

Indice generale

Introduzione	xiii
A chi è rivolto questo libro	xiii
Argomenti trattati	xiv
Per ottenere il massimo da questo libro	xiv
Il codice	xiv
Immagini a colori	xv
Convenzioni utilizzate	xv
L'autore	xv
I revisori	xvi
Nota relativa all'interfaccia italiana di Amazon SageMaker	xvi
Parte I Introduzione ad Amazon SageMaker.....	1
Capitolo 1 Presentazione di Amazon SageMaker	3
Requisiti tecnici	3
Le funzionalità di Amazon SageMaker.....	4
Le principali funzionalità di Amazon SageMaker	4
L'API Amazon SageMaker	6
Configurazione di Amazon SageMaker sul computer locale.....	9
Una parola sulle autorizzazioni di AWS.....	12
Configurazione di Amazon SageMaker Studio.....	13
Accesso ad Amazon SageMaker Studio.....	13
Distribuzione di soluzioni e modelli one-click con Amazon SageMaker JumpStart.....	18
Distribuzione di una soluzione.....	19
Distribuzione di un modello	21
Ottimizzazione di un modello	24
Riepilogo	26

Capitolo 2 Tecniche di preparazione dei dati27

Requisiti tecnici	27
Etichettatura dei dati con Amazon SageMaker Ground Truth	28
Utilizzo della forza lavoro	28
Creazione di una forza lavoro privata	29
Caricamento dei dati per l'etichettatura	32
Creazione di un lavoro di etichettatura.....	32
Etichettatura di immagini	35
Etichettatura di testi.....	37
Trasformazione dei dati con Amazon SageMaker Data Wrangler.....	40
Caricamento di un dataset in SageMaker Data Wrangler	40
Trasformazione di un dataset in SageMaker Data Wrangler.....	46
Esportazione di una pipeline SageMaker Data Wrangler	49
Esecuzione di job batch con Amazon SageMaker Processing	51
L'API Amazon SageMaker Processing	51
Elaborazione di un dataset con scikit-learn.....	51
Elaborazione di un dataset con il vostro codice.....	57
Riepilogo	57

Parte II Realizzazione e training dei modelli59**Capitolo 3 AutoML con Amazon SageMaker Autopilot61**

Requisiti tecnici	62
Introduzione ad Amazon SageMaker Autopilot.....	62
Analisi dei dati.....	63
Ingegnerizzazione delle feature	63
Ottimizzazione del modello.....	64
Utilizzo di Amazon SageMaker Autopilot in SageMaker Studio.....	64
Lancio di un job	64
Monitoraggio di un job	68
Confronto dei job.....	70
Distribuzione e chiamata di un modello.....	73
Utilizzo di SageMaker SDK Autopilot.....	76
Lancio di un job	76
Monitoraggio di un job	77
Pulizia	78
Più in profondità in SageMaker Autopilot	78
Gli artefatti del job	79
Il notebook di esplorazione dei dati	79
Il notebook di generazione dei candidati.....	81
Riepilogo	84

Capitolo 4 Training di modelli di machine learning85

Requisiti tecnici	85
Introduzione agli algoritmi built-in di Amazon SageMaker.....	86

Apprendimento con supervisione.....	86
Apprendimento senza supervisione	87
Una parola sulla scalabilità.....	87
Training e distribuzione di modelli creati con algoritmi built-in	88
Il flusso di lavoro end-to-end	88
Flussi di lavoro alternativi.....	89
Utilizzo di un'infrastruttura completamente gestita.....	89
Utilizzo di SageMaker SDK con gli algoritmi built-in	91
Preparazione dei dati.....	91
Configurazione di un job di training.....	93
Lancio di un job di training	95
Distribuzione di un modello	96
Pulizia	97
Utilizzo di altri algoritmi built-in.....	97
Regressione con XGBoost.....	97
Una raccomandazione sull'uso di Factorization Machines	99
Utilizzo dell'algoritmo Principal Component Analysis	105
Rilevamento delle anomalie con Random Cut Forest	107
Riepilogo	112

Capitolo 5 Training di modelli di visione artificiale113

Requisiti tecnici	113
Introduzione agli algoritmi built-in di visione artificiale con Amazon SageMaker.....	114
L'algoritmo di classificazione delle immagini	114
L'algoritmo di rilevamento degli oggetti.....	114
L'algoritmo di segmentazione semantica.....	115
Training con algoritmi per la visione artificiale.....	116
Preparazione dei dataset di immagini.....	116
Utilizzo di file di immagini	117
Utilizzo di file RecordIO	122
Utilizzo di file SageMaker Ground Truth	126
Uso degli algoritmi per la visione artificiale built-in	128
Training di un modello per la classificazione delle immagini... 128	
Ottimizzazione di un modello per la classificazione delle immagini.....	132
Training di un modello per il rilevamento degli oggetti.....	133
Training di un modello di segmentazione semantica.....	136
Riepilogo	140

Capitolo 6 Training di modelli per l'elaborazione del linguaggio naturale141

Requisiti tecnici	141
Introduzione agli algoritmi built-in per l'elaborazione del linguaggio naturale in Amazon SageMaker	142

L'algoritmo BlazingText.....	142
L'algoritmo LDA	143
L'algoritmo NTM	143
L'algoritmo seq2seq	144
Training con algoritmi per l'elaborazione del linguaggio naturale.....	145
Preparazione dei dataset	145
Preparazione dei dati per la classificazione con BlazingText.....	145
Preparazione dei dati per la classificazione con BlazingText, versione 2	148
Preparazione dei dati per vettori di parole con BlazingText	151
Preparazione dei dati per la modellazione degli argomenti con gli algoritmi LDA e NTM.....	152
Utilizzo di dataset etichettati con SageMaker Ground Truth ...	157
Utilizzo degli algoritmi built-in per l'elaborazione del linguaggio naturale.....	158
Classificazione dei testi con BlazingText.....	158
Calcolare i vettori di parole con BlazingText.....	160
Utilizzo di modelli BlazingText con FastText	160
Modellazione degli argomenti con l'algoritmo LDA.....	162
Modellazione degli argomenti con l'algoritmo NTM.....	165
Riepilogo	168

Capitolo 7 Estensione dei servizi tramite i framework

built-in	169
Requisiti tecnici	169
Introduzione ai framework built-in di Amazon SageMaker	170
Esecuzione di un primo esempio con XGBoost.....	171
Utilizzo dei container del framework	173
Training e distribuzione a livello locale	174
Training in modalità script.....	175
La distribuzione del modello.....	176
Gestire le dipendenze.....	178
In conclusione	179
Esecuzione del codice del framework su Amazon SageMaker	180
Implementazione della modalità script	180
Test in locale.....	181
Utilizzo della modalità locale	182
Utilizzo di un'infrastruttura gestita	183
Utilizzo dei framework built-in.....	183
Utilizzo di TensorFlow e Keras.....	183
Utilizzo di PyTorch	186
Utilizzo di Hugging Face.....	188
Utilizzo di Apache Spark	194
Riepilogo	199

Capitolo 8 Usare i vostri algoritmi e il vostro codice.....201

Requisiti tecnici	201
In quale modo SageMaker richiama il vostro codice	202
Il layout dei file all'interno di un container SageMaker.....	202
Le opzioni per il training personalizzato	203
Le opzioni per la distribuzione personalizzata.....	203
Personalizzazione di un container del framework preesistente	204
Configurazione dell'ambiente su EC2.....	204
Creazione di container di training e inferenza	205
Utilizzo di SageMaker Training Toolkit con scikit-learn.....	207
Creazione di un container completamente personalizzato per scikit-learn.....	209
Training con un container completamente personalizzato	209
Distribuzione di un container completamente personalizzato	210
Creazione di un container completamente personalizzato per R.....	213
Codifica con R e plumber	213
Creazione di un container personalizzato	214
Training e distribuzione di un container personalizzato su SageMaker	215
Training e distribuzione con il vostro codice in MLflow.....	216
Installazione di MLflow	216
Training di un modello con MLflow	216
Creazione di un container SageMaker con MLflow.....	218
Creazione di un container completamente personalizzato per SageMaker Processing.....	221
Riepilogo	222

Parte III Approfondimenti sul training.....223**Capitolo 9 Dimensionare i job di training.....225**

Requisiti tecnici	225
Capire quando e come dimensionare le cose	226
Che cosa significa ridimensionare	226
Adattare i tempi di training ai requisiti operativi.....	226
Corretto dimensionamento dell'infrastruttura di training.....	227
Decidere quando ridimensionare	228
Decidere come ridimensionare	228
Ridimensionamento di un job di training BlazingText	230
Monitoraggio e profiling dei job di training con Amazon SageMaker Debugger.....	232
Visualizzazione delle informazioni di monitoraggio e profiling in SageMaker Studio.....	232

Abilitazione del profiling in SageMaker Debugger.....	234
Risolvere i problemi di training	236
Streaming di dataset con la modalità pipe	238
Utilizzo della modalità pipe con gli algoritmi built-in	238
Utilizzo della modalità pipe con altri algoritmi e framework	239
Semplificazione del caricamento dei dati con MLIO	239
Training di macchine di fattorizzazione con la modalità pipe	240
Distribuzione dei job di training	241
Parallelismo dei dati e parallelismo del modello	241
Distribuzione del training per gli algoritmi built-in	241
Distribuzione del training per i framework built-in	241
Distribuzione del training per container personalizzati	242
Ridimensionamento di un modello per la classificazione delle immagini su ImageNet	242
Preparazione del dataset ImageNet.....	242
Definizione del job di training	244
Training su ImageNet.....	245
Aggiornamento della dimensione del batch.....	246
Aggiunta di altre istanze.....	247
Riassumendo.....	247
Training con i dati di SageMaker e le librerie di parallelismo	248
Training su TensorFlow con SageMaker DDP	249
Training su Hugging Face con SageMaker DDP	251
Training su Hugging Face con SageMaker DMP.....	251
Utilizzo di altri servizi di storage	252
Utilizzo di SageMaker e Amazon EFS.....	252
Utilizzo di SageMaker e Amazon FSx for Lustre	256
Riepilogo	258

Capitolo 10 Tecniche avanzate di training259

Requisiti tecnici	259
Ottimizzazione dei costi con il training spot gestito.....	260
Confronto dei costi.....	260
Le istanze spot Amazon EC2.....	261
Il training spot gestito	262
Utilizzo del training spot gestito con rilevamento degli oggetti.....	263
Utilizzo del training spot gestito e del checkpointing con Keras	264
Perfezionamento degli iperparametri con l'ottimizzazione automatica del modello.....	267
Che cos'è l'ottimizzazione automatica del modello	268
Utilizzo dell'ottimizzazione automatica del modello con il rilevamento degli oggetti.....	268

Utilizzo dell'ottimizzazione automatica del modello con Keras	271
Utilizzo dell'ottimizzazione automatica del modello per la ricerca dell'architettura	275
Esplorazione dei modelli con SageMaker Debugger	276
Debug di un job XGBoost.....	276
Ispezione di un job XGBoost.....	277
Debug e ispezione di un job Keras.....	279
Gestione delle feature e creazione di dataset con SageMaker Feature Store	282
Ingegnerizzazione delle feature con SageMaker Processing.....	283
Creazione di un gruppo di feature	284
Acquisizione delle feature	286
Esecuzione di query sulle feature per creare un dataset	286
Altre funzionalità di SageMaker Feature Store	287
Rilevamento delle distorsioni (bias) nei dataset e spiegazione delle predizioni con SageMaker Clarify	287
Configurazione di un'analisi del bias con SageMaker Clarify.....	288
Esecuzione di un'analisi del bias.....	289
Analisi delle metriche di bias.....	289
Esecuzione di un'analisi della spiegabilità	290
Contenimento del bias.....	291
Riepilogo	294

Parte IV Gestione dei modelli in produzione.....295

Capitolo 11 Distribuzione dei modelli di machine learning297

Requisiti tecnici	298
Gli artefatti dei modelli e l'esportazione di modelli.....	298
Esame ed esportazione di modelli built-in.....	298
Esame ed esportazione di modelli built-in di visione artificiale.....	300
Esame ed esportazione di modelli XGBoost.....	301
Esame ed esportazione di modelli scikit-learn	302
Esame ed esportazione di modelli TensorFlow.....	302
Esame ed esportazione di modelli Hugging Face.....	302
Distribuzione di modelli su endpoint real-time	303
Gestione degli endpoint con SageMaker SDK.....	304
Gestione degli endpoint con l'SDK boto3.....	308
Distribuzione di modelli su transformer batch	312
Distribuzione di modelli su pipeline di inferenza	313
Monitoraggio della qualità delle predizioni con Amazon SageMaker Model Monitor.....	314
Cattura dei dati.....	315
Creazione di un valore base	316

Impostazione di uno scheduling di monitoraggio	318
Invio di dati errati.....	318
Esame dei report delle violazioni	319
Distribuzione di modelli in servizi a container.....	320
Training su SageMaker e distribuzione su Amazon Fargate	321
Riepilogo	326

Capitolo 12 Automazione dei flussi di lavoro di machine learning327

Requisiti tecnici	328
Automazione con AWS CloudFormation.....	328
Scrittura di un template	329
Distribuzione di un modello su un endpoint real-time	330
Modifica di uno stack con un set di modifica	333
Aggiunta di una seconda variante di produzione all'endpoint	335
Implementazione della distribuzione canary	337
Implementazione della distribuzione blue-green	341
Automazione con AWS CDK	342
Installazione di CDK	342
Creazione di un'applicazione CDK.....	342
Scrittura di un'applicazione CDK	343
Distribuzione di un'applicazione CDK	345
Creazione di flussi di lavoro end-to-end con le Step Functions AWS.....	346
Configurazione delle autorizzazioni	346
Implementazione di un primo flusso di lavoro.....	347
Aggiunta dell'esecuzione parallela a un flusso di lavoro.....	351
Aggiunta di una funzione Lambda a un flusso di lavoro	353
Creazione di flussi di lavoro end-to-end con Amazon SageMaker Pipelines	356
Definizione dei parametri del flusso di lavoro	358
Elaborazione del dataset con SageMaker Processing.....	359
Acquisizione del dataset in SageMaker Feature Store con SageMaker Processing.....	360
Creazione di un dataset con Amazon Athena e SageMaker Processing.....	361
Training di un modello.....	361
Creazione e registrazione di un modello in SageMaker Pipelines.....	362
Creazione di una pipeline	363
Esecuzione di una pipeline.....	364
Distribuzione di un modello dal registro modelli.....	365
Riepilogo	367

Capitolo 13	Ottimizzazione dei costi e delle prestazioni delle predizioni	369
	Requisiti tecnici	369
	Autoscaling di un endpoint	370
	Distribuzione di un endpoint multi-modello	373
	Che cosa sono degli endpoint multi-modello	373
	Creazione di un endpoint multi-modello con scikit-learn	374
	Distribuzione di un modello con Amazon Elastic Inference	377
	Distribuzione di un modello con Amazon Elastic Inference	378
	Compilazione di modelli con Amazon SageMaker Neo	380
	Come funziona Amazon SageMaker Neo	381
	Compilazione e distribuzione di un modello per la classificazione delle immagini su SageMaker	382
	I modelli compilati con Neo	383
	Distribuzione di un modello per la classificazione delle immagini su un Raspberry PI	383
	Distribuzione di modelli su AWS Inferentia	385
	Creazione di una checklist per l'ottimizzazione dei costi	386
	Ottimizzazione dei costi di preparazione dei dati	386
	Ottimizzazione dei costi di sperimentazione	387
	Ottimizzazione dei costi di training del modello	388
	Ottimizzazione dei costi di distribuzione del modello	389
	Riepilogo	390
Indice analitico		391