Índice

PRÓLOGO **13**

PREFACIO **15**

ACERCA DE ESTE LIBRO **19**

A quién va dirigido el libro 19
Organización del libro 19
Acerca de los anglicismos usados en este libro 22
Página web del libro 23
Convenciones tipográficas usadas en este libro 23
Preparación del entorno de trabajo 25
Acerca de WATCH THIS SPACE 29

PARTE 1: FUNDAMENTOS DEL APRENDIZAJE POR REFUERZO **33**

1 INTRODUCCIÓN AL APRENDIZAJE POR REFUERZO **35**

1.1 Contexto **35**

Inteligencia artificial 36
Machine Learning 37
Deep Learning 39
Aprendizaje por refuerzo profundo 40

1.2 Aprendizaje por refuerzo **41**

Aprender interactuando 41
Piezas básicas en el aprendizaje por refuerzo 42

1.3 Modelización de un problema **46**

Frozen-Lake: un lago congelado donde patinar 47
Programación del entorno 49
Programación del agente 51
La clase Agent 55

1.4 En qué se diferencia el aprendizaje **57**

Aprendizaje por refuerzo versus aprendizaje supervisado 58
Aprendizaje por refuerzo versus aprendizaje no supervisado 58
¿Dónde están los datos en el aprendizaje por refuerzo? 59

2 FORMALIZACIÓN DEL APRENDIZAJE POR REFUERZO **61**

2.1 Proceso de decisión de Markov **61**

Recapitulación de conceptos 61
Markov Decision Processes 62

2.2 Piezas de un proceso de decisión de Markov **63**

Estados 63

Acciones ... 65
Función de transición ... 66
Recompensa ... 66
Factor de descuento ... 68
2.3 Entornos deterministas y estocásticos ... 69
Entorno determinista ... 69
Entorno estocástico ... 74
2.4 Configuración del problema a resolver ... 77
Episodio y trayectoria ... 78
Retorno con descuento ... 78
Política ... 80
3 FUNCIONES DE VALOR Y LA ECUACIÓN DE BELLMAN ... 83
3.1 Funciones de valor ... 83
Visión global ... 83
Función V: función de valor del estado ... 84
Función Q: función de valor de la acción ... 88
Función V vía función Q ... 88
3.2 La ecuación de Bellman ... 89
La ecuación de Bellman para la función V ... 89
La ecuación de Bellman para entornos estocásticos ... 90
La ecuación de Bellman para políticas estocásticas ... 91
La ecuación de Bellman para la función Q ... 92
3.3 La ecuación de Bellman óptima ... 92
Función V óptima ... 93
Función Q óptima ... 93
3.4 Resumen de la terminología ... 94
Notación matemática ... 94
Tabla resumen ... 95
PARTE 2: ALGORITMOS CLÁSICOS DE APRENDIZAJE POR REFUERZO .. 99
4 PROGRAMACIÓN DINÁMICA ... 101
4.1 Método Value Iteration ... 101
Programación dinámica ... 101
Cómo aprender recursivamente ... 103
Algoritmo Value Iteration ... 108
4.2 Implementación del método Value Iteration ... 109
La clase Agent ... 110
Bucle de entrenamiento ... 112
4.3 Análisis del comportamiento del agente ... 114
Política del agente ... 114
Episodios que genera el agente ... 119
Análisis con TensorBoard ... 121
Comportamiento del agente en un entorno más complejo ... 124

4.4 Estimación de la función de transición y recompensas 125
A mitad de camino entre Model-Based y Model-Free 126
Implementación de un agente que usa estimaciones 127
Evaluación del agente que usa estimaciones 130
Calcular directamente la función Q 132

5 EVALUACIÓN DE POLÍTICAS CON MONTE CARLO 135

5.1 Métodos Monte Carlo 135
Monte Carlo versus programación dinámica 135
Tareas de predicción y control 137
First-visit versus *Every-visit* 138

5.2 Algoritmo de predicción Monte Carlo 139
Visión global 139
Pseudocódigo 139

5.3 Métodos de predicción Monte Carlo para la función V 141
Caso de estudio: *Blackjack* 141
Reglas del juego *blackjack* 142
Entorno *blackjack* de la librería Gym 143
Bucle de aprendizaje 145
Análisis de resultados 149

5.4 Métodos de predicción Monte Carlo para la función Q 153
Algoritmo 154
Bucle principal de aprendizaje 155
Análisis de resultados 156
Limitaciones del método 157

6 OBTENCIÓN DE POLÍTICAS ÓPTIMAS 159

6.1 Dilema exploración-explotación 159
Exploración versus explotación 159
Monte Carlo para la tarea de control 161
Políticas épsilon-*greedy* 163
Hiperparámetro *decay rate* 164

6.2 Método de control Monte Carlo constant-alpha 165
Media incremental 165
Constant-alpha 167

6.3 Implementación del método de control Monte Carlo 168
Selección de hiperparámetros 169
Algoritmo 171
Análisis de resultados 174

6.4 Aprendizaje por diferencia temporal: SARSA y Q-Learning 181
Diferencia temporal como suma de Monte Carlo y programación dinámica 181
SARSA 183
Q-Learning 185
On-policy versus *off-policy* 187

PARTE 3: APRENDIZAJE PROFUNDO 189

7 INTRODUCCIÓN AL APRENDIZAJE PROFUNDO 191

7.1 Redes neuronales 191
Ejemplo básico 191
Una neurona artificial simple 194
Perceptrón multicapa 198
Función de activación *softmax* 199

7.2 Proceso de aprendizaje de una red neuronal 204
Visión general 204
Proceso iterativo de aprendizaje de una red neuronal 206
Función de pérdida 208
Optimización: descenso del gradiente 209

7.3 Hiperparámetros en redes neuronales 211
Hiperparámetros básicos relacionados con el algoritmo de aprendizaje 212
Funciones de activación 214

7.4 Tipos de redes neuronales 215
Arquitectura de una red neuronal 215
Redes neuronales convolucionales 216
Capa convolucional 217

8 PYTORCH BÁSICO 223

8.1 Introducción a PyTorch 223
¿Qué es PyTorch? 224
Componentes principales de la librería PyTorch 225
Tensores en PyTorch 226

8.2 Programación de una red neuronal con PyTorch 229
Importar las librerías requeridas 230
Cargar los datos 231
Preprocesado de datos 232
Definir un modelo de red neuronal en PyTorch 232

8.3 Configuración del entrenamiento de una red neuronal en PyTorch 235
Función de pérdida 236
Optimizador 237

8.4 Entrenamiento de una red neuronal en PyTorch 238
Bucle de entrenamiento 238
Monitorizar el proceso de entrenamiento 239
Evaluación del modelo en PyTorch 241

PARTE 4: APRENDIZAJE POR REFUERZO PROFUNDO 243

9 MÉTODOS *VALUE-BASED*: DEEP Q-NETWORK 245

9.1 Q-Learning con redes neuronales 245
Métodos *value-based* vs *policy-based* 245
Caso de estudio: el juego de Atari Pong 246

Arquitectura de la red neuronal 251
Wrappers de la librería Gym 252

9.2 *Experience replay* y *target network* 256
Experience replay 257
Target network 259

9.3 Algoritmo Deep Q-Learning 260
Pseudocódigo 260
Hiperparámetros y tiempo de ejecución 262
Programación del agente 264

9.4 Entrenamiento de un modelo Deep Q-Learning 266
Inicializaciones 266
Bucle de entrenamiento 267
Ejecución del código 273
Uso del modelo 275

10 MÉTODOS *POLICY-BASED*: REINFORCE 279

10.1 Métodos *policy-based* 279
Obtener directamente la política óptima 279
Entorno Cart-Pole 281
Redes neuronales para métodos *policy-based* 283
Los métodos *policy-based* son *on-policy* 284
Métodos *policy-gradient* 285

10.2 Método REINFORCE 286
Definiciones matemáticas 287
Ascenso del gradiente 288
Estimación del gradiente 288
Pseudocódigo del método REINFORCE 290

10.3 Programación del método REINFORCE 292
Red neuronal 292
El bucle de entrenamiento 293

10.4 Limitaciones de los métodos *policy-based* 298
Alta variancia de los gradientes 298
Se requieren muchas interacciones con el entorno 298
Métodos *actor-critic* 299

11 *FRAMEWORKS* DE APRENDIZAJE POR REFUERZO: RAY+RLLIB 301

11.1 Aprendizaje por refuerzo profundo 301
Conocimientos básicos del aprendizaje por refuerzo profundo 302
Siguientes pasos a partir de este libro 302

11.2 *Frameworks* de aprendizaje por refuerzo 304
Open source 304
Principales *frameworks* 305

11.3 Aprendizaje por refuerzo escalable con Ray 306
Nuevos requerimientos de computación 306
Pila de capas de *software* del sistema 306

Ray 307
RLlib 308
11.4 Resolviendo el entorno Cart-Pole con RLlib 309
Inicializaciones 309
Entrenamiento del modelo 310
CLAUSURA: SOMOS RESPONSABLES DE NUESTROS PROGRAMAS 313
APÉNDICE: TERMINOLOGÍA BÁSICA 317
Aprendizaje por refuerzo 317
Aprendizaje profundo 319
AGRADECIMIENTOS 323