Universidad Zaragoza

Minería de datos y texto

Grado en Ingeniería en Informática

Curso 2023-2024

Raquel Trillo Lado (raqueltl@unizar.es)
Carlos Telleria (telleria@unizar.es)
Fernando Tricas García (ftricas@unizar.es)
Dpto. Informática e Ingeniería de Sistemas

Universidad Zaragoza
1542

ÍNDICE

- Introducción
- Regresión
- Minería de patrones y reglas de asociación
- Agrupamiento (*clustering*)
- Clasificación
- Minería de textos

Universidad Zaragoza
1542

INTRODUCCIÓN A LA MINERÍA DE DATOS

DEFINICIÓN DE MINERÍA DE DATOS

"Knowledge Discovery in Databases"

Gregory Piatetsky-Shapiro
(Data Mining and Analytics Expert, President of KDnuggets)

*"Extracción **no trivial** de información que reside de manera implícita en los datos"*

Wikipedia

Universidad Zaragoza
1542

INTRODUCCIÓN A LA MINERÍA DE DATOS

DEFINICIÓN DE MINERÍA DE DATOS

"Torturar a los datos hasta que confiesen"

Anónimo

"Escarbar montañas de datos y encontrar pepitas de oro (o diamantes)"

Anónimo

Universidad Zaragoza
1542

INTRODUCCIÓN A LA MINERÍA DE DATOS

DEFINICIÓN DE MINERÍA DE DATOS

Descubrir, a partir de los datos, conocimiento…

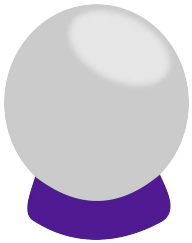
- No trivial
- Implícito
- Previamente desconocido
- Potencialmente útil
- En definitiva, interesante

INTRODUCCIÓN A LA MINERÍA DE DATOS

CONTEXTO DE LA MINERÍA DE DATOS

Según el objetivo general:

- Minería de datos predictiva



- Minería de datos descriptiva

Universidad Zaragoza

INTRODUCCIÓN A LA MINERÍA DE DATOS

ALGORITMOS POPULARES

Top 10:

1. *C4.5*
2. *K-Means*
3. *SVM*
4. *Apriori*
5. *EM*
6. *PageRank*
7. *AdaBoost*
8. *kNN*
9. *Naive Bayes*
10. *CART*

IEEE International Conference on Data Mining (ICDM), Diciembre 2006

Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. 2007. **Top 10 algorithms in data mining**. *Knowl. Inf. Syst.* 14, 1 (December 2007), 1-37. DOI=10.1007/s10115-007-0114-2.

https://link.springer.com/article/10.1007/s10115-007-0114-2

https://www.kdnuggets.com/2015/05/top-10-data-mining-algorithms-explained.html

Universidad Zaragoza
1542

INTRODUCCIÓN A LA MINERÍA DE DATOS

PASOS EN EL PROCESO DE MINERÍA DE DATOS

1. Aprender sobre el dominio de aplicación
   - Conocimiento del entorno y objetivos del análisis
2. Seleccionar los datos para analizar
   - Evitar sesgos, asegurar precisión

4. Preparar los datos (*cleaning* ⇢ fiabilidad)
   - Puede suponer en torno al 60% del esfuerzo total
5. Reducir y transformar los datos
   - Representación uniforme, detectar características de interés, reducción de la dimensionalidad

INTRODUCCIÓN A LA MINERÍA DE DATOS

PASOS EN EL PROCESO DE MINERÍA DE DATOS

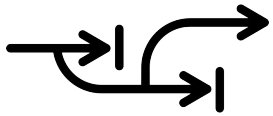
5. Escoger el objetivo de la minería de datos:
   - Agrupamiento
   - Clasificación
   - Asociación
   - …
6. Escoger un algoritmo de minería de datos apropiado
   - *No free lunch!*

Universidad Zaragoza

INTRODUCCIÓN A LA MINERÍA DE DATOS

PASOS EN EL PROCESO DE MINERÍA DE DATOS

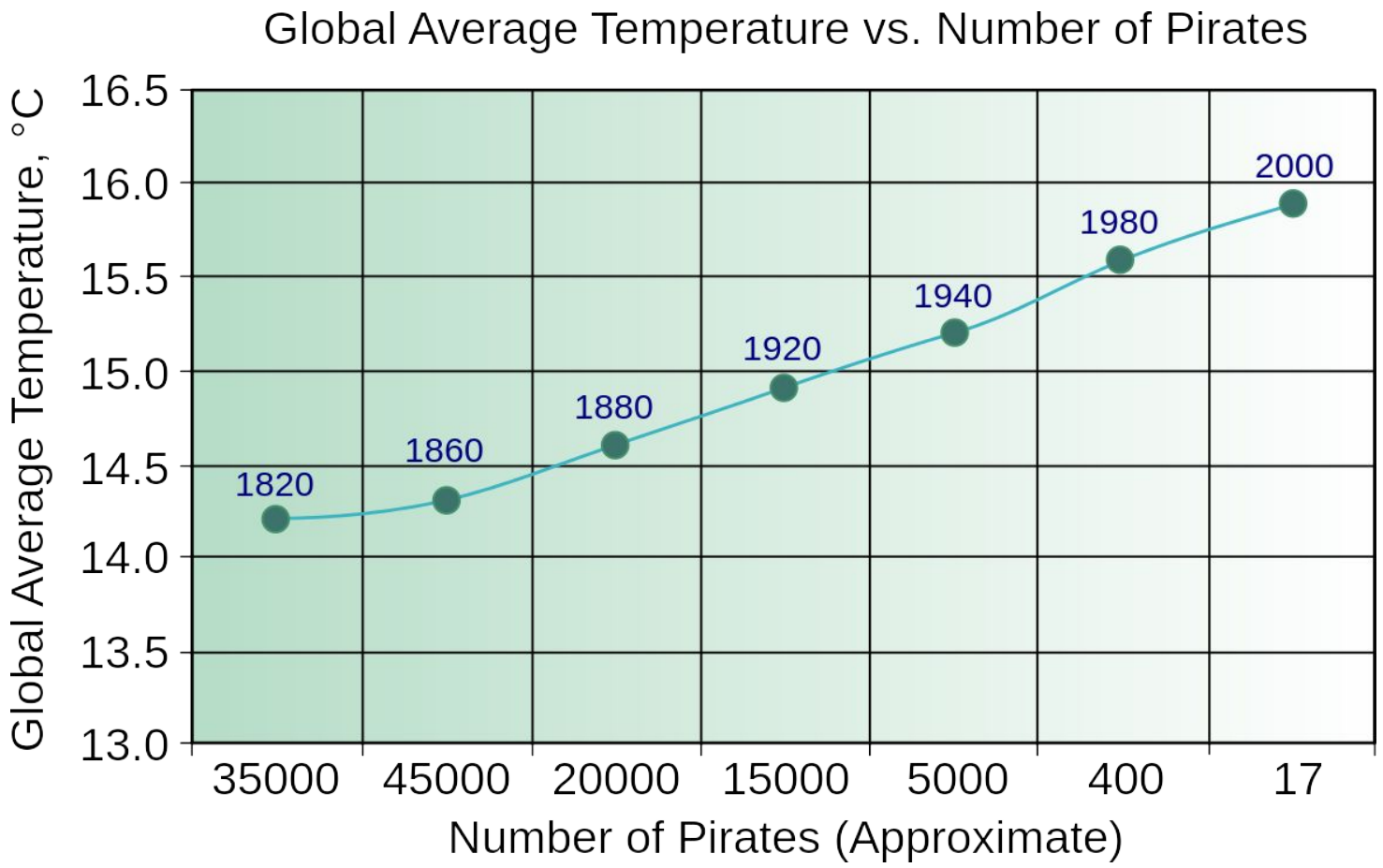
7. Analizar los resultados:
   - Utilización de herramientas de visualización de forma adecuada
   - Transformación de resultados, eliminación de patrones redundantes, etc.
   - Interpretación y extracción de conclusiones
     - No llegar a conclusiones precipitadas / erróneas
     - Peligro de la Estadística mal utilizada…
     - Correlación vs. causalidad



8. Explotación del conocimiento descubierto

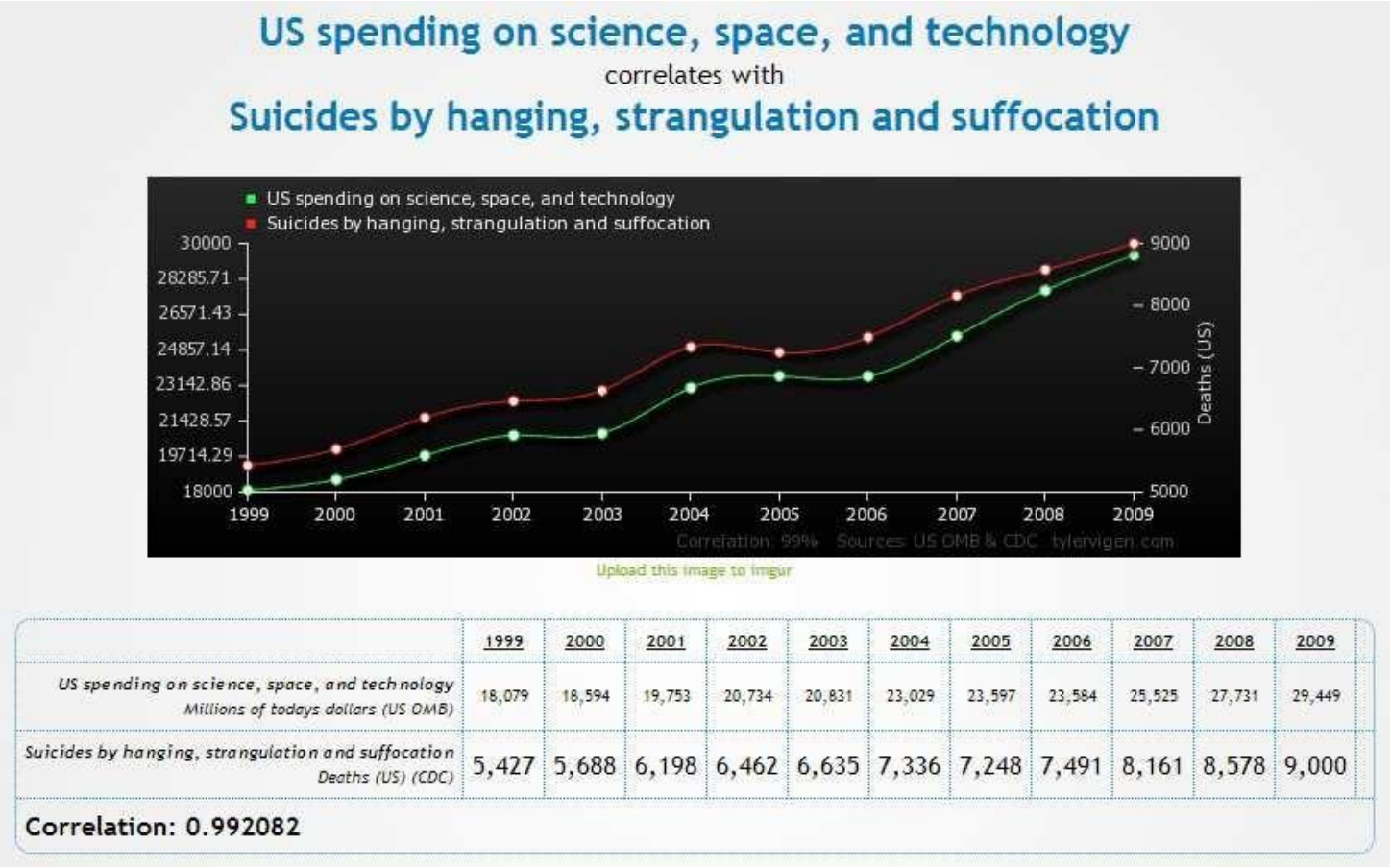
Universidad Zaragoza

CORRELACIÓN VS. CAUSALIDAD



Universidad Zaragoza
1542

CORRELACIÓN VS. CAUSALIDAD



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
US spending on science, space, and technology *Millions of todays dollars (US OMB)*	18,079	18,594	19,753	20,734	20,831	23,029	23,597	23,584	25,525	27,731	29,449
Suicides by hanging, strangulation and suffocation *Deaths (US) (CDC)*	5,427	5,688	6,198	6,462	6,635	7,336	7,248	7,491	8,161	8,578	9,000

Correlation: 0.992082

Spurious correlations:
http://www.tylervigen.com/

Universidad Zaragoza
1542

VISUALIZACIÓN

Basado en:
http://www.visualisingdata.com/2014/04/the-fine-line-between-confusion-and-deception/

Universidad Zaragoza

VISUALIZACIÓN

Basado en:
http://www.visualisingdata.com/2014/04/the-fine-line-between-confusion-and-deception/

Universidad Zaragoza

INTRODUCCIÓN A LA MINERÍA DE DATOS

EJEMPLOS DE HERRAMIENTAS

Herramientas gráficas:

- *Weka*
- *RapidMiner*
- *KNIME*

Librerías y entornos de trabajo:

- *Mahout*
- *Machine Learning Library (MLlib)*, para SPARK

Lenguajes de programación (con librerías):

- *R*
- *Python*

Universidad Zaragoza
1542

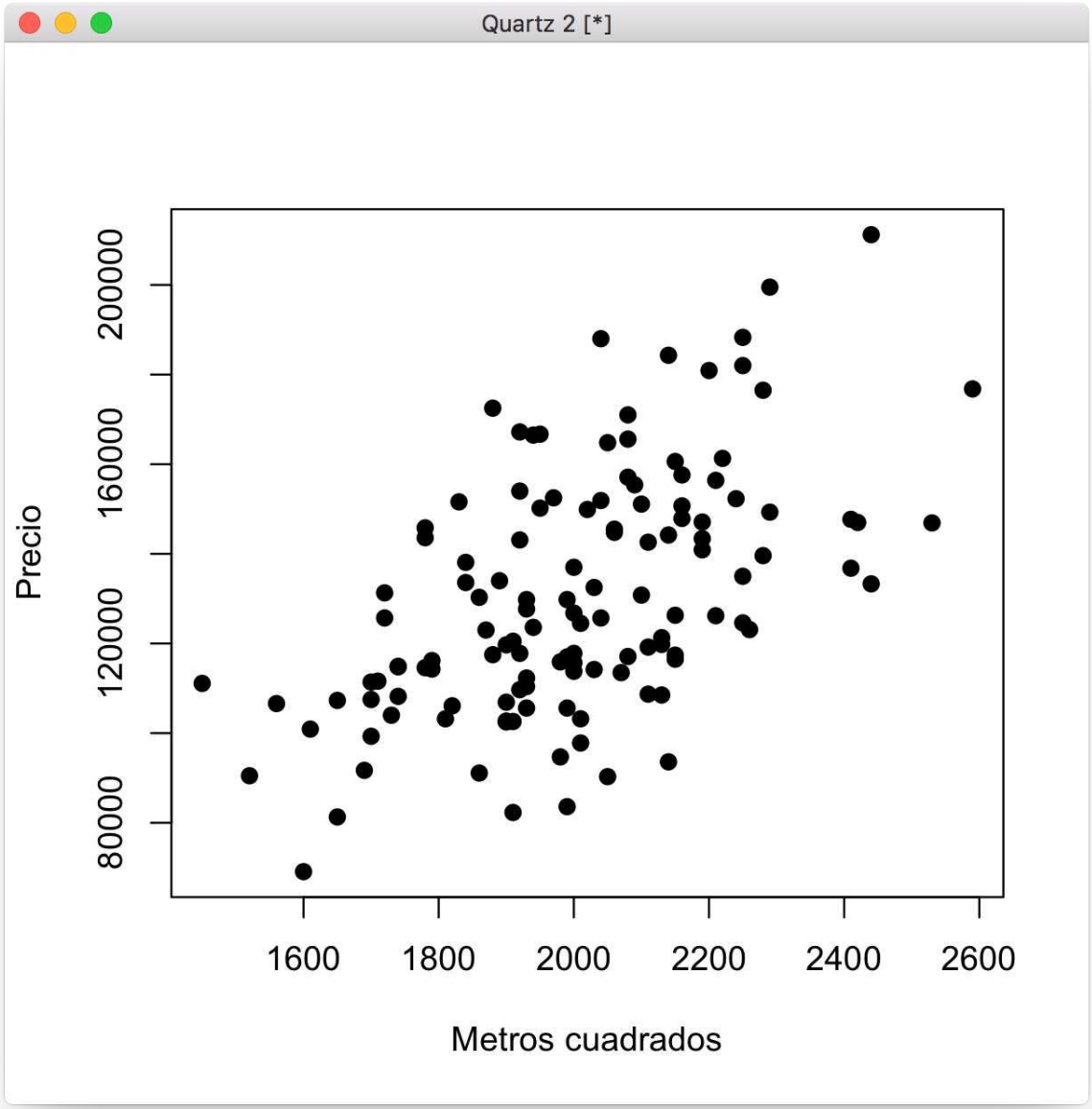
ÍNDICE

- Introducción
- Regresión
- Minería de patrones y reglas de asociación
- Agrupamiento (*clustering*)
- Clasificación
- Minería de textos

Universidad Zaragoza
1542

REGRESIÓN

Regresión Lineal



Universidad Zaragoza

REGRESIÓN

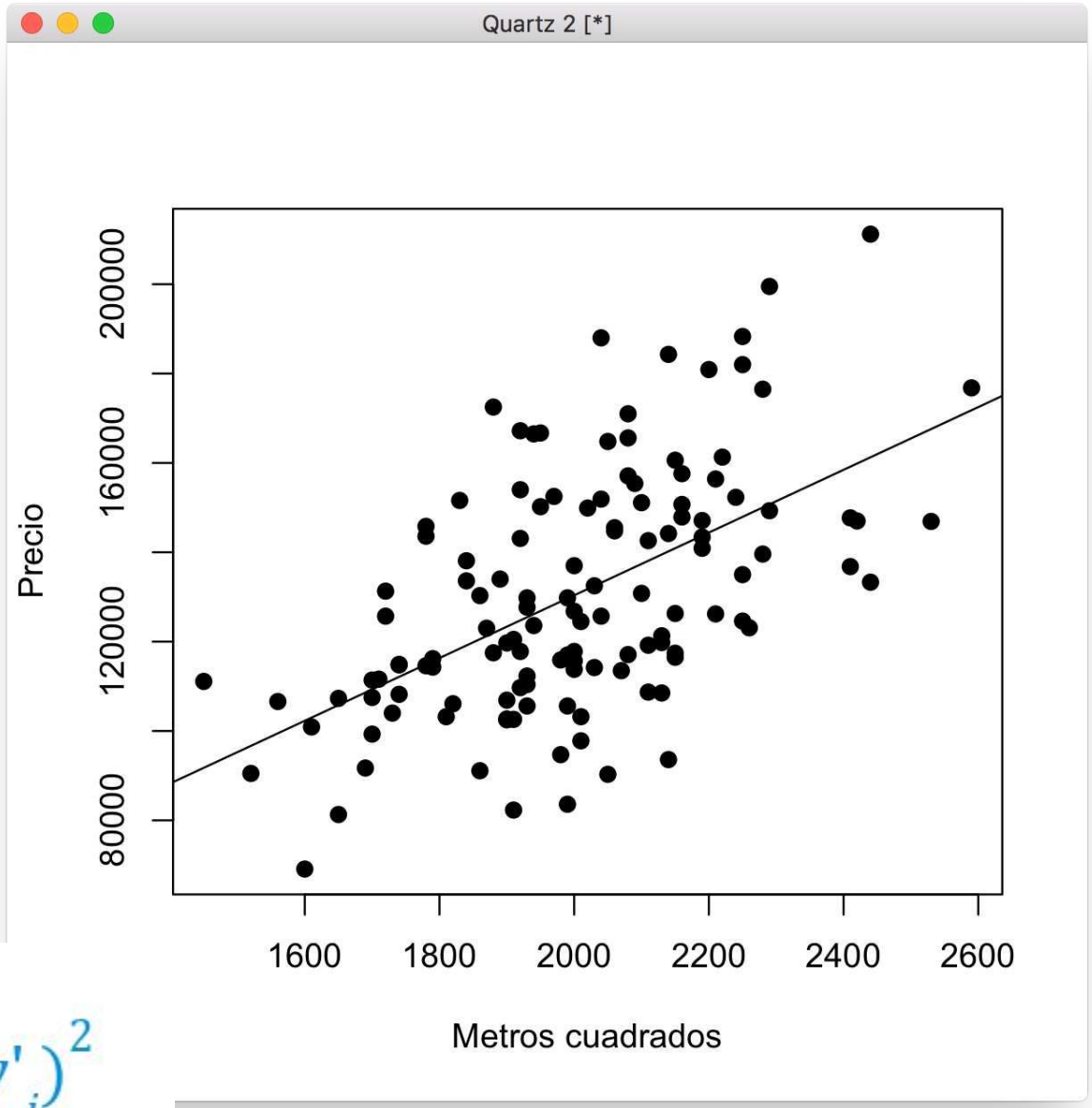
Regresión Lineal

Obtención de la línea que mejor se ajusta a los datos

$Y = \alpha + \beta X + \varepsilon$

$Y' = \alpha' + \beta' X$

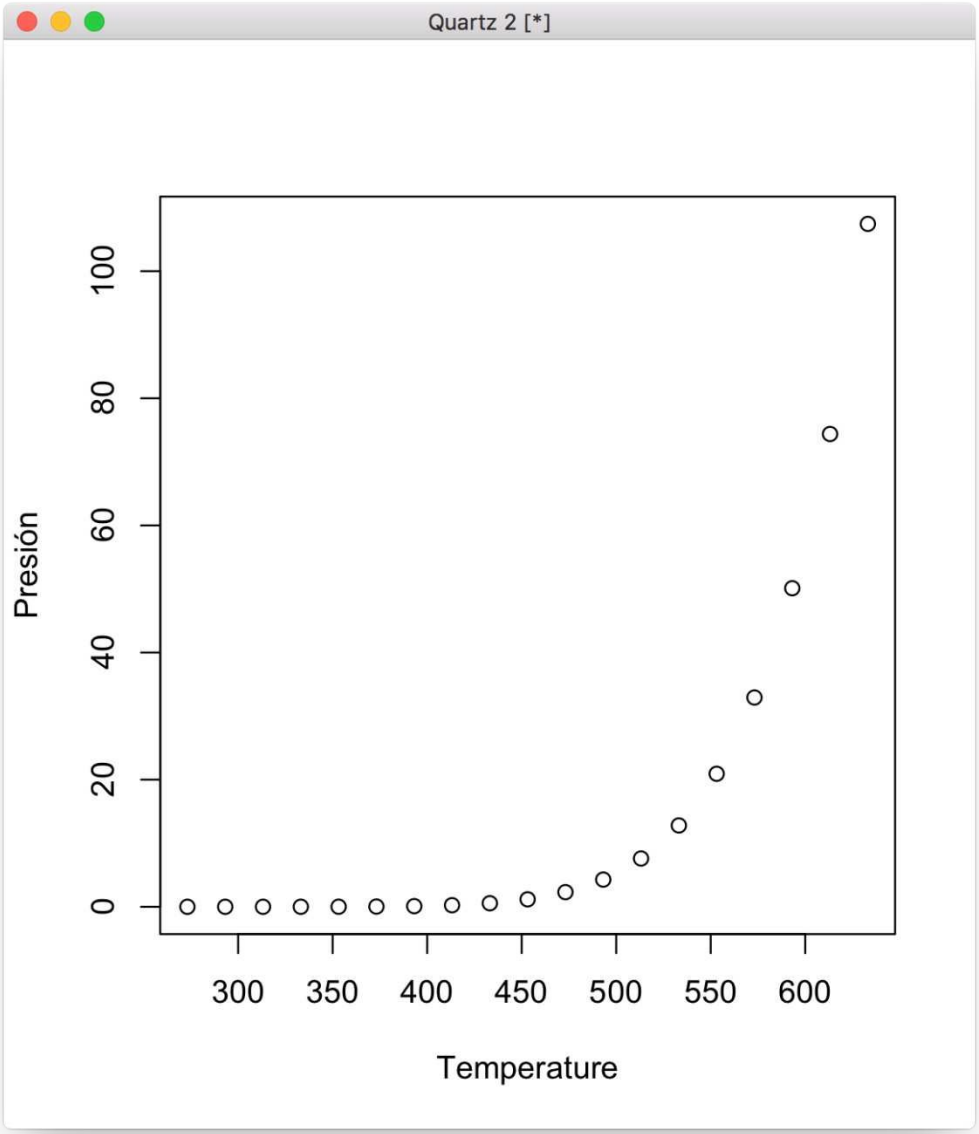
$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - y'_i)^2$$



Universidad Zaragoza
1542

REGRESIÓN

Regresión No Lineal

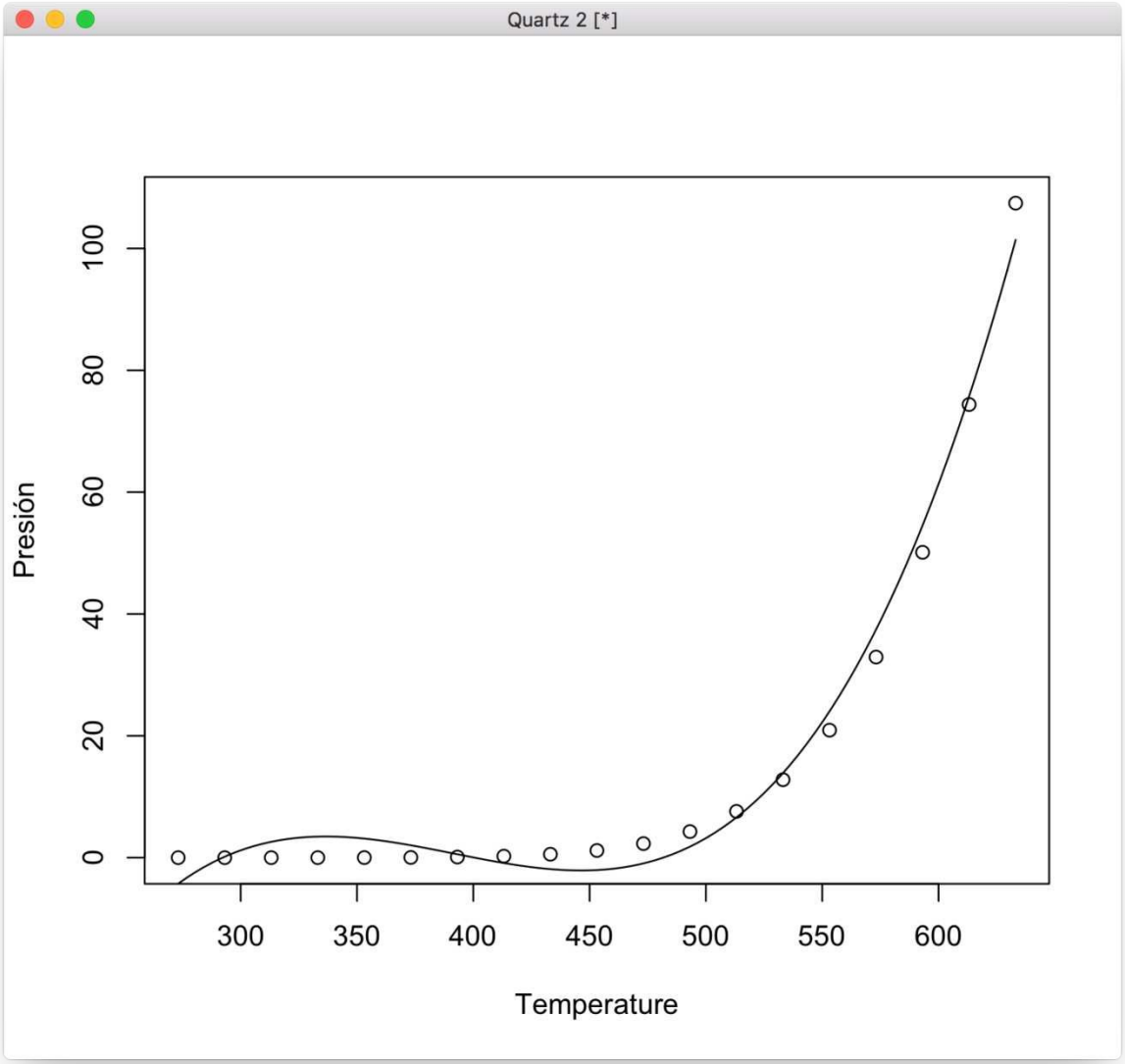


Universidad Zaragoza

REGRESIÓN

Regresión No Lineal

Obtención del polinomio (en este caso, cúbico) que mejor se ajusta a los datos



Universidad Zaragoza
1542

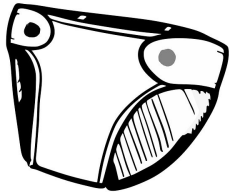
ÍNDICE

- Introducción
- Regresión
- Minería de patrones y reglas de asociación
- Agrupamiento (*clustering*)
- Clasificación
- Minería de textos

Universidad Zaragoza
1542

REGLAS DE ASOCIACIÓN

EJEMPLOS

{cerveza} ⇝ {pañales} (Ginebra + tónica, whisky + cocacola)

¿Limones?



{jamón, pan} ⇝ {aceite}



{orégano} ⇝ {espaguetis, salsa de tomate}

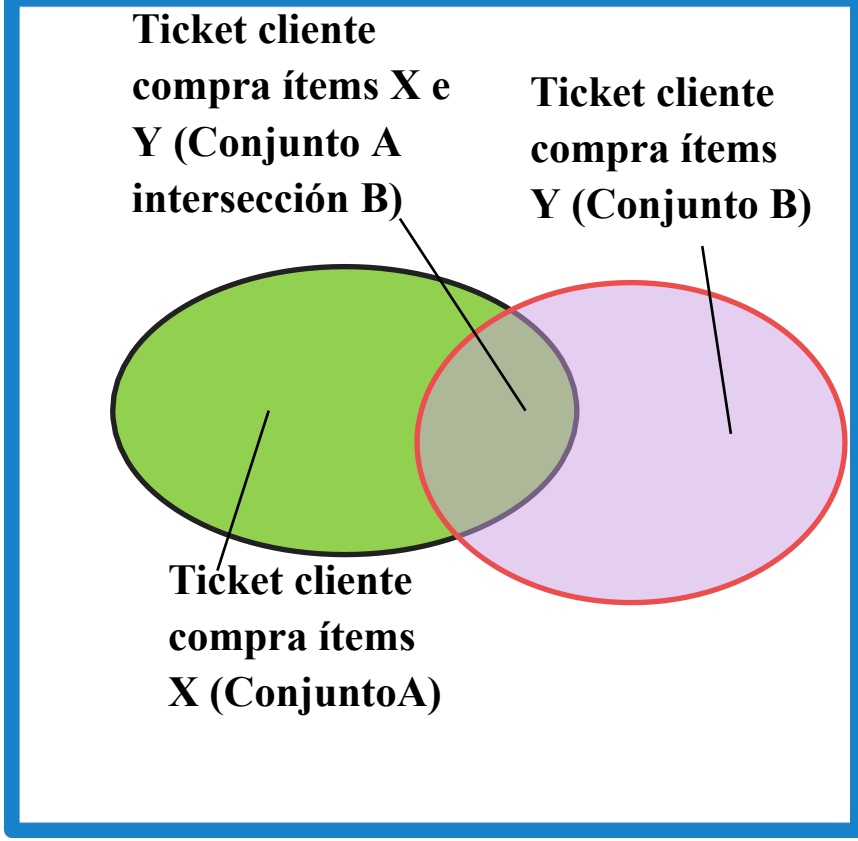
Universidad Zaragoza

REGLAS DE ASOCIACIÓN

DEFINICIÓN

Conjunto de ítems:

- $X = \{x_1, \ldots, x_k\}$ //conjunto de elementos del antecedente
- $Y = \{y_1, \ldots, y_l\}$ // conjunto de elementos del consecuente
- $A = \{ \text{evento} \mid \text{evento} \subset X \}$
- $B = \{ \text{evento} \mid \text{evento} \subset Y \}$
- $A \cap B = \{ \text{evento} \mid \text{evento} \subset (X \cup Y) \}$



Reglas $X \rightarrowtail Y$

- Si ocurre X entonces también ocurre Y

$A \cap B$ es el conjunto de eventos que cumple la regla $X \rightarrowtail Y$, es decir, que contienen el conjunto de items $X \cup Y = \{ x_1, \ldots, x_k, y_1, \ldots, y_l \}$

Ejemplo: Cada ticket es un evento. Los tickets que cumplen la regla "cerveza $\rightarrowtail$ pañales" son los tickets que contienen los ítems {cerveza, pañales}

REGLAS DE ASOCIACIÓN

DEFINICIÓN

Tres medidas importantes:

- Soporte (*support*)
  - s = probabilidad de que una transacción contenga $x_1 \ldots x_k$ e $y_1 \ldots y_l$
- Confianza (*confidence*)
  - c = probabilidad condicional de que una transacción que contenga X también contenga Y
- Elevación (*lift*)
  - Confianza / probabilidad no condicionada del consecuente de la regla

Universidad Zaragoza
1542

REGLAS DE ASOCIACIÓN

Soporte

Probabilidad de que una transacción tenga $X \cup Y = \{x_1, \ldots, x_k, y_1 \ldots y_l\}$

- soporte (X $\rightarrowtail\!\!\!\!\!\rightarrow$ Y) $= \frac{N_{X \cup Y}}{N}$

  N = Número total de instancias

  $N_{X \cup Y}$ = Número de instancias que contienen X e Y

  Valores entre 0 y 1 (0 = ningún soporte, 1=soporte total)

También podemos hablar del soporte del conjunto de items X

- soporte(X) = $\frac{N_X}{N}$ > soporte (X $\rightarrowtail\!\!\!\!\!\rightarrow$ Y) = soporte (X $\cup$ Y)

REGLAS DE ASOCIACIÓN

Soporte

- soporte({cerveza} → {pañales}) $= \frac{N_{\{cerveza, pañales\}}}{N}$
- soporte({jamón, pan} → {aceite}) $= \frac{N_{\{jamón, pan, aceite\}}}{N}$
- soporte({orégano} → {espaguetis, salsa de tomate}) $= \frac{N_{\{orégano, espaghettis, salsa\ de\ tomate\}}}{N}$

Universidad Zaragoza
1542

REGLAS DE ASOCIACIÓN

Confianza

Probabilidad condicional de que una transacción tenga X ∪ Y

- confianza (X ⇸ Y) $= \frac{N_{X \cup Y}}{N_X} = \frac{soporte(X \cup Y)}{soporte(X)}$

N = Número total de instancias

$N_{X \cup Y}$ = Número de instancias que contienen X e Y

N_X = Número de instancias que contienen X

Valores entre 0 y 1 (0 = ninguna confianza, 1 = confianza total)

REGLAS DE ASOCIACIÓN

Confianza

- confianza({cerveza} → {pañales}) = $\frac{N_{\{cerveza,pañales\}}}{N_{\{cerveza\}}}$
- confianza({jamón, pan} → {aceite}) = $\frac{N_{\{jamón,pan,aceite\}}}{N_{\{jamón,pan\}}}$
- confianza({orégano} → {espaguetis, salsa de tomate}) = $\frac{N_{\{orégano,espaghettis,salsa\ de\ tomate\}}}{N_{\{orégano\}}}$

Universidad Zaragoza
1542

REGLAS DE ASOCIACIÓN

Elevación (lift)

Tasa (ratio) entre el soporte y el producto de las probabilidades de cada conjunto de items por separado:

$$lift(X \to Y) = \frac{N_{X \cup Y}/N}{(N_X/N)*(N_Y/N)} = \frac{soporte(X \to Y)}{(N_X/N)*(N_Y/N)} = \frac{confianza(X \to Y)}{N_Y/N}$$

- proporción del soporte observado del conjunto de items sobre el soporte teórico asumiendo independencia (no asociación) entre los items. La elevación indica el incremento de la probabilidad de que ocurra el consecuente de la regla si se da el antecedente (frecuencia observada de una regla con la frecuencia esperada simplemente por azar).
- > 1 ⟶ correlación positiva (si se da X, es más probable que se dé Y)
- = 1 ⟶ sucesos independientes. Da igual si se da X o no para que se dé Y
- < 1 ⟶ correlación negativa (si se da X, es menos probable que se dé Y)

Cuando más se aleja de 1, mayor es la evidencias de que la regla no se debe a un artefacto aleatorio.

REGLAS DE ASOCIACIÓN

IMPORTANCIA DE ESTAS MEDIDAS

Sirven como criterio de calidad ⇝ interesa:

- Un soporte lo más alto posible
- Una confianza próxima a 1
- Una elevación > 1 (correlación positiva)

Se pueden utilizar estas medidas:

- Para filtrar reglas de asociación
- Para ordenar reglas de asociación

Universidad Zaragoza
1542

INTRODUCCIÓN

CUESTIÓN

¿Es $\{A\} \rightarrowtail\!\!\!\!\!\rightarrow \{B\}$ equivalente a $\{B\} \rightarrowtail\!\!\!\!\!\rightarrow \{A\}$?:

- Sí, ya que en ambos casos el conjunto de ítems es el mismo
- No, cambia el soporte de la regla
- No, cambia la confianza de la regla
- No, en el primer caso se produce A antes que B y en el segundo caso se produce B antes que A

Universidad Zaragoza

INTRODUCCIÓN

CUESTIÓN

¿Es $\{A\} \rightarrowtail\!\!\!\!\!\rightarrow \{B\}$ equivalente a $\{B\} \rightarrowtail\!\!\!\!\!\rightarrow \{A\}$?:

- Sí, ya que en ambos casos el conjunto de ítems es el mismo
- No, cambia el soporte de la regla
- No, cambia la confianza de la regla
- No, en el primer caso se produce A antes que B y en el segundo caso se produce B antes que A

Universidad Zaragoza
1542

ÍNDICE

- Introducción
- Regresión
- Minería de patrones y reglas de asociación
- Agrupamiento (*clustering*)
- Clasificación
- Minería de textos

Universidad Zaragoza
1542

INTRODUCCIÓN

ALGORITMOS DE AGRUPAMIENTO

Clúster / Agrupación / Grupo:

- Conjunto de entidades tales que hay:
  - Similitud **dentro** de la agrupación (cohesión): objetos dentro del mismo grupo son similares
  - Disimilitud entre agrupaciones **diferentes** (diferenciación): objetos de distintos grupos son diferentes
- Interesa :
  - **maximizar la similitud** dentro de la agrupación

    max. *intra-cluster similarity*
  - **maximizar la diferenciación** entre agrupaciones

    min. *inter-cluster similarity*

INTRODUCCIÓN

ALGORITMOS DE AGRUPAMIENTO

Se basan en una medida de distancia

Los grupos no están predefinidos

Si lo vemos como una técnica de aprendizaje:

- Sería aprendizaje no supervisado
- No hay clases predefinidas ⇝ no podemos tener muestras pre-etiquetadas

INTRODUCCIÓN

UTILIDADES

Ejemplos de aplicación:

- Para **describir y entender** los datos
- Como **paso previo** para otros algoritmos
- Descubrir **grupos** de clientes que puede ser **interesantes** considerar para lanzar campañas comerciales dirigidas a cada grupo
- **Identificar** conductores que comportan un **riesgo** especial para una aseguradora
- **Agrupar** viviendas similares (tipo, precio, localización)

INTRODUCCIÓN

UTILIDADES

Ejemplos de aplicación:

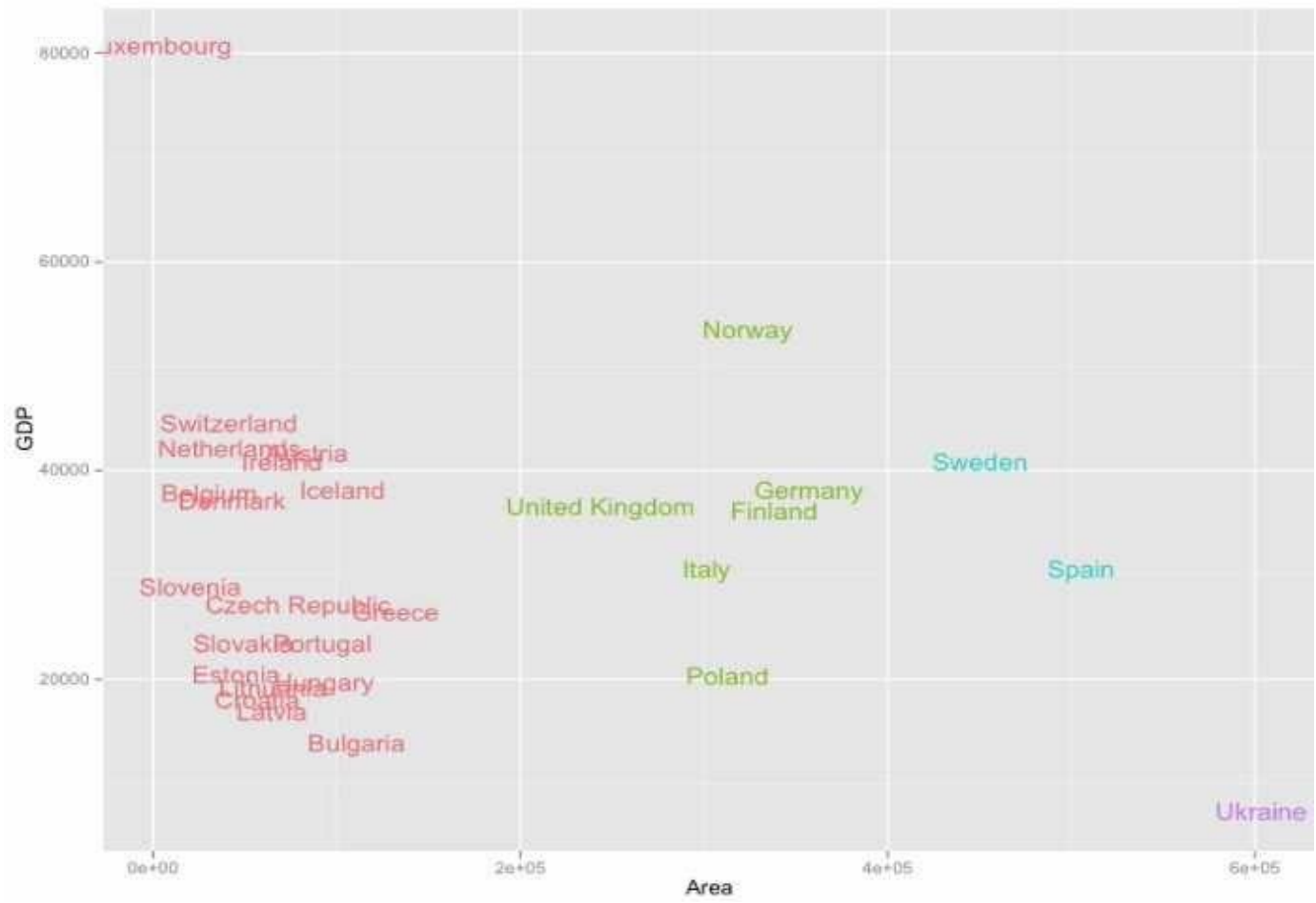
- Detección de outliers (valores atípicos)
  - Valores que difieren significativamente del resto

    - Ej. de utilidad: detección de **fraude** con tarjetas de crédito

Universidad Zaragoza
1542

K-MEANS

EJEMPLO (R)

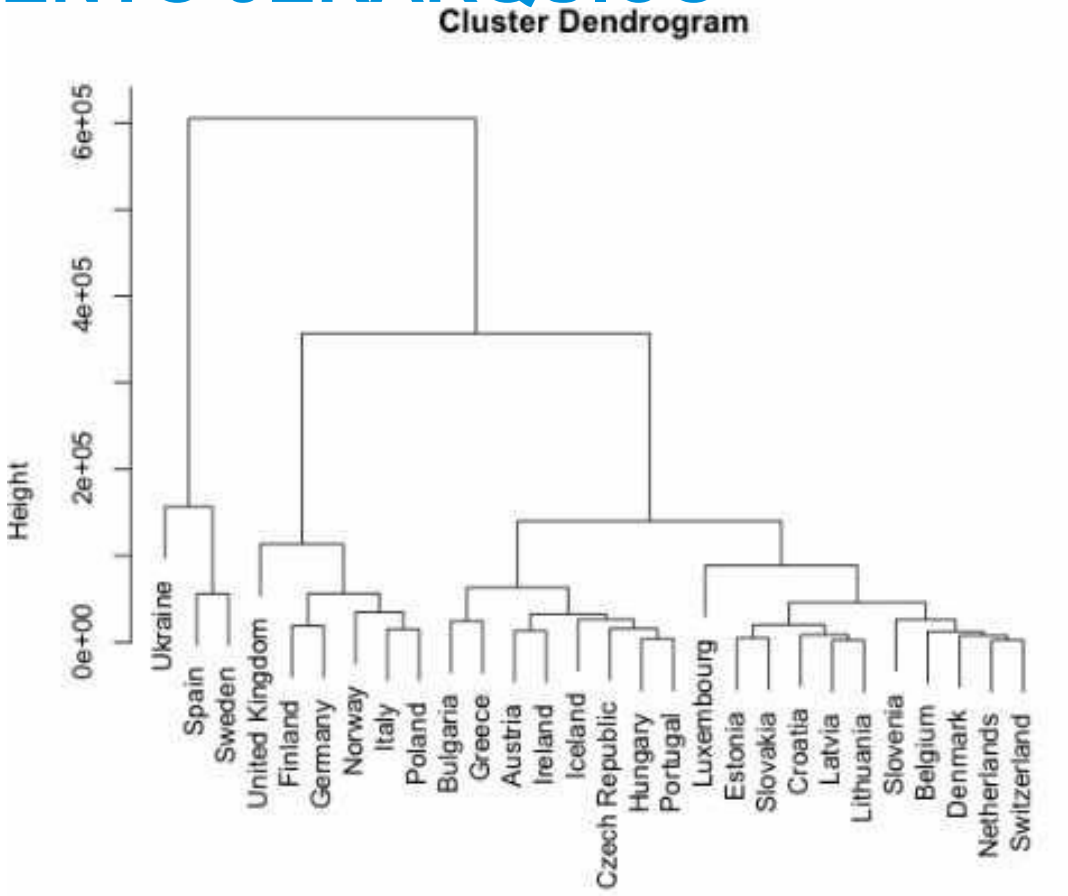


Ejemplo generado en R con el conjunto de datos *Europe* (http://www.instantr.com/wp-content/uploads/2013/01/europe.csv), el comando *kmeans*, y *qplot*

Universidad Zaragoza

AGRUPAMIENTO JERÁRQUICO

EJEMPLO (R)



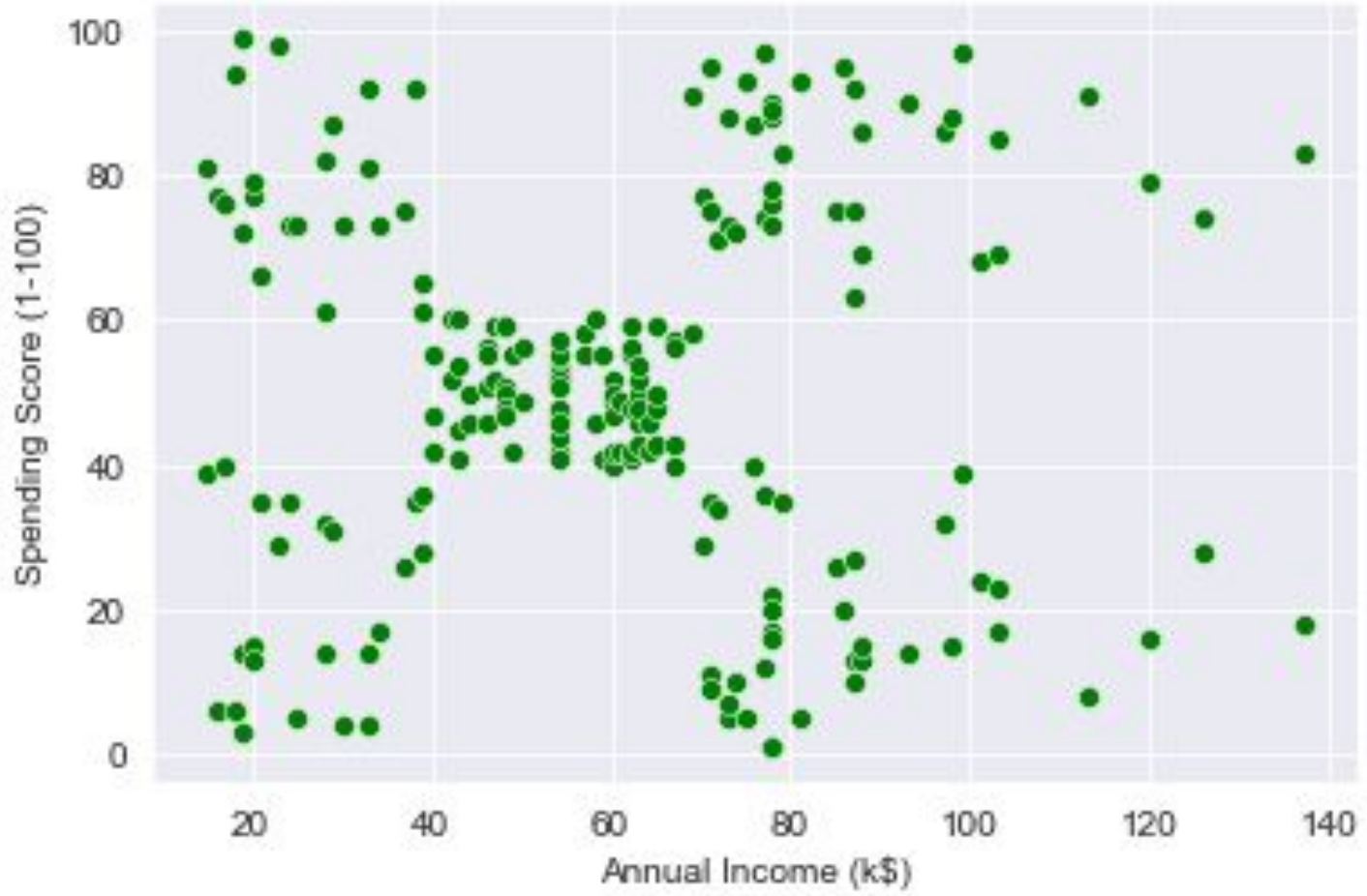
Ejemplo generado en R con el conjunto de datos *Europe,* el comando *hclust*, y *plot*

(http://www.instantr.com/wp-content/uploads/2013/01/europe.csv)

Universidad Zaragoza

K-MEANS

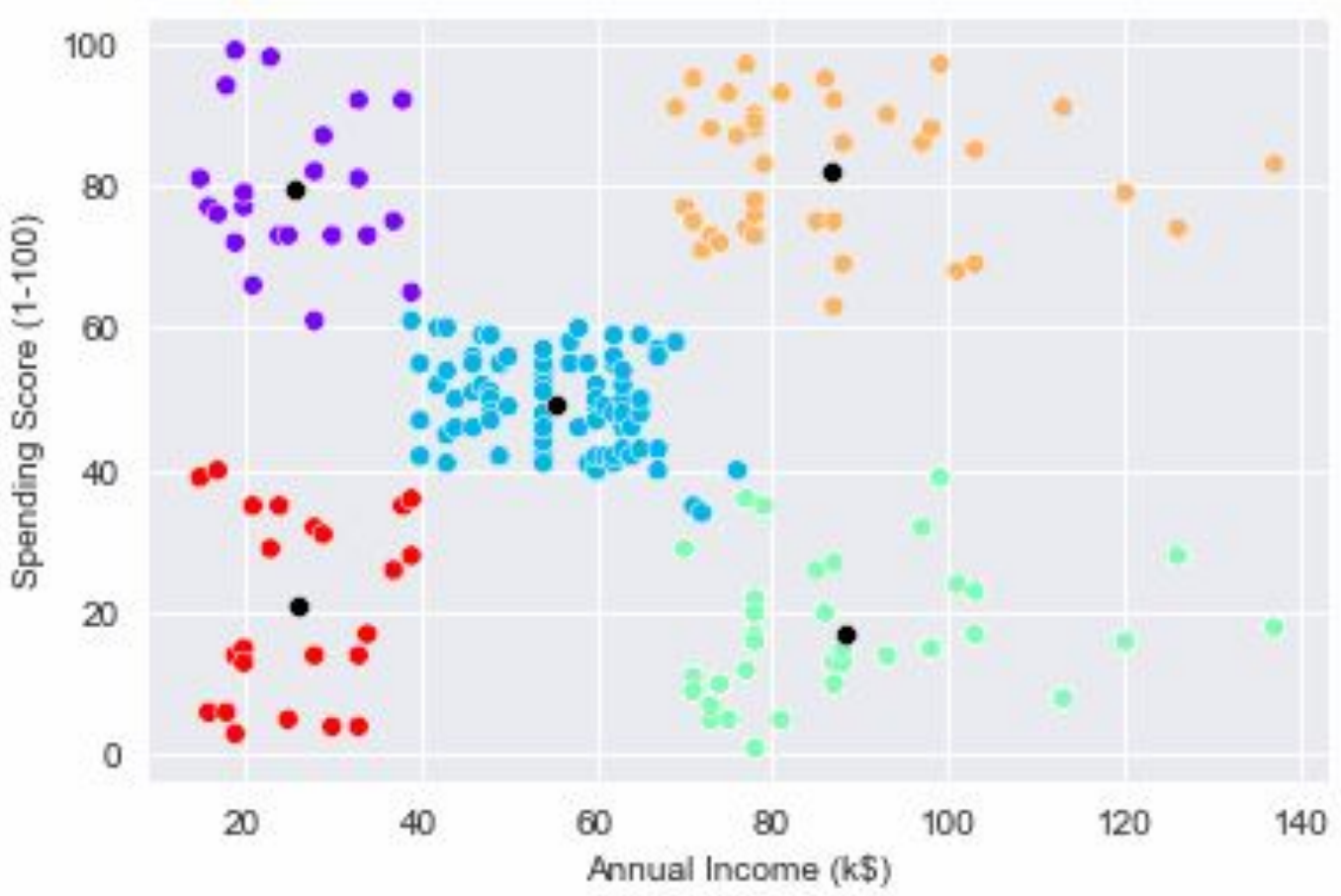
EJEMPLO (Python)



Universidad Zaragoza

K-MEANS

EJEMPLO (Python)



Ejemplo generado en Python.

https://wellsr.com/python/python-kmeans-clustering-with-scikit-lea

Universidad Zaragoza
1542

ÍNDICE

- Introducción
- Regresión
- Minería de patrones y reglas de asociación
- Agrupamiento (*clustering*)
- Clasificación
- Minería de textos

Universidad Zaragoza
1542

INTRODUCCIÓN

CONCEPTO DE CLASIFICACIÓN

- **Asociar** cada elemento del conjunto de datos a una de una serie de **categorías** definidas previamente

- Variables:
  - Hay una **variable respuesta** (*response variable*) o variable dependiente, que va a etiquetar cada instancia con la categoría correspondiente
  - El resto de **variables** se denominan **predictoras** (*predictor variables*) o variables independientes
  - Objetivo: explicar la variable dependiente en términos de las variables independientes

INTRODUCCIÓN

UTILIDADES

Ejemplos de utilidad:

- Clasificar el correo electrónico en **bueno** o correo **basura**
- Clasificar los clientes en **buenos**, **malos**, **medios**
- Clasificar automáticamente una noticia en la **sección** del periódico adecuada
- Clasificar ciudades en función de la **calidad de vida**
- …

Universidad Zaragoza

EVALUACIÓN DE LOS MÉTODOS

MATRICES DE CONFUSIÓN: CASO BINARIO

- Diagonal: clasificaciones correctas
- Elementos fuera de la diagonal: errores cometidos

		Clase detectada	
		Positiva	Negativa
Clase real	Positiva	tp	fn
	Negativa	fp	tn

- *tp* = *true positives* (Predicción correcta de positivo)
- *fp* = *false positives* (Predicción incorrecta de positivo)
- *fn* = *false negatives* (Predición incorrecta de negativo)
- *tn* = *true negatives* (Predicción correcta de un caso negativo)

Universidad Zaragoza

EVALUACIÓN DE LOS MÉTODOS

MATRICES DE CONFUSIÓN: CASO MULTICLASE

- Diagonal: clasificaciones correctas
- Elementos fuera de la diagonal: errores cometidos

		Clase detectada		
		A	B	C
Clase real	A	tp_A	e_{AB}	e_{AC}
	B	e_{BA}	tp_B	e_{BC}
	C	e_{CA}	e_{CB}	tp_C

- *tp = true positives*
- *e = errors*

CLASIFICACIÓN

MÉTRICAS TÍPICAS

- ***Precision (precisión)***
  - Capacidad para **no identificar incorrectamente** instancias como pertenecientes a una clase
  - Valor entre 0 y 1
  - Precision = tp/(tp + fp)
  - Para la clase A: $Precision_A = tp_A / (tp_A + e_{BA} + e_{CA})$
- ***Recall (exhaustividad)***
  - También llamada en ocasiones *sensitivity*: capacidad para **no dejarse** instancias de una clase **sin identificar correctamente** como pertenecientes a la misma
  - Valor entre 0 y 1
  - *Recall = Sensitivity = tp / (tp+fn)*
  - Para la clase A: $Recall_A = tp_A / (tp_A + e_{AB} + e_{AC})$
- ***F-measure (valor-F)***
  - Media armónica del precision y recall
  - F-measure = F1 = (2 * precision * recall) / (precision + recall)

Universidad Zaragoza
1542

CLASIFICACIÓN

MÉTRICAS TÍPICAS

- *Accuracy (exactitud)*
  - Medida de la corrección global del modelo
  - Número de clasificaciones correctas / número total de clasificaciones realizadas

Compromiso: precision vs. Recall

Universidad Zaragoza
1542

CLASIFICACIÓN

MÉTRICAS TÍPICAS: CLASIFICACIÓN BINARIA

- *FPR*
  - *False positive rate = false alarm rate = fall-out*
  - FPR= FP/N = FP / (FP + TN)
- *TPR*
  - *True positive rate = sensitivity*
  - TPR= TP/P = TP / (TP + FN)
- *FNR*
  - *False negative rate*
  - FNR = FN / (TP + FN) = 1 – TPR
- *TNR*
  - *True negative rate = specificity*
  - TNR = SPC = TN / N = TN / (TN + FP) = 1 - FPR

Universidad Zaragoza
1542

CLASIFICACIÓN

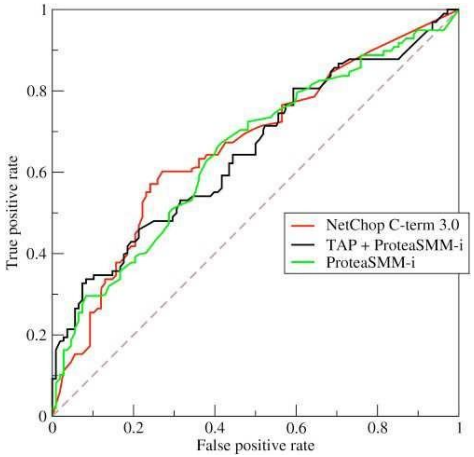
MÉTRICAS TÍPICAS: CLASIFICACIÓN BINARIA

- *Receiver Operator Characteristic (ROC) curve*, *AUC (Area Under the ROC Curve)*

Compromiso *true positives* - *false positives*

TPR = cuántos resultados positivos se detectan de entre todas las muestras positivas disponibles en el test

FPR = cuántos resultados positivos se detectan de forma incorrecta entre todas las muestras negativas disponibles en el test



Author: BOR at English Wikipedia Transferred from en.wikipedia to Commons https://en.wikipedia.org/wiki/File:ROC curves.png

Universidad Zaragoza
1542

EVALUACIÓN

IDEA ESENCIAL

Separar el conjunto de datos en:

- Conjunto de entrenamiento
- Conjunto de test (conjunto de validación)

Nunca hay que evaluar sobre el mismo conjunto de datos utilizado para entrenar (construir el modelo)

Capacidad de generalización vs. sobreajuste (*overfitting*)

EVALUACIÓN

EVALUACIÓN CRUZADA

Validación cruzada de k vías (*k-fold cross-validation*):

- Particionado de la muestra inicial en k muestras de igual tamaño
- 1 de las k muestras se utiliza como test y las k-1 restantes para entrenar
- Se repite k veces, de forma que cada una de las k muestras se utilizan 1 vez como test
- Se combinan los resultados de las k evaluaciones (promedios)
- Si k=n (número de muestras) => *leave-one out cross-validation*

Universidad Zaragoza
1542

INTRODUCCIÓN

CUESTIÓN

Supongamos que utilizamos el modelo de clasificación generado para predecir, tenemos un algoritmo “inteligente” si su tasa de acierto excede de:

- 50%
- 60%
- 80%
- 90%
- 100%
- No se puede determinar

Universidad Zaragoza
1542

INTRODUCCIÓN

CUESTIÓN

Supongamos que utilizamos el modelo de clasificación generado para predecir, tenemos un algoritmo “inteligente” si su tasa de acierto excede de:

- 50%
- 60%
- 80%
- 90%
- 100%
- No se puede determinar

- Por ejemplo, ¿cómo se comportaría un algoritmo trivial (ej., predecir la clase más popular, predecir de forma aleatoria)?

Universidad Zaragoza
1542

INTRODUCCIÓN

CUESTIÓN

- Modelo 1: identifica un 90% de instancias de la clase A y un 70% de instancias de la clase B
- Modelo 2: identifica un 20% de instancias de la clase A y un 100% de instancias de la clase B

¿Qué modelo es mejor?

- 1
- 2
- No se puede determinar

Universidad Zaragoza

INTRODUCCIÓN

CUESTIÓN

- Modelo 1: identifica un 90% de instancias de la clase A y un 70% de instancias de la clase B
- Modelo 2: identifica un 20% de instancias de la clase A y un 100% de instancias de la clase B

¿Qué modelo es mejor?

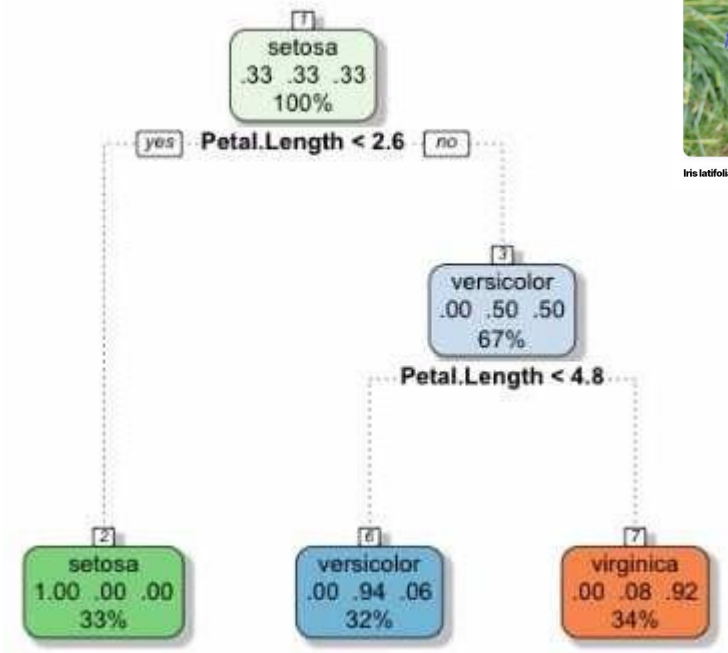
- 1
- 2
- No se puede determinar

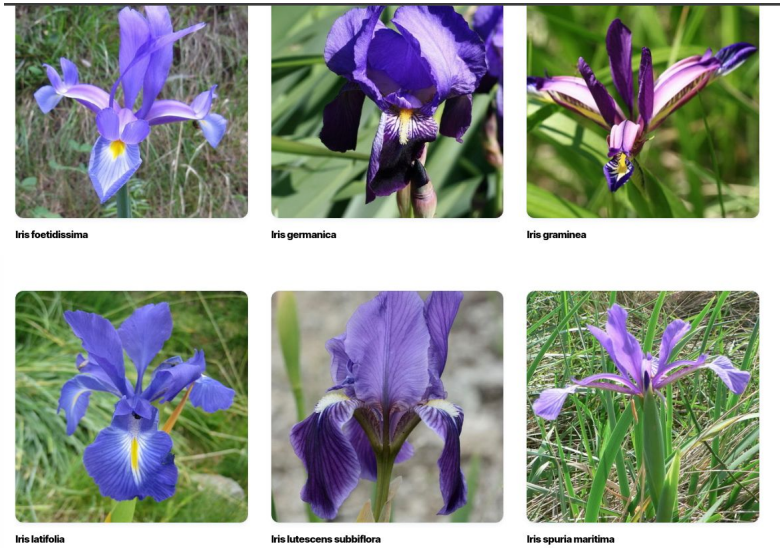
- Se necesita un modelo de costes y beneficios: puede que no todos los errores tengan la misma importancia (ej., falso positivo vs. falso negativo en medicina)

Universidad Zaragoza
1542

ÁRBOLES DE DECISIÓN

EJEMPLO (R)





Ejemplo generado en R con el conjunto de datos *Iris* y las bibliotecas *caret* y *rattle*

https://floressilvestresdearagon.es/lirios-2/

Universidad Zaragoza
1542

ÍNDICE

- Introducción
- Regresión
- Minería de patrones y reglas de asociación
- Agrupamiento (*clustering*)
- Clasificación
- Minería de textos

Universidad Zaragoza
1542

MINERÍA DE TEXTOS

INTRODUCCIÓN Y MOTIVACIÓN

- *Text mining* = *text analytics* = minería de datos textuales = minería de **datos no estructurados**
  - Proceso de derivar información de "alta calidad" a partir de fuentes de texto (no estructuradas o mínimamente estructuradas)
  - Estructurar la entrada (pre-procesamiento del texto) + minería de datos sobre los datos estructurados
  - Procesamiento del Lenguaje Natural (*Natural Language Processing –NLP–*)

Universidad Zaragoza
1542

MINERÍA DE TEXTOS

INTRODUCCIÓN Y MOTIVACIÓN

- Tareas típicas:
  - Clasificación de textos (**categorización**)
  - Agrupación de textos (extracción automática de **temas**)
  - Extracción de **información**
  - Análisis del **sentimiento** / minería de **opiniones**
  - Generación de **resúmenes**

Universidad Zaragoza
1542

MINERÍA DE TEXTOS

INTRODUCCIÓN Y MOTIVACIÓN

- Extracción de información:
  - **Reconocimiento** de entidades (*Named Entity Recognition –NER–)*
    - *Personas, fechas, localizaciones, organizaciones, etc.*
  - Extracción de **datos de interés**
  - Resolución de **correferencias**
    - Dos entidades reconocidas en el texto que hacen referencia a la misma entidad del mundo real
  - Extracción de **relaciones** entre entidades (ej., “vive en”)

MINERÍA DE TEXTOS

TÉCNICAS

- Ejemplos de tareas de pre-procesamiento de textos:
  - Reconocimiento de caracteres (*Optical Character Recognition* – *OCR*–)
  - Tokenización (tokens)
  - Eliminación de palabras 'poco interesantes' (comunes, conjunciones, preposiciones, …) -> *stopwords*
  - Lematización (canto, cantas, canta, cantamos, cantáis - niña, niño, niñita, niños, niñotes). Prefijos y sufijos usados comunmente.
  - *Stemming* (Universidad, Universidades, Universitario,...). Análisis morfológico.
  - Etiquetado gramatical (*part-of-speech tagging*)
  - Análisis sintáctico (*parsing*)
  - Desambiguación (*Word Sense Disambiguation*)

Universidad Zaragoza

MINERÍA DE TEXTOS

TÉCNICAS

- Stemming.

Form	Suffix	Stem
studies	-es	studi
studying	-ing	study
niñas	-as	niñ
niñez	-ez	niñ

- Lematización.

Form	Morphological information	Lemma
studies	Third person, singular number, present tense of the verb study	study
studying	Gerund of the verb study	study
niñas	Feminine gender, plural number of the noun niño	niño
niñez	Singular number of the noun niñez	niñez

https://www.bitext.com/blog/what-is-the-difference-between-stemming-and-lemmatization/

	D1	D2	D3	D4	D5	...
a	145	223	346	78	89	...
abandon	4	0	0	5	3	...
ability	5	10	0	4	7	...
able	31	35	64	3	5	...
about	64	68	89	24	9	...
above	4	5	8	0	0	...
abroad	0	0	1	0	0	...
absence	2	4	0	0	0	...
absent	0	0	1	0	0	...
absolute	3	1	5	0	1	...
abstract	5	1	2	1	0	...
abuse	0	1	0	0	0	...
academic	1	3	0	0	0	...
...	...	...	...	...	...	...

MINERÍA DE TEXTOS

TÉCNICAS

- Representación vectorial de documentos:
  - Los documentos se representan en un espacio vectorial multi-dimensional => bolsas de palabras (bags of words)
  - Los términos son las dimensiones del espacio
  - Los documentos son puntos o vectores en este espacio
  - El valor de cada componente del vector se determina a partir de la frecuencia del término en el documento y su frecuencia inversa
  - La similitud entre dos documentos puede calcularse midiendo el ángulo formado por sus vectores
  - Para mejorar el rendimiento, podrían seleccionarse únicamente las palabras más frecuentes

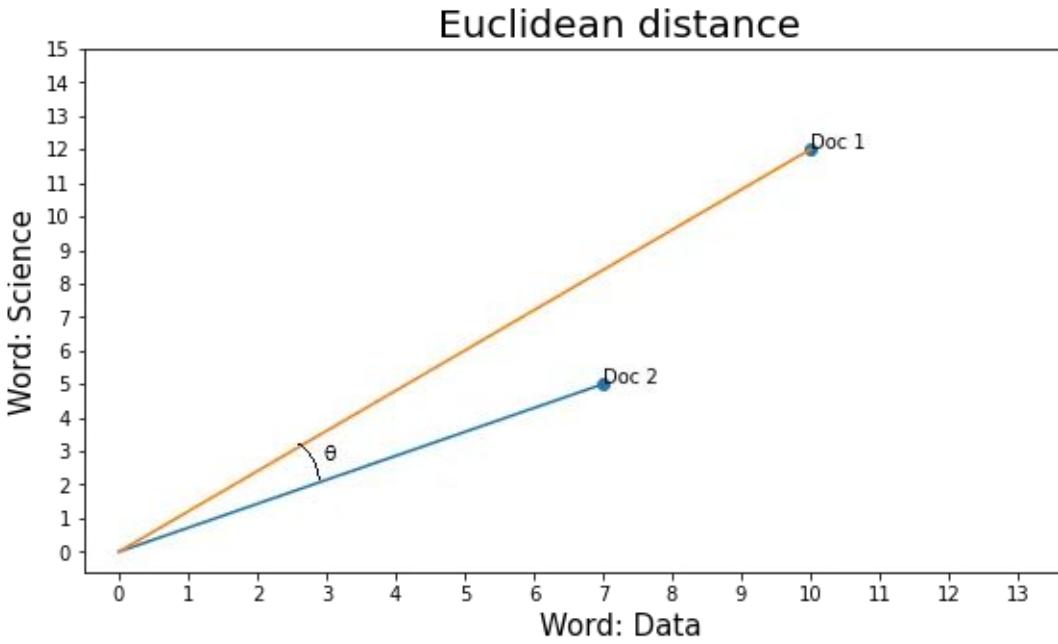
$$IDF = log\frac{N}{DF_t}$$

$IDF = \log \frac{N}{Dt}$

Universidad Zaragoza

MINERÍA DE TEXTOS

	D1	D2	D3	D4	D5	...
a	145	223	346	78	89	...
abandon	4	0	0	5	3	...
ability	5	10	0	4	7	...
able	31	35	64	3	5	...
about	64	68	89	24	9	...
above	4	5	8	0	0	...
abroad	0	0	1	0	0	...
absence	2	4	0	0	0	...
absent	0	0	1	0	0	...
absolute	3	1	5	0	1	...
abstract	5	1	2	1	0	...
abuse	0	1	0	0	0	...
academic	1	3	0	0	0	...
...	...	...	...	...	...	...



Universidad Zaragoza

MINERÍA DE TEXTOS

HERRAMIENTAS

- Freeling (http://nlp.lsi.upc.edu/freeling/)
- OpenNLP (https://opennlp.apache.org/)
- Gate (https://gate.ac.uk/)
- Stanford NLP Group's Software (http://nlp.stanford.edu/software/index.shtml)
- R (https://www.r-project.org/) con paquetes:
  - *tm: Text Mining Package* (https://cran.r-project.org/web/packages/tm/index.html)
  - *text2vec: Modern Text Mining Framework for R* (https://cran.r-project.org/web/packages/text2vec/)
- RapidMiner (https://rapidminer.com/)
- Weka (http://www.cs.waikato.ac.nz/ml/weka/)