



مقدمه‌ای بر داده کاوی

مؤلفین:
علی مشکاتی
عبدالرضا ناظمی

سرشناسه:	مشکاتی، علی، ۱۳۱۸ -
عنوان و نام پدیدآور:	مقدمه‌ای بر داده کاوی / مؤلفین علی مشکاتی، عبدالرضا ناظمی.
مشخصات نشر:	مشهد: علی مشکاتی، ۱۳۸۸.
مشخصات ظاهری:	۴۵۶ ص.: مصور.
شابک:	۷۵۰۰۰ ریال (ISBN: 978-964-04-3243-3)
وضعیت فهرست‌نویسی:	فیا.
موضوع:	داده کاوی.
شناسه افزوده:	ناظمی، عبدالرضا.
رده‌بندی کنگره:	۱۳۸۷ م ۲۹ / ۹۶۷۶ QA
رده‌بندی دیویی:	۰۰۶/۳۱۲
شماره کتابخانه ملی:	۱۶۷۰۳۶۵

مقدمه‌ای بر داده کاوی

تألیف

دکتر علی مشکاتی - عبدالرضا ناظمی

ناشر: علی مشکاتی، با همکاری دانشگاه آزاد اسلامی واحد نیشابور

وزیری، ۴۵۶ صفحه، ۱۰۰۰ نسخه، چاپ اول، بهار ۱۳۸۸

امور چاپ: مؤسسه چاپ و انتشارات دانشگاه فردوسی مشهد

بها: ۷۵۰۰۰ ریال

فهرست مطالب

۱۳	فصل اول: داده کاوی و مفاهیم آن
۱۴	۱-۱ مروری بر مفاهیم پایه
۱۵	۲-۱ داده کاوی
۱۹	۱-۳-۱ آمار و داده کاوی
۲۲	۲-۳-۱ آمار و یادگیری ماشین
۲۳	۱-۴-۱ بیان دیدگاه‌ها
۲۴	۲-۴-۱ دیدگاه دیگر
۲۷	۱-۵-۱ پایگاه داده‌ها یا بانک اطلاعاتی چیست؟
۲۹	۲-۵-۱ چرا باید از پایگاه داده‌ها استفاده کنیم؟
۳۰	۳-۵-۱ پیش پردازش داده‌ها
۳۱	۴-۵-۱ انبارداده چیست؟
۳۱	۱-۴-۵-۱ تعاریف انبارداده
۳۲	۲-۴-۵-۱ چهار خصوصیت اصلی انبارداده
۳۴	۳-۴-۵-۱ تفاوت انبارداده و پایگاه داده‌ها
۳۶	۴-۴-۵-۱ فعالیت‌های انبارداده
۳۷	۵-۵-۱ ساختار یک سیستم داده کاوی
۴۱	۶-۵-۱ وظایف و محدودیتهای داده کاوی
۴۱	۶-۱ نرم افزارهای داده کاوی
۴۲	۱-۷-۱ معرفی متن کاوی

۴۳	۲-۷-۱ تعاریف
۴۵	۳-۷-۱ کاربردهای متن کاوی
۴۸	۴-۷-۱ فرآیند متن کاوی
۵۰	۸-۱ روشهای متن کاوی
۵۰	۹-۱ یافتن روابط
۵۲	۱۰-۱- سؤالات و مسائل مروری
۵۲	۱۱-۱ منابع و مراجع برای مطالعه بیشتر
۵۵	فصل دوم: آماده سازی داده ها
۵۶	۱-۲ نمایش داده های خام
۶۱	۲-۲ مشخصه های داده های خام
۶۳	۳-۲ تبدیل داده های خام
۶۷	۴-۲ داده های گمشده
۷۰	۵-۲ داده های وابسته به زمان
۷۵	۶-۲ تحلیل نقاط دورافتاده
۷۹	۷-۲ سؤالات و مسائل مروری
۸۲	۸-۲ منابع و مراجع برای مطالعه بیشتر
۸۳	فصل سوم: تحویل داده ها
۸۴	۱-۳ ابعاد مجموعه های داده های حجیم
۸۷	۲-۳ تحویل صفات
۹۵	۳-۳ معیار آنترپی برای رتبه بندی صفات
۹۸	۴-۳ تحلیل مؤلفه اصلی
۱۰۱	۵-۳ تحویل مقادیر
۱۰۶	۶-۳ گسسته سازی صفت: فن کای مرچ
۱۱۱	۷-۳ تحویل موردها
۱۱۵	۸-۳ سؤالات و مسائل مروری

۱۱۷	۹-۳ منابع و مراجع برای مطالعه بیشتر
۱۱۹	فصل چهارم: آموختن از داده ها
۱۲۱	۱-۴ یادگیری ماشین
۱۲۸	۲-۴ نظریه یادگیری آماری
۱۳۶	۳-۴ انواع روشهای یادگیری
۱۳۸	۴-۴ وظایف مشترک یادگیری
۱۴۶	۵-۴ برآورد مدل
۱۵۳	۶-۴ سؤالات و مسائل مروری
۱۵۴	۷-۴ منابع و مراجع برای مطالعه بیشتر
۱۵۷	فصل پنجم: روشهای آماری
۱۵۸	۱-۵ استنباط آماری
۱۶۱	۲-۵ ارزیابی تفاوتها در مجموعه‌های داده‌ها
۱۶۴	۳-۵ استنباط بیزی
۱۶۸	۴-۵ رگرسیون پیشگو
۱۷۴	۵-۵ تحلیل واریانس
۱۷۷	۶-۵ رگرسیون لجستیکی
۱۷۹	۷-۵ مدل‌های لگ-خطی
۱۸۴	۸-۵ تحلیل تشخیص خطی
۱۸۷	۹-۵ سؤالات و مسائل مروری
۱۸۹	۱۰-۵ مراجع و منابع برای مطالعه بیشتر
۱۹۱	فصل ششم: تحلیل خوشه ای
۱۹۲	۱-۶ تعاریف
۱۹۶	۱-۱-۶ پیشنهادها
۱۹۷	۲-۱-۶ مشکلات

۲۰۱	۲-۶ ماتریس الگو
۲۰۲	۳-۶ ماتریس نزدیکی (مشابهت)
۲۰۳	۱-۳-۶ اندازه نزدیکی
۲۰۶	۴-۶ انواع رده‌بندی
۲۰۷	۱-۴-۶ رده‌بندی جدا از هم و رده‌بندی متداخل
۲۰۸	۲-۴-۶ رده‌بندی بدون ناظر و با ناظر
۲۰۸	۳-۴-۶ سلسله مراتبی و افرازی
۲۰۹	۵-۶ خوشه‌بندی سلسله مراتبی
۲۱۰	۱-۵-۶ دندروگرام
۲۱۴	۲-۵-۶ پیچیدگی زمانی و فضا
۲۱۴	۳-۵-۶ خوشه‌بندی سلسله مراتبی تجمعی اتصال تک
۲۱۶	۴-۵-۶ خوشه‌بندی سلسله مراتبی تجمعی اتصال کامل
۲۱۷	۵-۵-۶ خوشه‌بندی وارد (Ward) سلسله مراتبی تجمعی
۲۱۸	۶-۶ خوشه‌بندی افرازی
۲۲۱	۱-۶-۶ معیار خوشه‌بندی
۲۲۲	۲-۶-۶ کلیات خوشه‌بندی افرازی
۲۲۳	۳-۶-۶ پیچیدگی زمانی و فضا
۲۲۳	۴-۶-۶ اولین افراز (گام اول)
۲۲۳	۵-۶-۶ به روز کردن افراز (گام دوم)
۲۲۴	۶-۶-۶ محاسبه مرکز خوشه‌های جدید (گام سوم)
۲۲۶	۷-۶-۶ تعدیل تعداد خوشه‌ها (گام پنجم)
۲۲۷	۸-۶-۶ همگرایی (گام چهارم)
۲۲۹	۷-۶ سؤالات و مسائل مروری
۲۳۰	۸-۶ منابع و مراجع برای مطالعه بیشتر
۲۳۳	فصل هفتم: درخت تصمیم
۲۳۴	۱-۷ درخت تصمیم

۲۳۵	۲-۷ مثال ساده اعتبارسنجی مشتریان
۲۳۹	۳-۷ ساختن درخت‌های رده بندی
۲۳۹	۱-۳-۷ معیارهای ناخالصی و کیفیت یک افراز
۲۴۱	۱-۳-۷ خطای جانشینی مجدد
۲۴۳	۲-۱-۳-۷ شاخص جینی
۲۴۴	۳-۱-۳-۷ آنتروپی
۲۴۴	۴-۷ مجموعه افرازهای در نظر گرفته شده
۲۴۵	۱-۴-۷ افرازهای مربوط به متغیرهای عددی
۲۴۶	۲-۴-۷ افرازهای مربوط به متغیرهای رسته‌ای
۲۴۹	۵-۷ الگوریتم ساختن درخت
۲۵۱	۶-۷ هرس کردن درخت رده بندی
۲۵۱	۱-۶-۷ روش پیش هرس کردن
۲۵۱	۲-۶-۷ روش پس هرس کردن
۲۵۳	۳-۶-۷ هرس کردن با معیار نرخ پیچیدگی
۲۶۰	۷-۷ انتخاب درخت نهایی
۲۶۰	۱-۷-۷ آموزش و آزمون
۲۶۱	۲-۵-۷ اعتبارسنجی متقابل
۲۶۲	۸-۷ بررسی داده‌های گمشده در درخت‌های تصمیم
۲۶۵	۹-۷ استخراج قواعد رده بندی حاصل از درخت‌های تصمیم
۲۶۶	۱۰-۷ سؤالات و مسائل مروری
۲۶۸	۱۱-۷ مراجع و منابع برای مطالعه بیشتر
۲۶۹	فصل هشتم: کاوش قواعد پیوند در پایگاه داده‌های حجیم
۲۷۰	۱-۸ بازاریابی و فروش
۲۷۲	۲-۸ قواعد پیوند
۲۷۴	۱-۲-۸ تعاریف و مفاهیم
۲۷۵	۲-۲-۸ نمایش داده‌ها برای تحلیل سبد خرید

۲۷۶	۳-۲-۸ میزان پشتیبانی، میزان اطمینان و مجموعه کالاهای مکرر
۲۷۷	۱-۳-۲-۸ میزان پشتیبانی
۲۷۷	۲-۳-۲-۸ میزان اطمینان
۲۷۸	۳-۳-۲-۸ مجموعه‌های کالاها
۲۷۹	۴-۳-۲-۸ مجموعه کالاهای مکرر
۲۸۰	۳-۸ کاوش قواعد پیوند
۲۸۰	۱-۳-۸ دیدگاه دو مرحله‌ای two-step
۲۸۱	۲-۳-۸ دیدگاه Brute-Force
۲۸۳	۳-۳-۸ الگوریتم Apriori
۲۸۳	۱-۳-۳-۸ Apriori مشخصه
۲۸۳	۲-۳-۳-۸ الگوریتم Apriori چگونه کار می‌کند؟ (قسمت اول)
۲۸۶	۳-۳-۳-۸ الگوریتم Apriori چگونه کار می‌کند؟ (قسمت دوم)
۲۹۱	۴-۳-۸ قواعد تعمیم‌یافته استقرایی (GRI)
۲۹۱	۱-۴-۳-۸ معیار J-Measure
۲۹۲	۲-۴-۳-۸ چگونه می‌توان رفتار J-measure را توضیح داد؟
۲۹۳	۴-۸ چه مواقعی در استفاده از قواعد پیوند باید دقت کرد؟
۲۹۵	۱-۴-۸ ارزیابی
۲۹۶	۲-۴-۸ نسبت اطمینان
۲۹۷	۵-۸ سؤالات و مسائل مروری
۲۹۸	۶-۸ منابع و مراجع برای مطالعه بیشتر
۳۰۱	فصل نهم: شبکه‌های عصبی
۳۰۲	۱-۹ تاریخچه
۳۰۴	۲-۹ ساختار مغز
۳۰۸	۳-۹ شبکه عصبی چیست؟
۳۰۸	۱-۳-۹ قابلیت‌های شبکه عصبی
۳۰۹	۲-۳-۹ یادگیری و تعمیم

۳۱۱	۴-۹ شبکه‌های عصبی مصنوعی
۳۱۴	۱-۴-۹ نرون مصنوعی ساده
۳۱۵	۲-۴-۹ تشخیص خروجی
۳۱۶	۳-۴-۹ توابع فعال‌کننده
۳۱۹	۵-۹ ساختار شبکه‌های عصبی
۳۱۹	۱-۵-۹ پرسپترون تک لایه
۳۲۱	۲-۵-۹ پرسپترون چند لایه
۳۲۱	۳-۵-۹ الگوریتم یادگیری پرسپترون
۳۲۵	۴-۵-۹ لایه خروجی
۳۲۸	۶-۹ قاعدهٔ پس انتشار
۳۳۶	۷-۹ الگوریتم پرسپترون چند لایه
۳۳۷	۱-۷-۹ بیش برازش
۳۳۹	۲-۷-۹ توقف زود هنگام
۳۴۰	۸-۹ مقایسه طبقه‌بندی داده‌های انواع گل زنبق فیشر با استفاده از شبکه‌های عصبی و ...
۳۴۲	۹-۹ سؤالات و مسائل مروری
۳۴۲	۱۰-۹ منابع و مراجع برای مطالعه بیشتر
۳۴۳	فصل دهم: الگوریتم ژنتیک
۳۴۵	۱-۱۰ اصول الگوریتم ژنتیک
۳۴۹	۲-۱۰ بهینه‌سازی با استفاده از الگوریتم ژنتیک
۳۵۸	۳-۱۰ یک نمایش ساده از الگوریتم ژنتیک
۳۶۵	۴-۱۰ الگوریتم طراحی
۳۶۸	۵-۱۰ مسئله فروشنده دوره گرد
۳۷۲	۶-۱۰ یادگیری ماشین با استفاده از الگوریتم ژنتیک
۳۷۸	۷-۱۰ سؤالات و مسائل مروری
۳۷۹	۸-۱۰ منابع و مراجع برای مطالعه بیشتر

۳۸۱	فصل یازدهم: اعتبار سنجی مشتریان بانک
۳۸۲	۱-۱۱ اهداف تحلیل
۳۸۳	۲-۱۱ شرح داده‌ها
۳۸۶	۳-۱۱ تحلیل کاوشگرانه داده‌ها
۳۸۹	۴-۱۱ ساختن مدل
۳۹۰	۱-۴-۱۱ مدل‌های رگرسیون لجستیک
۳۹۳	۲-۴-۱۱ مدل‌های درخت رده‌بندی
۴۰۴	۴-۱۱ مقایسه مدل‌ها
۴۰۵	فصل دوازدهم: کاربردهای داده‌کاوی در بازاریابی و مدیریت رابطه با مشتری
۴۰۶	۱-۱۲ جذب کردن مشتریان احتمالی یا بالقوه
۴۰۷	۲-۱۲ شناسایی مشتریان احتمالی خوب
۴۰۸	۳-۱۲ انتخاب کانال ارتباطی
۴۰۹	۴-۱۲ داده‌کاوی برای انتخاب مکان درست برای تبلیغات
۴۲۴	۵-۱۲ مدل چگونه بر سوددهی تأثیر می‌گذارد؟
۴۳۱	۶-۱۲ استفاده از مشتریان فعلی جهت تعیین مشتریان احتمالی
۴۴۰	۷-۱۲ تعیین ارزش مشتری
۴۴۲	۸-۱۲ حفظ و قطع رابطه تجاری مشتری
۴۵۱	فهرست راهنما

پیشگفتار

دنیای مدرن در واقع دنیایی داده گراست. ما با داده‌های عددی و غیر عددی در بخش‌های صنعتی، بازرگانی، مدیریتی، مالی و علمی احاطه شده‌ایم که باید این داده‌ها تحلیل و پردازش شوند تا تبدیل به اطلاعاتی شوند که آگاهی‌بخش، آموزش دهنده و پاسخگو باشند یا به شکل دیگری به درک و تصمیم‌گیری ما کمک کنند. در واقع در عصر اینترنت، انبار داده‌ها و مراکز داده‌ای، تحلیل سنتی داده‌ها نیز مستعد و پذیرای تغییراند. مجموعه داده‌های بسیار بزرگ - گاهی اوقات صدها میلیون از رکوردهای انفرادی - در انبار داده‌های تمرکز یافته ذخیره شده‌اند که به تحلیل گران امکان استفاده از روش‌های کاوش داده‌ای جامع‌تر و توانمندتر را می‌دهند.

در سالهای اخیر رشد انفجار آمیزی از روش‌ها برای اکتشاف دانش جدید از داده‌های خام به وقوع پیوسته است. در پاسخ به این امر، رشته جدیدی از کاوش داده‌ها موسوم به داده کاوی به طور ویژه گسترش یافته است تا اطلاعات با ارزشی از چنین مجموعه داده‌های عظیم استخراج نماید. این مسئله با توجه به گسترش روزافزون رایانه‌های ارزان قیمت، مخابرات، فن پایگاه داده‌ها (برای جمع‌آوری و ذخیره سازی داده‌ها) بسیار هم باعث شگفتی نیست.

در سالهای اخیر فناوری داده کاوی به صورت موضوع داغی برای تصمیم سازان و تصمیم گیران درآمده است، زیرا این فن، اطلاعات نهفته با ارزش تجاری و علمی را از داده‌های حجیم ذخیره شده استخراج می‌کند. با این وجود، در اصل داده کاوی فناوری جدیدی محسوب نمی‌شود. استخراج اطلاعات و دانش از داده‌های ذخیره شده یک مفهوم کاملاً دیرینه در مطالعات علمی و پزشکی می‌باشد. آنچه که جدید است همگرایی و اشتراک چندین رشته و فناوری‌های متناظر آنها است که فرصت منحصر بفردی برای داده کاوی در دنیای علم، تجارت و اقتصاد ایجاد کرده است.

در واقع داده کاوی یا کشف دانش و معرفت از پایگاه داده‌ها (KDD) با تعریف فوق یک شاخه نسبتاً جدید علمی است که از انجام تحقیقات حداقل در رشته‌های آمار، یادگیری

ماشین، علوم رایانه، به ویژه پایگاه داده‌ها شکل گرفته است، در حالی که مرزهای آنها در داده کاوی مبهم است ولی می‌توان گفت که مهمترین این رشته‌ها آمار می‌باشد به طوریکه بدون آمار داده کاوی مفهومی نخواهد داشت.

در اصل هدف از نگارش این کتاب پدید آوردن یک مرجع و منبع برای استفاده دانشجویان عزیز است به ویژه که در دانشگاه‌های کشور ارایه موضوع داده کاوی به عنوان یک درس آغاز شده است. با وجود این به نظر می‌رسد که کتاب حاضر برای افرادی با زمینه‌های علمی و موقعیت‌های شغلی گوناگون گسترده که نیاز به تحلیل کاوشگرانه داده‌های خام بسیار بزرگ دارند به عنوان گردایه‌ای از مهمترین روش‌ها، ابزارها و الگوریتم‌ها در داده کاوی مؤثر و سودمند باشد. در پایان وظیفه خود می‌دانیم از جناب آقای دکتر علیرضا متولی زاده ریاست دانشگاه آزاد اسلامی واحد نیشابور که در امر چاپ و انتشار این کتاب از هیچ گونه پشتیبانی و کمکی دریغ نوزیده‌اند و همچنین از معاونت محترم پژوهشی که کمال همکاری را داشته‌اند صمیمانه سپاسگزاری و قدردانی نماییم.

مؤلفین کتاب بر خود لازم می‌دانند از همه عزیزانی که در این اثر دستی داشته‌اند بخصوص مسئولین محترم چاپ و انتشارات دانشگاه فردوسی مشهد نیز قدردانی نمایند.

علی مشکاتی و عبدالرضا ناظمی

بهار ۱۳۸۸