

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

رضایی تبار، وحید، ۱۳۶۳ -	سرشناسه:
داده کاوی با زبان برنامه نویسی R = Data Mining with R / مولفان: وحید رضایی تبار، شقایق خودکامه	عنوان و نام پدید آور:
ویراست ۲.	وضعیت ویراست:
تهران: دانشگاه علامه طباطبائی، ۱۴۰۳.	مشخصات نشر:
۵۲۰ص: مصور (بخشی رنگی)، جدول، نمودار، وزیری.	مشخصات ظاهری:
انتشارات دانشگاه علامه طباطبائی؛ [۲] ۵۷۴.	فروست:
۰۹-۰۶-۲۱۷-۹۶۴-۹۷۸: ۴۵۰۰۰ تومان	شابک:
فیپا	وضعیت فهرست نویسی:
واژه نامه.	موضوع:
کتابنامه.	موضوع:
داده کاوی -- راهنمای آموزشی (عالی)	موضوع:
Data Mining -- Study and Teaching (Higher)	موضوع:
آر (زبان برنامه نویسی کامپیوتر)	موضوع:
R (Computer Program Language)	موضوع:
خودکامه، شقایق، ۱۳۷۴ -	شناسه افزوده:
دانشگاه علامه طباطبائی	شناسه افزوده:
۱۴۰۳ د ۲ / ۶۷/۹ QA	رده بندی کنگره:
۰۰۶/۳۱۲	رده بندی دیوئی:
۹۸۱۶۴۸۸	شماره کتابخانه ملی:

داده کاوی با زبان برنامه نویسی R

مؤلفان:

دکتر وحید رضائی تبار

عضو هیئت علمی دانشگاه علامه طباطبائی

شقایق خودکامه

ویرایش دوم، چاپ اول



داده‌کاوی با زبان برنامه‌نویسی R

مؤلفان:

دکتر وحید رضائی تبار

شقایی خودکامه

زیر نظر معاونت پژوهشی دانشگاه

شابک: ۹۷۸-۹۶۴-۲۱۷-۶۰۹-۰۰

سروراستار: دکتر نازیلا فرمانی انوشه صفحه‌آرا: دکتر وحید رضائی تبار، شقایق خودکامه

طراح جلد: سمیرا حاجی‌گلدی ناظر فنی: عبدالرضا گودرزی

مرکز چاپ و انتشارات دانشگاه علامه طباطبائی؛ صندوق پستی: ۱۵۸۱۵/۳۴۸۷

فروشگاه آنلاین انتشارات book.atu.ac.ir

سایت مرکز چاپ و انتشارات press.atu.ac.ir

سایت فروشگاه الکترونیکی mybooket.com/atu

آدرس فروشگاه مرکزی: تهران انتهای بلوار دهکده المپیک، دانشگاه علامه طباطبائی

روبه‌روی ساختمان مرکزی جنب باجه بانک ملی تلفن: ۴۸۳۹۲۶۵۹ همراه: ۰۹۹۰۳۴۹۹۹۸۵

ویرایش دوم، چاپ اول ۱۴۰۳ شمارگان: ۱۰۰۰ قیمت: ۴۵۰۰۰۰ تومان

تقدیم به پدر و مادر عزیزم

و

تقدیم به همسر مهربانم و نور چشمانم ایلیا و آوای

عزیزم

وحید رضائی تبار

دانشیار گروه آمار

دانشگاه علامه طباطبائی

پیشگفتار مؤلفان

کتابی که در پیش رو دارید، دربردارنده‌ی توضیحات مفیدی در رابطه با دانش نوین داده‌کاوی است. در سال‌های اخیر و در مواجهه با حجم بسیار زیادی از داده‌ها در حوزه‌های مختلف آموزش، بانکداری، بازاریابی و غیره، استفاده از داده‌کاوی به‌عنوان یکی از الزامات مطرح شده است. همچنین زبان‌های برنامه‌نویسی زیادی برای پیاده‌سازی الگوریتم‌های داده‌کاوی نیز مطرح شده که در این کتاب با تأکید بر تحلیل مسائل از دید آماری، از زبان برنامه‌نویسی R استفاده شده است که یکی از بهترین زبان‌های برنامه‌نویسی برای پیاده‌سازی مسائل داده‌کاوی است.

در فصل اول این کتاب، تاریخچه علم داده‌کاوی و اهمیت مطالعه آن، نصب و راه‌اندازی زبان برنامه‌نویسی R ، دستورات پایه‌ای در R و توضیح انواع نمودارها ارائه شده است. از آن جا که آماده‌سازی داده‌ها از اهمیت خاصی برخوردار است و نتیجه نهایی هر آزمایش را تحت تأثیر قرار می‌دهد، در فصل دوم در خصوص آماده‌سازی داده‌ها صحبت شده است. در فصل سوم و چهارم به ترتیب به معرفی و تشریح روش‌های رده‌بندی کلاسیک و روش‌های خوشه‌بندی و انواع الگوریتم‌های آن‌ها پرداخته‌ایم. در فصل‌های پنجم و ششم شبکه بیزی، روش‌های یادگیری شبکه بیزی که شامل یادگیری ساختاری و پارامتری هستند، معرفی شده است. شبکه‌های بیزی نوع خاصی از مدل‌های گرافیکی به نام گراف‌های بدون دور سوار (DAG) هستند که برای تعیین رابطه علت و معلولی به کار می‌روند. در این کتاب، برای هر فصل تمرینات تشریحی مطرح شده است که حل آن‌ها به دانشجویان در درک کامل متن و موضوعات مربوطه کمک می‌کند.

این کتاب خصوصاً به نیازهای آموزشی دانشجویان رشته‌های مختلف آمار، علوم کامپیوتر و مهندسی صنایع نیز پاسخ می‌دهد.

امید است انتشار این کتاب مورد استفاده دانشجویان واقع شده و به‌عنوان یک منبع مفید درسی به آن‌ها کمک کند و در عین حال مورد قبول اساتید نیز واقع شود و گامی هر چند کوچک در اعتلای دانش داده‌کاوی در کشور محسوب شود. در انتها ضمن تشکر از مدیریت و دست‌اندرکاران نشر دانشگاهی علامه طباطبائی که در طی مراحل تکمیل این کتاب نهایت همکاری را با مؤلفان داشتند، از خوانندگان گرامی درخواست می‌کنیم که ما را از انتقادات و پیشنهادهای خود در جهت ارتقای کیفی کتاب بهره‌مند سازید و همچنین سپاسگزار خواهیم بود اگر نظرها، پیشنهادهای و سؤال‌های احتمالی خود را با ما به آدرس ایمیل vhrezaei@atu.ac.ir در میان بگذارید.

وحید رضائی تبار

شهریور ماه ۱۳۹۹

فهرست مطالب

فصل اول: مقدمه و معرفی زبان برنامه‌نویسی R	۲
۱-۱- مقدمه	۲
۲-۱- تاریخچه داده‌کاوی	۲
۳-۱- تعریف‌های داده‌کاوی	۲
۴-۱- روش‌های یادگیری مدل در داده‌کاوی	۵
۵-۱- روش‌های داده‌کاوی	۵
۶-۱- الگوریتم‌های داده‌کاوی	۹
۷-۱- کاربردهای داده‌کاوی	۱۱
۸-۱- آمار و داده‌کاوی	۱۳
۹-۱- محدودیت‌های داده‌کاوی	۱۶
۱۰-۱- ضرورت داده‌کاوی	۱۶
۱۱-۱- چند مثال درمورد مفهوم داده‌کاوی	۱۶
۱۲-۱- نرم‌افزارهای داده‌کاوی	۱۸
۱۳-۱- زبان برنامه‌نویسی R	۲۱

۲۲	۱-۱۳-۱- نصب و راه‌اندازی
۲۴	۱-۱۳-۱- زبان برنامه‌نویسی <i>RStudio</i>
۲۵	۱-۱۳-۲- نصب و روزآمدسازی بسته‌های نرم‌افزاری
۲۸	۱-۱۴- بسته‌های نرم‌افزاری داده‌کاوی به تفکیک فصل
۲۹	۱-۱۵- بسته نرم‌افزاری <i>Rattle</i>
۳۱	۱-۱۵-۱- فراخوانی مجموعه داده‌ها در <i>Rattle</i>
۳۲	۱-۱۶-۱- فراخوانی داده‌های پیش‌فرض <i>R</i>
۳۲	۱-۱۷- محاسبات پایه
۳۷	۱-۱۸-۱- دستورهای مهم زبان برنامه‌نویسی <i>R</i>
۳۷	۱-۱۸-۱- بردار
۳۷	۱-۱۸-۲- تابع‌ها در زبان برنامه‌نویسی <i>R</i>
۳۹	۱-۱۸-۳- <i>which()</i>
۳۹	۱-۱۸-۴- عملگر شرطی <i>if</i>
۴۰	۱-۱۸-۵- حلقه‌ی <i>for</i>
۴۳	۱-۱۸-۶- حلقه‌ی <i>while</i>
۴۴	۱-۱۸-۷- دستور <i>sapply</i>

- ۸-۱۸-۱- گراف ۴۴
- ۹-۱۸-۱- آرایه ۴۶
- ۱۰-۱۸-۱- ماتریس ۴۷
- ۱۱-۱۸-۱- ضرب ماتریسی ۴۸
- ۱۲-۱۸-۱- سایر عملگرهای ماتریسی ۴۸
- ۱۳-۱۸-۱- متغیرهای تصادفی ۵۱
- ۱۴-۱۸-۱- استانداردسازی داده‌ها ۵۲
- ۱۵-۱۸-۱- رسم نمودار ۵۳
- ۱۶-۱۸-۱- نمودار میله‌ای ۵۵
- ۱۷-۱۸-۱- نمودار دایره‌ای ۵۸
- ۱۸-۱۸-۱- نمودار نقطه‌ای ۶۱
- ۱۹-۱۸-۱- نمودارهای سه‌متغیری ۶۴
- ۲۰-۱۸-۱- نمودار جعبه‌ای ۶۷
- ۲۱-۱۸-۱- نمودار شاخه و برگ ۶۹
- ۲۲-۱۸-۱- نمودارهای سری زمانی ۷۰
- ۲۳-۱۸-۱- نمودارهای موازی ۷۱

۱-۱۸-۲۴- نمودارهای چندضلعی ۷۲

۱-۱۸-۲۵- نمودارهای ماتریسی ۷۳

۱-۱۸-۲۶- نمودار سه‌بُعدی ۷۴

۱-۱۸-۲۷- فلوچارت ۷۵

۱-۱۸-۲۸- رسم تابع‌های ریاضی ۷۸

۱-۱۸-۲۹- نمودار خطی ۸۲

۱-۱۸-۳۰- نقشه‌های جغرافیایی ۸۳

۱-۱۹- نکته‌ها ۸۵

۱-۲۰- خلاصه‌ی فصل ۹۴

تمرین ۹۵

فصل دوم: آماده‌سازی داده ۹۹

۲-۱- شناخت انواع داده و خصیصه‌های آن ۱۰۰

۲-۲- شاخص‌های مرکزی و پراکندگی ۱۰۰

۲-۲-۱- فراوانی ۱۰۱

۲-۲-۲- میانگین ۱۰۲

۲-۲-۳- میانه ۱۰۳

- ۲-۲-۴-مد ۱۰۴
- ۲-۲-۵-واریانس ۱۰۵
- ۲-۲-۶-انحراف معیار ۱۰۵
- ۲-۳-اهمیت آماده‌سازی داده‌ها ۱۰۶
- ۲-۴-کارهای عمده در آماده‌سازی و پیش‌پردازش داده‌ها ۱۰۷
- ۲-۴-۱-پاکسازی داده‌ها ۱۰۸
- ۲-۴-۱-۱-تحلیل داده‌های گمشده ۱۰۸
- ۲-۴-۱-۲-هموارسازی نوفه‌ها ۱۱۷
- ۲-۴-۱-۳-مشخص کردن داده‌های دور افتاده ۱۱۹
- ۲-۴-۱-۳-۱-روش دامنه‌ای ۱۲۱
- ۲-۴-۱-۳-۲-روش‌های آزمون آماری ۱۲۳
- ۱-آزمون کیو دیکسون ۱۲۳
- ۲-آزمون گراب ۱۲۴
- ۲-۴-۱-۴-رفع مشکل افزونگی داده‌ها ۱۴۰
- ۲-۴-۲-یکپارچه‌سازی داده‌ها ۱۴۱
- ۲-۴-۳-تبدیل داده‌ها ۱۴۱

۱۴۱	۲-۴-۳-۱- نرمالسازی داده‌ها
۱۴۶	۲-۴-۳-۲- تعمیم داده‌ها
۱۴۶	۲-۴-۳-۳- انبوهش
۱۴۷	۲-۴-۴- کاهش داده‌ها
۱۴۷	۲-۴-۴-۱- کاهش بُعد
۱۴۷	۲-۴-۴-۱- تحلیل مؤلفه‌های اصلی (PCA)
۱۶۷	۲-۴-۴-۲- فشرده‌سازی داده‌ها
۱۷۲	۲-۴-۴-۳- گسسته‌سازی
۱۷۳	۲-۵- خلاصه‌ی فصل
۱۷۳	تمرین
۱۷۵	فصل سوم: معرفی روش‌های رده‌بندی کلاسیک
۱۷۶	۳-۱- مقدمه
۱۷۶	۳-۲- رگرسیون لوژستیک
۱۷۹	۳-۳- درخت تصمیم
۱۷۹	۳-۳-۱- استقراء درخت تصمیم
۱۷۹	۳-۳-۲- مفهوم‌های اصلی درخت تصمیم

۱۸۰	 ۳-۳-۳ انواع درخت تصمیم
۱۸۰	 ۳-۳-۴ مزیت‌های درخت تصمیم
۱۸۱	 ۳-۳-۵ عیب‌های درخت تصمیم
۱۸۲	 ۳-۳-۶ نحوه نمایش درخت تصمیم
۱۸۳	 ۳-۳-۷ معیارهای انتخاب صفت خاص
۱۸۳	 ۳-۳-۷-۱ آنتروپی
۱۸۴	 ۳-۳-۷-۲ اطلاع سودمندی
۱۸۵	 ۳-۳-۷-۳ شاخص جینی
۱۸۶	 ۳-۳-۸ الگوریتم‌های درخت تصمیم
۱۸۷	 ۳-۳-۸-۱ الگوریتم ID_3
۱۸۷	 ۳-۳-۸-۲ الگوریتم ID_4
۱۸۸	 ۳-۳-۸-۳ الگوریتم ID_5
۱۸۸	 ۳-۳-۸-۴ الگوریتم ID_5_hat
۱۸۸	 ۳-۳-۸-۵ الگوریتم $C_{4.5}$
۱۸۹	 ۳-۳-۸-۶ الگوریتم $C_{0.5}$
۱۸۹	 ۳-۳-۸-۷ الگوریتم $CART$

- ۳-۳-۹- ارزیابی درخت تصمیم ۱۹۰
- ۳-۳-۱۰- هرس کردن درخت ۱۹۱
- ۳-۳-۱۰-۱- انواع روش‌های هرس کردن ۱۹۱
- ۳-۴- k - نزدیک‌ترین همسایه ۲۰۹
- ۳-۴-۱- الگوریتم k - نزدیک‌ترین همسایه (KNN) ۲۱۰
- ۳-۴-۲- مزیت و محدودیت الگوریتم k - نزدیک‌ترین همسایه در فرایندهای پیش‌بینی ۲۱۵
- ۳-۵- ماشین بردار پشتیبان ۲۴۷
- ۳-۵-۱- حالت جدایی‌پذیر در ماشین بردار پشتیبان ۲۵۰
- ۳-۵-۲- ماشین بردار پشتیبان در حالت‌های ناخطی و جدایی‌ناپذیر ۲۵۶
- ۳-۵-۲-۱- داده‌های جدایی‌ناپذیر ۲۵۶
- ۳-۵-۲-۲- ماشین‌های بردار پشتیبان ناخطی ۲۵۸
- ۳-۵-۳- ماشین بردار پشتیبان و شبکه عصبی ۲۶۰
- ۳-۵-۴- کاربردهای ماشین بردار پشتیبان ۲۶۱
- ۳-۵-۵- نقاط ضعف ماشین‌های بردار پشتیبان ۲۶۲
- ۳-۶- روش‌های رده‌بندی ترکیبی ۲۶۶
- ۳-۶-۱- الگوریتم انبوهش تصادفی ۲۶۶

۲۷۱	۳-۶-۲- الگوریتم بگینگ
۲۷۳	۳-۶-۳- الگوریتم بوستینگ
۲۷۵	۳-۷- خلاصه‌ی فصل
۳۰۵	تمرین
۳۰۷	فصل چهارم: خوشه‌بندی
۳۰۸	۴-۱- مقدمه
۳۱۰	۴-۲- نقاط قوت روش خوشه‌بندی
۳۱۰	۴-۳- نقاط ضعف روش خوشه‌بندی
۳۱۰	۴-۴- تعیین تعداد خوشه
۳۱۱	۴-۵- ارزیابی اعتبار در خوشه‌بندی
۳۱۳	۴-۶- ماتریس تشابه و فاصله
۳۱۳	۴-۶-۱- تابع تشابه
۳۱۴	۴-۶-۲- تابع فاصله
۳۱۴	۴-۶-۳- ماتریس تشابه
۳۱۵	۴-۶-۴- ماتریس فاصله
۳۱۵	۴-۷- روش‌های اصلی خوشه‌بندی

۳۱۵	 روش های افرازی	۱-۷-۴
۳۱۶	 $k - means$ الگوریتم	۱-۱-۷-۴
۳۲۳	 $k - medoids$ الگوریتم	۲-۱-۷-۴
۳۲۸	 روش های سلسله مراتبی	۲-۷-۴
۳۳۰	 <i>Birch</i> الگوریتم	۱-۲-۷-۴
۳۳۲	 <i>chameleon</i> الگوریتم	۲-۲-۷-۴
۳۳۴	 <i>AGNES</i> الگوریتم	۳-۲-۷-۴
۳۳۴	 <i>sin gel - link</i> خوشه بندی با روش	۱-۳-۲-۷-۴
۳۳۵	 <i>complete - link</i> خوشه بندی با روش	۲-۳-۲-۷-۴
۳۳۶	 <i>Average - link</i> خوشه بندی با روش	۳-۳-۲-۷-۴
۳۴۳	 مقایسه خوشه بندی سلسله مراتبی و غیر سلسله مراتبی	۴-۲-۷-۴
۳۴۳	 روش های مبتنی بر چگال	۳-۷-۴
۳۴۴	 <i>DBSCAN</i> الگوریتم	۱-۳-۷-۴
۳۴۵	 <i>DBSCAN</i> مزیت ها و عیب های الگوریتم	۱-۱-۳-۷-۴
۳۴۷	 <i>OPTICS</i> الگوریتم	۲-۳-۷-۴
۳۵۷	 روش مشبکی مبنا	۴-۷-۴

۳۵۷		STING الگوریتم	۴-۷-۱-۴
۳۶۲		خلاصه‌ی فصل	۴-۸
۳۶۳		تمرین	
۳۶۵		فصل پنجم: مقدمه‌ای بر شبکه‌های بیزی	
۳۶۶		مقدمه	۵-۱
۳۶۶		تاریخچه شبکه‌های بیزی	۵-۲
۳۶۸		معرفی شبکه‌های بیزی	۵-۳
۳۷۰		ساختار شبکه‌های بیزی	۵-۴
۳۷۱		ویژگی مارکوفی	۵-۴-۱
۳۷۱		استدلال با شبکه‌های بیزی	۵-۵-۵
۳۷۱		انواع استنباط	۵-۵-۱
۳۷۳		استدلال احتمال‌ها توسط توزیع احتمال مشترک (توأم)	۵-۵-۲
۳۸۰		انواع یادگیری در شبکه‌های بیزی	۵-۶-۱
۳۸۲		یادگیری ساختاری در شبکه‌های بیزی	۵-۶-۱-۱
۳۸۲		روش‌های مبتنی بر قید	۵-۶-۱-۱-۱
۳۸۳		روش‌های جست‌وجو- امتیازدهی	۵-۶-۱-۲

۳۸۶	۵-۶-۱-۲-۱- روش های جست و جو
۳۸۶	۵-۶-۱-۳- روش ترکیبی
۳۸۷	۵-۶-۲- یادگیری پارامتری شبکه بیزی
۳۸۸	۵-۷- نرم افزارهای مدل سازی شبکه بیزی
۳۸۹	۵-۸- خلاصه ی فصل
۳۸۹	تمرین
۳۹۱	فصل ششم: یادگیری ساختاری و پارامتری شبکه های بیزی
۳۹۲	۶-۱- مقدمه
۳۹۳	۶-۲- معرفی چند الگوریتم مهم مبتنی بر قید
۳۹۳	۱- الگوریتم <i>PC</i>
۳۹۵	۲- ساختن کالبد <i>DAG</i>
۴۰۱	۳- الگوریتم <i>GS</i>
۴۰۷	۴- الگوریتم <i>IAMB</i>
۴۰۹	۵- الگوریتم <i>Inter IAMB</i>
۴۱۱	۶-۳- الگوریتم های امتیازدهی
۴۱۳	۶-۳-۱- معرفی و طبقه بندی تابع های امتیاز شبکه بیزی

- ۶-۳-۲- تابع‌های امتیاز وابسته به توزیع پیشینی ۴۱۴
- ۶-۳-۳- مترهای مبتنی بر مفهوم‌های نظریه اطلاع ۴۱۹
- ۶-۳-۴- الگوریتم‌های مبتنی بر جست‌وجوی حریصانه در یادگیری ساختار ۴۲۰
- ۶-۳-۵- الگوریتم K_2 ۴۲۱
- ۶-۳-۶- الگوریتم جست‌وجوی ممنوع ۴۲۳
- ۶-۴-۴- الگوریتم‌های ترکیبی ۴۲۶
- ۶-۵-۵- یادگیری شبکه‌های بیزی ۴۳۲
- ۶-۵-۱- ساختار معلوم و داده‌های کامل ۴۳۳
- ۶-۵-۱-۱- برآورد بیشینه درست‌نمایی ۴۳۳
- ۶-۵-۱-۲- برآورد بیزی ۴۳۶
- ۶-۵-۲- ساختار معلوم و داده‌ها ناکامل ۴۳۸
- ۶-۵-۳- ساختار نامعلوم و داده‌ها کامل ۴۳۸
- ۶-۵-۴- ساختار نامعلوم و داده‌ها ناکامل ۴۳۸
- ۶-۶-۶- قضیه بیز ۴۳۹
- ۶-۷-۷- رده‌بند بیزی ساده ۴۴۰
- ۶-۷-۱- خصیصه‌های رده‌بندی بیزی ساده ۴۴۲

۶-۸- هموارسازی ۴۴۳

۶-۹- خلاصه‌ی فصل ۴۷۲

تمرین ۴۷۵

واژه‌نامه فارسی - انگلیسی ۴۷۷

نام‌نامه ۴۹۰

مرجع‌ها ۴۹۲