PRIMERA PARTE

CONCEPTOS y DEFINICIONES

Definición

“Data Science se refiere a un conjunto de algoritmos matemático/computacionales que permiten extraer información no trivial y potencialmente útil que reside implícitamente en los datos analizados.”

Data Science

Definición

“Data Science se refiere a un conjunto de algoritmos matemático/computacioneles que permiten extraer información no trivial y potencialmente útil que reside implícitamente en los datos analizados.”

“Machine Learning el área de estudio que le otorga a las computadoras la habilidad de aprender sin ser explícitamente programadas. “

Definición

“Data Science se refiere a un conjunto de algoritmos matemático/computacioneles que permiten extraer información no trivial y potencialmente útil que reside implícitamente en los datos analizados.”

“Machine Learning el área de estudio que le otorga a las computadoras la habilidad de aprender sin ser explícitamente programadas.”

La definición más moderna y descriptiva podría ser la siguiente:

“Un programa computacional se dice que ha aprendido de una experiencia E respecto a una tarea T y una medida de performance P, si su performance en T, medida por P, mejora con la experiencia E.”

Data Science

Definición

“Data Science se refiere a un conjunto de algoritmos matemático/computacioneles que permiten extraer información no trivial y potencialmente útil que reside implícitamente en los datos analizados.”

“Machine Learning el área de estudio que le otorga a las computadoras la habilidad de aprender sin ser explícitamente programadas.”

La definición más moderna y descriptiva podría ser la siguiente:

“Un programa computacional se dice que ha aprendido de una experiencia E respecto a una tarea T y una medida de performance P, si su performance en T, medida por P, mejora con la experiencia E.”

Términos Relacionados

Data Science
Data Analytics
Machine Learning
Knowledge Discovery
Big Data
Business Intelligence

Como usamos y nos afecta la Minería de Datos frecuentemente

-Cada vez que usamos un buscador como Google o Yahoo, decenas de algorítmos corren detrás del mismo para encontrar los mejores resultados

-Cuando configuramos un anti-spam en nuestra casillas de mail

-Cuando buscamos recomendaciones de libros, música o películas por Internet.

-Cuando escaneamos un texto con tecnología OCR

-Cuando solicitamos un crédito bancario o contratamos un seguro

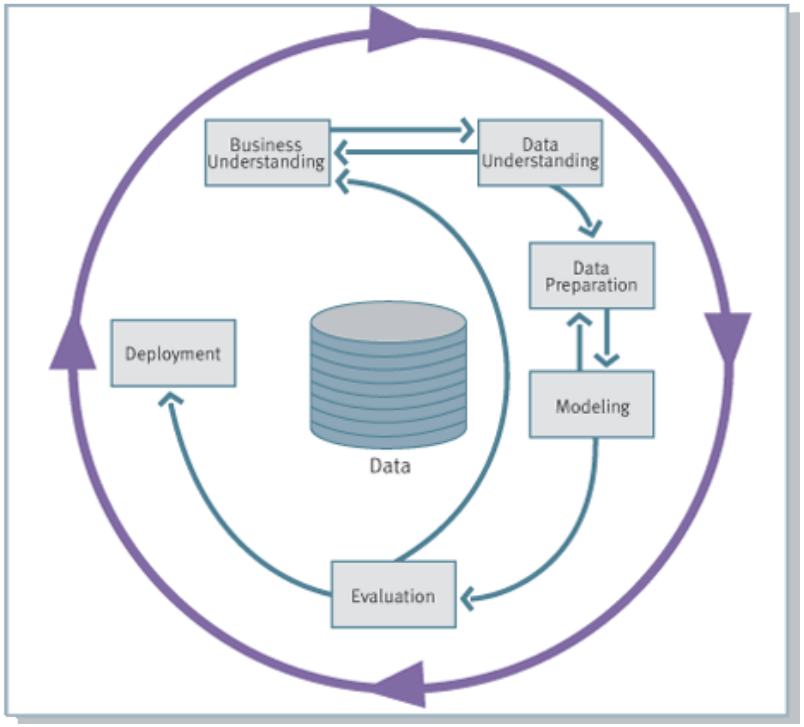
-Cuando empresas de marketing intentan contactarnos una y otra vez

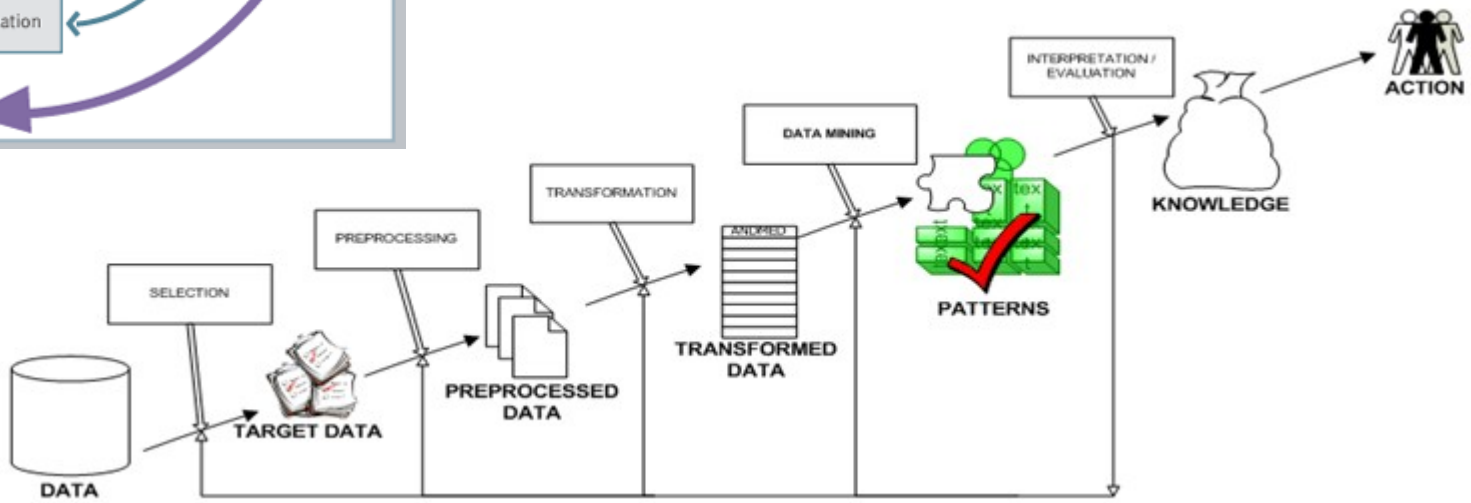
-Cuando empresas de salud monitorean nuestros comportamientos

-Cuando empresas gubernamentales o de seguridad también monitorean nuestros patrones y comportamiento.

-Cuando nos realizamos modernos estudios médicos de diagnóstico.

Modelo CRISP-DM
Cross Industry Standard Process for Data Mining





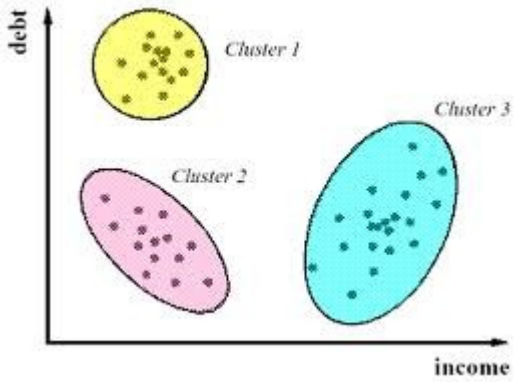
Clasificación de Modelos I

-**SUPERVISADOS:**
Existe una variable TARGET a predecir.
Conocemos el resultado de donde queremos llegar, pero no el camino de cómo conseguirlo Ej: Modelos de OCR en scanners.

-**NO-SUPERVISADOS**:
No existe ninguna variable a predecir.
Conocemos el camino y los datos históricos pero no conocemos el resultado futuro.
Ej: predicción de la demanda, predicción de la Bolsa de Comercio.

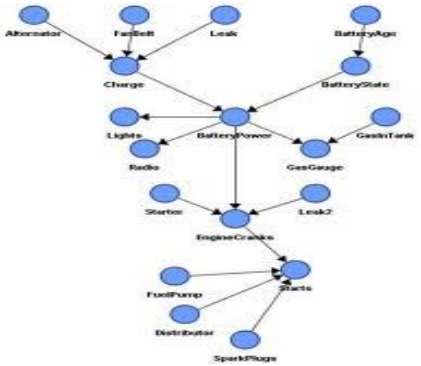
-**SEMI-SUPERVISADOS:**
Muchas veces se conoce la etiqueta (valor que toma la variable TARGET a predecir) de algunos casos en el set de datos. Pero no de otros.
Generalmente MUCHOS casos sin etiquetar, y POCOS casos con etiquetas.
De esta manera es posible clasificar aquellos casos sin etiquetas en base a aquellos pocos que se encuentran etiquetados.

Clasificación de Modelos II



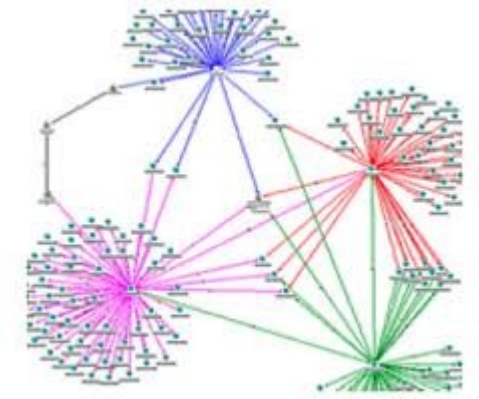
Cluster Analysis (Descriptive)

Segmenta elementos que son “similares en algún sentido
Se aplica ampliamente en Marketing y Empresas, que desean segmentar el comportamiento de sus clientes y productos en el mercado y realizar marketing directo y personalizado.



Classification y Regression (Predictive Analytics)

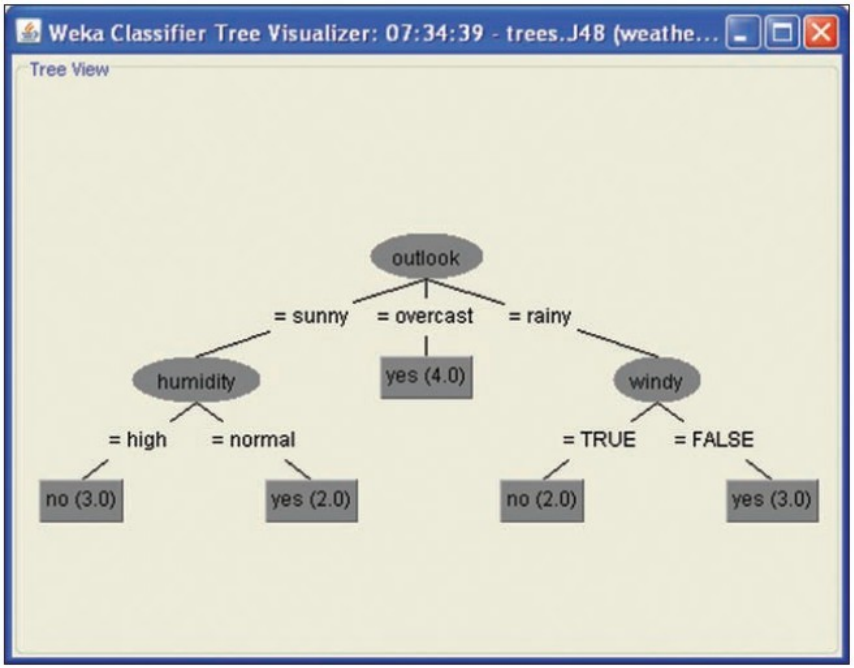
Es la tarea de generalizar una estructura de datos ya conocida, para luego aplicar nueva data a ser clasificada.
Todos los modelos y software de anti-spam conocidos aplican estas técnicas. Un mail se clasifica como SPAM o NO-SPAM



Association

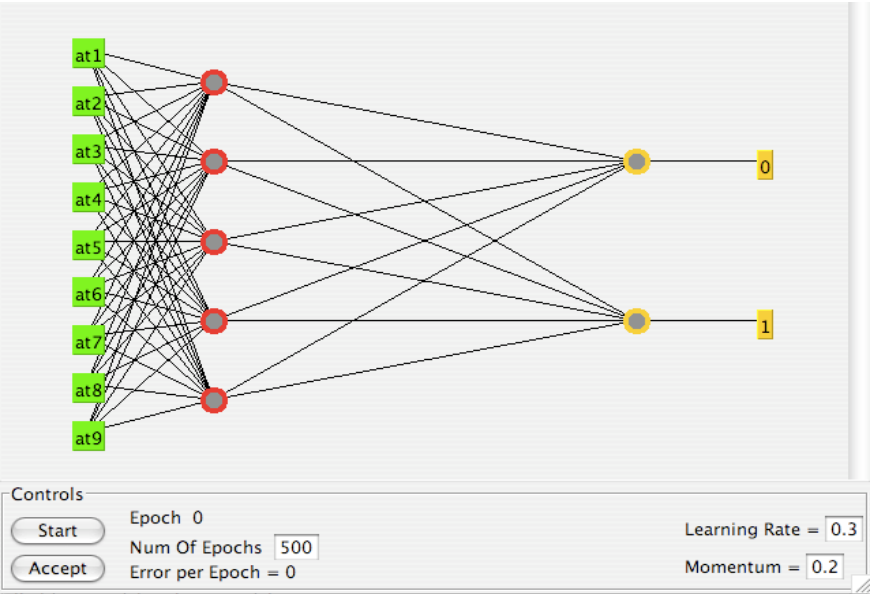
Método para descubrir nuevas relaciones entre variables de una gran base de datos.
Ampliamente utilizada en supermercados por ejemplo y en publicidad. Distribución de productos.

Algoritmos I



Árboles de Decisión

Es un modelo de predicción donde se construyen diagramas lógicos en formas de hojas, leyéndose de arriba hacia abajo, hasta llegar a una decisión.
En el ámbito empresarial, se puede decir que son diagramas de decisiones secuenciales que muestran posibles resultados.



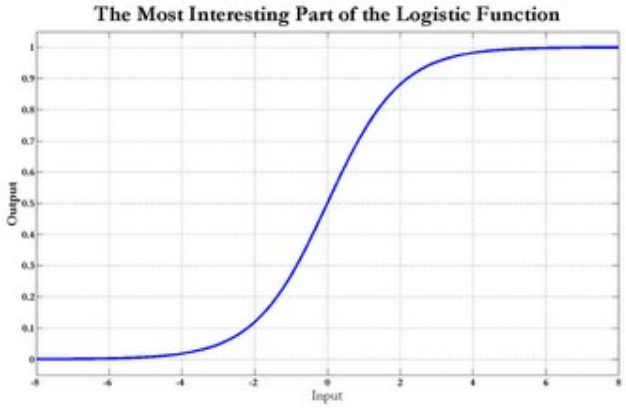
Redes Neuronales

Simulan en funcionamiento del cerebro humano.
Pueden encontrar patrones y segmentar base de datos.
Ejemplos conocidos: Modelos de OCR en scanners, detección de anomalías.

Algoritmos I

Inferencia Bayesiana

Generan modelos ampliamente utilizados en la Teoría de la Decisión.
En base a evidencias (nuestros datos) intenta predecir de que una hipótesis sea cierta.
Por ejemplo si alguien va a morir de un determinado tumor, si un cliente va a comprar un producto o no.
Se utiliza el Teorema de Bayes.



Regresión Logística

Ampliamente utilizada en Modelos de Scoring, específicamente en análisis de riesgos, sector crediticio y en Argentina lo utiliza el VERAZ.
Genera modelos de predicción basados en la probabilidad que suceda un evento, ajustando los datos a la función logística.

Que buscamos con el Análisis de Datos Predictivo? (Predictive Analytics)

a) Predecir con el máximo nivel de certeza casos no vistos previamente.

Ejemplos:
- Mañana las acciones de YPF van a subir o bajar
- Un medicamento a “Juan” le puede causar alergia o no
- Un cliente va a comprar un producto X o no

b) Entender como las variables predictores

Por ejemplos:
- Correlación entre el nivel de estudios alcanzado y el salario ganado
- Perfiles de clientes según región geográfica, rangos de edad que compran determinado producto.

La empresa no necesita un gran nivel de profundidad de entendimiento.

a) ← PREDICCION

b) ← INFERENCIA

Aspectos Importantes

- Es importante conocer cada uno de los algoritmos, y la naturaleza y estructura de los datos para saber que herramienta usar dependiendo del problema a resolver.

- Siempre es importante conocer los algoritmos mas simples y la manera de tratar los datos para ajustarlos al modelo menos complejo. A mayor complejidad, menos capacidad de interpretar el modelo existe. Muchas veces es tan importante la interpretación como el resultado.

- Es muy importante tener buenos indicadores y protocolos para la medición de performance. De no ser así, no es posible medir la certeza en las predicciones, lo cual significaría que un modelo no tiene ningún tipo de confiabilidad.
Por otro lado, sin buenos indicadores de performance, no es posible determinar que tanto es capaz de aprender un algoritmo.

PERFIL

DATA SCIENTIST

Data Science

Telecom Personal busca Data Scientist

Lugar de Trabajo: Córdoba o CABA

Responsabilidad:
Desarrollará sus actividades en tandem con un profesional de matemática, estadística y/o actuaría proveniente de Marketing. De manera conjunta construirán soluciones específicas para uso interno y externo a la organización
Construirá aplicaciones basadas en analytics de grandes volumenes de información, ya sea estructurada como no estructurada, interna y externa a la compañía
Abordará la resolución de problemas complejos de naturaleza diversas (sales, loyalty, eficiencia, monetización)

Formación Requerida:
Sólido background en matemática, estadística y/o actuaría
Curioso, autodidacta, decidido. Apasionado por la resolución de problemas complejos

Data Science

Competencias en:

Diseño y Construcción de APIs
Web Scraping
Data Extraction / Transformation
Data Analysis / Data Mining
Data Visualization
Predictive Analysis
Pattern recongnition
Machine Learning
Natural language processing
High performance computing

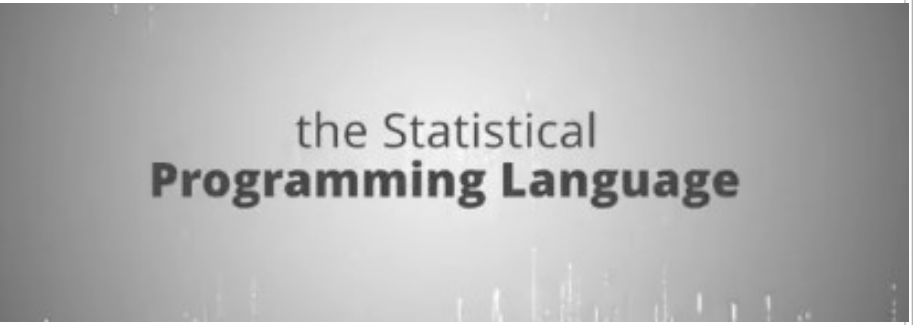
Conocimientos Específicos / Experiencia Laboral:

Python (Mandatoria)
Hadoop/Map Reduce/Spark/Pig/MongoDB/Cassandra (Mandatoria)
SAS o SPSS o R (Mandatoria)
Perl
Ruby
Java
C / C++
Scala
Django

SEGUNDA PARTE

R

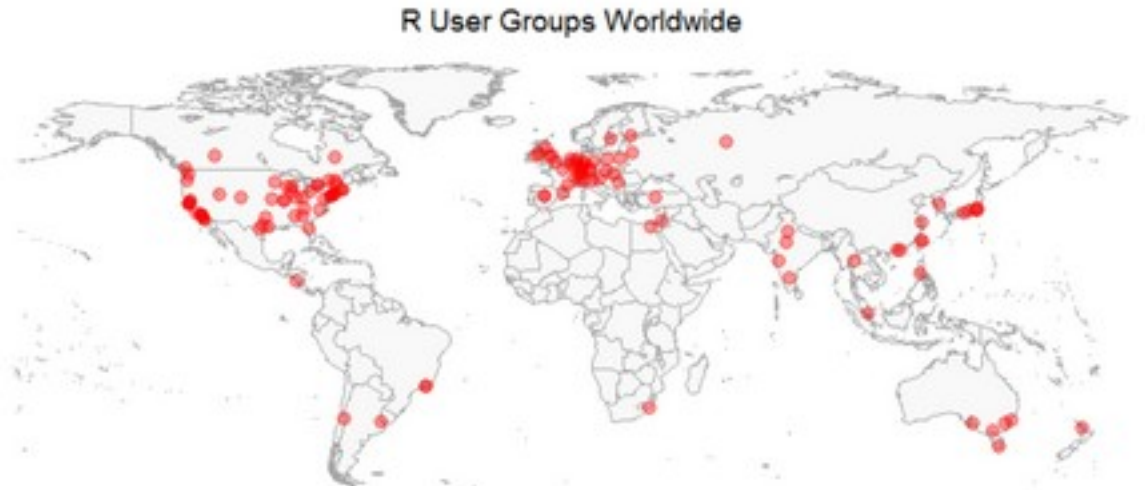
Es el lenguaje de Programación Estadística más utilizado del mundo.



Ampliamente aceptado en el área académica, científica y en sectores críticos de la industria y empresa.

What programming/statistics languages you used for an analytics / data mining / data science work in 2013? [713 votes total]

	% users in 2013	% users in 2012	% users in 2011
R (434 voters in 2013)	60.9%	52.5%	45.1%
Python (277)	38.8%	36.1%	24.6%
SQL (261)	36.6%	32.1%	32.3%
SAS (148)	20.8%	19.7%	21.2%
Java (118)	16.5%	21.2%	24.4%
MATLAB (89)	12.5%	13.1%	14.6%
High-level data mining suite (80)	11.2%	not asked in 2012	
Unix shell/awk/sed (79)	11.1%	14.7%	
C/C++ (66)	9.3%	14.3%	
Pig Latin/Hive/other Hadoop-based languages (57)	8.0%	6.7%	



Available Packages

Currently, the CRAN package repository features 5008 available packages.

Table of available packages, sorted by date of publication

Quiénes usan R ?

facebook Google

The New York Times NOVARTIS

citi LLOYD'S VISA

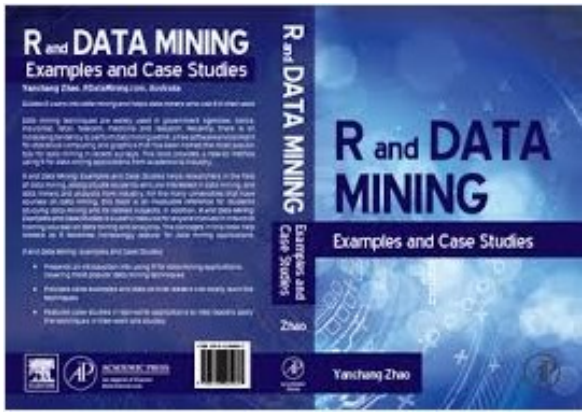
Software with R connections

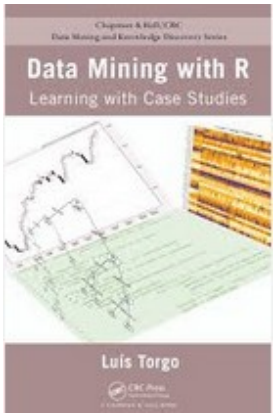
SPSS AN IBM® COMPANY sas

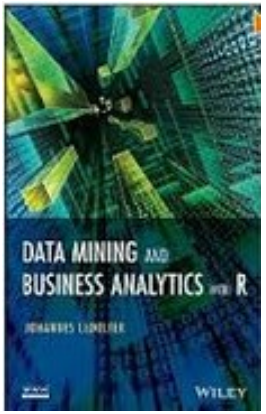
Spotfire ORACLE

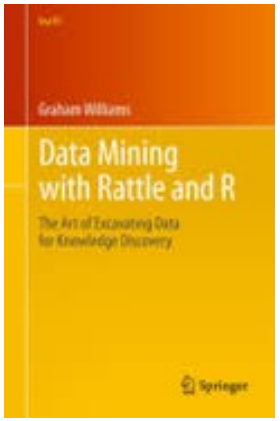
Data Mining con R

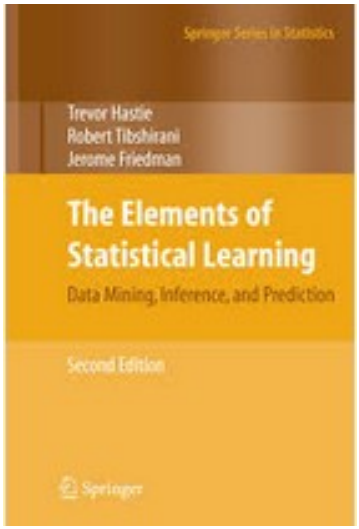
Amplia Documentación:

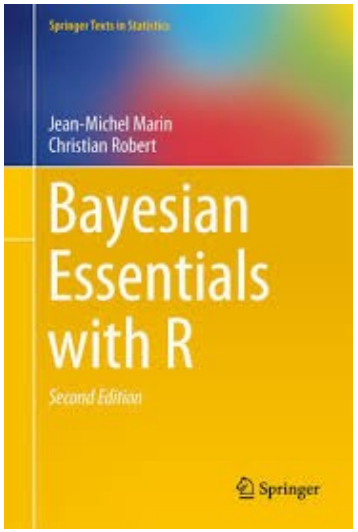


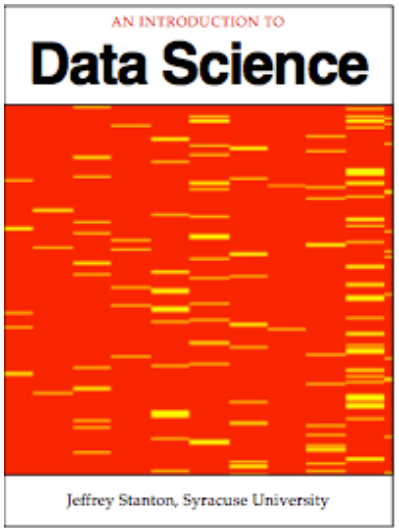


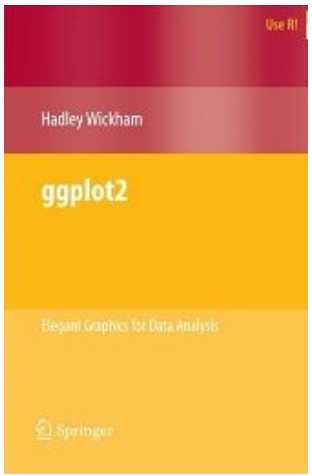












Data Mining con R

Tareas de Data Mining que se pueden realizar con R:

Reducción de Dimensionalidad
Pre-procesamiento de datos
Modelos Predictivos (Clasificación y Regresión)
Modelos Descriptivos (Clustering)
Visualización de Datos

Algoritmos para Data Mining R:

Árboles de Decisión: C4.5
Support Vector Machines
Naive Bayes
Neural Networks
Bagging and Busting
K-Means
A Priori
Regresión Logística

Algunos paquetes importantes:

CARET: Classification and Regression Training
CART: Classification and Regression Trees
e1071: Contiene el algoritmo de Naive Bayes y SVM.
Neuralnet: Contiene un Perceptron Multilayer
Ggplot2: Potente lenguaje de gráficos

TERCERA PARTE

Regresión Lineal con R

Regresión Lineal con R

En la Regresión Lineal exploramos la relación entre dos variables.
Ejemplo, set de datos:

Para este ejemplo, tomaremos las medidas en centímetros de la longitud (o el perímetro) de diversas circunferencias y las medidas de los diámetros que le corresponden a cada una de estas circunferencias

Diámetro en cm (x)	Longitud en cm (y)
2.10	6.50
5.50	17.10
4.00	12.50
3.80	12.00
6.00	18.90
3.50	11.00
4.60	14.40

Regresión Lineal con R

Defino las variables “diametro” y “longitud”

```
> diametro <- c(2.10,5.50,4.00,3.80,6.00,3.50,4.60)
> longitud <- c(6.50,17.10,12.50,12.00,18.90,11.00,14.40)
```

Grafico la distribución de puntos, llamando a la función “plot” para hacer un análisis
visual previo al modelo

```
> plot(longitud,diametro)
```

Deberia ser plot(diametro,longitud)

Creo el modelo de regresión
Para ello defino un objeto “resultado”, que es justamente donde se guardará el
resultado de la regresión lineal:

```
> resultado <- lm(longitud~diametro)
```

lm es la función que invoca el algoritmo de Regresión Lineal
Determino la variable dependiente e independiente
El símbolo “~” nos permite relacionar las variables en cuestión
Muestro el resultado por pantalla, invocando a la variable que guardó nuestro resultado:

```
> resultado
```

Regresión Lineal con R

```
Call:
lm(formula = longitud ~ diametro)

Coefficients:
(Intercept)     diametro
   -0.05266      3.14470
```

De aquí se desprende el formato de la fórmula y = mx + b

longitud = 3.14470 * diametro - 0.05266

Podemos graficar ahora la “línea de mejor ajuste”, llamando a la función “abline”

```
> plot(longitud~diametro)
> abline(resultado)
```

Predicción

Si ahora queremos hacer una predicción, por ejemplo, estimar la longitud que tendrá una circunferencia de 7.5 cm de diámetro:

longitud = 3.14470 * diametro - 0.05266

Longitud = 23.5326

Regresión Lineal con R

1. Realizar una Regresión Lineal con R, con los siguientes datos, los cuales representan la edad en meses de determinados niños, y su promedio de altura dependiendo de la edad que tienen:

Edad (en meses)	Promedio de Altura (en centímetros)
18	76.1
19	77
20	78.1
21	78.2
22	78.8
23	79.7
24	79.9
25	81.1
26	81.2
27	81.8
28	82.8
29	83.5

2. Estimar el promedio de altura de un niño con 27.5 meses de edad