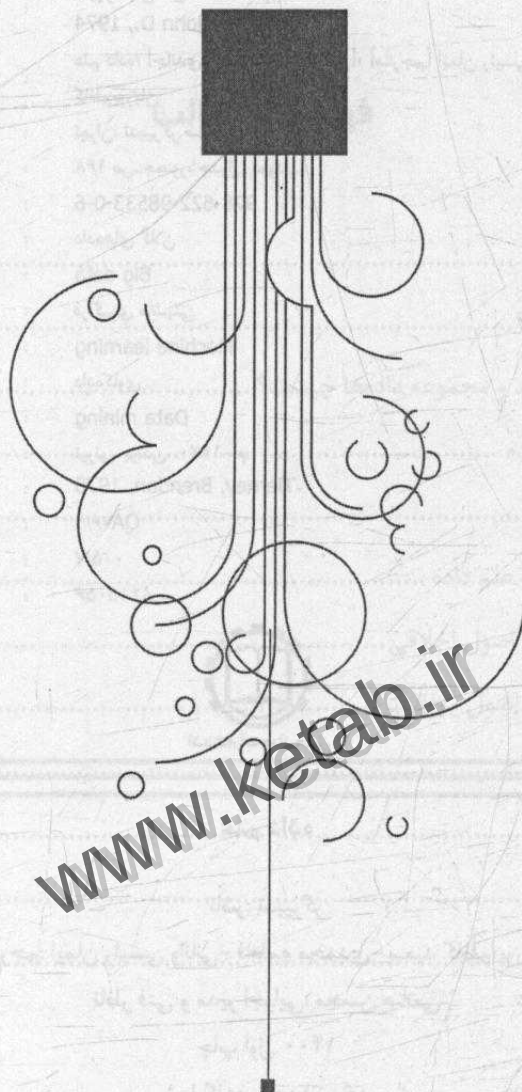




علم داده

ایمان رئیسی، فاطمه محمدی، سعید کاظم پوریان

سرشناسه	:	کلیهر، جان دی، ۱۹۷۴-م. Kelleher, John D., 1974
عنوان و نام پدیدآور	:	علم داده / [جان دی. کلیهر، برندن تیرنی]؛ [مترجم] ایمان رئیسی و انانی، فاطمه محمدی، سعید کاظم پوریان
مشخصات نشر	:	تهران: تدبیرگر صناعی، ۱۴۰۰.
مشخصات ظاهری	:	۱۶۸ ص.: مصور، جدول، نمودار.
شابک	:	978-622-98533-0-6
موضوع	:	داده‌های کلان
موضوع	:	Big data
موضوع	:	فراگیری ماشینی
موضوع	:	Machine learning
موضوع	:	داده کاوی
موضوع	:	Data mining
شناسه افزوده	:	تیرنی، برندن، ۱۹۷۰-م.
شناسه افزوده	:	-Tierney, Brendan, 1970
رده بندی کنگره	:	QAY۶/۹
رده بندی دیویی	:	۰۰۵/۷
شماره کتابشناسی ملی	:	۸۴۱۵۰۵۶



ادب و نشر صنعتی

علم داده

ناشر: تدبیرگر

مترجم: ایمان رئیسی و انانی- فاطمه محمدی- سعید کاظم پوریان

ناظر فنی و مدیر اجرایی: محسن صناعی

چاپ اول ۱۴۰۰

شمارگان: ۱۰۰۰ نسخه

قیمت: ۵۴۰۰۰ تومان

کلیه حقوق مادی و معنوی برای ناشر محفوظ است.

آدرس فروشگاه: تهران، خیابان انقلاب، بین ۱۲ فروردین و منیری جاوید، کتابفروشی صناعی، پ ۱۳۱۲

تلفن: ۶۶۴۰۵۳۸۵، ۶۶۴۰۹۹۲۴

آدرس دفتر: تهران، خیابان انقلاب، مابین ۱۲ فروردین و منیری جاوید، کتابسرای اندیشه، طبقه اول

تلفن: ۶۶۴۰۸۱۶۵، ۶۶۴۶۲۱۷۷

فروشگاه اینترنتی: www.saneibook.com

فهرست مطالب

پیشگفتار.....	۵
۱. علم داده چیست؟.....	۷
۲. داده‌ها چه هستند و مجموعه داده‌ها چیست؟.....	۲۹
۳. زیست‌بوم علم داده.....	۴۵
۴. یادگیری ماشین.....	۶۱
۵. کارهای استاندارد علم داده.....	۹۵
۶. حریم خصوصی و اصول اخلاقی.....	۱۱۱
۷. روندهای آینده و اصول موفقیت.....	۱۳۳
واژه‌نامه.....	۱۴۳
یادداشت‌ها.....	۱۵۳
مطالعه بیشتر.....	۱۵۹
منابع.....	۱۶۱

پیشگفتار

هدف علم داده، بهبود تصمیم‌گیری از طریق قرار دادن آن‌ها بر مبنای بینشی است که از مجموعه‌های بزرگ داده‌ها استخراج شده است. علم داده به عنوان یک حوزه فعالیت، دربرگیرنده مجموعه‌ای از اصول، تعاریف مسئله، الگوریتم‌ها و فرآیندهایی برای استخراج الگوهای نامشهود و مفید از مجموعه‌های بزرگ داده‌ها است. این علم، ارتباط نزدیکی با حوزه‌های داده‌کاوی و یادگیری ماشین دارد، اما قلمرو گسترده‌تری را شامل می‌شود. امروزه تقریباً در تمامی بخش‌های جوامع مدرن، تصمیم‌گیری‌ها توسط علم داده هدایت می‌شوند. علم داده ممکن است از راه‌های مختلفی بر زندگی روزمره شما تاثیر بگذارد: تعیین تبلیغات آنلاین مناسب برای شما، پیشنهادهای دوستی، توصیه فیلم و کتاب، ایمیل‌های قرارگرفته در پوشه هرنامه شما، پیشنهادهای ارائه‌شده هنگام تجدید خدمات تلفن همراه، هزینه حق بیمه سلامت، تعیین توالی و زمانبندی چراغ‌های راهنما در منطقه شما، نحوه طراحی داروهای مورد نیاز شما و مکان‌های هدف‌گذاری‌شده توسط پلیس در شهر شما.

رشد استفاده از علم داده در جوامع ما، برگرفته از ظهور کلان‌داده^۱ و رسانه‌های اجتماعی، افزایش قدرت محاسباتی، کاهش بسیار در هزینه حافظه کامپیوتری و توسعه روش‌های قدرتمندتر برای تحلیل داده‌ها و مدل‌سازی (مانند یادگیری عمیق) است. روی هم رفته، این عوامل نشان می‌دهند که جمع‌آوری، ذخیره و پردازش داده‌ها هرگز کار آسانی برای سازمان‌ها نبوده است. در عین حال، این نوآوری‌های فنی و کاربرد گسترده‌تر علم داده، فشار ناشی از چالش‌های اخلاقی مرتبط با استفاده از داده‌ها و حریم خصوصی فردی را بیش از گذشته کرده است. هدف این کتاب، ارائه مقدمه‌ای بر علم داده است که عناصر ضروری این حوزه را تا حدی پوشش می‌دهد که درکی اصولی از این حوزه ایجاد کند.

فصل اول، حوزه علم داده را معرفی می‌کند و تاریخچه مختصری از چگونگی توسعه و تکامل آن ارائه می‌دهد. در این فصل، علت اهمیت علم داده و برخی از عوامل موثر بر مقبولیت آن نیز مورد بررسی قرار می‌گیرند. این فصل با مرور و رد برخی از باورهای غلط مرتبط با علم داده به پایان می‌رسد. فصل دوم مفاهیم بنیادین مربوط به داده‌ها را معرفی می‌کند. هم‌چنین این فصل، مراحل استاندارد یک پروژه علم داده را تشریح می‌کند: درک کسب‌وکار، درک داده‌ها، آماده‌سازی داده‌ها، مدل‌سازی، ارزشیابی و استقرار. فصل سوم بر زیرساخت داده‌ها و چالش‌های ناشی از کلان‌داده و ادغام داده‌های چندین منبع متمرکز است. یکی از ابعاد چالش‌برانگیز یک زیرساخت معمول برای

داده‌ها، این است که داده‌های موجود در پایگاه‌های داده و انبارهای داده‌ها^۱ اغلب در سرورهایی متفاوت از سرورهای مورد استفاده برای تحلیل داده‌ها قرار دارند. در نتیجه، هنگامی که با مجموعه‌های بزرگ داده‌ها سروکار داریم، ممکن است زمان قابل توجهی صرف جابجایی داده‌ها میان سرورهایی که پایگاه داده یا انبار داده‌ها روی آن قرار دارد و سرورهای مورد استفاده برای تحلیل داده‌ها و یادگیری ماشین شود. فصل سوم با توصیفی از یک زیرساخت معمول علم داده برای یک سازمان و برخی از راهکارهای نوظهور برای چالش جابجایی مجموعه‌های بزرگ داده‌ها در یک زیرساخت داده‌ها آغاز می‌شود. این موضوع شامل استفاده از یادگیری ماشین در پایگاه داده، استفاده از هدوپ^۲ برای ذخیره‌سازی و پردازش داده‌ها و توسعه سیستم‌های ترکیبی پایگاه‌های داده می‌شود که نرم‌افزار پایگاه داده سنتی و راه‌حلهایی مانند هدوپ را به طور یکپارچه ترکیب می‌کنند. این فصل با تأکید بر برخی چالش‌ها در یکپارچه‌سازی داده‌های سرتاسر یک سازمان در قالب نمایشی یکپارچه که برای یادگیری ماشین مناسب است، به اتمام می‌رسد. فصل چهارم، حوزه یادگیری ماشین را معرفی می‌کند و برخی از معروف‌ترین الگوریتم‌ها و مدل‌های یادگیری ماشین^۳ از جمله شبکه‌های عصبی، یادگیری عمیق و مدل‌های درخت تصمیم را توضیح می‌دهد. فصل پنجم بر پیوند تخصص یادگیری ماشین با مسائل دنیای واقعی از طریق مرور طیفی از مسائل استاندارد کسب و کار، تشریح چگونگی امکان حل آن‌ها از طریق راهکارهای یادگیری ماشین تمرکز دارد. فصل ششم به مرور مفاهیم اخلاقی علم داده، پیشرفت‌های اخیر در قواعد داده‌ها و برخی از رویکردهای محاسباتی جدید جهت حفظ حریم خصوصی افراد در فرآیند علم داده می‌پردازد. نهایتاً فصل هفتم، برخی حوزه‌ها را تشریح می‌کند که علم داده، اثر قابل توجهی در آینده نزدیک بر آن‌ها خواهد گذاشت و برخی از اصول مهم در تعیین موفقیت یک پروژه علم داده را بیان می‌کند.