Preprocesado de Datos

Juan A. Botía

Departamento de Ingeniería de la Información y las Comunicaciones
Universidad de Murcia

Ingeniería Superior en Informática, UMU

Análisis Exploratorio de Datos

Pregunta

¿Por qué no olvidarnos de la preparación de los datos e ir directamente al grano, aplicando, directamente, técnicas de minería a los mismos?

Respuesta

Con casi toda seguridad, los datos van a estar dañados de alguna forma

¿Para qué el EDA?

En muchas aplicaciones, la preparación de los datos para el data mining puede suponer el 80% del esfuerzo total en un proyecto.

Posibles daños en los datos

Razones para que los datos puedan estar dañados ([1],[3],[2])

- Resultados espúreos
- Visión de los algoritmos de análisis como cajas negras
- Limitaciones de las técnicas
- Garantías en los resultados de los procesos de minería

Resultados espurios

Concepto

A menudo, los conjuntos de datos contienen información producida de manera artificial, que no forma parte de los datos genuinos del problema.Y esta información llega a estar tan presente que puede convertirse en un patrón.

Un ejemplo

Análisis de datos a una compañía de Telecomunicaciones grande

- Se detectaron patrones relacionados con errores en la comunicación en llamadas, relacionados con la transmisión de datos de señalización
- Se eliminaron esos errores
- Se realizó minería con los datos correctos

Algoritmos de análisis como cajas negras

Muchas veces se asume que los algoritmos de minería de datos son cajas negras

- El entender los entresijos de cada uno de los paradigmas y técnicas concretas es muy importante
- Ejemplo: clustering k-means
  - Utilizado, por ejemplo, describir targets en públicos de mercados típicos
  - Aparentemente fácil pero es difícil encontrar un k y tamaño relativo de clusters adecuados
  - E.g. si deseamos uno con el 90% de los datos y el otro con el 10% e inicializamos k a 2, encontraremos un gran grupo con casi todos los datos y un grupo muy pequeño con los outliers
  - Es más razonable (y laborioso!!) empezar con $k = 10$ y analizar de una forma iterativa, los grupos que se van encontrando

Limitaciones de las técnicas

A menudo, determinadas técnicas se usan simplemente porque son bien conocidas o disponemos de software que las implementa

- Se deben considerar sus limitaciones
  - Por ejemplo, los datos no son tan ideales como exigen las técnicas (i.e. distribución normal en las características de la muestra)
- La regresión lineal es un buen ejemplo ya que es fácil de entender y de interpretar
  - Ocurre que aun generándose un buen error rooted-square, el modelo no explique del todo los datos, precisamente por su sencillez (e.g. un cjto. de datos aleatorio)
- Muchos métodos son iterativos, lo que hace que, con un cjto. de datos grande, escalen mal

Garantías en los resultados del proceso de minería

Algunas técnicas per se no ofrece garantías del *goodness of fit*

- Este es el caso, a menudo, del clustering que optimiza la localización y estructura de los clusters según
  - cómo están distribuidos los clusters de manera relativa y
  - cómo de bien están los puntos distribuidos con respecto a cada centroide según una medida de distancia
- ¿y si queremos comprobar cómo replican dichos clusters (i.e. el vector de centroides) la estructura de los datos? (costoso en tiempo y dinero)
- Un caso real

> *Long Term Capital Management, que trabajan con bonos o valores muy estables (i.e. securities). En esta firma se trató de encontrar bonos que evolucionaran en direcciones opuestas, para comprar de ambos y así asegurar estabilidad en promedio para valores conjuntos. A largo plazo, se encontró que la aproximación del modelo no era la correcta y se perdieron miles de millones de dólares cuando esos bonos evolucionaron en el mismo sentido.*

EDA, primera aproximación basada en atributos singulares

El análisis exploratorio de datos puede ser extremadamente sencillo

- Podemos generar resúmenes breves sobre los datos
  - Justificación: conjuntos de datos no familiares y grandes $\longrightarrow$ las técnicas deben generar unos resultados inmediatamente interpretables y fáciles de computar
- Por tanto, nos dedicaremos a describir valores típicos de los atributos y la variación de esos valores para los mismos atributos.

Conceptos básicos

- La **muestra** es el cjto. de datos a partir de la cual vamos a extraer características regulares de un fenómeno que estamos estudiando.
- Modelamos cada atributo de la muestra como una **variable aleatoria**
  - El conjunto de valores que puede tomar se denomina **dominio** o soporte
  - Una variable aleatoria puede venir definida por su *función de distribución de probabilidad*
  - Si consideramos varias variables aleatorias simultáneamente, estudiamos cómo se interrelacionan las ocurrencias de las variables consideradas entre sí
  - Una **distribución de probabilidad multivariable** representa la probabilidad, para un conjunto de atributos, de que tomen un cjto. dato de valores de entre los de sus dominios respectivos

Estimaciones de elementos desconocidos

Dos ideas principales al respecto

No conocemos las v.a.

Con frecuencia no conocemos cómo funcionan las variables aleatorias con que modelamos los atributos del sistema. Es decir, las funciones de distribución de probabilidad, f, son desconocidas. Y habrá que estimarlas de alguna forma.

Selección del estimador

Estas estimaciones pueden ser tan sencillas como un valor promedio (e.g. una media) o como una regla de decisión como "si es hombre y tienen entre 10 y 50 años, preferirá una pelicula de acción".

Problema hipotético a modo de ejemplo

Población bajo estudio: animales mitológicos dispuestos en una región griega.

Especies identificadas: Grifos, Serpientes aladas y Unicornios

Características por ejemplar observado: Especie, edad, peso, volumen

Estimación de f

Va a ser una tarea compleja, realizaremos incrementalmente, análisis cada vez más complejos

1. Describir valores típicos para los atributos: e.g. una serpiente típica tiene 45 unidades de edad, pesa 10 unidades y ocupa 16 unidades de espacio.
2. Cuantificar cómo el resto de valores se desvían de los valores típicos: el 3% de los Grifos tiene un peso anormalmente grande.
3. Identificar diferencias entre diferentes grupos, para los mismos atributos: Las Serpientes y los Unicornios tienen una distribución de probabilidad diferente para el peso.
4. Generar hipótesis a contrastar: están la edad y el peso de las Serpientes Aladas correlados de alguna forma?
5. Caracterizar movimientos de atributos agredados a través del tiempo: los Unicornios que han ganado peso en tres unidades de tiempo consecutivas tienen más probabilidad de morir próximamente.

Definimos ahora el EDA

Exploratory Data Analysis

Es el descubrimiento de estructura en los datos por medio de métodos simples como parámetros de estadística descriptiva o técnicas de visualización.

- Es un paso previo al data mining típico
- Con el EDA elminaremos datos espúreos para aumentar la calidad de los datos
- Reforzaremos supuestos básicos que podamos hacer sobre los datos (e.g. el atributo X tiene una distribución normal o bien los atributos Y, Z, U están relacionados con X linealmente)

Más sobre el EDA

- Los métodos que vamos a utilizar generarán regularidades de interpretación trivial
- El coste computacional será muy bajo lo que permitirá obtener dichas regularidades de manera interactiva ya que
  - Nos enfrentamos a datos no conocidos (carecemos de información para seleccionar unos métodos frente a otros)
- Serán
  - De aplicación genérica
  - Facilmente adaptables a cambios los datos (i.e. una nueva especie de animal mitológico o un nuevo atributo)

Estimación paramétrica y no paramétrica

Estimación paramétrica

- Idea básica: asumimos que la distribución subyacente a los datos tiene forma de función paramétrica que podemos representar
- Ejemplo: una combinación lineal de variables

$$f() = ax_1 + bx_2 + cx_3 + d,$$

en la cual los parámetros a estimar son $\theta = (a, b, c, d)$

- Otro ejemplo: el atributo de la edad corresponde a una variable aleatoria con una distribución de probabilidad exponencial tal que

$$P(X \leq x) = \frac{1}{\theta} \int_0^x e^{-\frac{u}{\theta}} dx,$$

con θ el parámetro a estimar

Valores típicos de un atributo

Los valores típicos de un atributo son valores singulares, representativos del mismo (e.g. media, mediana, moda)

- Ya sabemos como se obtienen
- Una media es bastante sensible a valores fuera de rango
- Como complemento podemos usar una media recortada en los extremos (*trimmed mean*)
  - Ejemplo

$$(95, 90, 93, 98, 91, 90, 98, 97, 99, 9)$$

  - Media 86 (no representativa)
  - Si recortamos los extremos, media 94.6
  - Trimmed mean: se obtiene al eliminar entre un 2 y un 10% de los valores en los extremos (protegen frente a outliers)

Mediana

Recordemos una posible definición

Mediana de X

La mediana de la v.a. X, la denotamos con M, es el valor en el conjunto soporte de X tal que su valor acumulativo de probabilidad es 0.5, i.e.

$$P(X \leq M) = \int_{-\infty}^{M} f(u)d(u) = 0.5$$

Mediana

Detalles interesantes sobre la mediana

- Es posible calcularla de manera eficiente para grandes conjuntos de datos (i.e. con un error tolerable)
- Su estabilidad frente a cambios en el conjunto de datos.
  - M no cambia drásticamente del centro si no cambian la mitad de los datos (mientras que el orden se mantenga igual)
  - En el ejemplo anterior M se mantiene entre el valor 95 y el 97, aun con un 9, claramente fuera de rango.
- Nótese que si la media es considerablemente más grande que la mediana, la muestra está desplazada hacia la derecha.

Distribuciones de probabilidad, utilidad y formas

- La distribución de probabilidad de los atributos utilizados en la muestra, es importante en el EDM
- Es vital para caracterizar qué valores están dentro de los más probables o no
- Ejemplos extremos (formas picudas o aplanadas)
  - Si consideramos como v.a. la edad a la cual la gente empieza a conducir, la correspondiente distribución de probabilidad tendrá forma de pico
  - Si consideramos como v.a. la edad de los sujetos que viven en una gran ciudad, entonces la pdf tendrá una forma más aplananda

Varianza de variables aleatorias

Como mecanismos comunes de medir la dispersión de valores tenemos la varianza y su raíz cuadrada (i.e. la desviación estándar)

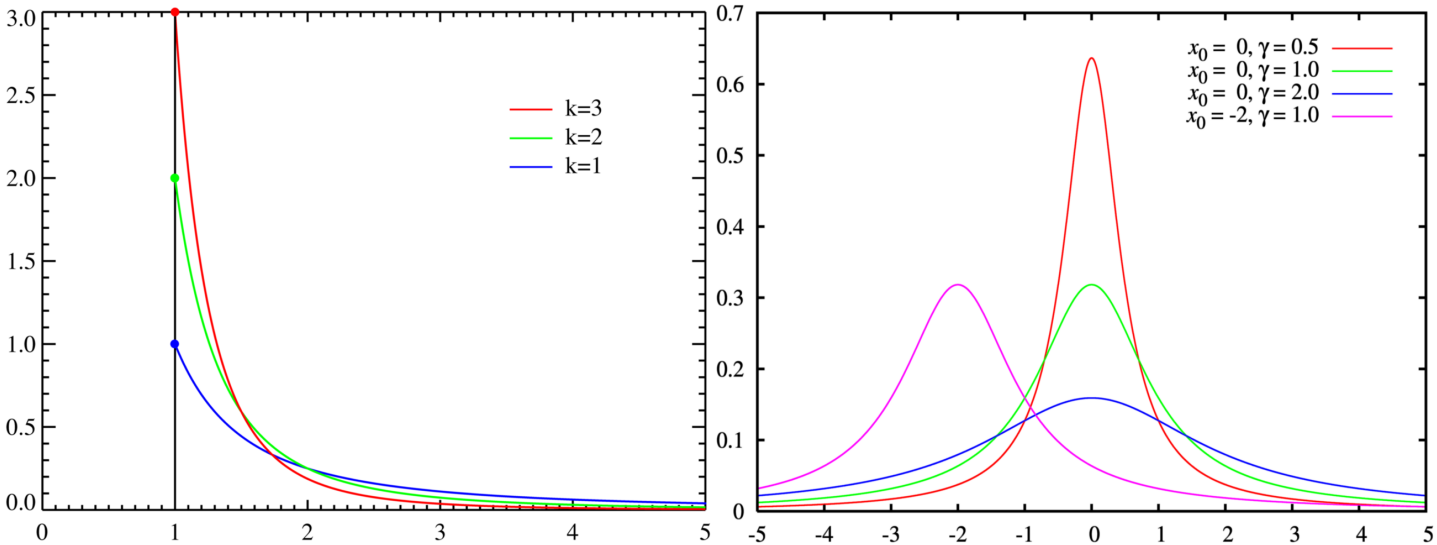
Varianza de una muestra

La varianza de una muestra es el promedio de las diferencias cuadráticas entre el valor de los puntos de la muestra y la media, tal que

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

No siempre es aplicable!!

Algunas distribuciones de probabilidad no tienen varianza (i.e. heavy-tailed pdfs), e.g. Pareto y Cauchy



http://en.wikipedia.org/wiki/Image:Pareto_distributionPDF.png

http://en.wikipedia.org/wiki/Image:Cauchy_distribution_pdf.png

Varianza para el caso multi-variable

La varianza generaliza bien para el caso multivariable

La mediana implicaba ordenar los valores para calcularla, lo que necesariamente desordenaba los valores del resto de variables

Para un par de variables **la matriz de dispersión** en la diagonal incluye la varianza por atributo y en las celdas inferiores la dispersión entre pares de atributos (i.e. entre pares de atributos).

Desviación absoluta

Idea principal: la media era muy sensible a valores fuera de rango $\rightarrow$ también las medidas de dispersión

- La desviación absoluta utiliza la mediana para calcular una medida de dispersión
- La MAD (*Median Absolute Deviation*) se obtiene según

$$MAD = mediana|X - \xi_f(0.5)|,$$

en donde $\xi(0.5)$ es la mediana del atributo X, según su distribucion de probabilidad f

Rangos

Rango de X

dado el soporte de la v.a. X, el rango de valores de X en la muestra, que se obtiene a partir de la expresión

$$R = max(X_i) - min(X_i),\ 1 \leq i \leq N.$$

No es fiable (recordemos la trimmed mean)

IQR (*Interquartile Range*)

Sea Q_3 un estimador del tercer cuartil (i.e. $\xi(0.75)$), basado en la muestra, y Q_1 un estimador del primer cuartil (i.e. $\xi(0.25)$). Si tenemos que

$$IQR = \xi(0.75) - \xi(0.25),$$

una estimación de IQR será entonces

$$\hat{IQR} = Q_3 - Q_1.$$

Ejemplo:

$$(1, 2, 2, 3, 4, 6, 6, 80)$$

El número de instancias es 8, el mínimo es 1 y el máximo 80, $Q_1 = 2$, $Q_3 = 6$. Tenemos entonces que

$$R = 80 - 1 = 97,\ IQR = 6 - 2 = 4.$$

Relaciones entre atributos

¿Para qué? Podemos averiguar si dos atributos están relacionados o son independientes

Si hablamos de relaciones lineales

- Dados dos atributos, la covarianza viene definida por

$$C(\hat{X}, Y) = \frac{\sum_{k=1}^{N} (X_k - \overline{X})(Y_k - \overline{Y})}{N - 1},$$

  siendo k el tamaño de la muestra.
- El coeficiente de correlación viene dado por

$$\rho = \frac{C(X, Y)}{\sigma_X \sigma_Y},$$

  donde las σ son las correspondientes desviaciones estándar de X e Y.

Interpretando los valores de ρ

- Valores absolutos grandes de ρ implican una alta correlación y próximos a cero una baja correlación.
- Si $\rho = 0$ no implica necesariamente independencia (no la hay lineal)
- Por ejemplo, la relación no lineal $Y = X^2$ da lugar a

$$(-1, 1), (1, 1), (0, 0), (2, 4), (-2, 4)$$

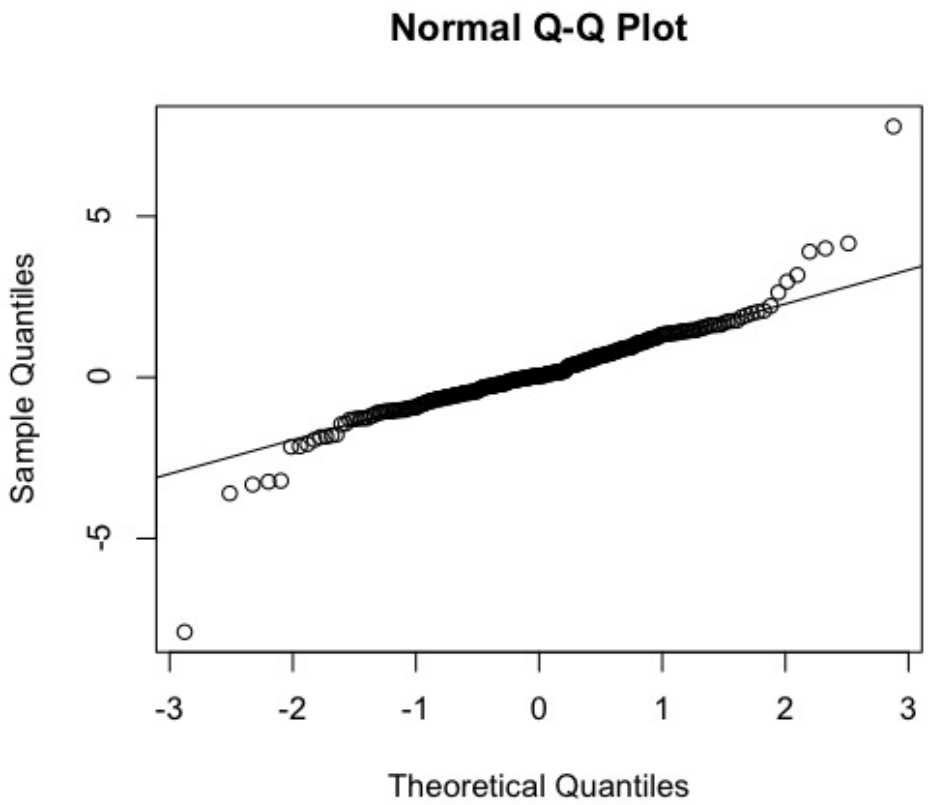
- Suelen producirse en la realidad

  Si consideramos el número de subscriptores a un servicio cerrado (teléfono o la TV por cable), se produce un crecimiento inicial en el número de servicios muy lento, pero cuando se alcanza una másа crítica de subscriptores, el número de subscriptores crece exponencialmente, seguido de un uso plano (i.e. sin crecimiento) cuando el servicio se satura (curva S, cjtos. de gran tamaño)

Curvas Q-Q

- Una buena herramienta visual para comprobar la forma de la distribución de probabilidad de nuestros atributos
- Representamos los cuantiles del atributo (usualmente escalado) contra los cuantiles de una determinada distribución de probabilidad
- Por ejemplo, una $N(0, 1)$
- La forma de la curva resultante, con respecto a una línea recta, ofrece indicaciones de la distribución del atributo.

Curvas Q-Q (II)



La parte izquierda queda por debajo de la normal, y la derecha es una cola que queda algo por encima.

Estimación no paramétrica

Es posible que partamos de problemas en los que no tenemos información alguna sobre la forma de la pdf. En ese caso, podremos alternativamente

- generar histogramas, encontrar valores que ocurran frecuentemente,
- localización de valores que co-ocurren frencuentemente (i.e. reglas de asociación),
- generar reglas que puedan aplicarse frecuentemente (i.e. árboles de decisión, etc), etc

Las dos últimas forman parte del resto de contenidos de la asignatura así que nos centraremos en el cjto. de técnicas que aparece en el primer item.

Cuando no sabemos nada...

- Necesitamos adentrarnos mucho más en los datos
- Necesitamos
  - Localizar relaciones no lineales entre atributos localizadas en franjas de valores
  - Delimitar variaciones localizadas de atributos y su localización en el espacio, etc.
- Y todo para responder a preguntas como
  - ¿cuánta gente en Murcia viaja en un carril particular de la autovía entre las 8:00AM y 10:00AM en días laborables?
  - ¿Qué artículos se compran frecuentemente de manera conjunta?

Empezaremos construyendo una tabla de frecuencias

Sean dos atributos, dividimos sus soportes en intervalos y tabulamos (utilizamos ej. anterior para la especie de los Grifos)

Edad	Peso					Total fila
	0-5	5-10	10-30	30-50	50+	
0-10	3	4	5	2	1	15
10-20	5	10	2	4	0	21
20-30	4	7	2	1	1	15
30+	1	2	0	1	0	4
Total Col.	13	23	9	8	2	55

Interpretación de la tabla de frecuencias

- Todo organismo que cae dentro del mismo intervalo será tratado igualmente al resto
- Los valores en la última fila/columna son totales marginales (estiman las probabilidades marginales del atributo)
- Por ejemplo,

$$P(E = e) = \sum_{w} P(E = e, W = w)$$

  es la probabidad marginal de que el atributo edad, E tiene como valor e, obtenido sumando todos los posibles valores del atributo peso, W
- Para estimar prob. marg. en, e.g. la edad

$$\hat{P}(E \in (10, 20)) = \frac{21}{55}.$$

Probabilidades conjunta y condicionada

Supongamos que queremos aproximar E para valores de W en 5-10 (i.e. probabilidad condicionada)

$$\hat{P}(E \in (10, 20) | W \in (5, 10)) = \frac{10}{23}.$$

Ahora la probabilidad conjunta de la edad en 10-20 y peso en 5-10,

$$\hat{P}(E \in (10, 20) \cap W \in (5, 10)) = \frac{10}{55}.$$

La relación entre probabiliad marginal, condicional y conjunta viene dada por

$$P(E \in I_1 | W \in I_2) = \frac{E \in I_1 \cap W \in I_2}{P(W \in I_2)},$$

para intervalos I_1 e I_2.

Probabilidad acumulada

Si nos interesara llevar la cuenta de los puntos que caen por debajo de un determinado valor de un atributo concreto...

Función de distribución acumulativa empírica (FDAE)

$$\hat{F(x)} = \frac{\text{número de puntos } \leq x}{\text{número total de puntos}},$$

$$= \frac{\sum_{i=1}^{n} I_i(x)}{N},$$

donde $I_i(x) = 1$, $X_i \leq x$, $I_i(x) = 0$, $X_i > x$, siendo X_i el punto de la muestra.

¿Para qué la FDAE?

- Si la representamos podemos ver en qué regiones es más probable que encontremos puntos de la muestra
- Si superponemos dos FDAE para un mismo X, podemos comparar grupos (e.g. Serpientes y Grifos)
  - Si son diferentes, paramos
  - Similitudes obligarían a análisis más detallados
- Algo similar para contrastar si un atributo sigue una determinada distribución de probabilidad.

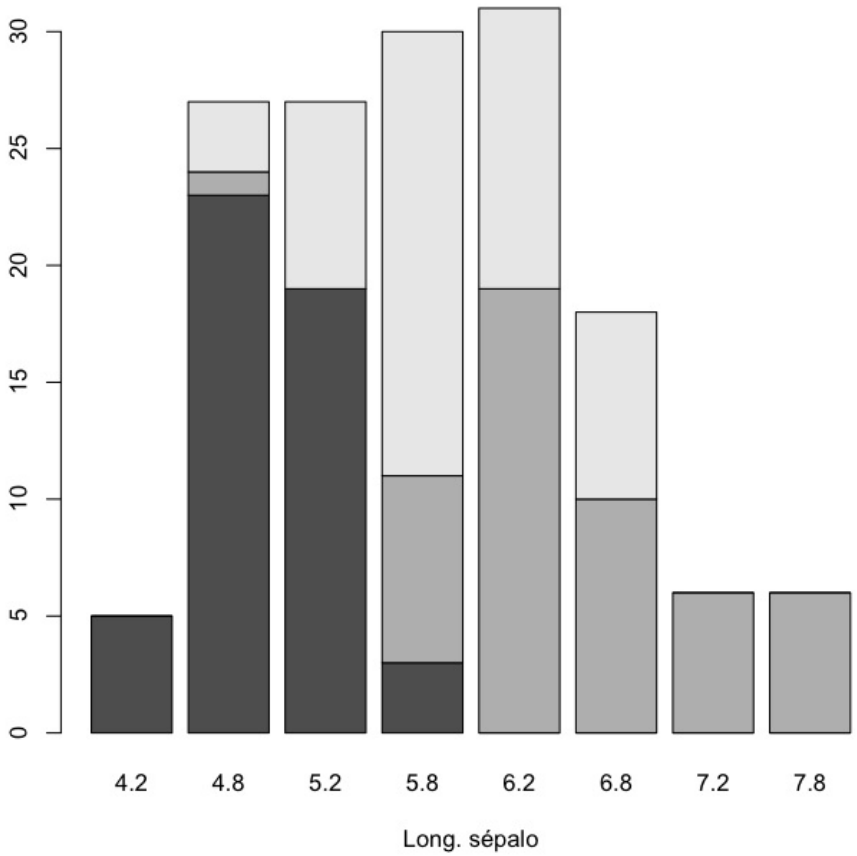
Histogramas

Son una representación que muestra la concentración de datos según determinados intervalos de valores

- ¿cuántas personas ha realizado una llamada internacional entre las 10:00 y las 11:00?
- Se construyen en dos fases
  1. dividir el soporte de cada atributo en intervalos y
  2. contar los puntos de datos que caen dentro de cada intervalo.

Ejemplo de histograma

Para el Iris



Libros aconsejables


Tamraparni Dasu and Theodore Johnson.
Exploratory Data Mining and Data Cleaning.
Wiley Interscience, 2003.


Jiawei Han and Micheline Kamber.
Data Mining. Concepts and Techniques.
Morgan Kauffman Publisher, 2001.


Dorian Pyle.
Data Preparation for Data Mining.
Morgan Kauffman Publisher, 1999.