLAP: datos a nivel de pueblos se agregan en zonas, etc.
- Agregación vertical Los datos de un atributo son muy detallados y es necesario agregar a un nivel superior. El ejemplo clasico es el tiempo. Otro la agregación semántica de términos según una ontología



Transformaciones en los datos

Agregación, resumen, discretización

Discretización

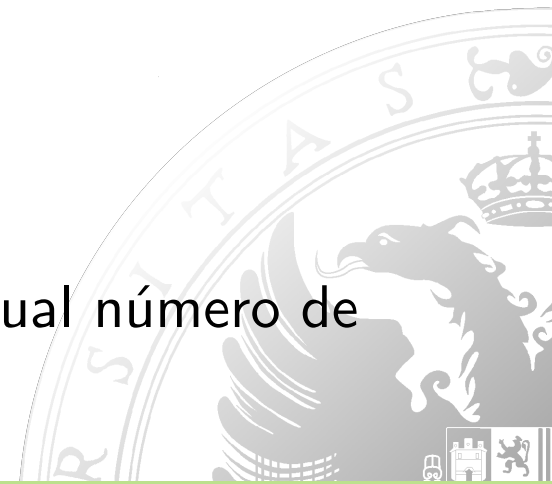
Discretizar datos consiste en sustituir un atributo numérico continuo por uno categórico. Es necesario en extracción de reglas asociación y en ciertos procesos de clasificación.

Proceso:

1. Se eligen k intervalos
2. Se asocia cada valor al punto medio del intervalo donde está situado y se renombra dicho punto medio.

Existen varios enfoques:

- Intervalos igualmente distribuidos Es lo standard
- Intervalos de igual frecuencia. Intenta que haya igual número de casos en cada intervalo.



Transformaciones en los datos

Agregación, resumen, discretización

Discretización

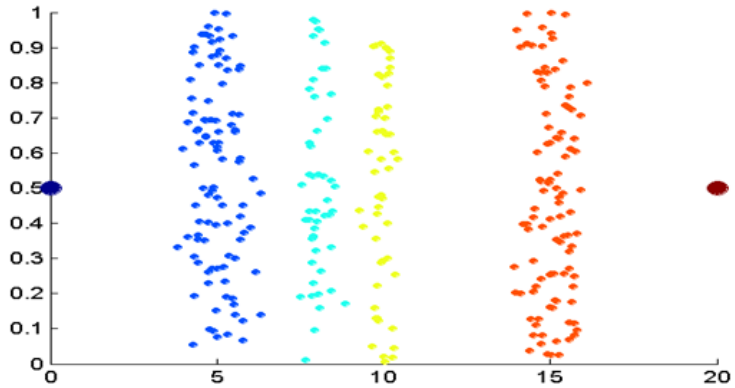
- Intervalos obtenidos por agrupamiento. Se aplica un agrupamiento particional (K-medias) considerando solo el atributo a discretizar. Los centroides nos dan los valores a sustituir.

  Existen métodos de discretización lingüística basados en agrupamiento difuso.

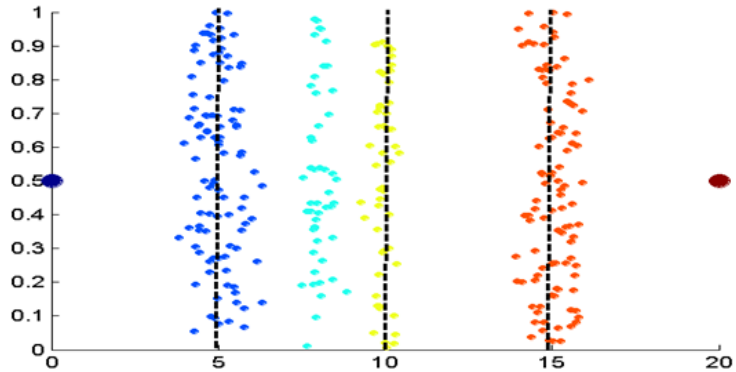
Transformaciones en los datos

Ejemplo de discretización

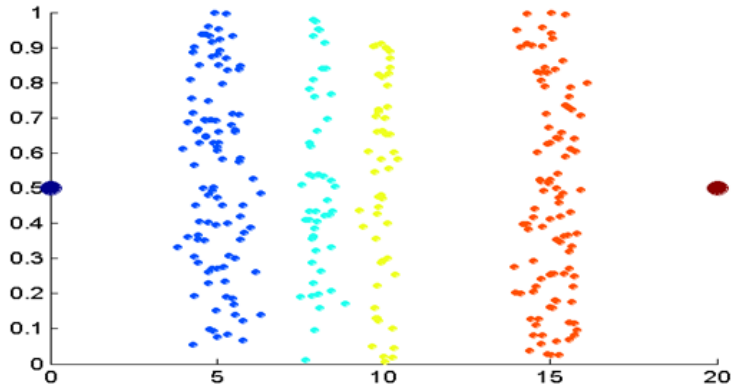
Datos



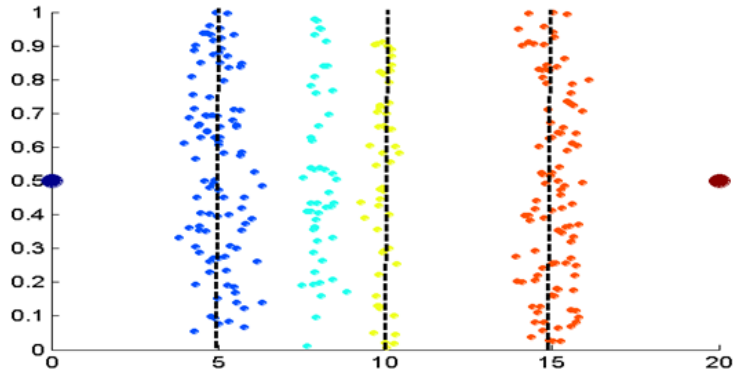
Intervalos Iguales



Frecuencia igual



Agrupamiento



Reducción de variables

Selección de variables

Se trata de reducir la dimensión del problema tomando un conjunto de variables menor

Técnicas

- Fuerza bruta Prueba ensayo/error
- Selección incluida La técnica de DM va eligiendo las variables más significativa (p.e. árboles de decisión)
- Filtrado Se seleccionan las variables previamente a la aplicación de las técnicas de DM (p.e. filtrado de términos en Text Mining)
- Selección por cobertura La bondad del resultado de aplicar técnica de DM sirve como criterio de selección.

Reducción de variables

Analisis de componentes principales

Problema a tratar

•*Ejemplo ilustrativo*

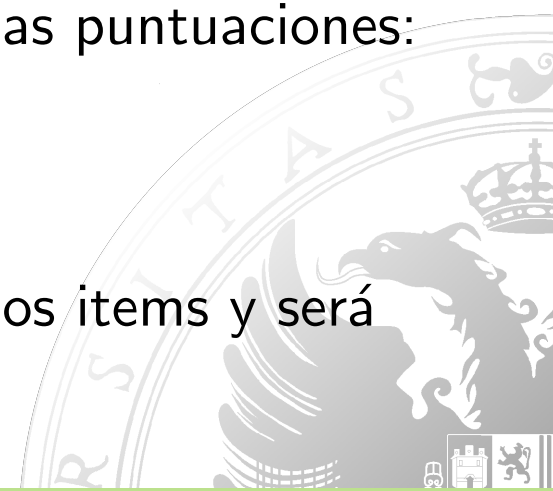
Consideremos un conjunto de alumnos para los que se dan las calificaciones obtenidas en cinco asignaturas, las dos primeros exámenes se han hecho sin apuntes y los otros tres con apuntes. Se quiere ordenar a los alumnos en función de su rendimiento.

Solución:

Encontrar una "combinación lineal normalizada" de las puntuaciones:

$$x_c = \sum l_j x_j \text{ tal que } \sum l_j^2 = 1$$

que recoja la máxima varianza, ya que "separará" a los items y será más fácil ordenarlos



Reducción de variables

Analisis de componentes principales

Problema a tratar

Otros ejemplos

- Ordenación de clientes bancarios
- Ordenación de items según preferencias
- En general obtención de resúmenes de atributos

Reducción de variables

Analisis de Componentes principales

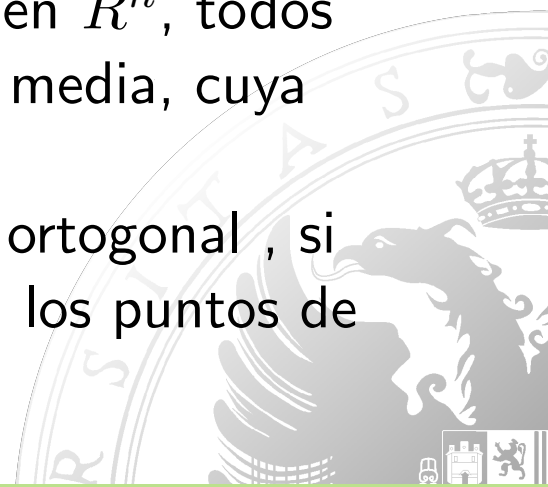
El modelo matemático

Consideremos un data set con datos numéricos reales, este puede verse como una variable N dimensional $\bar{x} = (x_1, ..x_N)$ y m valores de la misma $x_{ij}, i \in \{1, , M\}, j \in \{1, .., N\}$, queremos encontrar transformaciones lineales normalizadas (SLC) que "resuman" lo mejor posible los datos,capturando la mayor varianza de los mismos.

Idea Intuitiva

Si se consideran los items como una nube de puntos en R^n, todos ellos se pueden encerrar en un elipsoide, de centro la media, cuya matriz es la matriz de covarianzas.

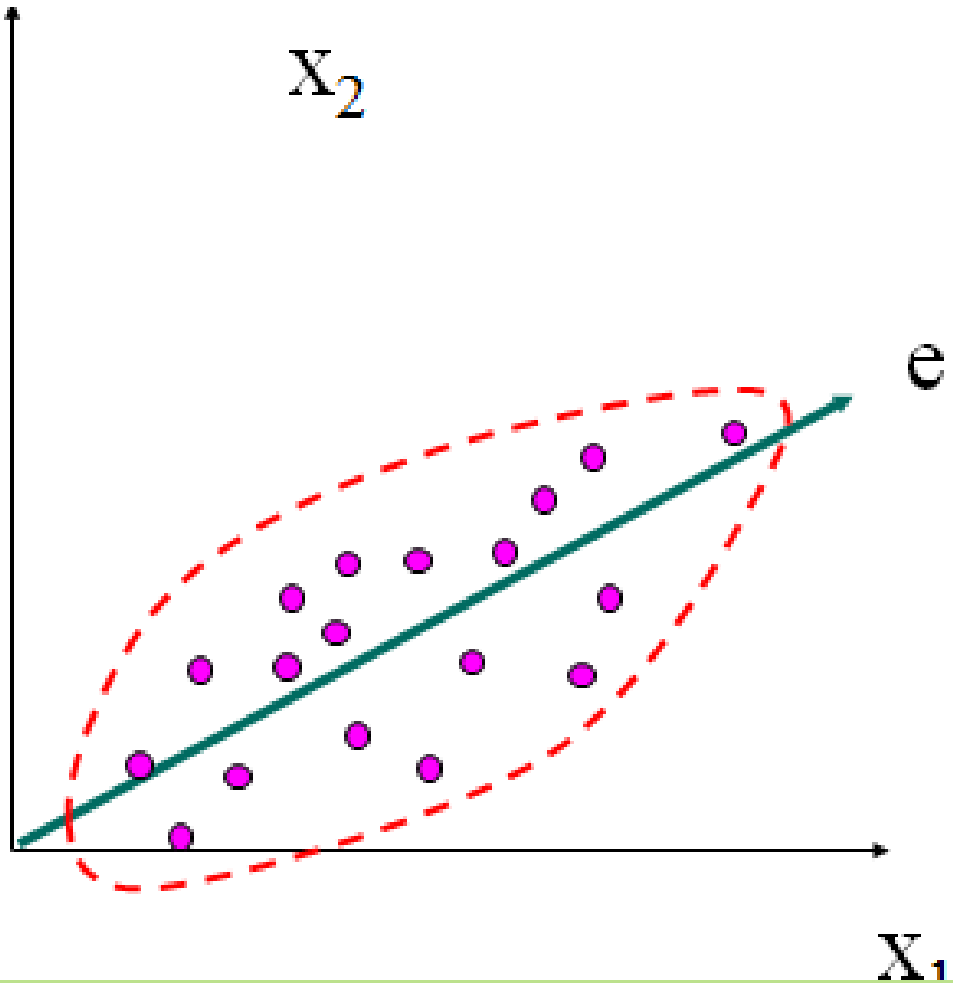
Los ejes del elipsoide son un sistema de coordenadas ortogonal , si realizamos un cambio a este sistema de coordenadas, los puntos de dispersan a lo largo de los ejes



Reducción de variables

Analisis de Componentes principales

Idea Intuitiva



Reducción de variables

Analisis de Componentes principales

El modelo matemático

Sea μ la media de $\bar{x}$ y Σ su matriz de covarianza, se trata de encontrar una transformación lineal $\bar{y} = \Gamma'(\bar{x} - \bar{\mu})$ tal que los nuevos ejes de coordenadas sean los ejes del elipsoide. Se prueba que Γ es una matriz tal que:

$$\Gamma'\Sigma\Gamma = \Lambda$$

donde,

$$\Lambda = diag(\bar{\lambda}) = \begin{pmatrix} \lambda_1 & & & & \\ & \ddots & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \lambda_n \end{pmatrix}$$

Reducción de variables

Analisis de Componentes principales

El modelo matemático

Los valores del vector $\bar{\lambda}$ verifican $\lambda_1 \geq, .. \geq \lambda_N$ y son los "autovalores" de la matriz de covarianza y asociado a cada uno de ellos λ_j existe un "autovector" $\bar{\gamma_j}$ que es la j-esima columna de la matriz Γ y verificándose:

$$\forall j \in \{1, , N\}\ y_j = \bar{\gamma_j}'(\bar{x} - \bar{\mu})$$

y_j se denomina j-esimo componente principal.

$$\forall k, j \in \{1, , N\}\ Cov(y_j, y_k) = 0\ Var(y_j) = \lambda_j$$

$$Var(y_1) \geq .. \geq Var(y_n)$$

y que dado $k \leq N$ no existe ninguna SLC que sea independiente de los k primeros componente principales y que tenga una varianza mayor que el $k+1$ componente principal.

Reducción de variables

Analisis de Componentes principales

El modelo matemático

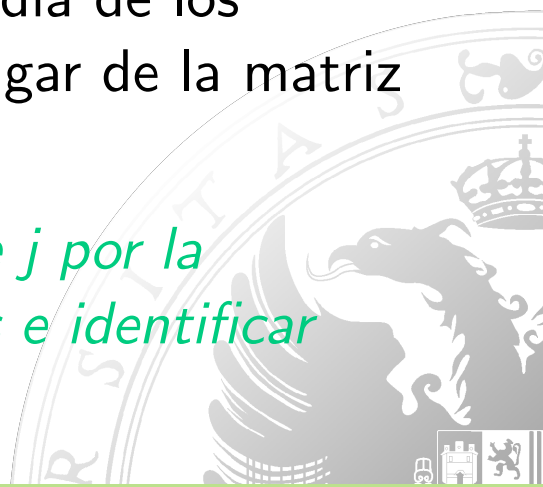
•*Proporción de varianza explicada*

⋆ La proporción de varianza explicada por k factores es $(\lambda_1 + .. + \lambda_k)/\lambda_1 + .. + \lambda_n)$ y nos permite reducir la dimensionalidad del espacio. Es decir expresar el fenómeno con menos variables.

¿Cuantas componentes tomar?:

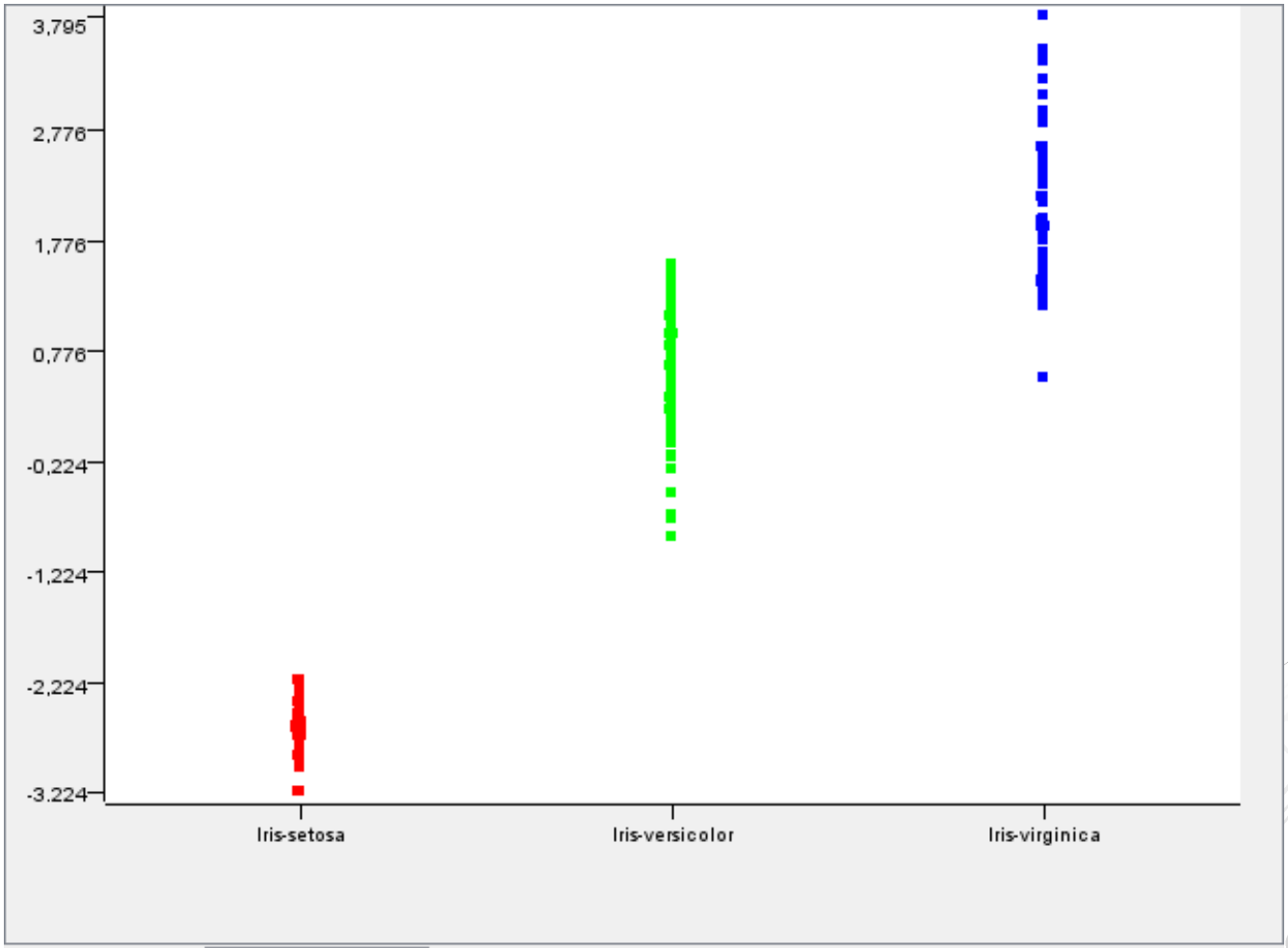
- Al menos el 90% de varianza explicada
- Todos los autovalores que sean mayores que la media de los mismos. Si se utiliza la matriz de correlación en lugar de la matriz de covarianzas autovalores mayores que 1.

La proporción de variación explicada de la variable j por la componente k permite identificar los componentes e identificar una semántica para ellos.



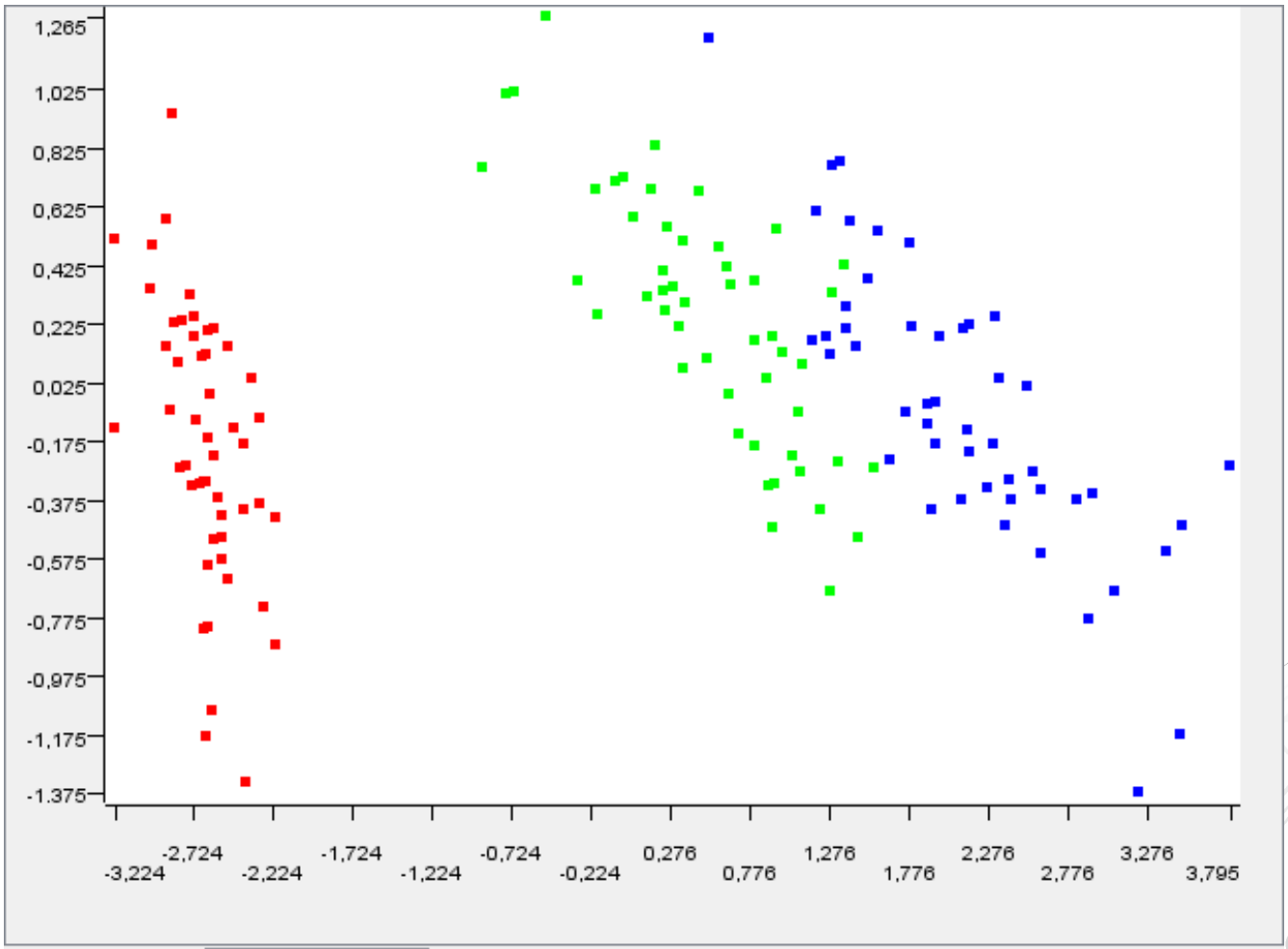
Componentes principales

Ejemplo: Iris (clases/primer factor)



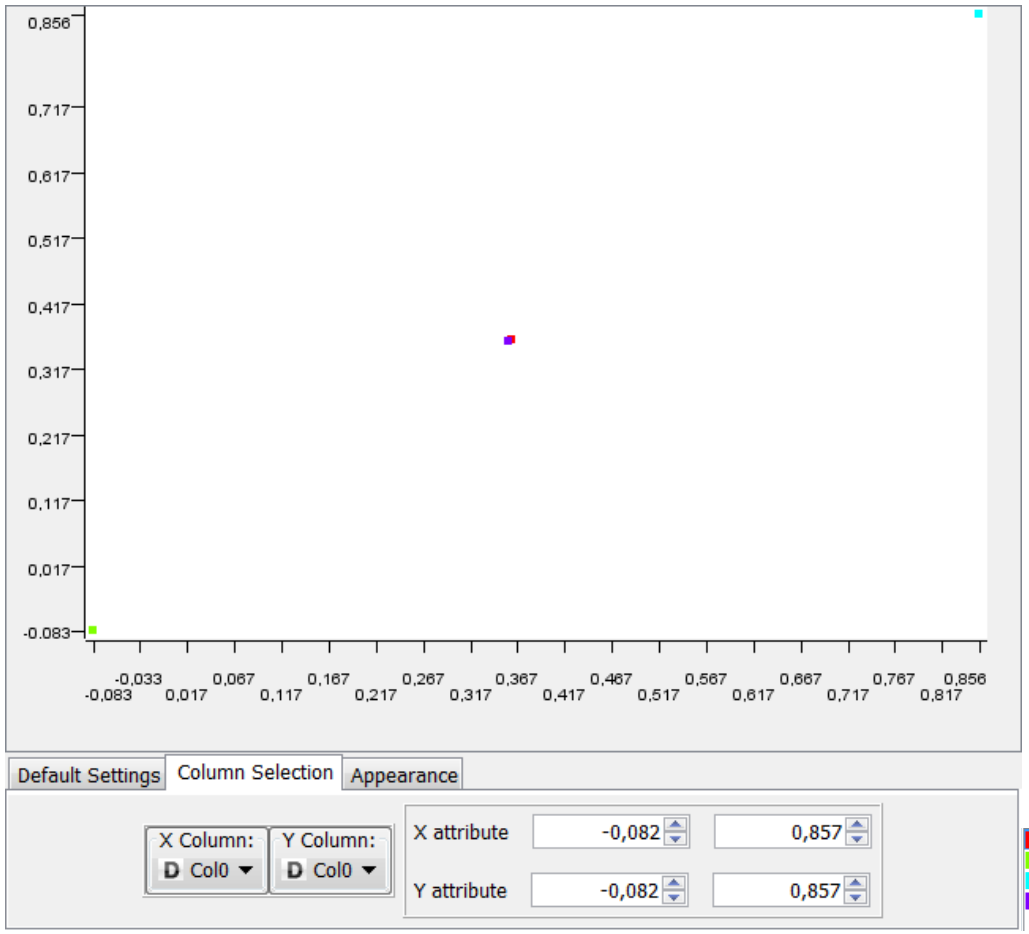
Componentes principales

Ejemplo: Iris (primer factor/segundo factor)



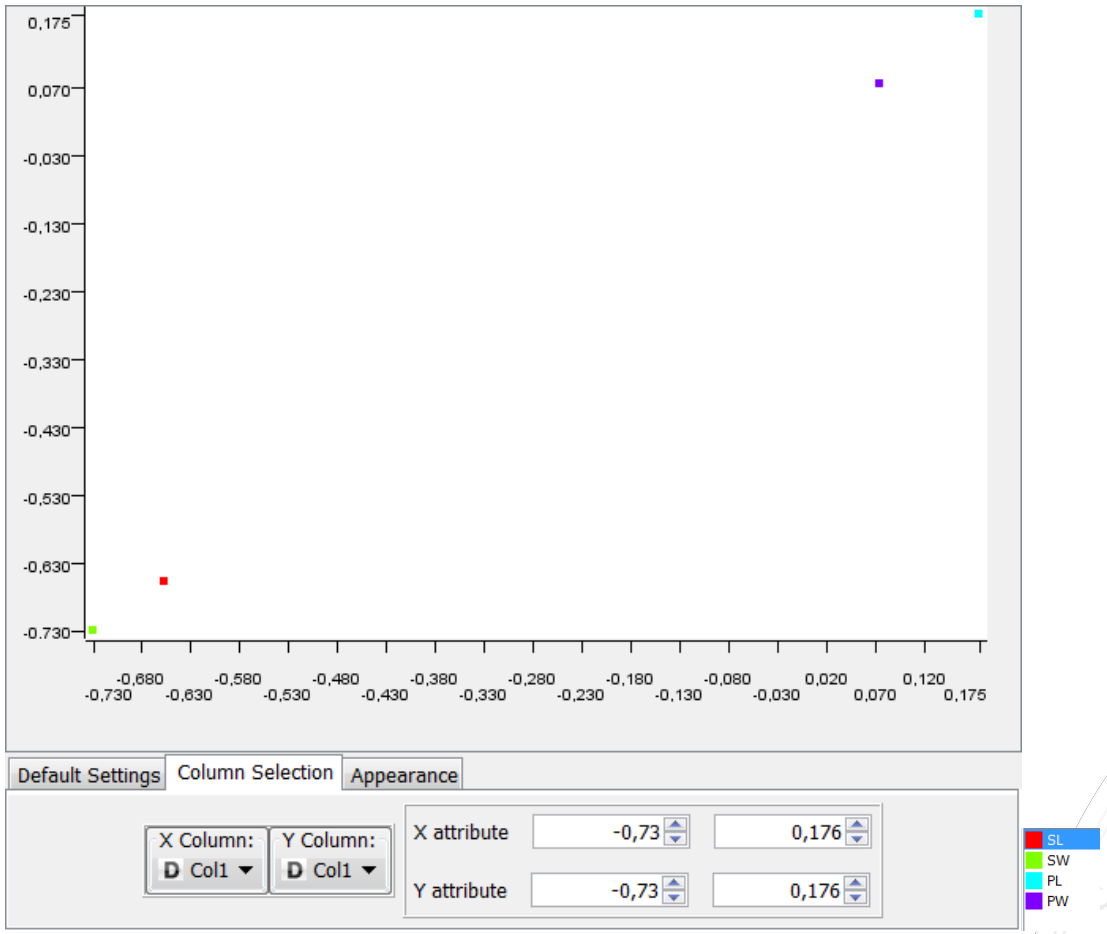
Componentes principales

Ejemplo: Iris (variables/primer factor



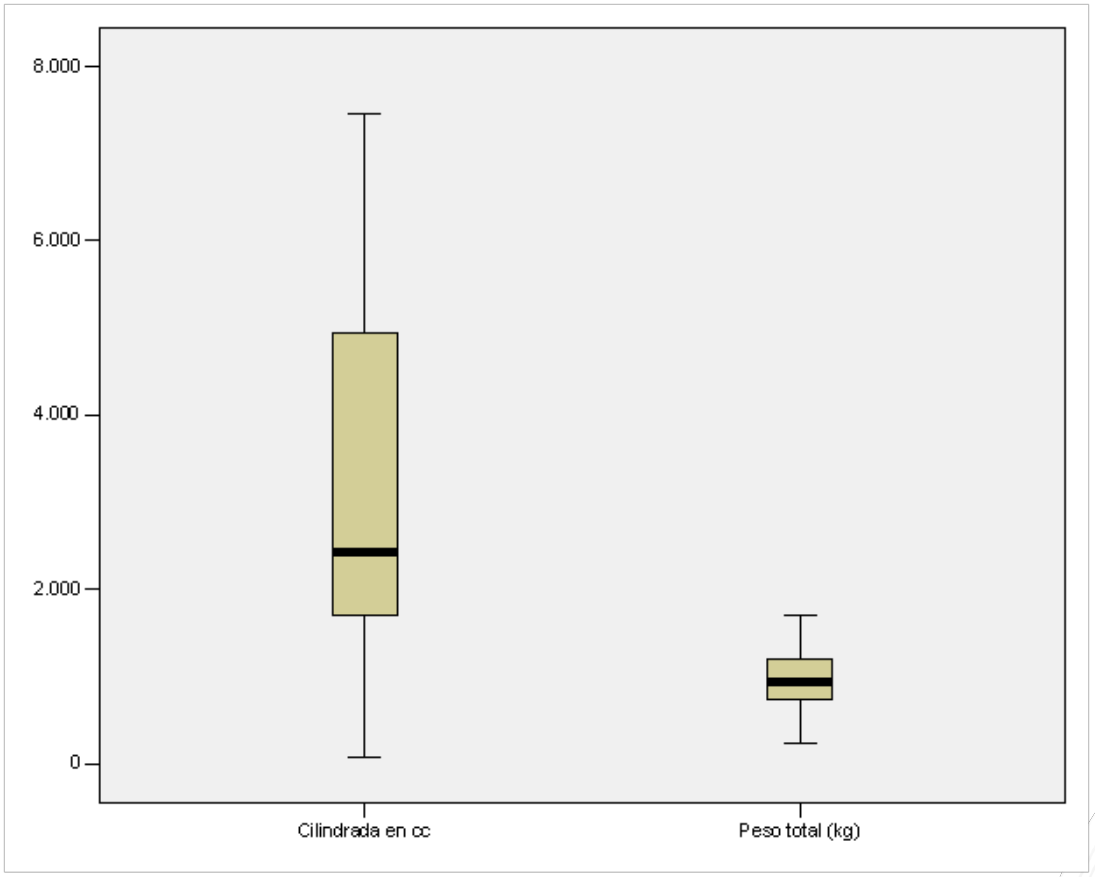
Componentes principales

Ejemplo: Iris (variables/segundo factor)



Problemas de cambio de escala

Motivación



Problemas de cambio de escala

Ideas básicas

Los valores de dos atributos numéricos son de escalas diferentes. Esto hace que no puedan ser tratados conjuntamente en temas tales como cálculo de distancias etc.

- *Algunas expresiones para normalizar*
  - Normalización en [0,1]

$$V = \frac{V_a - min_a}{max_a - min_a}$$

  - Tipificación

$$V = \frac{V_a - \bar{d}}{s}$$

  - Tipificación Robusta

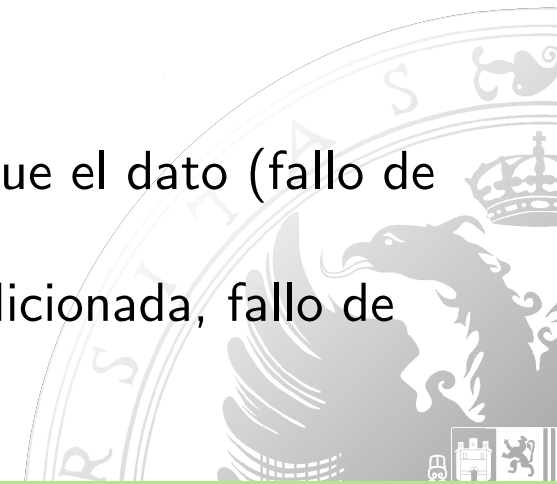
$$V = \frac{V_a - \text{mediana}}{\text{rango intercuartiles}}$$

Problemas de valores perdidos

Ideas básicas

Una variable tiene valores perdidos cuando no se conoce su valor para un dato concreto

- *Origen de los valores perdidos*
  - Desconocimiento del valor sin factores aleatorios. (Alguien no contesta algo en una encuesta)
  - Propiedad no aplicable. (Color de pelo en las ranas). No se puede identificar con el 0 o NO.
  - Error de origen aleatorio.
    - Totalmente aleatorio: sigue la misma distribución que el dato (fallo de origen desconocido de un sensor) (MCAR)
    - Aleatorio condicionado: sigue una distribución condicionada, fallo de un sensor que falla más cuando llueve (MAR)



Problemas de valores perdidos

Que hacer con los valores perdidos

Borrar registros Eliminando los datos perdidos

- Si se trata de una situación totalmente aleatoria.
- Si el volumen de datos no queda seriamente alterado

Sustituir Sustituir el dato perdido por un valor

- Se genera un nuevo valor dentro de un dominio cualitativo. (NS/NC, NA etc.)
- Se deduce utilizando el valor más frecuente según la clase del registro.
- Si es un dato cuantitativo se sustituye por:
  - La media si el error es completamente aleatorio
  - La media condicionada a la aparición de error si tenemos un tipo MAR
  - Una media de valores próximos (interpolación) si sabemos que existe una cierta dependencia temporal o espacial