

دانش‌یابی داده‌ها

مقدمه‌ای بر داده‌کاوی

نویسنده:

دانیل ت. لاروس

مدیر گروه داده‌کاوی دانشگاه کانک، تیکات آمریکا

برگردان:

علی زینل همدانی

استاد دانشکده مهندسی صنایع و سیستمها

دانشگاه صنعتی اصفهان

مهندس فرهاد ابراهیمیان

مهندس حدیث یعقوب‌زاده



دانش‌یابی داده‌ها مقدمه‌ای بر داده‌کاوی

پدیدآور.....	دانیل ت. لاروس
برگردان.....	علی‌زینل همدانی، فرهاد ابراهیمیان، حدیث یعقوب‌زاده
ویراستار ادبی.....	آتوسا سعادت‌ی
صفحه آبرو و طراحی جلد.....	زحل شیروانی
لیتوگرافی، چاپ و صحافی.....	چاپخانه دانشگاه صنعتی اصفهان
ناشر.....	انتشارات دانشگاه صنعتی اصفهان
چاپ دوم.....	تابستان ۱۳۹۷
شمارگان.....	۳۰۰ جلد
شابک.....	۹۷۸-۹۶۴-۸۴۷۶-۹۰-۳
قیمت.....	۱۵۰۰۰۰ ریال

سرشناسه	لارز، دانیل تی، ۱۹۵۰ - م. Daniel T. Larose
عنوان و نام پدیدآور	دانش‌یابی داده‌ها مقدمه‌ای بر داده‌کاوی
مشخصات نشر	اصفهان: دانشگاه صنعتی اصفهان، انتشارات، ۱۳۹۳.
مشخصات ظاهری	هشت، ۲۹۵ ص. + یک لوح فشرده.
شابک	978-964-8476-90-3
وضعیت فهرست نویسی	فهیپای مختصر
یادداشت	این مدرک در آدرس http://opac.nlai.ir قابل دسترسی است.
یادداشت	عنوان اصلی:
شناسه افزوده	همدانی، علی زینل، ۱۳۳۱ - مترجم
شناسه افزوده	ابراهیمیان، فرهاد، ۱۳۵۵ - مترجم
شناسه افزوده	یعقوب‌زاده، حدیث، ۱۳۶۶ - مترجم
شماره کتابشناسی ملی	۳۷۷۴۳۳۵

حق چاپ برای انتشارات دانشگاه صنعتی اصفهان محفوظ است.

اصفهان: دانشگاه صنعتی اصفهان - انتشارات - کدپستی ۸۳۱۱۱-۸۴۱۵۶، تلفن: ۳۳۹۱۲۵۰۹ (۰۳۱) دورنگار: ۳۳۹۱۲۵۲۸ برای خرید اینترنتی کلیه کتاب‌های منتشره انتشارات می‌توانید به وبگاه <http://publication.iut.ac.ir> مراجعه و یا مستقیماً از کتابفروشی انتشارات واقع در کتابخانه مرکزی دانشگاه صنعتی اصفهان (تلفن ۳۳۹۱۳۹۵۲) خریداری فرمائید.

پیشگفتار مترجمین

امروزه ثبت واقعیت با به صورت برنامه ریزی شده و با به روال عادت در به فعالیت های اقتصادی، اجتماعی، علمی، هنری و فرهنگی امری عادی محسوب می گردد و تجمع این به داده ها و به گردیده است که مسئولان و برنامه ریزان را ترغیب نماید تا از انبوه عموماً نامنظم آنها دانش استخراج شود که بتواند بسیاری از مسائل شناخته و واقعیت های پنهان شده در پدیده های مختلف را در قالب اطلاعات استخراج نمایند.

سال ها تلاش تحلیل داده ها و استخراج اطلاعات استفاده از روش های آماری بوده است که لزوماً با حجم زیادی از داده ها روبه روبرو بوده اند و حتی بعضاً برای دسترسی سریع به نتایج و بالا بودن توان تجزیه داده ها از به کار بردن حجم زیاد آنها به وسیله انتخاب نمونه خوداری می گردیده است.

روش‌های داده‌کاوی که با کمرشی نوبه مسند استخراج دانش نهفته در دل داده‌های پردازش در حال حاضر به عنوان مهم‌ترین فناوری جهت بهره‌برداری موثر از داده‌های حجیم محبوب گردیده و اهمیت آن به عنوان علم تجزیه و تحلیل داده‌ها و انکشاف دانش نهفته در آنها رو به فزونی است.

امروزه آموزش روش‌های داده‌کاوی و کشف دانش در کنار روش‌های آماری و بوش مصنوعی برای رشته‌های دانشگاهی ضرورت و اثربخشی این امر انگیزه‌ای شد تا ترجمه‌ای از مناسب‌ترین کتب منتشر شده در این زمینه را انجام دهیم به این امید که بتواند مورد استفاده محبت‌آمیز دانشجویان رشته‌های مختلف تحصیلی قرار گیرد.

از دیدیریت نشر شورای نشر و ترویج علم و به بخش انتشارات دانشگاه صنعتی اصفهان که قبول زحمت فرمودند و در جهت چاپ این کتاب ما را حمایت کردند. تشکر و قدردانی می‌شود. لازم است به طور خاص از داوران محترم و ویراستار ادبی این کتاب سرکار خانم سعادت و همچنین از سرکاران هم‌سایه‌ای که زحمت صفحه‌آرایی و تصحیح و طراحی جلد کتاب را کشیده‌اند، سپاسگزاری شود. از خوانندگان محترم تقاضا داریم که ضمن بنفش نظراتی که احتمالاً وجود دارد ما را در بهبود چاپ‌های بعدی کتاب راهنمایی فرمایند.

به امید توفیق الهی

مترجمین

فهرست مطالب

یک پیشگفتار

فصل ۱ : مقدمه‌ای بر داده‌کاوی

- ۳ داده‌کاوی چیست؟
- ۵ چرا داده‌کاوی؟
- ۶ نیاز داده‌کاوی به هدایت توسط انسان
- ۷ فرایند استاندارد میان صنعتی: CRISP-DM
- ۸ مراحل شش‌گانه CRISP-DM
- ۱۰ مطالعه موردی اول: تحلیل دغاوی ضمانت خودرو
- ۱۳ برداشت‌های غلط از داده‌کاوی
- ۱۴ داده‌کاوی چه کارهایی را می‌تواند انجام دهد؟
- ۱۴ تشریح
- ۱۵ برآوردیابی
- ۱۷ پیش‌بینی (پیشگویی)
- ۱۸ طبقه‌بندی
- ۲۰ خوشه‌بندی
- ۲۲ وابستگی

۲۳	 مطالعه موردی دوم: پیش‌بینی سود غیرعادی بازار بورس
۲۵	 مطالعه موردی سوم: استخراج قوانین انجمنی
۲۷	 مطالعه موردی چهارم: پیش‌بینی ورشکستگی شرکت‌ها
۳۰	 مطالعه موردی پنجم: توصیف بازار گردشگری
۳۱	 منابع
۳۳	 تمرین‌ها

فصل ۲: پیش‌پردازش داده‌ها

۳۶	 چرا نیاز به پیش‌پردازش داده‌ها هستیم؟
۳۶	 مرتب‌سازی داده‌ها
۳۹	 مدیریت داده‌های گمشده
۴۳	 شناسایی طبقه‌بندی‌های غلط
۴۴	 روش‌های گرافیکی شناسایی داده‌های پرت
۴۶	 تبدیل داده‌ها
۴۶	 روش نرمال‌سازی Min-Max
۴۸	 استانداردسازی به روش Z-Score
۴۹	 روش‌های عددی شناسایی داده‌های پرت
۵۱	 منابع
۵۱	 تمرین‌ها
۵۲	 تحلیل عملی

فصل ۳: تحلیل اکتشافی داده‌ها

۵۴	 شناخت مجموعه داده‌ها
۵۷	 پرداختن به متغیرهای همبسته
۵۹	 کاوش در متغیرهای رسته‌ای
۶۵	 استفاده از EDA جهت آشکارسازی فیلدهای غیرعادی
۶۶	 کاوش در متغیرهای عددی
۷۶	 کاوش در ارتباطات چند متغیره
۷۸	 انتخاب زیرمجموعه‌های مورد علاقه از داده‌ها، جهت بررسی بیشتر
۷۹	 دسته‌بندی
۸۰	 خلاصه
۸۱	 منابع
۸۱	 تمرین‌ها

۸۲	تحلیل عملی
----	------------------

فصل ۴: روش‌های آماری برای برآوردیابی و پیشگویی

۸۶	وظایف داده‌کاوی برای کشف دانش موجود در داده‌ها
۸۷	روش‌های آماری جهت برآوردیابی و پیشگویی
۸۷	روش‌های یک متغیره: معیارهای مرکزی و پراکندگی
۹۰	استنباط آماری
۹۲	تا چه میزان به برآوردهای خود می‌توان اطمینان کرد؟
۹۳	برآوردیابی با اطمینان
۹۵	روش‌های یک متغیره: رگرسیون خطی ساده
۱۰۰	خطرات برون‌یابی
۱۰۲	فواصل اطمینان برای y به شرط داده شدن x
۱۰۲	فواصل پیش‌بینی برای یک متغیره تصادفی y به شرط داده شدن x
۱۰۵	رگرسیون چندگانه
۱۰۸	صحه‌گذاری فرض‌های مدل
۱۱۲	منابع
۱۱۲	تمرین‌ها
۱۱۲	تحلیل عملی

فصل ۵: الگوریتم k نزدیک‌ترین همسایه

۱۱۶	روش‌های با ناظر و بدون ناظر
۱۱۷	متدلوژی مدل‌سازی با ناظر
۱۱۹	موازنه واریانس - اریبی
۱۲۱	عمل طبقه‌بندی
۱۲۳	الگوریتم k نزدیک‌ترین همسایه
۱۲۶	تابع فاصله
۱۲۹	تابع ترکیب
۱۲۹	رای‌گیری غیر وزن‌دار ساده
۱۳۰	رای‌گیری وزن‌دار
۱۳۳	ملاحظات پایگاه داده
۱۳۵	انتخاب k
۱۳۶	منابع
۱۳۶	تمرین‌ها

فصل ۶: درخت‌های تصمیم

۱۴۱	درخت‌های طبقه‌بندی و رگرسیون (CART)
۱۴۸	الگوریتم C4.5
۱۵۶	قوانین تصمیم
۱۵۷	مقایسه کاربرد الگوریتم‌های C4.5 و (CART) با استفاده از داده‌های واقعی
۱۶۳	منابع
۱۶۳	تمرین‌ها
۱۶۴	تحلیل عملی

فصل ۷: شبکه‌های عصبی

۱۷۱	کاربرد شبکه‌های عصبی در برآوردیابی و پیشگویی
۱۷۲	مثال ساده‌ای از یک شبکه عصبی
۱۷۴	تابع فعال‌سازی سیگموئید
۱۷۵	پس انتشار خطا
۱۷۶	روش کاهش گرادیان
۱۷۷	قوانین پس انتشار خطا
۱۷۸	مثالی از پس انتشار خطا
۱۸۰	شرایط توقف
۱۸۲	نرخ یادگیری
۱۸۳	اصطلاح اندازه حرکت
۱۸۵	تحلیل حساسیت
۱۸۶	کاربرد مدل‌سازی شبکه عصبی
۱۸۸	منابع
۱۸۸	تمرین‌ها
۱۸۹	تحلیل عملی

فصل ۸: خوشه‌بندی سلسله مراتبی و k-MEANS

۱۹۱	خوشه‌بندی
۱۹۵	روش‌های خوشه‌بندی سلسله مراتبی
۱۹۶	خوشه‌بندی ارتباط منفرد
۱۹۷	خوشه‌بندی ارتباط کامل
۱۹۹	خوشه‌بندی k-MEANS
۱۹۹	مثالی از کاربرد خوشه‌بندی k-MEANS

۲۰۶	انجام خوشه‌بندی k-MEANS با استفاده از نرم‌افزار
۲۰۹	پیش‌بینی رویگردانی مشتری با استفاده از عضویت در خوشه
۲۱۱	منابع
۲۱۱	تمرین‌ها
۲۱۲	تحلیل عملی

فصل ۹: شبکه‌های کوهن

۲۱۲	شبکه‌های خود سازمانده (SOM)
۲۱۶	شبکه‌های کوهن
۲۱۷	مثالی از خوشه‌بندی به وسیله شبکه‌های کوهن
۲۲۲	اعتبار خوشه
۲۲۳	کاربرد خوشه‌بندی به وسیله شبکه‌های کوهن
۲۲۵	تفسیر خوشه‌ها
۲۲۹	مشخصات خوشه
۲۳۱	استفاده از خروجی خوشه‌بندی، عنوان و ردی سایر مدل‌های داده‌کاوی
۲۳۲	منابع
۲۳۳	تمرین‌ها
۲۳۳	تحلیل عملی

فصل ۱۰: قوانین انجمنی

۲۳۵	تحلیل همسایگی و تحلیل سبد بازار
۲۳۸	نمایش داده‌ها برای تحلیل سبد بازار
۲۳۹	پشتیبان، اطمینان، مجموعه اقلام مکرر و الگوریتم استقرایی
۲۴۳	الگوریتم استقرایی چگونه کار می‌کند؟
۲۴۳	الگوریتم استقرایی چگونه کار می‌کند؟
۲۴۷	تعمیم داده‌های صفر یا یک به داده‌های رسته‌ای عمومی
۲۴۹	رویکرد نظریه اطلاعات: روش تعمیم یافته استنتاج قانون
۲۵۱	کاربرد روش تعمیم یافته استنتاج قانون
۲۵۳	چه موقع نباید از قوانین انجمنی استفاده کرد؟
۲۵۷	آیا قوانین انجمنی، یادگیری باناظر یا بدون ناظر را ارائه می‌کند؟
۲۵۷	الگوهای محلی در مقابل مدل‌های عمومی
۲۵۹	منابع
۲۵۹	تمرین‌ها

۲۵۹ تحلیل عملی

فصل ۱۱: روش‌های ارزیابی مدل

۳۶۴ تکنیک‌های ارزیابی مدل برای وظیفه تشریح

۲۶۵ تکنیک‌های ارزیابی مدل برای وظایف برآوردیابی و پیشگویی

۲۶۷ تکنیک‌های ارزیابی مدل برای وظیفه طبقه‌بندی

۲۶۷ نرخ خطا، مثبت‌های کاذب و منفی‌های کاذب

۲۷۰ تنظیم: زینه طبقه‌بندی غلط، برای انعکاس مشکلات دنیای واقعی

۲۷۲ تحلیل گزینه / سود تصمیم

۲۷۳ نمودارهای ترفیع و نمودارهای بهره

۲۷۷ ترکیب ارزیابی ساخت مدل

۲۷۸ تلاقی نتایج: کارگری یک مجموعه متناسب از مدل‌ها

۲۷۹ منابع

۲۷۹ تحلیل گروهی

۲۸۳ واژه‌نامه فارسی به انگلیسی

۲۸۹ واژه‌یاب