

Data Mining



Preprocesamiento

Tipos de variables, medidas correspondientes a cada tipo.
Limpieza de datos. Integración y preparación de los datos.

Outline

- ❑ Tipos de variables, medidas correspondientes a cada tipo.
- ❑ Limpieza de datos.
- ❑ Integración y preparación de los datos.

Entender los datos

Datos

- ❑ Los conjuntos de datos están formados por **objetos de datos**.
 - ❑ Un objeto de datos representa una entidad. (vector de características)
- ❑ Ejemplos:
 - ❑ base de datos de ventas: clientes, artículos de la tienda, ventas
 - ❑ base de datos médica: pacientes, tratamientos, estudios, ...
 - ❑ base de datos de la universidad: estudiantes, profesores, cursos
- ❑ Ejemplos, instancias, *data points*, objetos, tuplas.
 - ❑ Los objetos de datos se describen por atributos.

Filas de la base de datos → objetos de datos; columnas → atributos.

Atributos

Atributo (o dimensiones, *features*, variables): un campo de datos, que representa una característica objeto.



Data Warehouse

Machine Learning

Estadística

Por ejemplo:

`customer_ID, name, address`

Tipos:

- ☐ Nominal
- ☐ Binario
- ☐ Numérico: cuantitativo
 - ☐ Intervalos escalados
 - ☐ Ratio escalado

Tipos de atributos

❑ **Nominal:** categorías, estados o "nombres de cosas"

- ❑ Color de pelo = {castaño, negro, rubio, marrón, gris, rojo, blanco}
- ❑ estado civil, ocupación, números de identificación, códigos postales

❑ **Binario**

- ❑ Atributo nominal con solo 2 estados (0 y 1)
- ❑ **Binario simétrico:** ambos resultados son igualmente importantes
 - ❑ por ejemplo, género
- ❑ **Binario asimétrico:** los resultados no son igualmente importantes.
 - ❑ por ejemplo, examen médico (positivo vs. negativo)
- ❑ Convención: asigne 1 al resultado más importante (p. Ej., VIH positivo)

❑ **Ordinal**

- ❑ Los valores tienen un orden que tiene un sentido (*ranking*), pero se desconoce la magnitud entre los valores sucesivos.

Tamaño = {pequeño, mediano, grande}, calificaciones, rango militar

Tipos de atributos numéricos

- ❑ **Cuantitativos:** valores enteros o reales

- ❑ **Escalas de Intervalos**

- ❑ Mide una escala de iguales unidades

- ❑ Los valores tienen un orden:

- ❑ temperatura en C° o F°

- ❑ No hay un verdadero punto cero.

- ❑ **Escalas de razón (*ratio*)**

- ❑ Hay un punto cero

- ❑ Podemos hablar de un valor como un múltiplo (o razón) de otro valor.

- ❑ 10 K° es dos veces más alto que 5 K°.

- ❑ Ejemplos:

- ❑ temperatura en Kelvin, longitud/latitud, cantidades, montos de dinero

Atributos discretos vs continuos

Atributos discretos

- ❑ Tiene solo un conjunto de valores finito o infinito contable.
 - ❑ Por ejemplo: Id_cliente, Código Postal, profesión, términos en un documento.
- ❑ Algunas veces están representados por valores enteros
- ❑ Los atributos binarios son un caso particular de atributos discretos.

Atributos continuos

- ❑ Tiene números reales como valores de atributo
 - ❑ Por ejemplo: temperatura, altura, peso.
- ❑ Los valores reales solo se pueden medir y representar utilizando un número finito de dígitos.
- ❑ Se representan variables de punto flotante

Descripciones estadísticas básicas de datos

❑ Motivación:

- ❑ Tener una vista general de sus datos.

❑ Medidas de tendencia central:

- ❑ Dado un atributo ¿Donde caen la mayoría de sus valores?
 - ❑ Media, mediana, moda, rango medio

❑ Medidas de dispersión:

- ❑ ¿Cómo se distribuyen los datos?
 - ❑ Rango, cuartiles y rango intercuartil.
 - ❑ Desviación estándar y varianza
 - ❑ 5 números mágicos, boxplots...

❑ Análisis gráfico

- ❑ Histogramas, gráficos de dispersión, QQ, etc

Medidas de tendencia central

Media aritmética:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\mu = \frac{\sum x}{N}$$

Media aritmética pesada:

$$\bar{X} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

Los pesos reflejan la importancia o frecuencia de ocurrencia de sus respectivos valores.

Media recortada: Se quitan valores extremos (no más de un 2%)

Medidas de tendencia central

Mediana:

- ❑ Es el valor que divide en 50% y 50%. Se utiliza el valor central si la cantidad de observaciones es impar y sino el promedio de los dos centrales
- ❑ Puede ser estimada también a partir de datos agrupados.

$$\text{median} = L_1 + \left(\frac{n/2 - (\sum \text{freq})l}{\text{freq}_{\text{median}}} \right) \text{width}$$

<i>age</i>	<i>frequency</i>
1-5	200
6-15	450
16-20	300
21-50	1500
51-80	700
81-110	44

Moda:

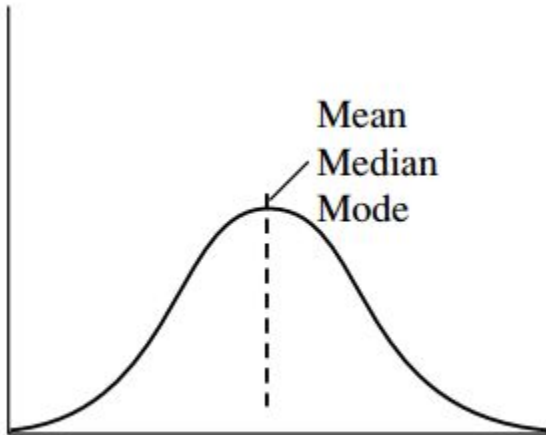
- ❑ El valor que ocurre con mayor frecuencia en una variable
- ❑ Unimodal **bimodal, trimodal** Multimodal
- ❑ Fórmula empírica:

$$\text{Media} - \text{moda} = 3x(\text{media} - \text{mediana})$$

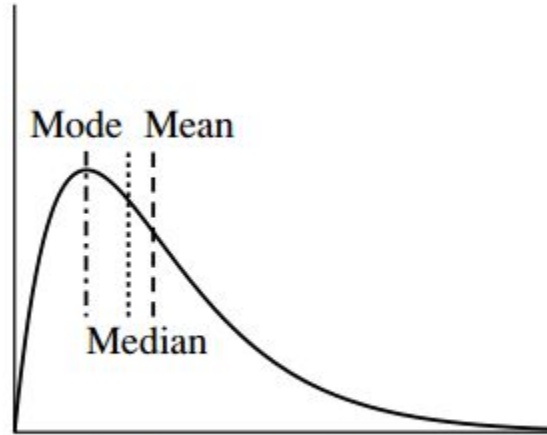
Simetría y datos sesgados

Los datos en la mayoría de las aplicaciones reales no son simétricos.

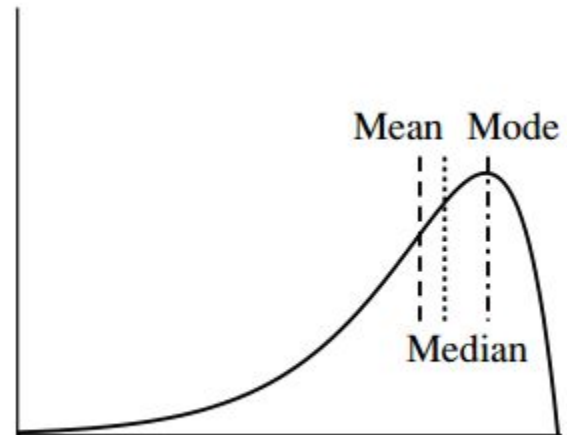
- ❑ Sesgo positivo, la moda es menor que la mediana
- ❑ Sesgo negativo, la moda es mayor que la mediana



(a) Symmetric data



(b) Positively skewed data



(c) Negatively skewed data

Medidas de dispersión

❑ Cuartiles, outliers y boxplots

- ❑ Cuartiles: $Q1$ (percentil 25), $Q3$ (percentil 75)
- ❑ Rango intercuartil: $IQR = Q3 - Q1$
- ❑ Resumen de 5 números: \min , $Q1$, median , $Q3$, \max
- ❑ Boxplot:
 - ❑ los extremos de la caja son los cuartiles;
 - ❑ la marca de la caja es la mediana; los bigotes están a $1.5 \cdot IQR$
- ❑ **Outlier**: generalmente son valores por encima o por debajo de $1.5 \cdot IQR$

❑ Varianza y Desviación estándar (muestral: s , población: σ)

❑ Varianza:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right]$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^n x_i^2 - \mu^2$$

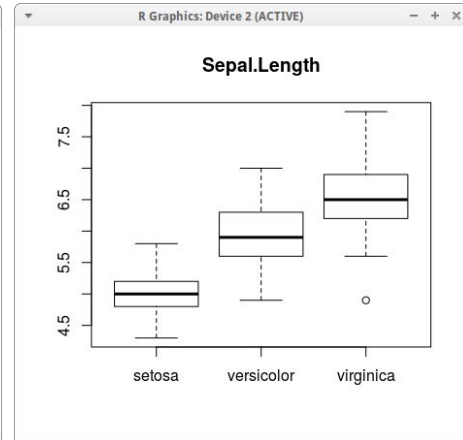
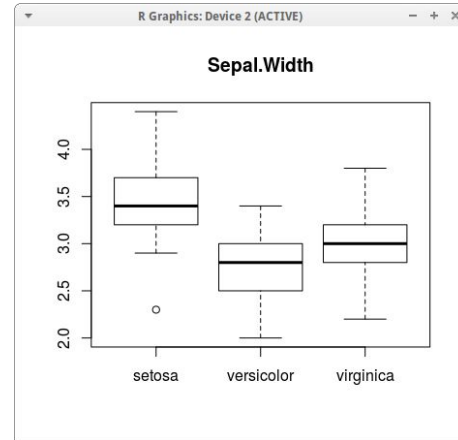
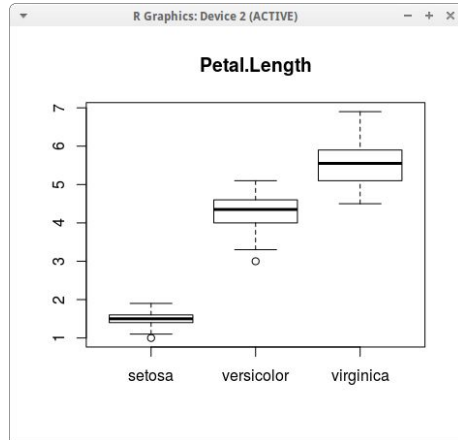
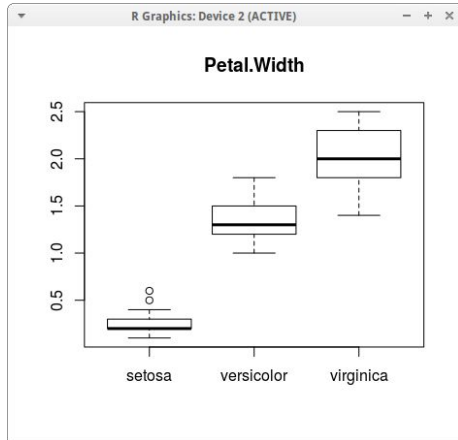
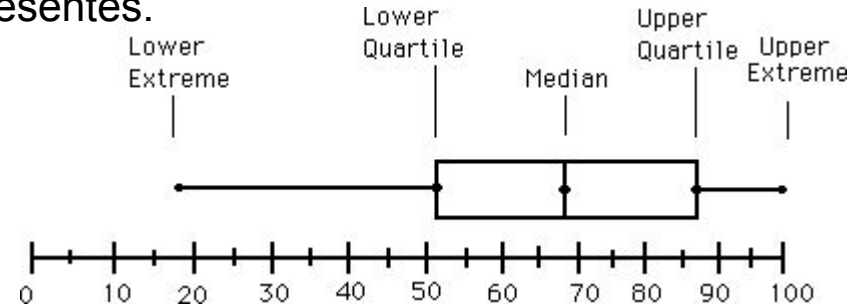
- ❑ Desviación estándar **s** (o **σ**) es la raíz cuadrada de s^2 (o σ^2)

Análisis del Boxplot

Los 5 números resumen de la distribución están presentes.

❏ Minimum, Q1, Median, Q3, Maximum

Análisis de separabilidad



Gráficos básicos de estadística descriptiva

- ❑ **Histogramas:** eje x representa los valores y el eje y las frecuencias.
- ❑ **Gráfico de Cuantiles:** Permite observar cuán cerca está la distribución de un conjunto de datos a alguna distribución ideal. `qqnorm()`
 - ❑ **QQplot:** Grafica los cuantiles de una distribución univariante contra los cuantiles correspondientes de otra.
- ❑ **Scatter plot:** cada par de valores es un par de coordenadas y se dibujan como puntos en el plano.

Histograma

Es un método gráfico para resumir la distribución de una variable.

El rango de valores para X se divide en sub-intervalos consecutivos disjuntos.

Los subintervalos, denominados cubos o contenedores, son subconjuntos disjuntos de la distribución de datos para X.

El rango de un clase se conoce como amplitud o intervalo de clase .

Por lo general, las clases son de igual ancho.

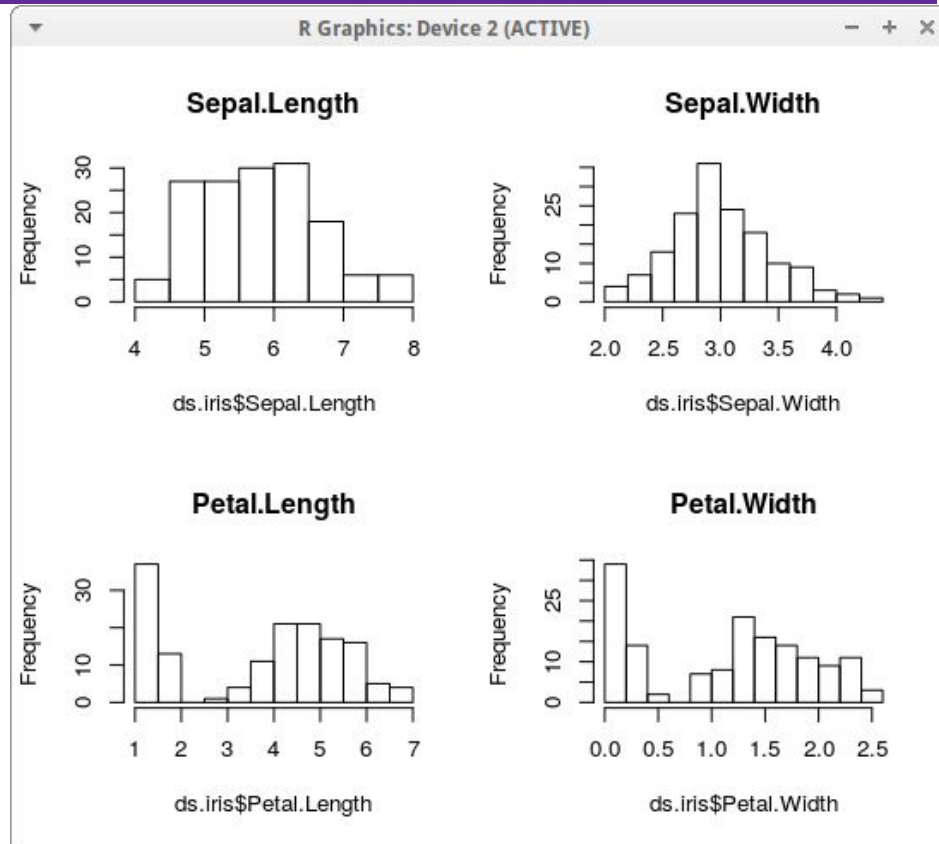
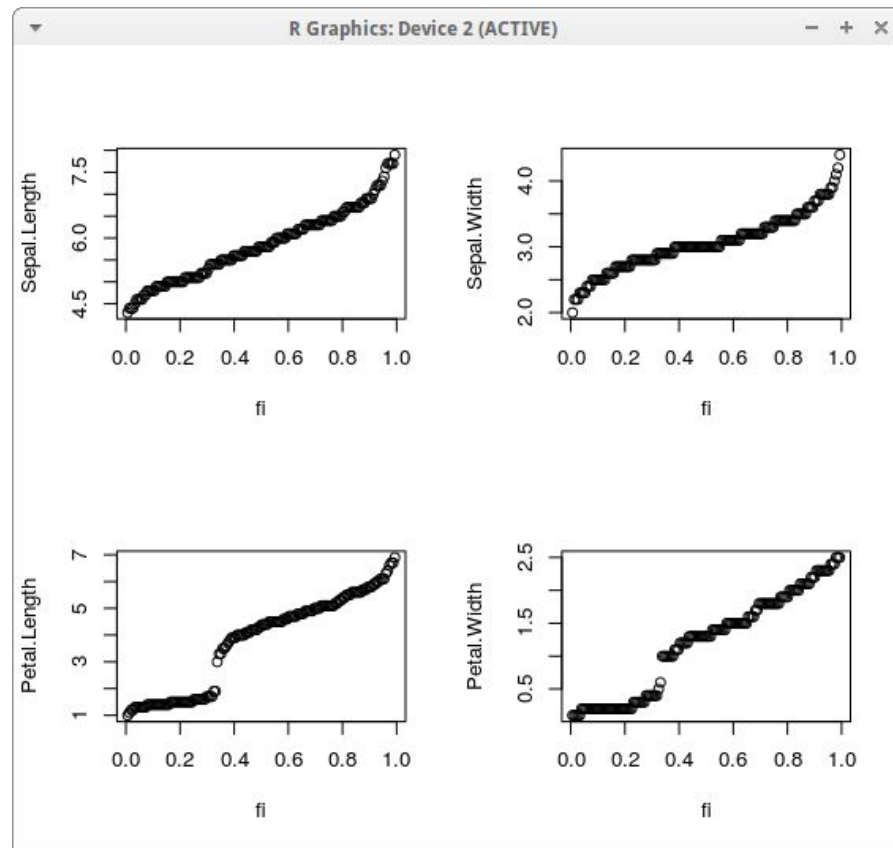


Gráfico de Cuantiles

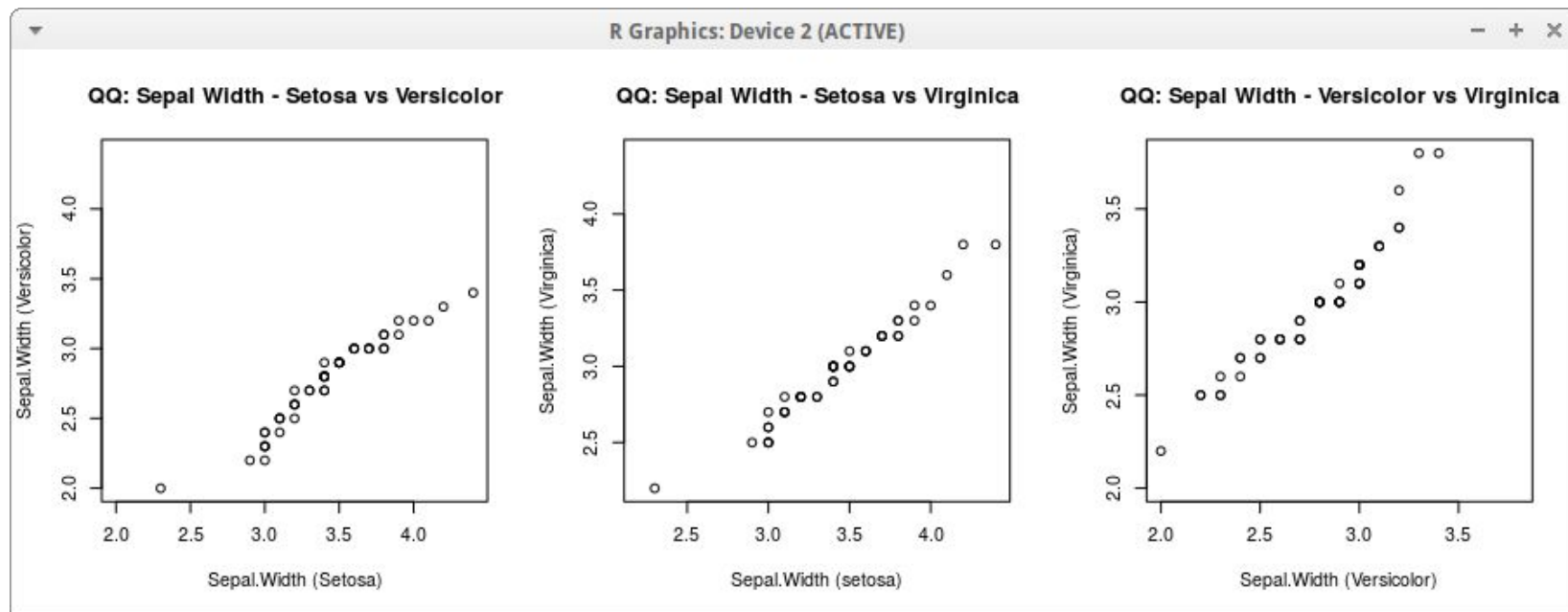
Muestra todos los datos (lo que permite al usuario evaluar tanto el comportamiento general como las ocurrencias inusuales)

Para datos de x_i ordenados en orden creciente, f_i indica que aproximadamente 100% de los datos son menores o igual al valor x_i



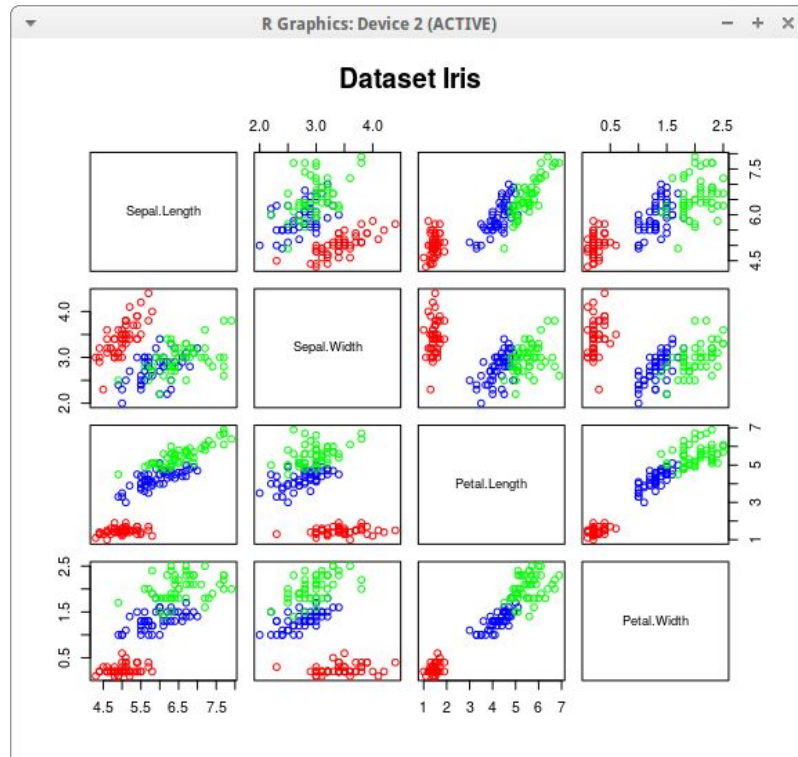
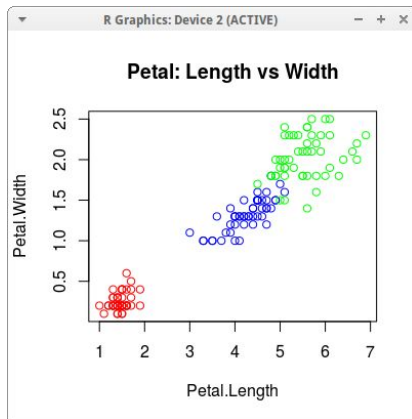
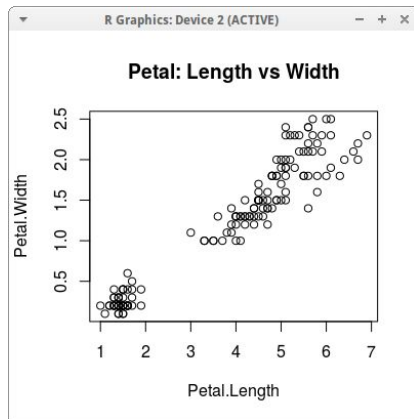
QQPlot

- ❑ Permite graficar los cuantiles de una distribución univariante contra los cuantiles correspondientes de otra.
- ❑ ¿Cómo está distribuido el ancho del sépalo entre las diferentes variedades de Iris?

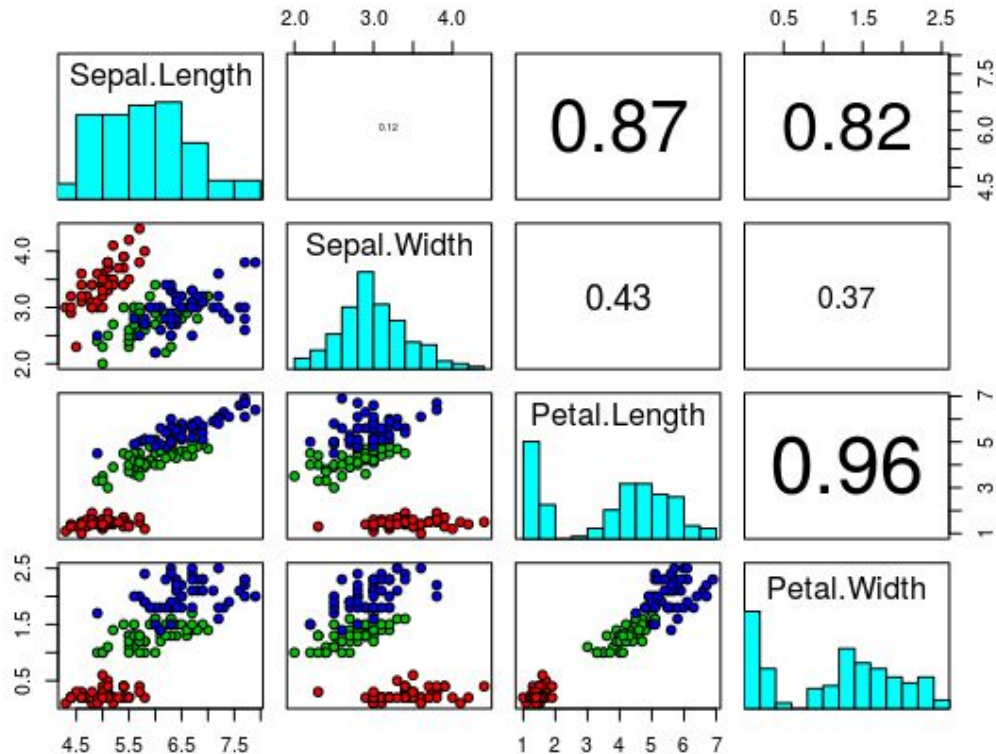


Scatter Plots

- ❑ Proporciona una primera mirada a los datos bivariados para ver **grupos de puntos**, **valores atípicos**, etc.
- ❑ Cada par de valores se trata como un par de coordenadas y se dibujan como **puntos en el plano**.

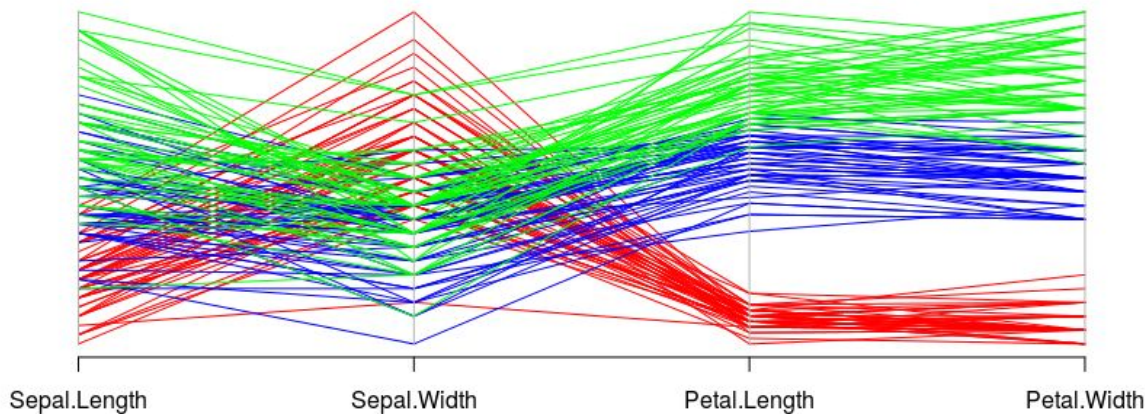


Matrices Mixtas



Coordenadas Paralelas

- ❑ N **ejes equidistantes** que son **paralelos** a uno de los ejes de pantalla y corresponden a los atributos
- ❑ **Los ejes se escalan** al [mínimo, máximo]: rango del atributo correspondiente.
- ❑ Cada ítem de datos corresponde a una **línea que cruza cada uno de los ejes** en el punto que corresponde al valor del atributo.



Caras de Chernoff

- ❑ Una forma de mostrar las variables en una superficie bidimensional, por ejemplo: x son las cejas inclinadas, y es tamaño de los ojos, z es la longitud de la nariz, etc.

La figura muestra rostros producidos usando 10 características:

- ❑ excentricidad de la cabeza,
- ❑ tamaño de ojo,
- ❑ espaciado de ojos,
- ❑ excentricidad de ojo,
- ❑ tamaño de pupila,
- ❑ inclinación de ceja,
- ❑ tamaño de nariz,
- ❑ forma de boca,
- ❑ tamaño de boca y
- ❑ apertura de boca



Preprocesamiento

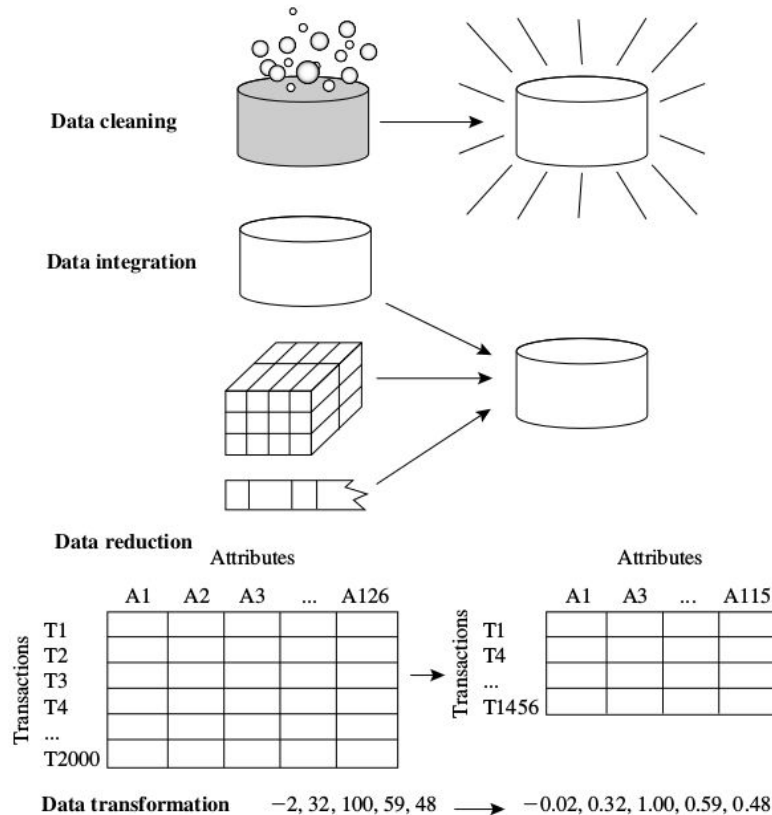
¿Por qué preprocesar?

- ❑ Muchos de los datos crudos contenidos en nuestras DB se encuentran incompletos y con ruido.
 - ❑ Campos obsoletos o redundantes,
 - ❑ Valores faltantes,
 - ❑ Outliers,
 - ❑ Los datos están en un formato no apto para los modelos de minería de datos,
 - ❑ Valores no consistentes con el sentido común o con las políticas de la organización.

Calidad de los datos

- ❑ Medidas para la calidad de los datos: una visión multidimensional
 - ❑ **Precisión:** correcta o incorrecta, precisa o no
 - ❑ **Integridad:** no registrado, no disponible, ...
 - ❑ **Consistencia:** algunos modificados pero otros no, ...
 - ❑ **Puntualidad:** ¿actualización oportuna? Tiene que ver con la caducidad de esos datos.
 - ❑ **Confiabilidad:** ¿qué tan confiables son los datos?
 - ❑ **Interpretabilidad:** ¿con qué facilidad se pueden entender los datos?

Tareas de preprocesamiento



Limpieza de datos

- ❑ Completar valores faltantes, eliminación de ruido, identificar o eliminar valores atípicos y corregir incoherencias.

Integración de datos

- ❑ Integración de múltiples bases de datos, cubos de datos o archivos

Reducción de datos

- ❑ Reducción de dimensionalidad
- ❑ Reducción de Numerosidad
- ❑ Compresión de datos

Transformación y discretización de datos

- ❑ Normalización
- ❑ Generación de jerarquía conceptual

Limpieza de datos

Los datos en el mundo real están sucios: datos incorrectos. Por ejemplo: error del instrumento, error humano o de la computadora, error de transmisión...

- ❑ **Incompleto:** carece de valores de atributo, carece de ciertos atributos de interés o contiene solo datos agregados
 - ❑ por ejemplo, `Ocupación = ""` (**datos faltantes**)
- ❑ **Ruidoso:** contiene ruido, errores o valores atípicos
 - ❑ por ejemplo, `Salario = "- 10"` (**un error**)
- ❑ **Inconsistente:** contiene discrepancias en códigos o nombres, por ejemplo:
 - ❑ `Edad = "42"`, `cumpleaños = "03/07/2010"`
 - ❑ Estaba clasificando "1, 2, 3", ahora clasificando "A, B, C"
 - ❑ Discrepancia entre registros duplicados
- ❑ **Intencional** (por ejemplo, datos faltantes disfrazados)
 - ❑ ¿1 de enero como el cumpleaños de todos?

Ejemplo: Listado del Curso DM

titulo

<fct>

- 1 Ingenieria en Sistemas
- 2 Ingenieria en sistemas de informacion
- 3 Ingenieria en Sistemas de Informacion
- 4 Ing en sistemas de informacion
- 5 Ingeniería en Sistemas
- 6 lic en sistemas de informacion de las organizaciones
- 7 Licenciado en sistemas de Informacion

titulo

<fct>

- 1 Analista Universitario de Computacion
- 2 licenciado en ciencias de la computacion
- 3 Licenciado en cs de la computacion
- 4 Licenciatura en Cs de la Computacion

56 títulos de grado

Licenciatura en Cs Químicas
Licenciatura en Cs de la Computacion
Licenciatura en Composicion
Licenciatura en Ciencias Fisicas
Licenciatura en Ciencias Biológicas
Licenciatura en Administración
Licenciado en sistemas de Informacion
licenciado en publicidad
Licenciado en Economía empresarial
licenciado en economía
Licenciado en economía
Licenciado en Cs. Fisicas
Licenciado en cs de la computacion
licenciado en ciencias de la computacion
licenciado en administracion
Licenciada en Sociología
licenciada en bioinformatica
Lic.en matematica
Lic.en Informatica
lic en sistemas de informacion de las organizaciones

titulo

<fct>

- 1 Ingenieria en Informatica
- 2 ingeniero en informatica
- 3 Ingeniería en Informática
- 4 Lic.en Informatica

titulo

<fct>

- 1 Licenciatura en Economía
- 2 Licenciatura en Economía
- 3 Economía
- 4 Licenciado en economía
- 5 "licenciado en economía "
- 6 Licenciado en Economía empresarial

Economia
bioquímica
Analista Universitario de Computacion
Abogacia

0

1

2

3

4

5

n

Datos faltantes

- ❑ Los datos no siempre están disponibles.
 - ❑ Muchas tuplas no tienen ningún valor registrado para varios atributos.
 - ❑ ¿Cuál es el ingreso de un cliente?
- ❑ Los datos faltantes pueden deberse a:
 - ❑ Mal funcionamiento del un equipo.
 - ❑ Tiene inconsistencias con otros datos grabados y, por lo tanto, es eliminado.
 - ❑ Datos no ingresados debido a malentendidos.
 - ❑ Ciertos datos pueden no ser considerados importantes en el momento de la entrada
 - ❑ No se registra un historial o cambios de los datos

Las opciones al enfrentarnos con datos faltantes son: eliminar registros o trabajar con métodos de imputación.

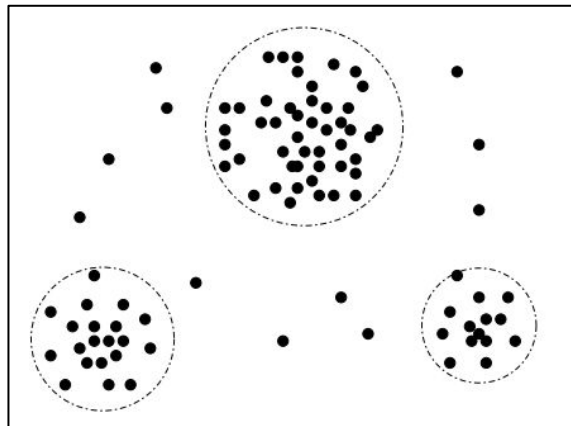
Ruido en los datos

El **ruido** es un error aleatorio o varianza en una variable medida.

- ❑ Los valores incorrectos de un atributo pueden deberse a:
 - ❑ Instrumentos de recolección de datos defectuosos
 - ❑ Problemas de entrada de datos
 - ❑ Problemas de transmisión de datos
 - ❑ Limitación de tecnología
 - ❑ Incoherencia en la convención de nombres
- ❑ Otros problemas de datos que requieren limpieza de datos
 - ❑ registros duplicados
 - ❑ datos incompletos
 - ❑ datos inconsistentes

¿Cómo manejar el ruido en los datos?

- ❑ **Binning:**
 - ❑ Primero ordenar los datos y particionarlos en **bins** (de igual frecuencia)
 - ❑ Luego uno puede suavizar por media, mediana, por límites, etc.
- ❑ **Regresión:** Suavizado ajustando una función de regresión lineal, a partir de otra de las variables del dataset.
- ❑ **Clustering:** Es una herramienta que permite identificar **outliers**. Los valores similares son agrupados y los que quedan aislados pueden ser considerados outliers.



Binning: Ejemplo

Dada una variable Precio (\$): [45, 26, 21, 15, 9, 64, 24, 21, 8, 28, 4, 25]

Ordenada de menor a mayor:

4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 45, 64

Particionar los bins (igual-frecuencia):

Bin 1: 4, 8, 9, 15

Bin 2: 21, 21, 24, 25

Bin 3: 26, 28, 45, 64

Suavizado por medias:

Bin 1: 9, 9, 9, 9

Bin 2: 23, 23, 23, 23

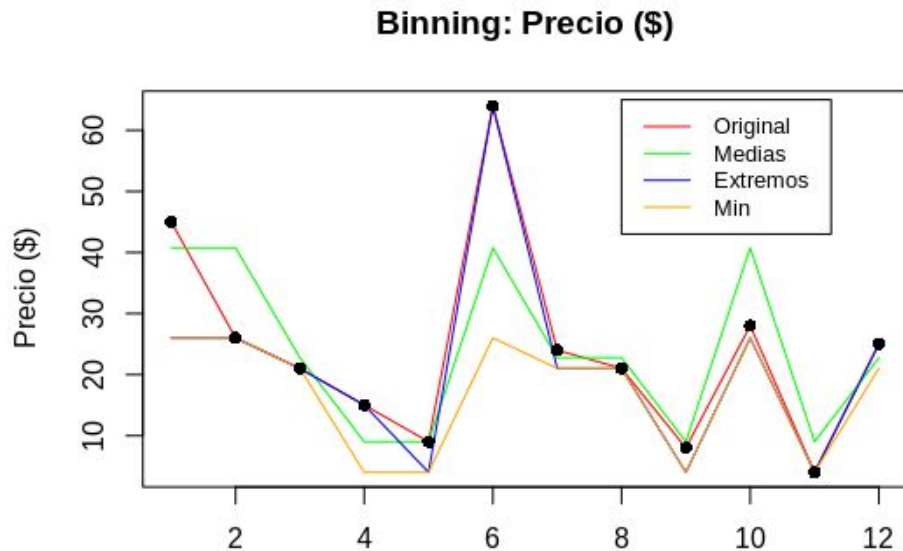
Bin 3: 40.75, 40.75, 40.75, 40.75

Suavizado por límites:

Bin 1: 4, 4, 4, 15

Bin 2: 21, 21, 21, 25

Bin 3: 26, 26, 26, 64



Ruido en el contexto del ML

Ruido en aprendizaje supervisado

- ❑ **Robustez** es la capacidad de un algoritmo para construir modelos que son insensibles a **datos corruptos** y **sufren menos por el impacto del ruido**.
- ❑ Cuanto más robusto es un algoritmo, más similares son los modelos construidos a partir de datos limpios y ruidosos.
- ❑ La robustez es tan importante como la precisión cuando trabajamos con datos ruidosos, sobretodo en los casos en que se **desconocen las características del ruido**.

Robust Learners: Son técnicas que se caracterizan por ser menos influenciadas por datos ruidosos. C4.5 es un ejemplo

Data polishing methods: Su objetivo es corregir instancias ruidosas antes de entrenar. Esta opción solo es viable cuando los conjuntos de datos son pequeños.

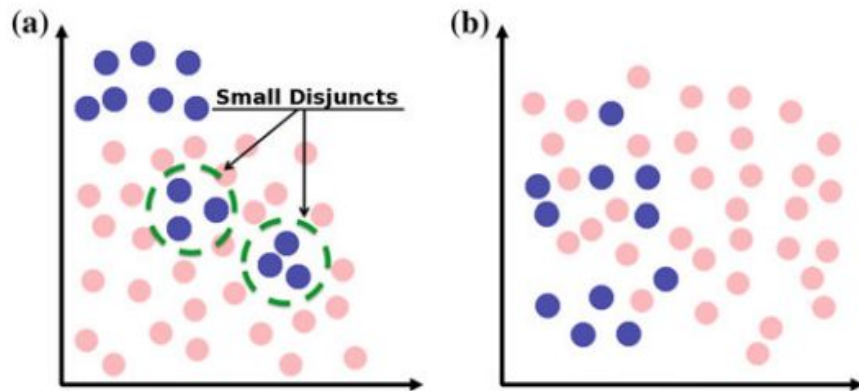
Noise Filters: Se identifican instancias con ruido que puedan ser eliminadas. Hay algoritmos que son sensibles a datos ruidosos.

Ruido y degradación de performance

Los **límites complejos y no lineales entre clases** son problemas que pueden obstaculizar el rendimiento de los clasificadores y, a menudo, es difícil distinguir entre dicha superposición y la presencia de ejemplos ruidosos.

Presencia de pequeñas disyunciones: La clase minoritaria se puede descomponer en muchos subgrupos con muy pocos ejemplos en cada uno, rodeados de ejemplos de clase mayoritaria. Es un gran desafío para los algoritmos de aprendizaje detectar subconceptos.

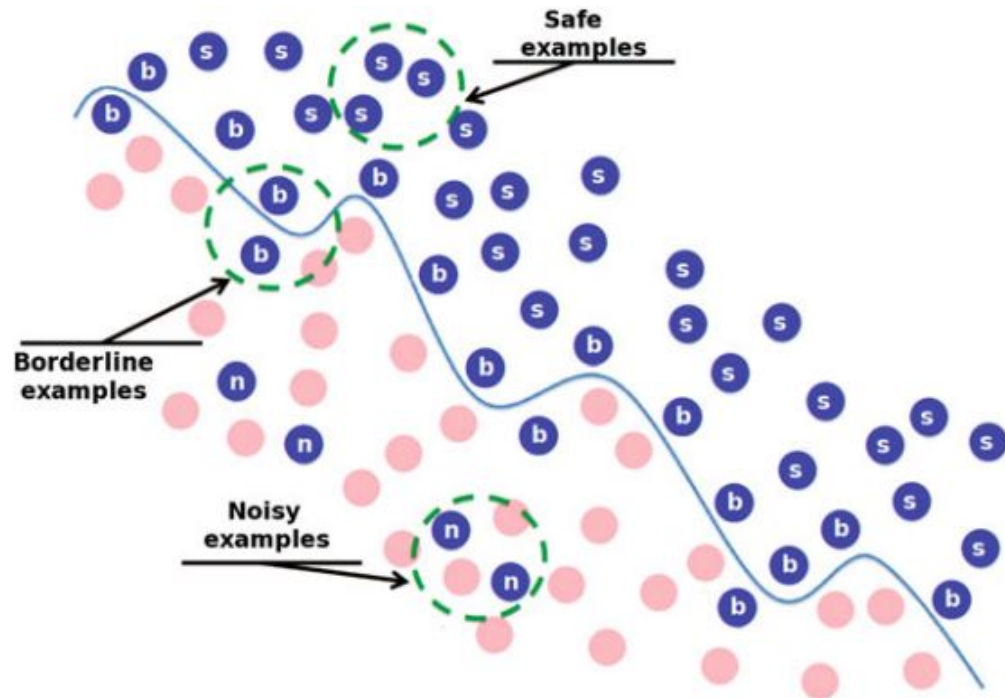
Solapamiento entre clases: Hay ejemplos de diferentes clases pero con *features* que son parecidas, que están cerca al límite de decisión de la clase.



Ruido y degradación de performance

La clasificación errónea ocurre a menudo cerca de los límites de la clase donde la superposición generalmente ocurre también y es difícil encontrar una solución.

La degradación del rendimiento del clasificador se verá fuertemente afectada por la cantidad de ejemplos cerca de la línea de separación y la presencia de otros ejemplos ruidosos ubicados más allá de la región solapada.




Tipos de ruidos: Clase y Atributos

- ❑ **Ruido en la clase:** Las etiquetas de la clase pueden no estar asignadas correctamente. Las causas pueden ser:
 - ❑ **Subjetividad durante el etiquetado**, errores de un *data entry*, uso de información inadecuada para etiquetar los ejemplos.
 - ❑ Los tipos de ruido en la clase pueden ser:
 - ❑ **Ejemplos contradictorios:** Hay ejemplos duplicados con diferentes clases.
 - ❑ **Ejemplos mal clasificados:** Los ejemplos fueron etiquetados con una clase diferente a la verdadera.

Integración de datos

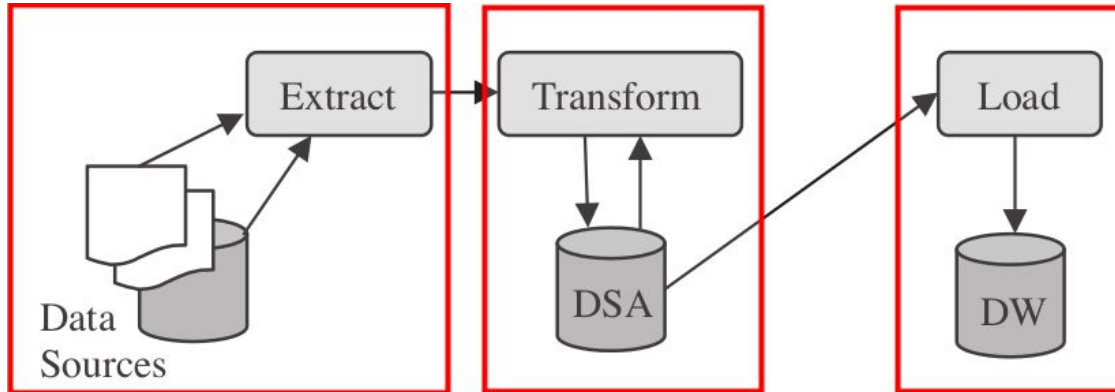
Integración de datos

- ❑ Integración de datos:
 - ❑ Combina datos de múltiples fuentes en un único conjunto coherente
- ❑ **Esquema de Integración:** por ejemplo, `A.cust-id` `B.cust-#`
 - ❑ Integra metadatos de diferentes fuentes
- ❑ Problema de **identificación de entidades:**
 - ❑ Identificar entidades del mundo real a partir de múltiples fuentes de datos.
 - ❑ Por ejemplo, Bill Clinton = William Clinton 
- ❑ Detectar y resolver conflictos de valores de datos:
 - ❑ Para la **misma entidad del mundo real**, los valores de los atributos de diferentes fuentes son diferentes
 - ❑ **Posibles razones:** diferentes representaciones, diferentes escalas, por ejemplo, unidades métricas vs. británicas



Herramientas de integración

- ❑ En el contexto del Data Warehousing existen técnicas para realizar la integración de una forma ordenada.
- ❑ Las herramientas de **Extraction–Transformation–Loading** (ETL) son piezas de software responsables de la extracción de datos desde varias fuentes, su limpieza, puesta a punto, re formato, integración e inserción en un Data Warehouse.



Se puede implementar cada parte de este proceso o utilizar herramientas como Pentaho Data Integration (Kettle) u otras.



Redundancia en la integración de datos

- ❑ Los **datos redundantes** ocurren a menudo cuando se integran múltiples bases de datos
 - ❑ **Identificación de objetos:** el mismo atributo u objeto puede tener diferentes nombres en diferentes bases de datos
 - ❑ **Datos derivables:** un atributo puede ser "derivado" en otra tabla, por ejemplo, ingresos anuales
- ❑ Los atributos redundantes pueden ser detectados por **análisis de correlación y análisis de covarianza**
- ❑ La integración cuidadosa de los datos de múltiples fuentes puede ayudar a reducir / evitar redundancias e inconsistencias y mejorar la velocidad y la calidad de la minería.

Análisis de Correlación (Datos Nominales)

- ❑ Prueba de independencia χ^2 (chi-cuadrado)

H_0 : X e Y son independientes

H_1 : X e Y no son independientes

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

- ❑ Cuanto mayor es el valor de χ^2 , es más probable que la hipótesis nula sea rechazada.
- ❑ Las celdas que más contribuyen al valor de χ^2 son aquellas cuyo recuento real es muy diferente del recuento esperado.
- ❑ La correlación no implica causalidad
 - ❑ **# de hospitales y # de robo de automóvil** en una ciudad están correlacionados
 - ❑ Ambos están relacionados causalmente con la tercera variable: población

Prueba de χ^2 : Ejemplo

Verificamos si las variables **play_chess** = Si/No y **like_science_fiction** = Si/No están correlacionadas.

Armamos la tabla de contingencia

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

Calculamos los valores esperados.
Son los que están entre paréntesis.

$$e_{1,1} = \frac{\text{count}(\text{PlayChess}) * \text{count}(\text{fiction})}{N} = \frac{300 * 450}{1500} = 90$$

Calculamos el estimador de χ^2

$$\begin{aligned}\chi^2 &= \frac{(250 - 90)^2}{90} + \frac{(50 - 210)^2}{210} + \frac{(200 - 360)^2}{360} + \frac{(1000 - 840)^2}{840} \\ &= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.\end{aligned}$$

- ❑ Para la tabla de 2 x 2 tenemos $(2 - 1)(2 - 1) = 1$ grados de libertad.
- ❑ Nivel de significancia es 0.001

$$\chi^2 = 10,82$$

Se rechaza la hipótesis que son independientes y decimos que *play chess* y *like science fiction* están fuertemente correlacionados.

Análisis de Correlación (Datos Numéricos)

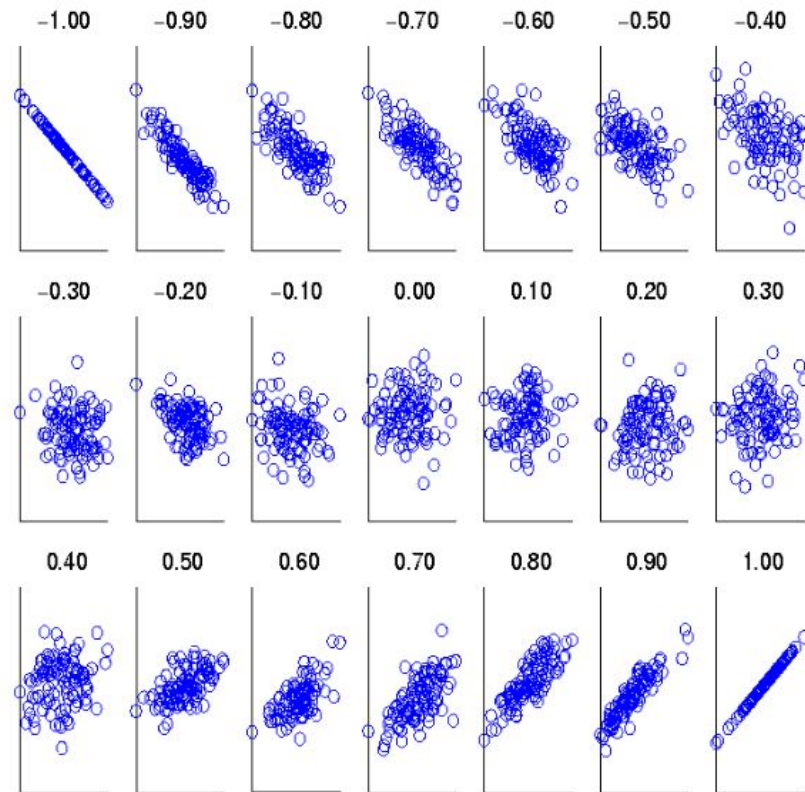
Coeficiente de correlación de Pearson

$$r_{A,B} = \frac{\sum (a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A\sigma_B}$$

Si $r_{A,B} > 0$, A y B están positivamente correlacionadas.

$r_{A,B} = 0$: no hay relación

$r_{AB} < 0$: relacionadas de forma negativa.



Covarianza (Datos Numéricos)

$$\text{Cov}(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{n}.$$

- ❑ Si $\text{Cov}(A, B) > 0$, entonces A y B tienden a ser más grandes que sus valores esperados.
- ❑ Si $\text{Cov}(A, B) < 0$, entonces, si A es mayor que su valor esperado, B es probable que sea menor que su valor esperado.
- ❑ Si $\text{Cov}(A, B) = 0$

Bibliografía

- ❑ Jiawei Han, Micheline Kamber, Jian Pei. 2012. Tercera edición. Data Mining: Concepts and Techniques. Cap. 2 y Cap. 3
- ❑ García, S., Luengo, J., & Herrera, F. (2016). *Data preprocessing in data mining*. Springer. (Cap. 5)
- ❑ Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, 1(1), 3–14. [[pdf](#)]