

Contenidos

Prefacio	xi
1. Análisis exploratorio de datos	1
Elementos de datos estructurados.....	2
Lecturas complementarias.....	4
Datos rectangulares.....	4
Marcos de datos e índices.....	6
Estructuras de datos no rectangulares	6
Lecturas complementarias.....	7
Estimación de la localización	7
Media.....	9
Estimación de medianas robustas.....	10
Ejemplo: estimaciones de localización de la población y tasas de homicidios.....	12
Lecturas complementarias.....	13
Estimación de la variabilidad	13
Desviación estándar y estimaciones relacionadas	15
Estimación basada en percentiles.....	17
Ejemplo: estimaciones de variabilidad de la población estatal	18
Lecturas complementarias.....	19
Exploración de la distribución de datos.....	19
Percentiles y diagramas de caja	20
Tablas de frecuencias e histogramas	22
Diagrama y estimación de la curva de densidad.....	24
Lecturas complementarias.....	26
Exploración de datos binarios y categóricos	27
Moda	29
Valor esperado	29
Probabilidad	30
Lecturas complementarias.....	30
Correlación	30
Diagramas de dispersión	34
Lecturas complementarias.....	36
Exploración de dos o más variables.....	36
Agrupación hexagonal y contornos (representación numérica frente a datos numéricos).....	36

Dos variables categóricas	39
Datos categóricos y numéricos	41
Visualización de varias variables	43
Lecturas complementarias	45
Resumen	45
2. Distribuciones de datos y muestreo	47
Muestreo aleatorio y sesgo de la muestra	48
Sesgo	50
Selección aleatoria	51
Tamaño frente a calidad: ¿cuándo importa el tamaño?	52
Media muestral frente a media poblacional	53
Lecturas complementarias	53
Sesgo de selección	54
Regresión a la media	55
Lecturas complementarias	57
Distribución muestral del estadístico	57
Teorema del límite central	60
Error estándar	60
Lecturas complementarias	61
Bootstrap	61
Remuestreo frente a bootstrapping	65
Lecturas complementarias	65
Intervalos de confianza	65
Lecturas complementarias	68
Distribución normal	69
Normal estándar y diagramas QQ	70
Distribuciones de cola larga	72
Lecturas complementarias	74
Distribución t de Student	74
Lecturas complementarias	77
Distribución binomial	77
Lecturas complementarias	79
Distribución chi cuadrado	79
Lecturas complementarias	80
Distribución F	81
Lecturas complementarias	81
La distribución de Poisson y distribuciones relacionadas	81
Distribución de Poisson	82
Distribución exponencial	83
Estimación de la tasa de fallos	83
Distribución de Weibull	84
Lecturas complementarias	85
Resumen	85

3. Experimentos estadísticos y pruebas significativas	87
Prueba A/B.....	88
¿Por qué tener un grupo de control?.....	90
¿Por qué solo A/B? ¿Por qué no C, D, ...?	91
Lecturas complementarias.....	92
Pruebas de hipótesis	93
La hipótesis nula.....	94
Hipótesis alternativa	95
Pruebas de hipótesis unidireccionales o bidireccionales.....	95
Lecturas complementarias.....	96
Remuestreo	96
Prueba de permutación	97
Ejemplo: adherencia de la web.....	98
Pruebas de permutación exhaustiva y de bootstrap	102
Pruebas de permutación: el resultado final de la ciencia de datos	102
Lecturas complementarias.....	103
Significación estadística y valores p.....	103
Valor p	106
Alfa	107
Errores de tipo 1 y 2	108
Ciencia de datos y valores p.....	109
Lecturas complementarias.....	109
Pruebas t.....	109
Lecturas complementarias.....	111
Pruebas múltiples	111
Lecturas complementarias.....	115
Grados de libertad	115
Lecturas complementarias.....	117
ANOVA.....	117
Estadístico F.....	120
ANOVA bidireccional	122
Lecturas complementarias.....	123
Prueba de chi cuadrado.....	123
Prueba de chi cuadrado: enfoque de remuestreo	124
Prueba de chi cuadrado: teoría estadística.....	126
Prueba exacta de Fisher	127
Relevancia para la ciencia de datos	129
Lecturas complementarias.....	130
Algoritmo Multi-Arm Bandit.....	130
Lecturas complementarias.....	134
Potencia y tamaño de la muestra.....	134
Tamaño de la muestra	135
Lecturas complementarias.....	138
Resumen	138

4. Regresión y pronóstico	139
Regresión lineal simple	139
La ecuación de regresión.....	141
Valores ajustados y residuos.....	143
Mínimos cuadrados.....	145
Pronóstico frente a explicación (elaboración de perfiles)	146
Lecturas complementarias.....	147
Regresión lineal múltiple	147
Ejemplo: datos de las viviendas del condado de King	148
Evaluación del modelo	149
Validación cruzada	151
Selección del modelo y regresión escalonada	152
Regresión ponderada	156
Lecturas complementarias.....	157
Pronóstico mediante la regresión	157
Los peligros de la extrapolación.....	158
Intervalos de confianza y de pronóstico	158
Variables de tipo factor en la regresión	160
Representación de variables ficticias	161
Variables de tipo factor con muchos niveles	163
Variables de tipo factor ordenadas.....	165
Interpretación de la ecuación de regresión	166
Predictoras correlacionadas.....	166
Multicolinealidad	168
Variables de confusión	168
Interacciones y efectos principales	170
Diagnósticos de regresión	172
Valores atípicos	173
Valores influyentes.....	175
Heterocedasticidad, anormalidad y errores correlacionados	177
Diagramas de residuos parciales y falta de linealidad	180
Regresión polinomial y por spline	183
Polinomial	183
Splines	185
Modelos aditivos generalizados.....	187
Lecturas complementarias.....	189
Resumen	189
5. Clasificación.....	191
Bayes ingenuo	192
Por qué la clasificación bayesiana exacta no es práctica	193
La solución ingenua.....	194
Variables predictoras numéricas.....	196
Lecturas complementarias.....	197

Análisis discriminante	197
Matriz de covarianza	198
Discriminante lineal de Fisher.....	199
Un ejemplo sencillo	200
Lecturas complementarias.....	203
Regresión logística.....	203
Función de respuesta logística y logit	204
Regresión logística y GLM	206
Modelos lineales generalizados	207
Valores pronosticados de regresión logística	208
Interpretación de los coeficientes y de la razón de oportunidades.....	208
Regresión lineal y logística: similitudes y diferencias	210
Evaluación del modelo	211
Lecturas complementarias.....	214
Evaluación de modelos de clasificación	215
Matriz de confusión	216
El problema de las clases raras	218
Precisión, exhaustividad y especificidad	218
Curva ROC	219
AUC.....	221
Sustentación.....	222
Lecturas complementarias.....	224
Estrategias para datos que no están equilibrados	224
Submuestreo	225
Sobremuestreo y aumento/disminución de la ponderación	226
Generación de datos	228
Clasificación basada en los costes.....	228
Exploración de pronósticos	229
Lecturas complementarias.....	230
Resumen.....	230
6. Aprendizaje automático estadístico	231
K-vecinos más cercanos.....	232
Un pequeño ejemplo: pronóstico del incumplimiento de préstamos ...	233
Métricas de distancia	235
Codificador One-Hot	236
Estandarización (normalización, puntuación z)	237
Elección de K	240
KNN como motor de características	241
Modelos de árbol.....	243
Un ejemplo sencillo	244
Algoritmo de partición recursiva.....	246
Medición de la homogeneidad o la impureza.....	248
Detención del crecimiento del árbol.....	249
Pronóstico de un valor continuo	251

Cómo se utilizan los árboles.....	252
Lecturas complementarias.....	253
Métodos de bagging y bosque aleatorio.....	253
Bagging.....	254
Bosque aleatorio.....	255
Importancia de la variable.....	259
Hiperparámetros.....	262
Boosting.....	263
El algoritmo boosting.....	264
XGBoost.....	265
Regularización: evitación del sobreajuste.....	268
Hiperparámetros y validación cruzada.....	272
Resumen.....	275
7. Aprendizaje no supervisado.....	277
Análisis de componentes principales.....	278
Un ejemplo sencillo.....	279
Cálculo de los componentes principales.....	282
Interpretación de componentes principales.....	282
Análisis de correspondencias.....	285
Lecturas complementarias.....	287
Agrupación K-means.....	287
Un ejemplo sencillo.....	288
Algoritmo K-means.....	290
Interpretación de los grupos.....	291
Selección del número de grupos.....	293
Agrupación jerárquica.....	296
Un ejemplo sencillo.....	296
El dendrograma.....	297
El algoritmo de aglomeración.....	299
Medidas de disimilitud.....	299
Agrupación basada en el modelo.....	301
Distribución normal multivariante.....	301
Mezclas de distribuciones normales.....	303
Selección del número de grupos.....	305
Lecturas complementarias.....	308
Variables categóricas y escalado.....	308
Escalado de variables.....	309
Variables dominantes.....	310
Datos categóricos y distancia de Gower.....	311
Problemas con la agrupación de datos mixtos.....	314
Resumen.....	316
Bibliografía.....	317