

داده کاوی

مفاهیم، مدل‌ها، روش‌ها و الگوریتم‌ها

ویراست سوم

مؤلف

محمد کانتارزیک

مترجم

ایوب ترکیان

نیاز دانش

فهرست مطالب

شماره صفحه	عنوان
۹	فصل ۱ / مفاهیم داده‌کاوی
۹	۱.۱ مقدمه
۱۲	۲.۱ ریشه‌های داده‌کاوی
۱۵	۳.۱ فرایند داده‌کاوی
۲۰	۴.۱ از جمع‌آوری داده تا پردازش داده
۲۶	۵.۱ مخازن داده برای داده‌کاوی
۳۰	۶.۱ از داده حجیم تا علم داده
۳۵	۷.۱ جوانب بیزینسی داده‌کاوی
۳۹	۸.۱ سازمان کتاب
۴۱	۹.۱ سوالات مروری و مسایل
۴۵	فصل ۲ / آماده‌سازی داده‌ها
۴۵	۱.۲ نمایش داده‌های خام
۵۱	۲.۲ خصوصیات داده‌های خام
۵۳	۳.۲ تبدیل داده‌های خام
۵۳	۱.۳.۲ نرمال‌سازی
۵۵	۲.۳.۲ هموارسازی داده‌ها
۵۵	۳.۳.۲ تفاوت‌ها و نسبت‌ها
۵۶	۴.۲ داده‌های ناموجود
۵۹	۵.۲ داده‌های وابسته به زمان
۶۴	۶.۲ تحلیل داده‌های پرت
۷۲	۷.۲ سوالات مروری و مسایل
۷۷	فصل ۳ / کاهش داده‌ها
۷۸	۱.۳ ابعاد مجموعه داده‌های بزرگ
۸۰	۲.۳ کاهش ویژگی‌ها

۸۳	انتخاب ویژگی	۱.۲.۳
۸۹	استخراج ویژگی	۲.۲.۳
۹۳	الگوریتم نجات	۳.۳
۹۶	سنجه انتروپی برای ویژگی‌های درجه‌بندی	۴.۳
۹۸	تحلیل مؤلفه اصلی	۵.۳
۱۰۲	کاهش مقدار	۶.۳
۱۰۶	گسسته‌سازی ویژگی: تکنیک تلفیق Chi	۷.۳
۱۱۰	کاهش مورد	۸.۳
۱۱۴	سوالات مروری و مسایل	۹.۳

فصل ۴ / یادگیری از داده‌ها

۱۱۷	ماشین یادگیری	۱.۴
۱۲۴	تئوری یادگیری آماری	۲.۴
۱۳۱	انواع روش‌های یادگیری	۳.۴
۱۳۳	فعالیت‌های یادگیری متداول	۴.۴
۱۳۹	ماشین‌های بردار پشتیبان	۵.۴
۱۵۴	ماشین‌های بردار پشتیبان نیمه‌نظارتی (SVM)	۶.۴
۱۵۹	KNN: طبقه‌گر نزدیکترین همسایه	۷.۴
۱۶۴	انتخاب مدل در مقابل تعمیم	۸.۴
۱۶۸	تخمین مدل	۹.۴
۱۷۸	طبقه‌بندی داده غیرمتوازن	۱۰.۴
۱۸۳	۹۰٪ صحت... حالا چی؟	۱۱.۴
۱۸۳	۱.۱۱.۴ آشکارسازی تقلب بیمه	
۱۸۶	۲.۱۱.۴ بهبود مراقبت قلبی	
۱۸۷	سوالات مروری و مسایل	۱۲.۴

فصل ۵ / روش‌های آماری

۱۹۱	استنباط آماری	۱.۵
۱۹۴	برآورد تفاوت‌ها در مجموعه داده‌ها	۲.۵
۱۹۸	استنباط بیزی	۳.۵
۲۰۱	رگرسیون پیش‌بینانه	۴.۵
۲۰۸	آنالیز واریانس	۵.۵
۲۱۱	رگرسیون لجیستیک	۶.۵
۲۱۳	مدل‌های لگاریتم-خطی	۷.۵
۲۱۸	تحلیل افتراقی خطی	۸.۵
۲۲۰	سوالات مروری و مسایل	۹.۵

۲۲۵	فصل ۶ / درختان تصمیم و قواعد تصمیم
۲۲۷	۱.۶ درختان تصمیم
۲۳۰	۲.۶ الگوریتم C _{4.5} : تولید درخت تصمیم
۲۳۸	۳.۶ مقادیر نامعلوم مشخصه
۲۴۳	۴.۶ هرس درختان تصمیم
۲۴۵	۵.۶ الگوریتم C _{4.5} : تولید قواعد تصمیم
۲۴۹	۶.۶ الگوریتم CART و نمایه جینی
۲۵۴	۷.۶ محدودیت‌های درختان تصمیم و قواعد تصمیم
۲۵۶	۸.۶ سوالات مروری و مسایل

۲۶۳	فصل ۷ / شبکه‌های عصبی مصنوعی
۲۶۵	۱.۷ مدل نورون مصنوعی
۲۶۹	۲.۷ معماری‌های شبکه عصبی مصنوعی
۲۷۱	۳.۷ فرایند یادگیری
۲۷۵	۴.۷ فعالیت‌های یادگیری با استفاده از ANN
۲۷۶	۱.۴.۷ تلازم الگو
۲۷۶	۲.۴.۷ شناسایی (تشخیص) الگو
۲۷۷	۳.۴.۷ تخمین تابع
۲۷۷	۴.۴.۷ کنترل
۲۷۸	۵.۴.۷ فیلتر کردن
۲۷۸	۶.۴.۷ پیش‌بینی
۲۷۹	۵.۷ پرسپترون‌های چندلایه
۲۸۹	۶.۷ شبکه‌های رقابتی و یادگیری رقابتی
۲۹۴	۷.۷ نگاشت‌های خودسازمان‌ده
۳۰۰	۸.۷ یادگیری عمیق
۳۰۶	۹.۷ شبکه‌های عصبی کانولوشن (CNN-ها)
۳۱۱	۱۰.۷ سوالات مروری و مسایل

۳۱۵	فصل ۸ / یادگیری آنسمل
۳۱۶	۱.۸ متدولوژی‌های یادگیری آنسمل
۳۲۲	۲.۸ شیماهای ترکیب برای یادگیرهای چندگانه
۳۲۳	۳.۸ سببندی و تقویت
۳۲۶	۴.۸ ADABOOST
۳۲۷	۵.۸ سوالات مروری و مسایل

۳۳۳	فصل ۹ / تحلیل خوشه‌بندی	
۳۳۳	مفاهیم خوشه‌بندی	۱.۹
۳۳۷	سنجش‌های مشابهت	۲.۹
۳۴۵	خوشه‌بندی سلسله‌مراتبی تجمیعی	۳.۹
۳۴۹	خوشه‌بندی تفکیکی	۴.۹
۳۵۳	خوشه‌بندی تدریجی	۵.۹
۳۵۷	الگوریتم DBSCAN	۶.۹
۳۶۰	الگوریتم BIRCH	۷.۹
۳۶۴	اعتبارسنجی خوشه‌بندی	۸.۹
۳۷۰	سوالات مروری و مسایل	۹.۹

۳۷۷	فصل ۱۰ / قواعد تلازم	
۳۷۸	تحلیل سبد بازار	۱.۱۰
۳۸۰	الگوریتم پیشینی	۲.۱۰
۳۸۳	از مجموعه آیتم کراری تا قواعد تلازم	۳.۱۰
۳۸۵	بهبود راندمان الگوریتم پیشینی	۴.۱۰
۳۸۷	روش رشد الگوی کراری	۵.۱۰
۳۹۰	روش تلازمی - طبقه‌بندی	۶.۱۰
۳۹۴	کاوش قاعده تلازم چندبُعدی	۷.۱۰
۳۹۶	سوالات مروری و مسایل	۸.۱۰

۴۰۱	فصل ۱۱ / وب‌کاوی و متن‌کاوی	
۴۰۱	وب‌کاوی	۱.۱۱
۴۰۳	کاوش محتوا، ساختار، و استفاده وب	۲.۱۱
۴۰۷	الگوریتم‌های HITS و LOGSOM	۳.۱۱
۴۱۴	کاوش الگوهای طی مسیر	۴.۱۱
۴۱۷	الگوریتم PageRank	۵.۱۱
۴۲۰	سیستم‌های توصیه‌گر	۶.۱۱
۴۲۲	متن‌کاوی	۷.۱۱
۴۲۷	تحلیل سمانتیک نهفته	۸.۱۱
۴۳۳	سوالات مروری و مسایل	۹.۱۱

۴۳۷	فصل ۱۲ / پیشرفت‌های داده‌کاوی	
۴۳۷	۱.۱۲ گراف‌کاوی	
۴۵۴	۲.۱۲ گراف‌کاوی زمانی	
۴۵۶	۱.۲.۱۲ نمایش داده زمانی	
۴۶۲	۲.۲.۱۲ سنجش‌های مشابهت بین دنباله‌ها	
۴۶۵	۳.۲.۱۲ مدل‌سازی داده زمانی	
۴۶۷	۴.۲.۱۲ دنباله‌کاوی	
۴۷۲	۳.۱۲ داده‌کاوی مکانی	
۴۷۷	۴.۱۲ داده‌کاوی توزیعی	
۴۸۴	۱.۴.۱۲ خوشه‌بندی DBSCAN توزیعی	
۴۸۸	۵.۱۲ عدم تداعی علیت از هم‌بستگی	
۴۸۹	۱.۵.۱۲ شبکه‌های بیزی	
۴۹۶	۶.۱۲ حریم خصوصی، امنیت، و جوانب حقوقی	
۵۰۳	۷.۱۲ محاسبه کلاود مبتنی بر HADOOP و MAP/REDUCE	
۵۰۹	۸.۱۲ یادگیری تقویتی	
۵۱۵	۹.۱۲ سوالات مروری و مسایل	

۵۱۹	فصل ۱۳ / الگوریتم‌های ژنتیک	
۵۲۰	۱.۱۳ مبانی الگوریتم‌های ژنتیک	
۵۲۳	۲.۱۳ بهینه‌سازی با استفاده از الگوریتم‌های ژنتیک	
۵۲۴	۱.۲.۱۳ شیماهای رمزگذاری و آغازگری	
۵۲۵	۲.۲.۱۳ ارزیابی برازندگی	
۵۲۵	۳.۲.۱۳ انتخاب	
۵۲۷	۴.۲.۱۳ آمیزش	
۵۲۸	۵.۲.۱۳ جهش	
۵۳۰	۳.۱۳ شرح ساده الگوریتم ژنتیک	
۵۳۰	۱.۳.۱۳ نمایش	
۵۳۱	۲.۳.۱۳ جمعیت اولیه	
۵۳۱	۳.۳.۱۳ ارزیابی	
۵۳۲	۴.۳.۱۳ تناوب	
۵۳۳	۵.۳.۱۳ عملگرهای ژنتیک	
۵۳۵	۶.۳.۱۳ ارزیابی (تکرار دوم)	
۵۳۶	۴.۱۳ نقشه‌ها (شیماها)	
۵۳۹	۵.۱۳ مسئله فروشنده دوره‌گرد	
۵۴۱	۶.۱۳ یادگیری ماشین با الگوریتم‌های ژنتیک	
۵۴۵	۱.۶.۱۳ RuleExchange	
۵۴۵	۲.۶.۱۳ RuleGeneralization	
۵۴۶	۳.۶.۱۳ RuleSpecialization	

۵۴۶RuleSplit	۴۶.۱۳
۵۴۶الگوریتم‌های ژنتیک برای خوشه‌بندی	۷.۱۳
۵۵۰سوالات مروری و مسایل	۸.۱۳

فصل ۱۴ / مجموعه‌های فازی و منطق فازی ۵۵۳

۵۵۳مجموعه‌های فازی	۱.۱۴
۵۶۰عملیات مجموعه فازی	۲.۱۴
۵۶۶اصل بسط و روابط فازی	۳.۱۴
۵۷۱منطق فازی و سیستم‌های استنباط فازی	۴.۱۴
۵۷۶ارزیابی چندفاکتوری	۵.۱۴
۵۷۷۱.۵.۱۴ مسئله انتخاب لباس	
۵۷۸۲.۵.۱۴ مسئله ارزیابی تدریس	
۵۷۹استخراج مدل‌های فازی از داده‌ها	۶.۱۴
۵۸۵داده‌کاوی و مجموعه‌های فازی	۷.۱۴
۵۸۸سوالات مروری و مسایل	۸.۱۴

فصل ۱۵ / روش‌های مصورسازی ۵۹۱

۵۹۱ادراک و مصورسازی	۱.۱۵
۵۹۳مصورسازی علمی و مصورسازی اطلاعات	۲.۱۵
۶۰۱مختصات موازی	۳.۱۵
۶۰۴مصورسازی شعاعی	۴.۱۵
۶۰۷مصورسازی با نگاشت‌های خودسازمان‌ده	۵.۱۵
۶۰۹سیستم‌های مصورسازی برای داده‌کاوی	۶.۱۵
۶۱۶سوالات مروری و مسایل	۷.۱۵

فصل ۱

مفاهیم داده کاوی

۱.۱ مقدمه

علوم و مهندسی نوین مبتنی بر استفاده از مدل‌های قاعده-اول (فرمول پایه) برای شرح سیستم‌های فیزیکی، بیولوژیکی، و اجتماعی هستند. این گونه رویکرد با یک مدل علمی پایه، نظیر قوانین نیوتن حرکت یا معادلات ماکسول در الکترومغناطیس، شروع شده، و آنگاه کاربردهای مختلف را در مهندسی مکانیک و مهندسی برق بنا می‌نهد. در این رویکرد، داده‌های تجربی برای صحت‌گذاری مدل‌های فرمول پایه نهشته شده و تخمین بعضی از پارامترهایی که مشکل بوده یا در بعضی مواقع برای اندازه‌گیری مستقیم محال است، استفاده می‌شوند. با این حال، در بسیاری از زمینه‌ها، فرمول پایه نهشته معلوم نبوده، یا سیستم‌های تحت بررسی برای ساختاردهی ریاضی بیش از حد پیچیده هستند. با رشد استفاده از کامپیوترها، مقدار زیادی داده توسط این گونه سیستم‌ها در حال تولید است. در غیاب مدل‌های فرمول پایه، این گونه داده‌های به سادگی موجود را می‌توان برای استخراج مدل‌ها با تخمین روابط مفید بین متغیرهای یک سیستم (یعنی، وابستگی‌های ورودی-خروجی نامعلوم)، استفاده کرد. بدین طریق، در حال حاضر یک شیفت پارادایمی از مدل‌سازی و تحلیل‌های مبتنی بر قواعد فرمول پایه به توسعه مدل‌ها و تحلیل‌های متناظر از داده‌ها در حال وقوع است.

به تدریج به این واقعیت عادت کرده‌ایم که حجم بالایی از داده‌های اشغال‌کننده کامپیوترها، شبکه‌ها، و زندگی ما وجود دارد. دوایر دولتی، نهادهای علمی، و بیزینس‌ها، منابع عظیمی را برای