

Ciencia de datos desde cero

Principios básicos con Python

2.^a Edición

Joel Grus



Índice

Agradecimientos

Sobre el autor

Prefacio a la segunda edición

Convenciones empleadas en este libro

Uso del código de ejemplo

Sobre la imagen de cubierta

Prefacio a la primera edición

Ciencia de datos o data science

Partir de cero

1. Introducción

El ascenso de los datos

¿Qué es la ciencia de datos o data science?

Hipótesis motivadora: DataSciencester

Localizar los conectores clave

Científicos de datos que podría conocer

Salarios y experiencia

Cuentas de pago

Temas de interés

Sigamos adelante

2. Un curso acelerado de Python

El zen de Python

Conseguir Python

- El modelo
- Utilizar descenso de gradiente
- Estimación por máxima verosimilitud
- Para saber más

15. Regresión múltiple

- El modelo
- Otros supuestos del modelo de mínimos cuadrados
- Ajustar el modelo
- Interpretar el modelo
- Bondad de ajuste
- Digresión: el bootstrap
- Errores estándares de coeficientes de regresión
- Regularización
- Para saber más

16. Regresión logística

- El problema
- La función logística
- Aplicar el modelo
- Bondad de ajuste
- Máquinas de vectores de soporte
- Para saber más

17. Árboles de decisión

- ¿Qué es un árbol de decisión?
- Entropía
- La entropía de una partición
- Crear un árbol de decisión
- Ahora, a combinarlo todo
- Bosques aleatorios
- Para saber más

18. Redes neuronales

Perceptrones
Redes neuronales prealimentadas
Retropropagación
Ejemplo: Fizz Buzz
Para saber más

19. Deep learning (aprendizaje profundo)

El tensor
La capa de abstracción
La capa lineal
Redes neuronales como una secuencia de capas
Pérdida y optimización
Ejemplo: XOR revisada
Otras funciones de activación
Ejemplo: FizzBuzz revisado
Funciones softmax y entropía cruzada
Dropout
Ejemplo: MNIST
Guardar y cargar modelos
Para saber más

20. Agrupamiento (clustering)

La idea
El modelo
Ejemplo: Encuentros
Elegiendo k
Ejemplo: agrupando colores
Agrupamiento jerárquico de abajo a arriba
Para saber más

21. Procesamiento del lenguaje natural

Nubes de palabras
Modelos de lenguaje n-Gram

Gramáticas

Un inciso: muestreo de Gibbs

Modelos de temas

Vectores de palabras

Redes neuronales recurrentes

Ejemplo: utilizar una RNN a nivel de carácter

Para saber más

22. Análisis de redes

Centralidad de intermediación

Centralidad de vector propio

Multiplicación de matrices

Centralidad

Grafos dirigidos y PageRank

Para saber más

23. Sistemas recomendadores

Método manual

Recomendar lo que es popular

Filtrado colaborativo basado en usuarios

Filtrado colaborativo basado en artículos

Factorización de matrices

Para saber más

24. Bases de datos y SQL

CREATE TABLE e INSERT

UPDATE

DELETE

SELECT

GROUP BY

ORDER BY

JOIN

Subconsultas

- Índices
- Optimización de consultas
- NoSQL
- Para saber más

25. MapReduce

- Ejemplo: Recuento de palabras
- ¿Por qué MapReduce?
- MapReduce, más general
- Ejemplo: Analizar actualizaciones de estado
- Ejemplo: Multiplicación de matrices
- Un inciso: Combinadores
- Para saber más

26. La ética de los datos

- ¿Qué es la ética de los datos?
- No, ahora en serio, ¿qué es la ética de datos?
- ¿Debo preocuparme de la ética de los datos?
- Crear productos de datos de mala calidad
- Compromiso entre precisión e imparcialidad
- Colaboración
- Capacidad de interpretación
- Recomendaciones
- Datos sesgados
- Protección de datos
- En resumen
- Para saber más

27. Sigamos haciendo ciencia de datos

- IPython
- Matemáticas
- No desde cero
 - NumPy

bien la forma en la que yo hackeo cosas, que no tiene por qué ser necesariamente la suya. También conocerán bastante bien algunas de las herramientas que utilizo, que no han de ser obligadamente las mejores para ellos. Y entenderán bien el modo en que yo abordo los problemas de datos, que tampoco tiene por qué ser el mejor modo para ellos. La intención (y la esperanza) es que mis ejemplos les inspiren a probar las cosas a su manera. Todo el código y los datos del libro están disponibles en GitHub³ para que puedan ponerse manos a la obra.

De forma similar, la mejor manera de aprender matemáticas es haciendo matemáticas. Este no es rotundamente un libro de mates, y en su mayor parte no estaremos “haciendo matemáticas”. Sin embargo, no se puede hacer ciencia de datos de verdad sin ciertos conocimientos de probabilidad, estadística y álgebra lineal. Esto significa que, donde corresponda, profundizaremos en ecuaciones matemáticas, intuición matemática, axiomas matemáticos y versiones caricaturizadas de grandes ideas matemáticas. Espero que los lectores no teman sumergirse conmigo.

A lo largo de todo el libro también espero dar a entender que jugar con datos es divertido porque, bueno, ¡jugar con datos realmente lo es! (especialmente si lo comparamos con algunas alternativas, como hacer la declaración de la renta o trabajar en una mina).

Partir de cero

Hay muchísimas librerías de ciencia de datos, *frameworks*, módulos y kits de herramientas que implementan de forma eficaz los algoritmos y las técnicas de ciencia de datos más conocidas (así como las menos habituales). Si alguno de mis lectores llega a ser científico de datos, acabará estando íntimamente familiarizado con NumPy, scikit-learn, pandas y todas las demás librerías existentes. Son fabulosas para hacer ciencia de datos, pero también suponen una buena forma de empezar a hacer ciencia de datos sin realmente comprender lo que es.

En este libro nos acercaremos a la ciencia de datos desde el principio de los principios. Esto significa que crearemos herramientas e implementaremos