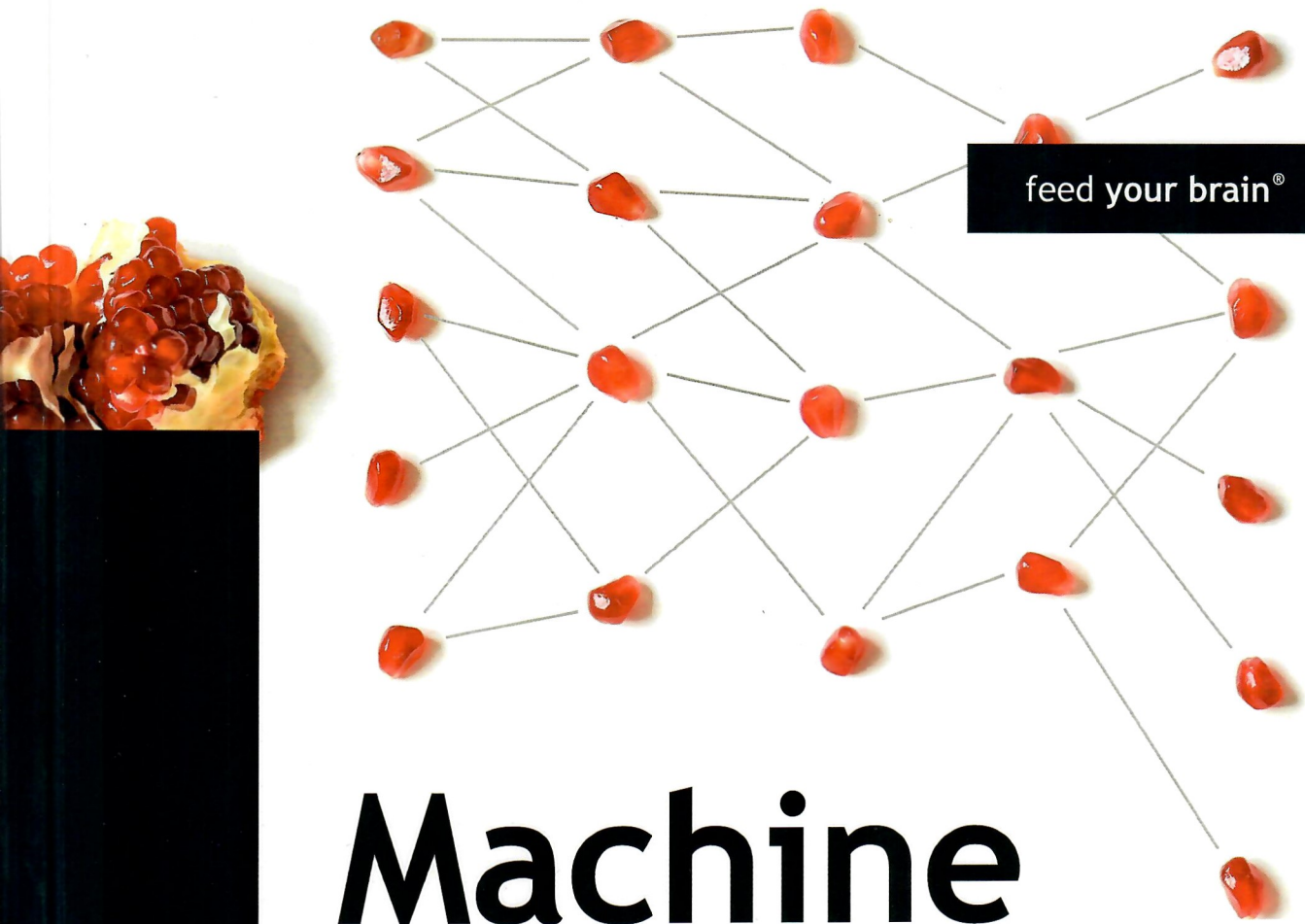

feed your brain®

Machine Learning y Ciencia de Datos

con Python y R

Francisco Charte
David Charte

krasis PRESS

MACHINE LEARNING Y CIENCIA DE DATOS CON PYTHON Y R

Krasis Consulting, S. L.U.
www.Krasis.com

ISBN: 978-84-945822-5-7
Depósito Legal: VG 52-2021

Impreso en España-Printed in Spain

Contenido

AGRADECIMIENTOS III

CONTENIDO V

AUTORES XIX

CAPÍTULO 1: CONCEPTOS FUNDAMENTALES 21

1.- Introducción 21
1.1.- Estrategias de aprendizaje 21
1.2.- Definición de aprendizaje automático 22
1.2.1.- Ejemplos 22
1.3.- Ideas relacionadas 23
1.3.1.- Inteligencia artificial 23
1.3.2.- Estadística 23
1.3.3.- Métodos numéricos 24
1.3.4.- Métodos computacionales 24
1.3.5.- Comparación de los distintos métodos 24
2.- Paradigmas de aprendizaje y tipologías de problemas 25
2.1.- Paradigmas de aprendizaje 25
2.2.- Tipos de problemas 26
2.2.1.- Tareas de aprendizaje supervisado 26
2.2.2.- Tareas de aprendizaje no supervisado 27
2.2.3.- Otros tipos de tareas de aprendizaje 28
3.- Ejemplos de casos prácticos de aprendizaje automático 29
3.1.- Bases de datos 29
3.2.- El conjunto de datos iris 30
3.3.- El conjunto de datos titanic 32
3.4.- Aprendizaje supervisado 33
3.4.1.- Clasificación de iris con árboles de decisión 33
3.4.2.- Clasificación de titanic con los vecinos más cercanos 35
3.5.- Aprendizaje no supervisado 36
3.5.1.- Agrupamiento 37
3.5.2.- Reglas de asociación 39
4.- Herramientas y lenguajes 41
4.1.- Python y Jupyter 41
4.1.1.- Instalación en Windows 43
4.1.2.- Instalación en macOS 44
4.1.3.- Instalación en Linux 44
4.2.- R y RStudio 45
4.2.1.- Instalación de R 46

4.2.2.- Instalación de RStudio 47
4.3.- Uso de las herramientas mediante Docker 48
5.- El entorno de trabajo para R 49
5.1.- Inicio de RStudio 49
5.2.- Paneles en el entorno de trabajo 51
5.3.- Edición de guiones R 52
5.4.- Trabajo con *notebooks* 54
5.5.- Cómo realizar algunas tareas comunes 55
5.5.1.- Acceso a la documentación integrada 55
5.5.2.- Establecer la ruta de trabajo 56
5.5.3.- Guardar y recuperar la sesión de trabajo 58
6.- Introducción al lenguaje R 58
6.1.- El lenguaje R 59
6.2.- Operar con valores simples 60
6.2.1.- Tipos de datos básicos en R 61
6.2.2.- Operadores básicos 61
6.3.- Tipos de datos complejos 63
6.3.1.- Vectores 63
6.3.2.- Matrices 64
6.3.3.- Listas 65
6.3.4.- Conjuntos de datos 67
6.4.- Estructuras condicionales y de repetición 68
6.4.1.- Condicionales simples 68
6.4.2.- Bucles 69
6.4.3.- Condicionales vectorizados 70
6.5.- Reutilización de código R 70
6.5.1.- Funciones 71
6.5.2.- *Scripts* y *notebooks* 72
6.5.3.- Paquetes 72
7.- El entorno de trabajo para Python 74
7.1.- Acceso a Jupyter Lab 74
7.2.- Componentes de Jupyter Lab 75
7.2.1.- Intérprete interactivo 77
7.2.2.- Edición de cuadernos 77
7.2.3.- Composición de guiones y ejecución interactiva 79
8.- Introducción al lenguaje Python 79
8.1.- El lenguaje Python 80
8.2.- Variables y operaciones simples 81
8.2.1.- Tipos de dato básicos en Python 82
8.2.2.- Asignación de variables 82
8.2.3.- Operadores básicos 82
8.3.- Tipos de dato complejos 83
8.3.1.- Listas y tuplas 83
8.3.2.- Diccionarios 85
8.4.- Estructuras de control 85
8.4.1.- Condicionales 85
8.4.2.- Repetición 86

8.5.- Reutilización de código 87
8.5.1.- Funciones 87
8.5.2.- Clases y métodos 88
8.5.3.- Instalación de paquetes y utilización de módulos 88
9.- Prácticas propuestas 89

CAPÍTULO 2: CARGA Y EXPORTACIÓN DE DATOS 91

1.- Introducción 91
2.- Carga de datos 92
2.1.- Conjuntos de datos integrados 93
2.1.1.- R 93
2.1.2.- Python 95
2.2.- Carga de datos alojados en archivos 96
2.2.1.- Lectura de archivos CSV 96
R 97
Python 98
2.2.2.- Lectura de documentos Excel 99
R 100
Python 101
2.2.3.- Lectura de archivos ARFF 101
R 102
Python 103
2.3.- Recuperación de datos desde el portapapeles 103
2.3.1.- R 105
2.3.2.- Python 106
2.4.- Obtención de datos desde páginas web 106
2.4.1.- URL como ruta de archivo 106
2.4.2.- Proceso de web *scraping* 108
2.4.3.- Obtención del selector de datos 108
R 110
Python 111
3.- Exportación de datos 111
3.1.- Exportación a formatos de texto 112
3.1.1.- R 112
Exportación como CSV 112
Exportación como ARFF 114
3.1.2.- Exportación como CSV con Python 114
3.2.- Exportación a formatos binarios - R 115
3.2.1.- Formatos binarios propios de R 116
3.2.2.- Trabajar con archivos HDF5 desde R 117
3.3.- Exportación a formatos binarios - Python 119
3.3.1.- Formatos binarios propios de Python 119
3.3.2.- Trabajar con archivos HDF5 desde Python 119
4.- Prácticas propuestas 120

CAPÍTULO 3: ANÁLISIS EXPLORATORIO DE DATOS 123

1.- Introducción 123
1.1.- Herramientas para el análisis exploratorio de datos 124
2.- Obtención de estadísticos básicos 124
2.1.- Estadísticos de posición 125
2.1.1.- Definiciones y ejemplos 125
Media aritmética 126
Media geométrica y armónica 127
Mediana 128
Moda 128
Percentiles y cuartiles 129
2.2.- Cálculo de estadísticos de posición con R 130
2.2.1.- Media aritmética 130
2.2.2.- Medias geométrica y armónica 131
2.2.3.- Mediana y moda 132
2.2.4.- Cuantiles 133
2.3.- Cálculo de estadísticos de posición con Python 134
2.3.1.- Media aritmética 134
2.3.2.- Medias armónica y geométrica 136
2.3.3.- Mediana y moda 137
2.3.4.- Cuantiles 138
2.4.- Estadísticos de dispersión 138
2.4.1.- Rangos 139
2.4.2.- Varianza y desviación típica 140
2.5.- Cálculo de estadísticos de dispersión 142
2.5.1.- Cálculo con R 142
2.5.2.- Cálculo con Python 143
2.6.- Estadísticos de forma 144
2.6.1.- Definiciones y ejemplos 145
Coeficientes de asimetría 146
Coeficiente de curtosis 147
2.7.- Cálculo de estadísticos de forma 148
2.7.1.- Cálculo con R 148
2.7.2.- Cálculo con Python 149
2.8.- Distribuciones fundamentales 150
2.8.1.- Distribuciones de datos con R 151
2.8.2.- Distribuciones de datos con Python 152
3.- Prácticas propuestas 153
4.- Visualización de los datos 153
4.1.- Representaciones gráficas: importancia y utilidad 154
4.1.1.- Importancia de la visualización de datos 154
Análisis con estadísticos 155
Análisis con visualización básica 156
4.1.2.- Utilidad de las representaciones gráficas 157
4.2.- Introducción a la generación de gráficas con R 158
4.2.1.- Paquetes para gráficas en R 159

4.2.2.- El lienzo de dibujo de la gráfica 159
4.2.3.- Agregar capas visibles sobre el lienzo 160
4.2.4.- Visualización de múltiples gráficas 161
4.2.5.- Gráficas a partir de funciones 162
4.3.- Introducción a la generación de gráficas con Python 163
4.3.1.- Paquetes para gráficas en Python 163
4.3.2.- El lienzo de dibujo de la gráfica 164
4.3.3.- Agregar capas visibles sobre el lienzo 164
4.3.4.- Visualización de múltiples gráficas y gráficas a partir de funciones 165
4.4.- Gráficas de nubes de puntos 167
4.4.1.- Python 167
4.4.2.- R 168
4.5.- Gráficas de barras 171
4.5.1.- Python 171
4.5.2.- R 172
4.6.- Histogramas 176
4.6.1.- R 176
4.6.2.- Python 179
4.7.- Gráficas de líneas 181
4.7.1.- R 181
4.7.2.- Python 184
Versión básica 184
Líneas que conectan puntos 185
Personalización de las gráficas de líneas 187
4.8.- Diagramas de cajas y bigotes 190
4.8.1.- R 190
4.8.2.- Python 195
5.- Prácticas propuestas 196

CAPÍTULO 4: LIMPIEZA DE DATOS 199

1.- Introducción 199
2.- Exploración de los datos a limpiar 201
2.1.- La variable Edad 203
2.2.- La variable Sexo 204
2.3.- La variable Embarazos 205
2.4.- Variables Sistolica y Diastolica 206
2.5.- La variable Masa 207
3.- Selección de muestras 207
3.1.- Eliminación de instancias duplicadas 207
3.1.1.- R 208
3.1.2.- Python 209
3.2.- Detección de filas con datos incorrectos 211
3.2.1.- Python 212
3.2.2.- R 213
4.- Eliminación de valores anómalos y datos perdidos 215
4.1.- Tratamiento de valores anómalos (*outliers*) 215
4.1.1.- R 216

4.1.2.- Python 217
4.2.- Borrado de muestras con valores perdidos 218
4.2.1.- Python 218
4.2.2.- R 219
5.- Filtrado de variables 221
5.1.- Conversión de tipo de una variable 221
5.1.1.- R 222
5.1.2.- Python 224
6.- Tratamiento de variables redundantes 228
6.1.- Python 228
6.2.- R 230
7.- Imputación de valores perdidos 234
7.1.- Imputación de valores perdidos en variables numéricas 234
7.1.1.- Python 235
7.1.2.- R 237
7.2.- Imputación de valores perdidos en variables nominales 240
7.2.1.- R 241
7.2.2.- Python 242
7.3.- Tratamiento de otros valores especiales 243
8.- Planificación de las operaciones de limpieza 244
8.1.- El orden es importante 244
8.2.- Directrices generales 244
8.3.- Limpieza con R 245
8.4.- Limpieza con Python 246
9.- Prácticas propuestas 247

CAPÍTULO 5: PREDICCIÓN DE VALORES NUMÉRICOS 249

1.- Introducción 249
1.1.- Problemas de regresión 249
1.2.- Interpolación y extrapolación 250
1.3.- Regresión lineal 250
1.3.1.- Caso de estudio 250
1.3.2.- Ajuste del modelo 252
1.4.- Otros métodos de regresión 253
2.- Predicción del precio de la vivienda 254
2.1.- Carga del conjunto de datos boston 255
2.2.- Estructura y variables 255
2.3.- Análisis de valores perdidos y extremos 256
2.4.- Análisis de correlaciones 258
2.5.- Estudio de la variable objetivo de la predicción 260
3.- Introducción a los árboles de decisión 261
3.1.- Características básicas de los DT 261
3.2.- Funcionamiento intuitivo de un DT 261
3.3.- Tipos de nodos en un DT 263
3.4.- Impureza de los nodos hoja 263
3.5.- Selección de las variables en nodos de decisión 265

3.6.- Criterios de detención de crecimiento 265
3.7.- Poda de un DT 267
4.- Solución con árboles de decisión 267
4.1.- Carga de paquetes y datos 268
4.2.- Generación y análisis del DT 268
4.3.- Representación visual del DT 269
4.4.- Predicción de valores a partir del DT 271
4.5.- Configuración del proceso de generación del DT 272
5.- Evaluación de modelos predictivos 274
5.1.- Particionamiento de los datos 275
5.2.- Técnicas de evaluación 276
5.3.- Evaluación de múltiples modelos alternativos 277
5.4.- Un caso práctico 278
6.- Introducción al diseño de redes neuronales 282
6.1.- Qué son las redes neuronales 282
6.1.1.- Neuronas artificiales 282
6.1.2.- Funciones de activación 283
6.1.3.- Capas de neuronas 285
7.- Implementación de redes neuronales en Python 286
7.1.- Capa de entrada 287
7.2.- Capa de salida 287
7.3.- Capas ocultas 287
7.4.- Compilación y ajuste del modelo 288
7.5.- Predicción de nuevos valores 288
8.- Solución con redes neuronales 289
8.1.- Preparación de los datos 290
8.2.- Definición del modelo 290
8.3.- Entrenamiento del modelo 292
8.4.- Evaluación del modelo 293
9.- Prácticas propuestas 295

CAPÍTULO 6: AGRUPAMIENTO DE DATOS 297

1.- Introducción 297
1.1.- Objetivo 297
1.2.- Técnicas de agrupamiento 299
2.- Cómo medir la distancia entre dos muestras 301
2.1.- Distancia euclídea 302
2.2.- Distancia Manhattan 303
2.3.- Distancia del máximo 304
2.4.- Generalización: distancias de Minkowski 304
2.5.- Otras distancias: Levenshtein y coseno 305
2.5.1.- Distancia de Levenshtein 305
2.5.2.- Similitud del coseno 306
3.- Agrupamiento por particionamiento - K-medias 307
3.1.- Estandarización de magnitudes 307
3.2.- Cómo medir la similitud-distancia de grupos 308
3.3.- El algoritmo K-medias 309

4.- Agrupamiento con K-medias en R 309
4.1.- Un primer acercamiento 310
4.2.- Visualización del agrupamiento 311
5.- K-medias: estimación del número de grupos 313
5.1.- Estimación con la función pamk() 314
5.2.- Estimación analizando el SSE 316
6.- Problemas del algoritmo K-medias 318
6.1.- Número de grupos y aleatoriedad 319
6.2.- Fallos al tratar con datos que tienen estructura compleja 319
6.2.1.- Conjuntos de datos en el paquete mlbench 320
6.2.2.- Resultados de K-medias sobre datos complejos 321
6.3.- Influencia de datos extremos 323
7.- Agrupamiento por densidad 323
7.1.- Conceptos previos 324
7.2.- El algoritmo DBSCAN 325
7.3.- Agrupamiento con DBSCAN en R 326
7.3.1.- Uso básico de dbscan() 326
7.3.2.- Análisis de los grupos generados 328
7.4.- Limitaciones de DBSCAN 330
8.- Agrupamiento jerárquico 330
8.1.- Interpretación de dendrogramas 331
8.2.- Construcción de la jerarquía 332
8.3.- Uso del agrupamiento jerárquico en Python 333
9.- Prácticas propuestas 336

CAPÍTULO 7: CLASIFICACIÓN BINARIA 339

1.- Introducción 339
1.1.- Tipos de clasificación 340
2.- Clasificación mediante técnicas de regresión 341
2.1.- Transformar iris en un problema de clasificación binaria 341
2.2.- Particionamiento de los datos 342
2.3.- Análisis exploratorio de los datos 342
2.4.- Creación del modelo y uso para clasificación 343
2.5.- Evaluación del clasificador 344
2.6.- Visualización de la frontera de división 345
3.- Clasificación con técnicas de agrupamiento 346
3.1.- Creación del modelo 346
3.2.- Cómo predecir la clase a partir de los grupos 348
3.3.- Evaluación del clasificador 349
4.- Clasificación automática de correo basura 350
4.1.- Generación de un conjunto de datos a partir de mensajes de correo 350
4.2.- El conjunto de datos *Spambase* 352
4.2.1.- Variables predictoras 352
4.2.2.- Etiquetas de clase 353
4.3.- Análisis exploratorio del conjunto de datos 353
4.3.1.- Distribución de las clases 353

4.3.2.- Uso de mayúsculas según tipo de mensaje 354
4.3.3.- Aparición de ciertos símbolos según tipo de correo 355
4.3.4.- Frecuencia de las palabras dependiendo del tipo de correo 356
4.4.- Agrupamiento de las muestras con K-Medias 358
5.- Introducción a las máquinas de vectores soporte 359
5.1.- Características básicas de las SVM 359
5.1.1.- Margen de separación máximo 360
5.1.2.- Margen de separación flexible 362
5.1.3.- Los vectores de soporte 362
5.1.4.- Proceso de predicción 362
5.2.- Funcionamiento intuitivo de una SVM 363
5.2.1.- Paso 1 363
5.2.2.- Paso 2 364
5.2.3.- Paso 3 364
5.2.4.- Paso 4 366
5.3.- Búsqueda de separabilidad en dimensiones superiores 366
5.3.1.- Funciones linealmente separables en un plano: AND y OR 367
5.3.2.- Funciones no linealmente separables en un plano: XOR 369
5.3.3.- Aplicación de un *kernel* para proyectar en un espacio de más alta dimensionalidad 369
5.4.- Ventajas y desventajas de las SVM 370
5.4.1.- SVM para clasificación no binaria 371
5.4.2.- Resumiendo 372
6.- Solución usando máquinas de vectores soporte 373
6.1.- Particionamiento de los datos 373
6.2.- Entrenamiento del modelo 376
6.2.1.- Configuración de validación interna 376
6.2.2.- Información sobre el modelo 377
6.3.- Predicción para nuevos datos 378
6.3.1.- Obtención de predicciones 379
6.3.2.- Evaluación del rendimiento predictivo 379
6.4.- Ajuste de parámetros con búsqueda en cuadrícula 380
6.4.1.- Cómo definir la matriz de combinaciones 380
6.4.2.- Entrenamiento y evaluación 381
6.4.3.- Comparación de diferentes configuraciones 383
6.5.- Normalización y optimización usando otras métricas 384
6.5.1.- Normalización de los datos 384
6.5.2.- Optimizar el modelo usando otras métricas 385
6.6.- Comparativa de los modelos sobre datos de test 388
7.- Evaluación del rendimiento predictivo de un clasificador (I) 389
7.1.- Cómo medir el rendimiento de un clasificador 390
7.2.- Evaluación de clasificadores binarios con salida discreta 391
7.2.1.- La matriz de confusión 391
7.3.- Exactitud, precisión y sensibilidad 392
7.3.1.- Exactitud 392
7.3.2.- Precisión 393
7.3.3.- Sensibilidad (TPR) 394

7.3.4.- Equilibrio entre precisión y sensibilidad 394
7.4.- Otras métricas obtenidas de la matriz de confusión 396
7.4.1.- Especificidad 397
7.4.2.- False Positive Rate (FPR) 397
7.4.3.- Kappa 398
7.5.- Uso combinado de múltiples métricas 399
8.- Entrenamiento de redes neuronales 400
8.1.- Optimización de una red 401
8.2.- Selección de pesos 401
8.3.- Selección de hiperparámetros 402
9.- Solución usando una red neuronal 403
9.1.- Diseño del modelo 404
9.2.- Entrenamiento 404
9.3.- Evaluación del modelo 406
10.- Evaluación del rendimiento predictivo de un clasificador (II) 408
10.1.- Umbralización de la salida 408
10.1.1.- El umbral de corte trivial 409
10.1.2.- Ajuste del umbral de corte 409
10.2.- Curva ROC y AUC 410
11.- Prácticas propuestas 411

CAPÍTULO 8: CLASIFICACIÓN MULTICLASE 413

1.- Introducción 413
2.- Clasificación de dígitos numéricos manuscritos 414
2.1.- Origen del conjunto de datos MNIST 414
2.2.- Obtención de MNIST 415
2.2.1.- Desde R 415
2.2.2.- Desde Python 416
2.3.- Estructura de MNIST 416
2.3.1.- Visualización de los dígitos 417
2.3.2.- Análisis exploratorio 419
3.- Introducción a kNN 421
3.1.- Modelos predictivos perezosos 421
3.2.- Funcionamiento de kNN 422
3.2.1.- El algoritmo básico kNN 423
3.2.2.- Implementación básica de kNN en R 423
3.3.- Aspectos a considerar al usar kNN 428
3.3.1.- Selección del valor k 428
3.3.2.- Métricas de distancia entre muestras 429
3.3.3.- Otras aplicaciones de kNN 429
3.3.4.- Ventajas e inconvenientes de kNN 430
4.- Solución con vecinos más cercanos en R 430
4.1.- Carga de paquetes y preparación de datos 431
4.1.1.- Selección de particiones de entrenamiento y test 432
4.2.- Entrenamiento básico de kNN 432
4.3.- Optimización de otros parámetros de kNN 434

4.3.1.- Métricas de distancia 434
4.3.2.- Predicción ponderando las distancias 435
4.3.3.- kNN con optimización de todos los parámetros 435
4.4.- Comparativa de los dos modelos 437
4.5.- Análisis de rendimiento sobre datos de test 438
4.5.1.- Matriz de confusión multiclase 438
4.5.2.- Representación gráfica de la matriz de confusión 440
4.6.- Paralelización al usar caret 442
4.6.1.- Configuración de una ejecución en múltiples núcleos 443
4.6.2.- Análisis de la diferencia de rendimiento 444
5.- Evaluación del rendimiento predictivo de un clasificador (III) 445
5.1.- Matriz de confusión extendida 446
5.1.1.- Conteo de casos por clase 447
5.2.- Cálculo de métricas en problemas multiclase 449
5.2.1.- Conteo de casos por clase 449
5.2.2.- Procedimientos de cálculo de las métricas 450
Macro-averaging 451
Micro-averaging 452
5.2.3.- Qué métricas usar para evaluar un clasificador 453
6.- La operación de convolución 453
6.1.- Formulación 454
6.2.- Ejemplo 455
7.- De la convolución a las redes convolucionales 457
7.1.- Otras operaciones de las redes convolucionales 458
7.2.- Redes convolucionales 460
7.3.- Arquitecturas populares 460
8.- Solución con una red neuronal convolucional 461
8.1.- Diseño de la red 461
8.2.- Entrenamiento de la red 463
8.3.- Evaluación 464
8.4.- Comparativa 466
9.- Prácticas propuestas 467

CAPÍTULO 9: SISTEMAS DE RECOMENDACIÓN 469

1.- Introducción 469
2.- Conceptos sobre sistemas de recomendación (SR) 470
2.1.- Análisis de la cesta de la compra 470
2.2.- Tipos de SR 471
2.2.1.- Filtrado de contenido 471
2.2.2.- Filtrado colaborativo 472
2.2.3.- Sistemas híbridos 472
3.- El conjunto de datos Groceries 473
3.1.- El formato de archivo *basket* 473
3.2.- Estructura del conjunto de datos 475
3.2.1.- Carga del conjunto de datos 475
3.2.2.- Dimensiones del conjunto de datos 475
3.2.3.- Enumeración de las columnas 477

3.2.4.- Exploración de transacciones 477
3.2.5.- Frecuencias de los productos 478
4.- Introducción a las reglas de asociación 480
4.1.- ¿Qué es una regla de asociación? 480
4.2.- Métricas de evaluación de reglas de asociación 481
4.2.1.- Soporte de un *itemset* 481
4.2.2.- Confianza de una regla 482
4.2.3.- *Lift* de una regla 483
5.- Minería de reglas de asociación 485
5.1.- Búsqueda de reglas mediante fuerza bruta 485
5.1.1.- Generación de todos los *itemset* posibles 486
5.1.2.- Búsqueda de combinaciones de pares de *itemset* 486
5.1.3.- Evaluación y filtrado de las reglas 487
5.2.- El algoritmo Apriori 487
5.2.1.- Configuración de umbrales 488
5.2.2.- Extracción de *itemset* frecuentes 488
6.- Minería de reglas de asociación con R 489
6.1.- Configuración y extracción de reglas 491
6.1.1.- Ejecución con parámetros por defecto 491
6.1.2.- Configuración del soporte y confianza mínimos 492
6.2.- Exploración de las reglas 493
6.2.1.- Resumen estadístico de características de las reglas 493
6.2.2.- Inspección de las reglas 494
6.3.- Evaluación de calidad de las reglas 496
6.3.1.- Ordenar las reglas por una métrica 497
6.3.2.- Filtrado de reglas 498
6.3.3.- Representaciones gráficas adicionales 499
7.- Obtención de reglas para el sistema de recomendación 501
7.1.- Extracción de reglas con una apariencia concreta 502
7.2.- Construcción del sistema de recomendación 503
7.2.1.- Función para obtención de las reglas 503
7.2.2.- Extracción de los productos recomendados 504
7.2.3.- Probando el sistema de recomendación 505
8.- Prácticas propuestas 506

CAPÍTULO 10: PROBLEMAS AVANZADOS - CASOS DE ESTUDIO 509

1.- Introducción 509
2.- Uso de múltiples modelos para mejorar los resultados 510
2.1.- Tipos de *ensembles* 511
2.1.1.- Aspectos generales 511
Diversidad en los modelos 512
Estrategia de combinación de salidas 512
Ensembles homogéneos vs heterogéneos 513
2.1.2.- *Bagging* 513
2.1.3.- *Boosting* 514
2.1.4.- *Random Forest* 514

2.2.- *Ensembles* en R 515
2.2.1.- Uso de *ensembles* con el paquete caret 515
Ensembles homogéneos 516
Ensembles heterogéneos 517
3.- Eliminación de ruido en imágenes 517
3.1.- Qué es un *autoencoder* 519
3.1.1.- Representaciones y códigos 520
3.1.2.- Arquitectura de un *autoencoder* 520
3.1.3.- Limpieza de imágenes con *autoencoders* 521
3.2.- Diseño de *autoencoders* con Keras 522
4.- Predicción de series temporales 525
4.1.- ¿Qué es una serie temporal? 525
4.2.- Tipos de series temporales 525
4.3.- Funcionamiento de la predicción de series temporales 526
4.4.- Componentes de una serie temporal 526
4.4.1.- Nuestra serie temporal de ejemplo 527
4.4.2.- Descomposición de la serie 528
Tendencia de la serie 529
Estacionalidad de la serie 529
Componente aleatoria 530
¿Qué componentes tiene mi serie temporal? 530
4.5.- Predicción de series temporales con ARIMA 530
4.5.1.- ¿Qué es ARIMA? 531
4.5.2.- Parámetros de entrada a ARIMA 532
4.5.3.- Validación de la predicción 533
5.- Detección de objetos en imágenes 533
5.1.- Conceptos fundamentales 534
5.2.- *Transfer learning* 534
5.3.- Redes neuronales para detección de objetos 535
5.3.1.- Modelos multietapa 536
5.3.2.- Modelos de una etapa 536
YOLO 537
SSD 537
CenterNet 538
5.4.- Detección de objetos en Python 538
5.4.1.- Instalación de la API de detección de objetos 538
5.4.2.- Descarga del modelo 540
5.4.3.- Uso de la API de detección de objetos 540
6.- Búsqueda de soluciones a problemas de optimización difíciles 542
6.1.- Problemas clásicos de optimización 543
6.1.1.- Problema del viajante de comercio 543
6.1.2.- Problema de la mochila 543
6.1.3.- Problemas de optimización continua 544
6.2.- Métodos de optimización 545
6.3.- Técnicas de optimización evolutivas 546
6.3.1.- Funcionamiento general de las técnicas evolutivas 546

6.3.2.- Algoritmos de optimización basados en principios evolutivos/biológicos ... 547
6.3.3.- Equilibrio entre explotación y exploración ... 549
6.4.- Algoritmos genéticos ... 549
6.4.1.- Pasos esenciales del algoritmo ... 550
6.4.2.- Representación de los individuos ... 551
6.4.3.- El mecanismo de selección ... 552
6.4.4.- El operador de mutación ... 552
6.4.5.- El operador de cruce ... 553
7.- Modelado generativo de datos ... 553
7.1.- Redes neuronales generativas ... 554
7.1.1.- Redes adversariales ... 554
7.1.2.- *Autoencoders* generativos ... 556
Autoencoder variacional ... 556
Autoencoder adversarial ... 557
7.2.- Implementación de redes generativas en Python ... 558
7.2.1.- Componentes de un *autoencoder* adversarial ... 558
7.2.2.- Construcción del modelo ... 559
7.2.3.- Entrenamiento y predicción ... 560
8.- Prácticas propuestas ... 561

La inteligencia artificial (AI) y el aprendizaje automático (ML) **forman parte de nuestras vidas**, incluso sin que lo notemos: las recomendaciones que recibimos en muchos sitios online, los asistentes personales, la detección de fraudes o un "simple" buscador efectivo. La utilizas cada día sin saberlo. Y con la Internet de las cosas (IoT) y la hiperconectividad que trae el 5G, **cada vez se incorporará a más procesos**, como los coches autónomos, la gestión de la energía, la gestión de la producción en las fábricas, la definición de estrategias de marketing y ventas, el reconocimiento de voz y los flujos de atención al cliente...

El **aprendizaje automático** o *machine learning* es la ciencia de **conseguir que una computadora haga cosas para las que no está explícitamente programada.** Es decir, en lugar de programarlo de una determinada manera, logramos que aprenda de forma autónoma para lograr el objetivo perseguido. La **ciencia de datos**, también llamada "descubrimiento de conocimiento" involucra las **técnicas y modelos para extraer conocimiento no evidente a partir de datos**: localizar patrones ocultos, correlaciones relevantes o conclusiones complejas.

Diseñado por dos experimentados científicos de datos, **con este libro**, profundo, pero al mismo tiempo claro, **aprenderás desde cero todo el proceso de trabajo para ciencia de datos y *machine learning***, incluyendo los conceptos fundamentales subyacentes y la práctica necesaria para ponerlos a trabajar sin problema según tus propias necesidades o las de tu empresa. Al mismo tiempo, aprenderás lo necesario para utilizar los dos lenguajes más comunes en ciencia de datos: **Python** y **R**. Además de estos lenguajes, utilizarás entre otras las herramientas: ggplot, caret, arules, Keras, Numpy, Matplotlib, Tensorflow, Pandas, SciPy...

Nivel: Iniciación - Intermedio - Avanzado

Francisco Charte
Dr. Ingeniero Informático

Autor de numerosos libros y centenares de artículos en revistas nacionales e internacionales, Francisco es ingeniero informático y Doctor en tecnologías de la información y la comunicación.
Está especializado en ciencia de datos y ML, con amplia experiencia en proyectos reales.

David Charte
Ingeniero Informático

Ingeniero informático y matemático apasionado por la divulgación del conocimiento, David, trabaja y hace investigación en el campo de la ciencia de datos y ML.
Ha participado en multitud de proyectos para varios sectores dentro de estos campos.

krasis PRESS

campus MVP.es

ISBN 978-84-945822-5-7