

نویسنده: Bala Deshpande | Vijay Kotu

ویراست دوم

علوم داده

مفاهیم و شیوه عمل

به همراه آموزش نرم افزار RapidMiner

مترجمان:

دکتر رامین مولاناپور

(مدرس دانشگاه های تهران)

مهندس عادل قسمتی

علوم داده «مفاهیم و شیوه عمل»

به همراه آموزش نرم‌افزار RapidMiner

مترجمان: دکتر رامین مولاناپور، مهندس عادل قسمتی

ناشر: انتشارات آتی‌نگر

ناشر همکار: انتشارات وینا

طراحی جلد و صفحه‌آرایی: همتا بیداریان

چاپ اول، ۱۴۰۲

شمارگان: ۱۰۰ نسخه

قیمت: ۳.۴۰۰.۰۰۰ ریال

شابک: ۹۷۸-۶۲۲-۷۵۷۱-۴۸-۶

ISBN: 978-622-7571-48-6

حق چاپ برای انتشارات آتی‌نگر محفوظ است.

نشانی دفتر فروش: خیابان جمالزاده جنوبی، روبه‌روی کوچه رشتچی، پلاک ۱۴۴، واحد ۱

نمابر: ۶۶۵۶۵۳۳۷

تلفن: ۸-۶۶۵۶۵۳۳۶



www.ati-negar.com * info@ati-negar.com

سرشناسه: کوتو، ویجی، Kotu, Vijay

علوم داده «مفاهیم و شیوه عمل» به همراه آموزش نرم‌افزار RapidMiner: نویسندگان: ویجی کوتو، بالا دشیپند/

مترجمان: رامین مولاناپور، عادل قسمتی

تهران: آتی‌نگر، وینا ۱۴۰۲

۵۷۶ ص.: مصور، جدول، نمودار.

ISBN: 978-622-757-48-6

فینیا.

یادداشت: عنوان اصلی کتاب: Data science: concepts and practice, Second edition, 2019

موضوع: داده‌کاوی - Data mining

موضوع: مصرف‌کنندگان - رفتار - Consumer behavior - داده‌پردازی - Electronic data processing

شناسه‌افزوده: دشیپند، بالا، Bala, Deshpande

شناسه‌افزوده: مولاناپور، رامین، ۱۳۵۲ - مترجم

شناسه‌افزوده: قسمتی، عادل، ۱۳۶۹ - مترجم

شناسه‌افزوده: بیداریان، همتا، ۱۳۶۱

رده‌بندی کنگره

رده‌بندی دیویی

شماره کتابشناسی ملی

QAY۶/۹

۰۰۶/۳۱۲

۹۱۸۴۴۴۳

فهرست مطالب

۱۱	پیشگفتار
۱۵	مقدمه
۱۹	قدردانی‌ها
۲۱	فصل اول: مقدمه
۲۲	۱-۱ هوش مصنوعی، یادگیری ماشین و علوم داده
۲۵	۱-۲ علوم داده چیست؟
۲۵	۱-۲-۱ استخراج الگوهای معنادار
۲۵	۱-۲-۲ ایجاد مدل‌های معرف
۲۶	۱-۲-۳ ترکیب آمار، یادگیری ماشین و رایانش
۲۷	۱-۲-۴ الگوریتم‌های یادگیری
۲۷	۱-۲-۵ زمینه‌های مرتبط
۲۹	۱-۳ دلایل مستدل در رابطه علوم داده
۲۹	۱-۳-۱ حجم
۲۹	۱-۳-۲ ابعاد
۳۰	۱-۳-۳ سوالات پیچیده
۳۱	۱-۴ طبقه‌بندی علوم داده
۳۳	۱-۵ الگوریتم‌های علوم داده
۳۵	۱-۶ نقشه راه برای این کتاب
۳۵	۱-۶-۱ شروع کار با علوم داده
۳۶	۱-۶-۲ تمرین با استفاده از RapidMiner
۳۶	۱-۶-۳ الگوریتم‌های اصلی
۴۱	منابع

فصل دوم: فرایند علوم داده

۴۳	
۴۶	۲-۱ دانش پیشین
۴۶	۲-۱-۱ هدف
۴۷	۲-۱-۲ حوزه موضوعی
۴۷	۲-۱-۳ داده‌ها
۴۹	۲-۱-۴ علیت در برابر همبستگی

۴۹	۲-۲ آماده‌سازی داده‌ها
۵۰	۲-۲-۱ کاوش داده‌ها
۵۰	۲-۲-۲ کیفیت داده‌ها
۵۱	۲-۲-۳ مقادیر ناموجود
۵۲	۲-۲-۴ انواع داده‌ها و تبدیل
۵۲	۲-۲-۵ تبدیل
۵۲	۲-۲-۶ نقاط پرت
۵۳	۲-۲-۷ انتخاب ویژگی
۵۳	۲-۲-۸ نمونه‌برداری داده‌ها
۵۴	۲-۳ مدل‌سازی
۵۴	۲-۳-۱ مجموعه داده‌های آموزشی و آزمایشی
۵۶	۲-۳-۲ الگوریتم‌های یادگیری
۵۷	۲-۳-۳ ارزیابی مدل
۵۹	۲-۳-۴ مدل‌سازی جمعی
۵۹	۲-۴ کاربرد
۵۹	۲-۴-۱ آمادگی تولید
۶۰	۲-۴-۲ یکپارچه‌سازی فنی
۶۰	۲-۴-۳ زمان پاسخگویی
۶۱	۲-۴-۴ نوسازی مدل
۶۱	۲-۴-۵ همگون‌سازی
۶۱	۲-۵ دانش
۶۲	منابع

۶۵

فصل سوم: کاوش داده‌ها

۶۶	۳-۱ اهداف کاوش داده‌ها
۶۶	۳-۲ مجموعه داده‌ها
۶۸	۳-۲-۱ انواع داده‌ها
۶۹	۳-۳ آمار توصیفی
۷۰	۳-۳-۱ کاوش تک متغیره
۷۳	۳-۳-۲ کاوش چندمتغیره
۷۵	۳-۴ مصورسازی داده‌ها
۷۶	۳-۴-۱ مصورسازی تک متغیره
۸۱	۳-۴-۲ مصورسازی چندمتغیره
۸۴	۳-۴-۳ مصورسازی داده‌های با ابعاد بالا

۳-۵ نقشه راه برای کاوش داده‌ها..... ۹۰
منابع ۹۱

فصل چهارم: طبقه‌بندی

۹۳

۴-۱ درختان تصمیم ۹۴
۴-۱-۱ نحوه کار آن ۹۴
۴-۱-۲ نحوه پیاده‌سازی ۱۰۳
۴-۱-۳ نتیجه‌گیری ۱۱۶
۴-۲ استنتاج قوانین ۱۱۷
۴-۲-۱ نحوه کار آن ۱۲۱
۴-۲-۲ نحوه پیاده‌سازی ۱۲۴
۴-۲-۳ نتیجه‌گیری ۱۲۸
۴-۳ الگوریتم k- نزدیک‌ترین همسایه ۱۲۸
۴-۳-۱ نحوه کار آن ۱۳۰
۴-۳-۲ نحوه پیاده‌سازی ۱۳۸
۴-۳-۳ نتیجه‌گیری ۱۴۱
۴-۴ بیز ساده ۱۴۲
۴-۴-۱ نحوه کار آن ۱۴۴
۴-۴-۲ نحوه پیاده‌سازی ۱۵۳
۴-۴-۳ نتیجه‌گیری ۱۵۵
۴-۵ شبکه‌های عصبی مصنوعی ۱۵۶
۴-۵-۱ نحوه کار آن ۱۶۰
۴-۵-۲ نحوه پیاده‌سازی ۱۶۳
۴-۵-۳ نتیجه‌گیری ۱۶۶
۴-۶ ماشین‌های بردار پشتیبان ۱۶۷
۴-۶-۱ نحوه کار آن ۱۷۱
۴-۶-۲ نحوه پیاده‌سازی ۱۷۳
۴-۶-۳ نتیجه‌گیری ۱۸۰
۴-۷ یادگیرنده‌های جمعی ۱۸۱
۴-۷-۱ نحوه کار آن ۱۸۳
۴-۷-۲ نحوه پیاده‌سازی ۱۸۵
۴-۷-۳ نتیجه‌گیری ۱۹۴
منابع ۱۹۵

فصل پنجم: روش‌های رگرسیون

۱۹۹

۲۰۰	۵-۱ رگرسیون خطی
۲۰۱	۵-۱-۱ نحوه کار آن
۲۰۷	۵-۱-۲ نحوه پیاده‌سازی
۲۱۴	۵-۱-۳ نقاط واریسی
۲۲۰	۵-۲ رگرسیون لجستیک
۲۲۲	۵-۲-۱ نحوه کار آن
۲۲۸	۵-۲-۲ نحوه پیاده‌سازی
۲۳۱	۵-۲-۳ خلاصه نکات
۲۳۱	۵-۳ نتیجه‌گیری
۲۳۲	منابع

فصل ششم: تحلیل وابستگی

۲۳۳

۲۳۵	۶-۱ قوانین کاوش وابستگی
۲۳۷	۶-۱-۱ مجموعه اقلام
۲۴۰	۶-۱-۲ تولید قانون
۲۴۱	۶-۲ الگوریتم Apriori
۲۴۲	۶-۲-۱ نحوه کار آن
۲۴۶	۶-۳ الگوریتم رشد الگوهای پرتکرار
۲۴۶	۶-۳-۱ نحوه کار آن
۲۵۰	۶-۳-۲ نحوه پیاده‌سازی
۲۵۵	۶-۴ نتیجه‌گیری
۲۵۵	منابع

فصل هفتم: خوشه‌بندی

۲۵۷

۲۶۳	۷-۱ خوشه‌بندی K-MEANS
۲۶۵	۷-۱-۱ نحوه کار آن
۲۷۱	۷-۱-۲ نحوه پیاده‌سازی
۲۷۶	۷-۲ خوشه‌بندی DBSCAN
۲۷۷	۷-۲-۱ نحوه کار آن
۲۸۱	۷-۲-۲ نحوه پیاده‌سازی
۲۸۴	۷-۳ نقشه‌های خودسازمان‌دهنده
۲۸۶	۷-۳-۱ نحوه کار آن
۲۸۹	۷-۳-۲ نحوه پیاده‌سازی

منابع ۲۹۷

فصل هشتم: ارزیابی مدل ۲۹۹

۳۰۱	۸-۱ ماتریس درهم‌ریختگی.....
۳۰۳	۸-۲ ROC و AUC.....
۳۰۶	۸-۳ منحنی‌های بُرا.....
۳۱۰	۸-۴ نحوه پیاده‌سازی.....
۳۱۵	۸-۵ نتیجه‌گیری.....
۳۱۵	منابع.....

فصل نهم: متن‌کاوی ۳۱۷

۳۲۰	۹-۱ نحوه کار آن.....
۳۲۰	۹-۱-۱ تکرار عبارت- معکوس تکرار در سند.....
۳۲۲	۹-۱-۲ مجموعه اصطلاحات.....
۳۲۷	۹-۲ نحوه پیاده‌سازی.....
۳۲۷	۹-۲-۱ پیاده‌سازی ۱: خوشه‌بندی کلمات کلیدی.....
۳۳۲	۹-۲-۲ پیاده‌سازی ۲: پیش‌بینی جنسیت نویسندگان وبلاگ.....
۳۴۱	۹-۳ نتیجه‌گیری.....
۳۴۲	منابع.....

فصل دهم: یادگیری ژرف ۳۴۳

۳۴۷	۱۰-۱ زمستان هوش مصنوعی.....
۳۵۳	۱۰-۲ نحوه کار آن.....
۳۵۴	۱۰-۲-۱ مدل‌های رگرسیون به‌عنوان شبکه‌های عصبی.....
۳۵۶	۱۰-۲-۲ گرادیان نزولی.....
۳۶۰	۱۰-۲-۳ نیاز به پسانتشار.....
۳۶۱	۱۰-۲-۴ طبقه‌بندی بیش از ۲ دسته: تابع بیشینه هموار (Softmax).....
۳۶۳	۱۰-۲-۵ شبکه‌های عصبی کانولوشن.....
۳۷۰	۱۰-۲-۶ لایه متراکم.....
۳۷۱	۱۰-۲-۷ لایه حذف تصادفی.....
۳۷۱	۱۰-۲-۸ شبکه‌های عصبی تراجعی.....
۳۷۳	۱۰-۲-۹ خودرمزگذارها.....
۳۷۴	۱۰-۲-۱۰ مدل‌های هوش مصنوعی مرتبط.....
۳۷۵	۱۰-۳ نحوه پیاده‌سازی.....

۳۸۱.....	۱۰-۴ نتیجه‌گیری
۳۸۱.....	منابع.....

۳۸۳ فصل یازدهم: موتورهای پیشنهاددهنده

۳۸۷.....	۱۱-۱ مفاهیم موتور پیشنهاددهنده
۳۹۲.....	۱۱-۱-۱ انواع موتورهای پیشنهاددهنده.....
۳۹۵.....	۱۱-۲ پالایش مشارکتی
۳۹۶.....	۱۱-۲-۱ روش‌های مبتنی بر همسایگی
۴۰۹.....	۱۱-۲-۲ تجزیه عاملی ماتریس.....
۴۱۶.....	۱۱-۳ پالایش مبتنی بر محتوا.....
۴۱۹.....	۱۱-۳-۱ محاسبه نمایه کاربر
۴۲۷.....	۱۱-۳-۲ مدل‌های یادگیری یا ناظر.....
۴۳۳.....	۱۱-۴ پیشنهاددهنده‌های ترکیبی
۴۳۵.....	۱۱-۵ نتیجه‌گیری
۴۳۸.....	منابع.....

۴۴۱ فصل دوازدهم: پیش‌بینی سری‌های زمانی

۴۴۶.....	۱۲-۱ تجزیه سری‌های زمانی
۴۴۹.....	۱۲-۱-۱ تجزیه کلاسیک.....
۴۵۰.....	۱۲-۱-۲ نحوه پیاده‌سازی.....
۴۵۳.....	۱۲-۲ روش مبتنی بر هموارسازی.....
۴۵۳.....	۱۲-۲-۱ روش‌های پیش‌بینی ساده.....
۴۵۵.....	۱۲-۲-۲ هموارسازی نمایی.....
۴۵۸.....	۱۲-۲-۳ نحوه پیاده‌سازی.....
۴۶۰.....	۱۲-۳ روش‌های مبتنی بر رگرسیون.....
۴۶۱.....	۱۲-۳-۱ رگرسیون
۴۶۲.....	۱۲-۳-۲ رگرسیون با تغییرات فصلی.....
۴۶۵.....	۱۲-۳-۳ میانگین متحرک یکپارچه خودهمبسته
۴۷۲.....	۱۲-۳-۴ ARIMA فصلی
۴۷۵.....	۱۲-۴ روش‌های یادگیری ماشین
۴۷۶.....	۱۲-۴-۱ پنجره‌بندی.....
۴۸۲.....	۱۲-۴-۲ شبکه عصبی خودهمبسته
۴۸۵.....	۱۲-۵ ارزیابی عملکرد
۴۸۵.....	۱۲-۵-۱ مجموعه داده اعتبارسنجی.....

۴۸۸ اعتبارسنجی پنجره کشویی..... ۱۲-۵-۲
۴۸۸ نتیجه‌گیری..... ۱۲-۶
۴۸۹ پیش‌بینی به‌روش‌ها..... ۱۲-۶-۱
۴۹۰ منابع.....

فصل سیزدهم: تشخیص ناهنجاری

۴۹۱	
۴۹۲ مفاهیم ۱۳-۱.....
۴۹۲ علل نقاط پرت..... ۱۳-۱-۱
۴۹۵ تکنیک‌های تشخیص ناهنجاری..... ۱۳-۱-۲
۴۹۸ تشخیص نقطه پرت بر اساس فاصله..... ۱۳-۲
۴۹۹ نحوه کار آن..... ۱۳-۲-۱
۴۹۹ نحوه پیاده‌سازی..... ۱۳-۲-۲
۵۰۳ تشخیص نقطه پرت بر اساس چگالی..... ۱۳-۳
۵۰۳ نحوه کار آن..... ۱۳-۳-۱
۵۰۳ نحوه پیاده‌سازی..... ۱۳-۳-۲
۵۰۶ عامل پرت محلی..... ۱۳-۴
۵۰۶ نحوه کار آن..... ۱۳-۴-۱
۵۰۷ نحوه پیاده‌سازی..... ۱۳-۴-۲
۵۰۹ نتیجه‌گیری..... ۱۳-۵
۵۱۰ منابع.....

فصل چهاردهم: انتخاب ویژگی

۵۱۱	
۵۱۳ طبقه‌بندی روش‌های انتخاب ویژگی..... ۱۴-۱
۵۱۴ تحلیل مؤلفه اصلی..... ۱۴-۲
۵۱۵ نحوه کار آن..... ۱۴-۲-۱
۵۱۷ نحوه پیاده‌سازی..... ۱۴-۲-۲
۵۲۲ پالایش مبتنی بر نظریه اطلاعات..... ۱۴-۳
۵۲۴ پالایش مبتنی بر خی‌دو..... ۱۴-۴
۵۲۷ انتخاب ویژگی نوع روکش..... ۱۴-۵
۵۲۹ حذف پسرو..... ۱۴-۵-۱
۵۳۲ نتیجه‌گیری..... ۱۴-۶
۵۳۳ منابع.....

فصل پانزدهم: شروع کار با RapidMiner

۵۳۵

- ۱-۱۵ رابط کاربر و مجموعه اصطلاحات ۵۳۶
- ۲-۱۵ ابزارهای وارد کردن و خروجی گرفتن از داده‌ها ۵۴۲
- ۳-۱۵ ابزارهای مصورسازی داده‌ها ۵۴۵
- ۴-۱۵ ابزارهای تبدیل داده ۵۴۸
- ۵-۱۵ ابزارهای نمونه‌برداری و مقدار ناموجود ۵۵۳
- ۶-۱۵ ابزارهای بهینه‌سازی ۵۵۷
- ۷-۱۵ یکپارچگی با R ۵۶۴
- ۸-۱۵ نتیجه‌گیری ۵۶۵
- منابع ۵۶۵

مقایسه الگوریتم‌های علوم داده‌ها

۵۶۷

- رگرسیون: پیش‌بینی یک متغیر هدف کمی رسته‌ای ۵۶۷
- رگرسیون: پیش‌بینی یک متغیر هدف کمی عددی ۵۶۹
- تحلیل وابستگی: فرایند بدون ناظر برای یافتن روابط بین ارقام ۵۶۹
- خوشه‌بندی: فرایندی بدون ناظر برای یافتن گروه‌های با معنی در داده‌ها ۵۷۰
- تشخیص ناهنجاری: تکنیک‌های با ناظر و بدون ناظر برای یافتن نقاط پرت در داده‌ها ۵۷۱
- یادگیری ژرف: آموزش با استفاده از لایه‌های متعدد بازنمایی داده‌ها ۵۷۲
- پیشنهاددهنده‌ها: یافتن ارجحیت کاربر برای یک قلم ۵۷۳
- پیش‌بینی سری‌های زمانی: پیش‌بینی مقدار آتی یک متغیر ۵۷۴
- انتخاب ویژگی: انتخاب مهم‌ترین صفات خاصه ۵۷۵