

۲۰۵۳۹۸۸

اصول و روش‌های یادگیری

علم داده

تألیف:

Sinan Ozdemir

مترجمان:

دکتر حامد تابش

الهام نظری، پرنیان عسگری، مرضیه افکن پور
مهران آغه میری، محبوبه اسلامی، شیوا قادری

انتشارات پندار پارس

سرشناسه	: ازدمیر، سینان Ozdemir, Sinan
عنوان و نام پدیدآور	: اصول و روشهای یادگیری علم داده/ تالیف سینان ازدمیر؛ مترجمان حامد تابش... [و دیگران].
مشخصات نشر	: تهران: پندار پارس، ۱۳۹۸.
مشخصات ظاهری	: ۴۱۰ ص: مصور، جدول، نمودار.
شابک	: 978-600-8201-72-4
وضعیت فهرست نویسی	: فیبا
یادداشت	: عنوان اصلی: Principles of Data Science: Learn the techniques and math you need to start making sense of your data, 2016
یادداشت	: مترجمان حامد تابش، الهام نظری، پرنیان عسگری، مرضیه افکن پور، مهران آغمیری، محبوبه اسلامی، شیوا قادری.
موضوع	: داده‌کاری
موضوع	: Data mining
موضوع	: کسب و کار -- داده‌پدازی
موضوع	: Business -- Data processing
شنامه افزوده	: تابش، حامد، ۱۳۵۶ -- مترجم
رده بندی کنگره	: ۹/۲ QA
رده بندی دیویی	: ۳ / ۱۰۰۶
شماره کتابشناسی ملی	: ۱۳۰۰۰

انتشارات پندار پارس



دفتر فروش: انقلاب، ابتدای کارگر جنوبی، کوی رشاجی، پلاک ۴، واحد ۱۶ www.pendarepars.com
 تلفن: ۶۶۵۷۲۳۳۵ - تلفکس: ۶۶۹۲۶۵۷۸ همراه: ۰۲۴۵۲۳۴۸ info@pendarepars.com

نام کتاب	: اصول و روشهای یادگیری علم داده
ناشر	: انتشارات پندار پارس
تالیف	: سینان ازدمیر
ترجمه	: حامد تابش، الهام نظری، پرنیان عسگری، مرضیه افکن پور، مهران آغمیری، محبوبه اسلامی، شیوا قادری
چاپ نخست	: شهریور ۹۸
شمارگان	: ۵۰۰ نسخه
طرح جلد	: رامین شکرالهی
چاپ، صحافی	: فرشویه

شابک: ۹۷۸-۶۰۰-۸۲۰۱-۷۲-۴

: ۸۰۰۰۰ تومان

هرگونه کپی برداری، تکثیر و چاپ کاغذی یا الکترونیکی از این کتاب بدون اجازه ناشر تخلف بوده و پیگرد قانونی دارد *

فهرست

فصل ۱: چگونه به عنوان یک متخصص علوم داده به نظر برسیم؟	۵
علم داده چیست؟	۷
اصطلاحات پایه	۷
چرا علم داده؟	۹
مثال - تکنولوژی‌های Sigma	۹
نمودار وزن علم داده	۱۱
ریاضی	۱۳
مثال - مدل‌های spawn-reruit	۱۳
برنامه‌نویسی کامپیوتر	۱۵
چرا Python؟	۱۵
پایتون در عمل	۱۶
مثال پایه از پایتون	۱۸
مثال - تجزیه یک توئیت واحد	۱۹
حوزه دانش	۲۰
برخی اصطلاحات بیشتر	۲۱
مطالعات موردی در علوم داده	۲۳
مطالعه موردی - خودکار و اتومازیسیون کردن فرم‌های کاغذی دولتی	۲۴
نادیده گرفتن جنبه انسانی، آیا درست است؟	۲۵
مطالعه موردی - دلارهای بازاریابی	۲۵
بودجه‌های تبلیغاتی	۲۶
نمودار بودجه‌های تبلیغاتی	۲۷

- ۲۷ مطالعه موردی - چه چیزهایی در توصیف یک شغل استفاده می‌شود؟
- ۲۸ یک مثال از لیست کارهای متخصصان علوم داده
- ۳۳ فصل ۲: انواع داده
- ۳۳ طعم و مزه داده‌ها
- ۳۴ چرا باید به این تمایز نگاه کنیم؟
- ۳۵ داده‌های ساختاریافته در برابر داده‌های بدون ساختار
- ۳۶ مثالی از پیش‌پردازش داده‌ها
- ۳۷ شمارش کلمه / عبارت
- ۳۷ وجود برخی کاراکترهای خاص
- ۳۷ طول نسبی متن
- ۳۸ انتخاب عنوان‌ها (موضوعات)
- ۳۹ داده‌های کمی در برابر داده‌های کیفی
- ۳۹ مثال - داده کافی‌شاپ
- ۴۰ دو مورد مهم برای یادآوری
- ۴۳ بررسی و کاوش عمیق‌تر
- ۴۳ مسیر تاکنون پیموده شده
- ۴۴ چهار سطح داده
- ۴۴ سطح اسمی
- ۴۵ عملیات ریاضی مجاز
- ۴۵ اندازه مرکز
- ۴۶ چه داده‌هایی در سطح اسمی است
- ۴۶ سطح رتبه‌ای
- ۴۶ مثال‌ها

۴۷.....	عملیات ریاضی مجاز
۴۷.....	اندازه مرکز
۴۹.....	بررسی و بازنگری سریع
۴۹.....	سطح فاصله‌ای
۴۹.....	مثال
۵۰.....	عملیات ریاضی مجاز
۵۰.....	اندازه مرکز
۵۱.....	اندازه تغییرات
۵۱.....	اندراف مینر
۵۲.....	سطح نسبی
۵۲.....	مثال‌ها
۵۴.....	اندازه مرکز
۵۴.....	چالش‌های سطح نسبی
۵۵.....	داده‌ها در مقابل چشمان بیننده است!!
۵۷.....	فصل ۳: مراحل پنج‌گانه علوم داده
۵۷.....	معرفی علم داده
۵۷.....	بررسی پنج مرحله
۵۸.....	پرسیدن یک سوال جالب
۵۸.....	به دست آوردن داده
۵۸.....	بررسی داده
۵۹.....	مدل‌سازی داده
۵۹.....	برقراری ارتباط و بصری سازی نتایج
۵۹.....	بررسی داده

۶۰	سؤالات اساسی برای اکتشاف داده
۶۱	مجموعه داده ۱ - Yelp
۶۴	فرمت داده‌ای
۶۴	سری‌ها
۶۵	نکات اکتشافی برای داده‌های کیفی
۶۷	فیلتر کردن در pandas
۶۹	سطح‌های مرتبه‌ای
۷۱	مجموعه داده ۲ - titanic
۷۷	فصل ۴: ریاضیات پایه
۷۷	ریاضیات به راز جریسته
۷۸	اصطلاحات و نمادها - پایه
۷۸	بردارها و ماتریس‌ها
۸۱	تمرین
۸۱	نمادهای علم حساب
۸۱	مجموع (جمع)
۸۲	تناسب
۸۳	حاصل ضرب نقطه‌ای
۸۶	نمودارها
۸۷	لگاریتم و نما
۹۰	نظریه مجموعه
۹۵	جبر خطی
۹۵	ضرب ماتریس‌ها
۹۵	نکاتی در رابطه با ضرب ماتریس‌ها

فصل ۵: غیرممکن یا غیرمحمتم - مقدمه‌ای ساده بر احتمال	۱۰۱
تعاریف پایه	۱۰۲
احتمال	۱۰۲
بیزین در مقابل فریکوئنسیست	۱۰۴
رویکرد فریکوئنسیست	۱۰۴
مثال - آمار بازاریابی	۱۰۵
انون اعداد بزرگ	۱۰۵
رویدادهای ترکیبی	۱۰۷
احتمال شرطی	۱۱۰
قوانین ابرسان	۱۱۱
قانون افزودن	۱۱۱
انحصار متقابل	۱۱۳
قانون ضرب	۱۱۳
استقلال	۱۱۵
رویدادهای تکمیلی	۱۱۵
کمی عمیق‌تر بنگریم	۱۱۶
فصل ۶: احتمال پیشرفته	۱۱۹
مجموعه رویدادهای جامع	۱۱۹
ایده‌های بیزی بازبینی‌شده	۱۲۰
قاعده بیز	۱۲۰
کاربردهای بیشتر قضیه بیز	۱۲۴
مثال - تایتانیک	۱۲۵
مثال - آزمایش‌های پزشکی	۱۲۶

- ۱۲۸..... منغیرهای تصادفی
- ۱۲۹..... متغیرهای تصادفی گسسته
- ۱۳۵..... انواع متغیرهای تصادفی گسسته
- ۱۳۵..... متغیرهای تصادفی دوجمله‌ای
- ۱۳۶..... مثال - جلسات جمع‌آوری کمک مالی
- ۱۳۶..... مثال - افتتاح رستوران
- ۱۳۷..... مثال - گروه خونی
- ۱۳۸..... متغیرهای تصادفی هندسی
- ۱۳۹..... مثال - آب و هوا
- ۱۴۰..... متغیر تصادفی پواسون
- ۱۴۱..... مثال - مرکز تلفن
- ۱۴۲..... متغیرهای تصادفی پیوسته
- ۱۴۷..... فصل ۷: آمار پایه
- ۱۴۷..... آمار چیست؟
- ۱۴۹..... چگونه داده‌ها را به دست آوریم و نمونه بگیریم
- ۱۴۹..... به دست آوردن اطلاعات
- ۱۴۹..... مشاهدات
- ۱۵۰..... تجربی
- ۱۵۲..... داده‌های نمونه‌گیری
- ۱۵۲..... نمونه‌گیری احتمالی
- ۱۵۳..... نمونه‌گیری تصادفی
- ۱۵۴..... نمونه‌گیری احتمالی نابرابر
- ۱۵۵..... چگونه می‌توانیم آمار را اندازه‌گیری کنیم؟

۱۵۵.....	اندازه‌گیری مرکز.....
۱۵۶.....	اندازه‌گیری متغیرها.....
۱۶۱.....	تعریف.....
۱۶۱.....	مثال - حقوق کارمندان.....
۱۶۲.....	اندازه‌گیری مقادیر نسبی.....
۱۶۸.....	بخش تفصیلی - همبستگی داده‌ها.....
۱۷۰.....	قاعده تجربی.....
۱۷۲.....	مثال - نمرات امتحان.....
۱۷۳.....	فصل ۸: آمار پیشرفته.....
۱۷۳.....	برآورد نقطه‌ای.....
۱۷۸.....	توزیع نمونه‌گیری.....
۱۸۱.....	فاصله اطمینان.....
۱۸۴.....	آزمون فرضیه.....
۱۸۵.....	انجام آزمون فرضیه.....
۱۸۷.....	آزمون t تک نمونه‌ای.....
۱۸۷.....	مثالی از آزمون t تک نمونه‌ای.....
۱۸۸.....	فرضیه‌های یک نمونه آزمون t.....
۱۹۱.....	خطای نوع اول و نوع دوم.....
۱۹۱.....	آزمون فرضیه برای متغیرهای دسته‌ای.....
۱۹۲.....	آزمون نیکویی برازش کای اسکوئر.....
۱۹۲.....	فرضیه‌هایی از آزمون نیکویی برازش کای اسکوئر.....
۱۹۳.....	مثالی از آزمون نیکویی برازش کای اسکوئر.....
۱۹۵.....	آزمون کای اسکوئر برای وابسته / مستقل.....

- ۱۹۵..... فرضیه آزمون مستقل کای اسکوئر
- فصل ۹: به اشتراک‌گذاری داده ۱۹۹
- ۲۰۰..... چرا به اشتراک‌گذاری مهم است؟
- ۲۰۰..... تشخیص بصری‌سازی مؤثر و غیرمؤثر
- ۲۰۱..... نمودار پراکنندگی
- ۲۰۳..... نمودارهای خطی
- ۲۰۴..... نمودار میله‌ای
- ۲۰۶..... هیستوگرام
- ۲۰۸..... نمودار جعبه‌ای
- ۲۱۱..... هنگامی که نمودارها را با دروغ می‌گویند
- ۲۱۱..... همبستگی در مقابل علیت
- ۲۱۴..... پارادوکس سیمپسون
- ۲۱۵..... اگر همبستگی دو متغیر به معنی علت و معلول بودن آنها نباشد چه کار کنیم؟
- ۲۱۶..... ارتباط کلامی
- ۲۱۶..... گفتن یک داستان
- ۲۱۷..... ارائه برای مکان‌های رسمی‌تر
- ۲۱۸..... استراتژی "چرا، چگونه، چه چیزی"، برای ارائه دادن
- فصل ۱ یادگیری ماشین ۲۲۱
- ۲۲۲..... یادگیری ماشین چیست؟
- ۲۲۳..... مثال- تشخیص چهره
- ۲۲۴..... یادگیری ماشین کامل نیست
- ۲۲۵..... یادگیری ماشین چگونه کار می‌کند؟
- ۲۲۶..... مروری بر مدل‌های یادگیری ماشین

- ۲۲۶.....انواع مختلف یادگیری ماشین
- ۲۲۷.....یادگیری تحت نظارت
- ۲۲۷.....مثال - پیش‌بینی حمله قلبی
- ۲۳۰.....انواع مختلف مدل‌های یادگیری تحت نظارت
- ۲۳۰.....رگرسیون
- ۲۳۱.....طبقه‌بندی
- ۲۳۱.....مثال - رگرسیون
- ۲۳۲.....داده در چشم‌های بیننده است
- ۲۳۲.....یادگیری بدون نظارت
- ۲۳۴.....یادگیری تقویتی
- ۲۳۵.....مروری بر انواع یادگیری ماشین
- ۲۳۷.....چگونه مدل‌سازی آماری در همه این مدل‌ها تأثیر دارد؟
- ۲۳۷.....رگرسیون خطی
- ۲۴۲.....اضافه کردن پیشگوه‌های بیشتر
- ۲۴۵.....معیارهای رگرسیون
- ۲۵۳.....رگرسیون لجستیک
- ۲۵۴.....احتمال، شانس و لگاریتم شانس
- ۲۵۸.....محاسبات ریاضی رگرسیون لجستیک
- ۲۶۱.....متغیرهای ساختگی
- ۲۶۷.....فصل ۱۱: آیا می‌توان از طریق درختان پیش‌بینی‌ها را انجام داد؟
- ۲۶۷.....طبقه و کلاسه‌بندی بیزین ساده
- ۲۷۶.....درخت تصمیم
- ۲۷۸.....چگونه یک کامپیوتر یک درخت رگرسیون ایجاد می‌کند؟

- چگونه رایانه مناسب یک درخت طبقه‌بندی است؟ ۲۷۹
- یادگیری بدون نظارت ۲۸۴
- چه موقعی از یادگیری بدون نظارت استفاده می‌کنیم ۲۸۴
- خوشه‌بندی K-means ۲۸۵
- یک مثال روشن - نقاط داده‌ای ۲۸۷
- مثال - دلستر ۲۹۲
- انتخاب یک شماره بهینه برای k و اعتبارسنجی خوشه ۲۹۵
- اثر کم‌ری Silhouette ۲۹۵
- استخراج ویژگی و تبدیل مؤلفه اصلی ۲۹۷
- فصل ۱۲: فراتر از نیاز ۳۰۹**
- توازن بین واریانس/بایاس ۳۱۰
- خطای ناشی از بایاس ۳۱۰
- خطای ناشی از واریانس ۳۱۰
- مثال - مقایسه وزن بدن و مغز پستانداران ۳۱۱
- نمودار پراکندگی وزن بدن و مغز پستانداران ۳۱۲
- همان نمودار پراکندگی قبلی با نمایش رگرسیون خطی در آن ۳۱۴
- نمودار پراکندگی برای نمونه‌های ۱ و ۲ ۳۱۵
- استفاده از چندجمله‌ای درجه چهار برای اهداف رگرسیون ۳۱۷
- نمودار پراکندگی با استفاده از چندجمله‌ای درجه چهار به‌عنوان تخمین دهنده ما ۳۱۸
- دو حالت نهایی از توازن واریانس/بایاس ۳۱۸
- کم برآزش ۳۱۸
- بیش برآزش ۳۱۹
- چگونگی تأثیر بایاس/واریانس در تابع‌های خطا ۳۱۹

۳۲۱.....	اعتبارسنجی متقاطع K فولد
۳۲۵.....	نمودار خطای KNN در مقابل پیچیدگی KNN
۳۲۵.....	جست‌وجوی توری
۳۲۹.....	بصری کردن خطای آموزشی در مقابل خطای اعتبارسنجی متقاطع
۳۳۱.....	روش‌های انسمبل
۳۳۳.....	جنگل تصادفی
۳۳۸.....	مقایسه جنگلهای تصادفی با برشتهای تصمیم
۳۳۹.....	شبکه‌های عصبی
۳۳۹.....	سازار اساسی
۳۴۷.....	فصل ۱۳: مطالعات موردی
۳۴۷.....	مطالعه موردی نخست: پیش‌بینی قیمت سهام بر اساس رسانه‌های اجتماعی
۳۴۷.....	آنالیز احساسات متن
۳۴۸.....	تجزیه و تحلیل داده‌های اکتشافی
۳۵۸.....	روش رگرسیون
۳۶۰.....	روش طبقه‌بندی
۳۶۳.....	فراتر از این مثال رفتن
۳۶۳.....	مطالعه موردی دوم: چرا برخی از مردم، همسران خود را قریب می‌دهند؟
۳۷۲.....	مطالعه موردی ۳ - استفاده از tensorflow
۳۷۷.....	Tensorflow و شبکه‌های عصبی
۳۸۵.....	ضمایم و پیوست‌ها
۳۸۵.....	واژگان مهم و ضروری کتاب

مقدمه

موضوع این کتاب علم داده است که در زمینه تحقیق و کاربرد علم داده توضیحات بسیار ارزشمندی را عنوان می‌کند. علم داده در چند دهه گذشته به سرعت در حال رشد و توسعه بوده است. به عنوان یک زمینه رو به رشد، توجه زیادی در رسانه‌ها و همچنین در بازار کار به دست آورده است. توجه به علم داده پس از ظهور شرکت‌های فناوری مدل‌سازی از اهمیت ویژه‌ای برخوردار گردید و اخیراً شروع به استخدام تیم‌های داده‌کاوی متخصص کرده‌اند.

این کتاب تلاش خواهد کرد تا شکاف بین تخصص ریاضی / برنامه‌نویسی / داده‌کاوی را متوقف کند. امروزه اکثر مردم حداقل یک (یا شاید دو) تخصص دارند، اما علم داده‌کاوی به بیش از سه تخصص نیاز دارد. به موضوعاتی از هر سه تخصص وارد شدیم و مشکلات پیچیده را قابل حل ساخته‌ایم.

برای ارزیابی نتایج علم داده، دقیقاً داده‌ها را تمیز، کشف و تحلیل خواهیم کرد. یادگیری ماشین و روش‌های یادگیری عمیق را برای حل وظایف پیچیده داده مورد استفاده قرار می‌گیرد را بیان می‌کنیم.

این کتاب چه مواردی را پوشش می‌دهد؟!!

فصل ۱. چگونه به عنوان یک متخصص داده با نظر برسیم: مقدمه‌ای بر اصطلاحات اساسی استفاده شده توسط متخصصان داده و نگاهی به انواع مشکلات که در سراسر این کتاب حل خواهد شد.

فصل ۲. انواع داده‌ها: به سطوح مختلف و نحوه دستکاری انواع داده نگاه اساسی می‌کند.

این فصل شروعی برای بررسی ریاضیات مورد نیاز برای علم داده است.

فصل ۳. پنج مرحله علوم داده: پنج مرحله اساسی در انجام سون داده از جمله دستکاری و تمیز کردن داده‌ها را باز می‌کند، و نمونه‌هایی از هر مرحله را بیان می‌کند.

فصل ۴. ریاضیات پایه: به ما کمک می‌کند تا اصول پایه ریاضی را کشف کنیم که عملکردهای متخصصان داده را با دیدن و حل نمونه‌هایی در حساب، جبر خطی و غیره هدایت می‌کنند.

فصل ۵. غیرممکن یا غیرقابل پیش‌بینی-مقدمه‌ای ساده درباره احتمال: یک نگاه مبتدی به نظریه احتمال و نحوه استفاده از آن برای به دست آوردن درک جهان تصادفی ما است.

فصل ۶. احتمال پیشرفته: از اصول فصل قبلی استفاده می‌کند و قضیه‌ها و نظریه‌هایی نظیر قضیه بایاس را به کار می‌گیرد و امید به کشف معنای پنهان در دنیای داده‌ها دارد.

فصل ۷، آمار پایه: در پی بررسی انواع مشکلی است که استنتاج آماری تلاش می‌کند با استفاده از اصول آزمایشی، نرمال‌سازی و نمونه‌گیری تصادفی توضیح دهد.

فصل ۸، آمار پیشرفته: از آزمون فرضیه‌ها و فاصله اطمینان استفاده می‌کند تا بینشی را از آزمایش‌های ما به دست آورد. داشتن توانایی انتخاب تست مناسب و نحوه تفسیر مقادیر p و سایر نتایج نیز بسیار مهم است.

فصل ۹، ارتباطات داده‌ها: توضیح می‌دهد که چگونه ارتباطات و علیت بر تفسیر ما از داده تأثیر می‌گذارد. همچنین با استفاده از تجسم و بصری‌سازی به دنبال به اشتراک گذاشتن نتایج خود با جهان پیرامونمان است.

فصل ۱۰، مربوط به ملزومات یادگیری ماشین است: بر تعریف یادگیری ماشین متمرکز است و به مثال‌های واقعی در مورد آموزش و زمان استفاده از یادگیری ماشین می‌پردازد.

فصل ۱۱، پیش‌بینی‌ها و درخت تصمیم: به مدل‌های پیچیده‌تر یادگیری ماشین، مانند درخت تصمیم‌گیری و پیش‌بینی‌های منشی بر بیزینس، به منظور حل وظایف مربوط به تحلیل پیچیده‌تر داده‌ها نگاه می‌کند.

فصل ۱۲، فراتر از ملزومات: برخی از اصولی که امروز را که علوم داده‌ها را هدایت می‌کنند معرفی می‌کند، از جمله بایاس و واریانس، در این فصل شبکه‌های عصبی به‌عنوان روش یادگیری عمیق مدرن معرفی می‌شوند.

فصل ۱۳، مطالعات موردی: از مجموعه‌ای از مطالعات موردی استفاده می‌کند تا ایده‌های علوم داده را تقویت کند. همچنین نمونه‌ها و مثال‌های مختلف، از جمله پیش‌بینی قیمت سهام و تشخیص دستخط، بررسی خواهد شد.

آنچه شما در این کتاب نیاز دارید:

این کتاب از پایتون برای تمام نمونه‌های کد و حل مثال‌ها استفاده می‌کند. یک نسخه کامپیوتر با سیستم‌عامل لینوکس / مک / ویندوز با دسترسی به ترمینال یونیکس با پایتون ۲.۷ نصب‌شده موردنیاز است.

نصب و راه‌اندازی توزیع Anaconda نیز توصیه می‌شود، بسیاری از بسته‌های مورد استفاده نمونه‌ها در کتاب از آن استفاده کرده است.