

Índice general

Prefacio	5
----------------	---

Capítulo 1

Introducción	13
1.1 Introducción.....	13
1.2 Los datos.....	19
1.3 Etapas en los procesos de <i>big data</i>	20
1.4 Minería de datos.....	21
1.5 Estructura de un proyecto de análisis de datos..	22
1.6 Aplicaciones	25
1.6.1 <i>Marketing</i>	25
1.6.2 Compañías de seguros.....	26
1.6.3 Banca	26
1.6.4 Telecomunicaciones	27
1.6.5 Medicina	27
1.6.6 Industria farmacéutica.....	27
1.6.7 Biología	28
1.6.8 Minería de textos	28
1.6.9 Minería de datos web	29
1.6.10 Redes sociales.....	30
1.7 Modelos y tareas	31
1.7.1 Tareas descriptivas	32
1.7.1.1 Agrupamiento	32
1.7.1.2 Correlaciones y factorizaciones	32
1.7.1.3 Reglas de asociación.....	33
1.7.1.4 Dependencias funcionales.....	33
1.7.2 Tareas predictivas.....	34
1.7.2.1 Clasificación.....	34
1.7.2.2 Clasificación suave	35
1.7.2.3 Categorización	35
1.7.2.4 Preferencias o priorización.....	35
1.7.2.5 Regresión	36
1.8 Métodos y técnicas.....	36
1.8.1 Técnicas algebraicas y estadísticas.....	36
1.8.2 Técnicas bayesianas	37
1.8.3 Técnicas basadas en conteos de frecuencias y tablas de contingencia	37

1.8.4 Técnicas basadas en árboles de decisión y sistemas de aprendizaje de reglas.....	37
1.8.5 Técnicas relacionales, declarativas y estructurales.....	37
1.8.6 Técnicas basadas en redes neuronales artificiales	37
1.8.7 Técnicas basadas en núcleo y máquinas de soporte vectorial	38
1.8.8 Técnicas estocásticas y difusas	38
1.8.9 Técnicas basadas en casos, en densidad o distancia	38

Capítulo 2

Análisis estadístico de datos.....39

2.1 Introducción.....	39
2.2 Análisis de una variable. Estadística descriptiva e inferencia.....	40
2.2.1 Estadísticos de variable continua.....	41
2.2.2 Histograma.....	42
2.2.3 Estadísticos de variables nominales	44
2.3 Contrastes de hipótesis	46
2.3.1 Distribuciones de probabilidad.....	46
2.3.1.1 Distribución normal	47
2.3.2 Inferencia	49
2.3.3 Evaluación de hipótesis.....	51
2.4 Análisis de relaciones entre variables. Evaluación de hipótesis.....	54
2.4.1 Relación entre variables nominales-nominales.....	54
2.4.2 Relaciones numéricas-nominales	56
2.4.2.1 Comparación de dos medias.....	56
2.4.2.2 Análisis de la varianza	58
2.4.3 Relaciones numéricas-numéricas.....	61

Capítulo 3

Introducción al lenguaje R. Lectura, procesamiento y visualización de datos:

***data wrangling*.....63**

3.1 Carga y transformaciones de datos.....	63
3.1.1 Estructura básica de datos	64
3.1.2 Lectura de fichero.....	65

3.2 Estadística descriptiva.....	68
3.2.1 Variables categóricas.....	84
3.2.2 Correlación.....	88
3.2.2.1 Visualización.....	88
3.2.3 Test de hipótesis.....	97
3.2.4 Representación de datos.....	100

Capítulo 4

Predicción y clasificación con técnicas

numéricas.....117

4.1 Técnicas numéricas de predicción.....	117
4.1.1 Regresión lineal.....	117
4.1.1.1 Regresión lineal simple.....	118
4.1.1.2 Regresión lineal múltiple.....	119
4.1.1.3 Regresión lineal ponderada localmente.....	121
4.1.1.4 Atributos nominales.....	123
4.1.2 Evaluación del modelo de regresión.....	124
4.1.2.1 Error de regresión y selección de variables.....	126
4.1.3 Regresión no lineal.....	130
4.1.3.1 Transformaciones sencillas.....	131
4.1.3.2 Otras transformaciones.....	133
4.1.4 Ejemplos de regresión lineal.....	133
4.2 Técnicas numéricas de clasificación.....	136
4.2.1 Clasificación mediante regresión lineal.....	138
4.2.2 Clasificación mediante regresión logística.....	139
4.2.3 Clasificación bayesiana.....	141
4.2.3.1 Clasificación bayesiana de atributos numéricos.....	141
4.2.3.2 Clasificación bayesiana con atributos nominales.....	146
4.2.4 Ejemplos de clasificación bayesiana.....	147

Capítulo 5

Predicción y clasificación con R153

5.1 Regresión.....	153
5.1.1 Regresión lineal.....	153
5.1.2 Selección de atributos.....	165
5.1.3 Regresión no lineal.....	169
5.1.4 Regresión de atributos no continuos.....	172
5.1.5 Modelos lineales generalizados.....	179

5.2 Algoritmos de clasificación.....	185
5.2.1 Detección de valores atípicos.....	186
5.2.2 LDA, <i>Linear Discriminant Analysis</i>	195
5.2.3 Clasificadores probabilísticos.....	200
5.2.3.1 <i>Naive</i> bayesiano.....	201
5.2.3.2 Redes bayesianas	202

Capítulo 6

Técnicas de minería de datos.....	205
6.1 Técnicas de minería de datos	205
6.2 <i>Clustering</i>	207
6.2.1 <i>Clustering</i> numérico (k-medias).....	209
6.2.2 <i>Clustering</i> conceptual (COBWEB)	210
6.2.3 <i>Clustering</i> probabilístico (EM)	214
6.3 Reglas de asociación	217
6.4 Predicción numérica	220
6.4.1 Predicción no lineal con árboles de regresión.....	220
6.4.2 Estimador de núcleos.....	225
6.4.2.1 Aplicación a problemas multivariantes.....	228
6.4.2.2 Aplicación a problemas de clasificación.....	229
6.5 Clasificación.....	231
6.5.1 Tabla de decisión	231
6.5.2 Árboles de decisión.....	233
6.5.3 Reglas de clasificación	245
6.5.4 Clasificación bayesiana.....	251
6.5.5 Aprendizaje basado en ejemplares	257
6.5.5.1 Algoritmo de los k-vecinos más próximos.....	258
6.5.5.2 Algoritmo k-estrella.....	260
6.5.5.3 Probabilidad de transformación para los atributos permitidos.....	261
6.5.5.4 Combinación de atributos	262
6.5.5.5 Selección de los parámetros aleatorios.....	262
6.5.5.6 Clasificación de un ejemplo	264
6.5.6 Máquinas de vectores de soporte (SVM)	265
6.5.6.1 SVM lineal.....	266
6.5.6.2 SVM lineal de margen blando (<i>soft margin</i>)	270
6.5.6.3 SVM no lineal. Funciones <i>kernel</i>	272
6.5.6.4 Clasificación multiclase	276
6.5.7 Redes de neuronas	277
6.5.7.1 Estructura de las redes de neuronas	278
6.5.7.2 Proceso de entrenamiento (retropropagación).....	279
6.5.8 Lógica borrosa (<i>fuzzy logic</i>)	281
6.5.9 Técnicas genéticas: algoritmos genéticos (<i>genetic algorithms</i>)	282

Capítulo 7

Técnicas de minería de datos en R.....	285
7.1 Agrupamiento. <i>Clustering</i>	285
7.1.1 Agrupamiento jerárquico.....	286
7.1.2 Número óptimo de agrupaciones.....	289
7.1.3 Agrupamiento por particionamiento.....	299
7.1.4 Agrupamiento basado en modelos.....	305
7.1.5 Agrupamiento borroso (<i>fuzzy</i>).....	312
7.1.6 Otras técnicas de agrupamiento.....	314
7.1.7 Representación y análisis de las clases.....	323
7.1.8 Validación de resultados.....	326
7.2 Clasificación.....	329
7.2.1 Selección de atributos.....	329
7.2.2 Reducción de la dimensionalidad.....	340
7.2.3 Árboles de decisión.....	347
7.2.3.1 RPART (<i>Recursive Partitioning and Regression Trees</i>).....	347
7.2.3.2 Árboles de inferencia condicional, CTREE.....	350
7.2.3.3 C5.0.....	351
7.2.4 Metaalgoritmos.....	354
7.2.4.1 AdaBoost (<i>ADAPtative BOOSTing</i>).....	355
7.2.4.2 GBM (<i>Gradient Boosting Machine</i>).....	356
7.2.4.3 <i>Random forest</i>	358
7.2.5 SVM, máquinas de vectores de soporte.....	359
7.2.6 K vecinos próximos. k-NN (<i>k-Nearest Neighbors</i>).....	361
7.2.7 Redes de neuronas artificiales.....	363

Capítulo 8

Internet de las cosas y análisis de series temporales.....	369
8.1 Internet de las cosas.....	369
8.2 Thingier.io IoT.....	371
8.2.1 <i>Hardware</i>	372
8.2.2 Configuración de la plataforma.....	374
8.2.3 <i>Software</i> del dispositivo.....	376
8.2.4 Visualización y exportación de la información.....	379
8.3 Series temporales.....	381
8.3.1 Predicción con series temporales.....	382
8.3.1.1 Predicción lineal (autorregresión).....	382
8.3.1.2 Error de predicción.....	384
8.3.1.3 Predicción no lineal.....	385
8.3.2 Análisis y descomposición de series.....	386
8.3.2.1 Tendencia y estacionariedad.....	386
8.3.2.2 Modelos ARMA/ARIMA.....	389

8.4 Análisis de series con R.....	390
8.4.1 Componentes de la serie temporal.....	392
8.4.2 Modelos de predicción.....	400
8.4.3 Detección de anomalías.....	408

Capítulo 9

Análisis de datos espaciales.....411

9.1 Introducción.....	411
9.1.1 Datos de tipo espacial.....	412
9.1.2 Latitud, longitud.....	413
9.1.3 La clase de datos Spatial en RStudio.....	414
9.1.4 Datos.....	414
9.2 Tipos de datos.....	416
9.2.1 Creación de objetos SpatialPoints.....	416
9.2.2 Creación de objetos SpatialGrid.....	420
9.3 Visualización de datos espaciales.....	421
9.4 Análisis estadístico (interpolación).....	423
9.4.1 Análisis exploratorio de datos.....	424
9.4.1.1 Crear <i>grid</i> Ávila-Madrid.....	424
9.4.2 Interpolación IDW (<i>Inverse Distance Weighted</i>).....	427
9.4.2.1 Presentación de resultados IDW.....	428
9.4.3 Correlación espacial (variograma).....	428
9.4.3.1 Selección modelo de variograma.....	428
9.4.3.2 Presentación de resultados.....	430

Bibliografía.....431