اصول و روش‌های یادگیری

علم داده

تألیف:
Sinan Ozdemir

مترجمان:
دکترحامد 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	:	٤١٠ ص.: مصور، جدول، نمودار.
شابک	:	978-600-8201-72-4
وضعیت فهرست نویسی	:	فیپا
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 افکن‌پور، مهران آغه‌میری، محبوبه اسلامی، شیوا قادری.
موضوع	:	داده‌کاوی
موضوع	:	Data mining
موضوع	:	کسب و کار -- داده‌پردازی
موضوع	:	Business -- Data processing
شناسه افزوده	:	تابش، حامد، ۱۳۵۶-، مترجم
رده بندی کنگره	:	۹/۷۶QA
رده بندی دیویی	:	۳۱۲/۰۰۶
شماره کتابشناسی ملی	:	۵۸۳۴۶۳۴

انتشارات پندارپارس

دفتر فروش: انقلاب، ابتدای کارگرجنوبی، کوی رشتچی، شماره ۱۴، واحد ۱۶ **www.pendarepars.com**

تلفن: ۶۶۵۷۲۳۳۵ - تلفکس: ۶۶۹۲۶۵۷۸همراه: ۰۹۱۲۲۴۵۲۳۴۸ **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 افکن‌پور، مهران آغه میری، محبوبه اسلامی، شیوا قادری**
چاپ نخست	:	**شهریور ۹۸**
شمارگان	:	**۵۰۰ نسخه**
طرح جلد	:	**رامین شکرالهی**
چاپ، صحافی	:	**فرشیوه**
قیمت	:	**۸۰۰۰۰ تومان** **شابک : ۴- ۷۲-۸۲۰۱-۶۰۰-۹۷۸**

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

فهرست

مقدمه

موضوع این کتاب علم داده است که درزمینه تحقیق و کاربرد علم داده توضیحات بسیار ارزشمندی را عنوان می‌کند. علم داده در چند دهه گذشته به‌سرعت در حال رشد و توسعه بوده است. به‌عنوان 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 اخیرا شروع به استخدام تیم‌های داده کاو 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. ما به موضوعاتی از هر سه تخصص وارد شدیم و مشکلات پیچیده را قابل‌حل ساخته‌ایم.

برای ارزیابی نتایج علمی و دقیق، داده‌ها را تمیز، کشف و تحلیل خواهیم کرد. یادگیری ماشین و روش‌های یادگیری عمیق که برای حل وظایف پیچیده داده مورداستفاده قرار می‌گیرد را بیان می‌کنیم.

این کتاب چه مواردی را پوشش می‌دهد؟!!

فصل ۱، چگونه به‌عنوان یک متخصص داده به 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 موردنیاز برای علم داده است.

فصل ۳، پنج مرحله علوم داده: پنج مرحله اساسی در انجام علوم داده، ازجمله 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 نحوه و زمان استفاده از یادگیری ماشینی می‌پردازد.

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

فصل ۱۲، فراتر از ملزومات: برخی از نیروهای مرموز را که علوم داده‌ها را هدایت می‌کنند معرفی می‌کند، ازجمله 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، ازجمله 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 دستگاه 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 مورداستفاده نمونه‌ها در کتاب از آن استفاده کرده است.

این کتاب برای چه کسی است؟

این کتاب برای افرادی است که به دنبال درک و استفاده از شیوه‌های اساسی تحلیل علم داده برای هر حوزه کاری هستند.

خواننده باید با ریاضیات پایه (جبر، شاید احتمالات) نسبتا آشنا باشد و باید کمی با R / Python آشنایی داشته باشد .

انتظار نمی‌رود که خواننده درزمینه داده پیش‌تر کارکرده باشد بااین‌حال، باید تمایل به یادگیری و اعمال تکنیک‌های ارائه‌شده در این کتاب در مجموعه داده‌های خود و یا ارائه و تمرین مثال‌هایی از این قبیل را داشته باشد و تلاش خود را در راه یادگیری روش‌های علم تحلیل و بررسی داده همواره ادامه دهد تا نتایج ارزشمندی را به دست آورد و از آن‌ها در بالا بردن بهره‌وری کار و امور زندگی خود استفاده کند.

فصل ۱

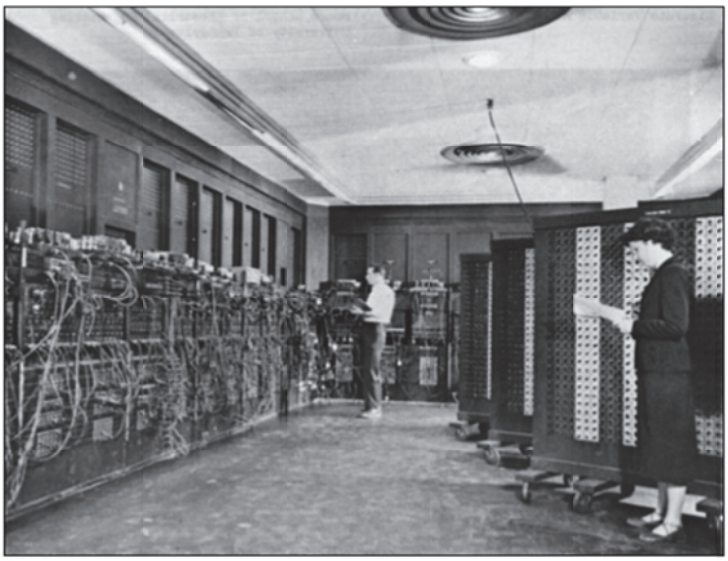
چگونه به‌عنوان یک متخصص علوم داده به نظر برسیم؟

مهم نیست شما در کدام صنعت کار می‌کنید؛ فناوری اطلاعات، مد، غذا یا سرمایه‌گذاری. شکی نیست که داده بر زندگی و کار شما تأثیر می‌گذارد. در این هفته، درزمانی خاص، شما یا مکالمه‌ای در مورد داده‌ها دارید یا می‌شنوید. رسانه‌های خبری داستان‌های بیشتر و بیشتری درباره فراوانی داده‌ها، جرائم اینترنتی و نحوه دسترسی به داده در زندگی ما را پوشش می‌دهند. اما چرا حالا؟ چه چیزی باعث می‌شود که این دوران چنین دورانی برای صنایع مرتبط با داده باشد؟

در قرن نوزدهم، جهان درگیر عصر صنعتی بود. بشر جایگاه خود در صنعت را کنار صنعت اختراعات ماشینی غول‌پیکر پیدا کرد. کاپیتان صنعت، مانند هنری فورد، فرصت‌های بزرگ بازار را در دست این ماشین‌ها به رسمیت شناخت و قادر به دستیابی به سود غیرقابل پیش‌بینی بود. البته عصر صنعتی دارای مزایا و معایبی هم بود. درحالی‌که تولید انبوه کالا را در اختیار مصرف‌کنندگان بیشتری قرارداد، در این زمان نبرد ما با آلودگی نیز شروع شد.

در قرن بیست و یکم، ما در ساخت ماشین‌آلات بسیار حرفه‌ای بودیم؛ هدف این بود که آن‌ها را در ابعاد کوچک‌تر و با سرعت بیشتر بسازیم. عصر صنعتی به پایان رسید و با آنچه به‌عنوان عصر اطلاعات معرفی می‌شد جایگزین شد. ما شروع به استفاده از ماشین‌هایی برای جمع‌آوری و ذخیره اطلاعات (داده‌ها) درباره خود و محیط‌زیست کردیم تا بتوانیم جهان پیرامون خود را بهتر درک کنیم.

از دهه ۱۹۴۰، ماشین‌آلاتی مانند ENIAC (به‌عنوان یکی از، نه اولین، کامپیوترها) محاسبه معادلات ریاضی و اجرای مدل‌ها و شبیه‌سازی‌ها را انجام می‌دادند که هرگز قبل از آن مانند این کامپیوترها نبوده است.



The ENIAC, http://ftp.arl.mil/ftp/historic-computers/

درنهایت یک دستیار آزمایشگاه مناسب داشتیم که می‌توانست اعداد را بهتر از ما اجرا کند! همانند عصر صنعتی، عصر اطلاعات نیز برای ما هم مزایا و هم معایبی داشت. از مزیت‌های این عصر، قطعات فوق‌العاده‌ای از فناوری، ازجمله تلفن‌های همراه و تلویزیون بود. معایب در این مورد به بدی آلودگی در سراسر جهان نیست، اما هنوز هم در قرن بیست و یکم با یک مشکل روبه‌رو شده‌ایم که آن وجود داده‌های زیاد است.

درست است که عصر اطلاعات، در تلاش برای تهیه داده‌ها، تولید داده‌های الکترونیکی را گسترده کرده است. برآوردها نشان می‌دهد که در سال ۲۰۱۱ حدود ۱.۸ تریلیون گیگابایت اطلاعات ایجاد کردیم (تنها یک‌لحظه فکر کنید که این مقدار چقدر است). فقط یک سال بعد، در سال ۲۰۱۲، ما بیش از ۲.۸ تریلیون گیگابایت اطلاعات ایجاد کردیم! گسترش این تعداد ادامه می‌یابد تا به ۴۰ تریلیون گیگابایت تخمین زده‌شده در سال ۲۰۲۰ تنها در یک سال برسد. هر زمان افراد یک (tweet) در فیس‌بوک به اشتراک بگذارند، هر بار که بخواهند، یک رزومه جدید را در نرم‌افزار MS Word ذخیره کنند، یا فقط یک عکس از طریق پیام متنی برای مادرشان ارسال کنند، این کار را انجام می‌دهند.

نه‌تنها داده‌ها را با سرعت بی‌سابقه‌ای ایجاد می‌کنیم، آن‌ها را با سرعت نیز مصرف می‌کنیم. تنها سه سال پیش، در سال ۲۰۱۳، کاربری با استفاده در حد متوسط از تلفن همراه در ماه، کمتر از ۱ گیگابایت اطلاعات استفاده می‌کرد. امروزه این تعداد در ماه به بیش از ۲ گیگابایت تخمین زده‌شده است. آنچه ما دنبال آن هستیم فهم (درک) است. (که از همه این اطلاعات در بیرون، فقط برخی از آن‌ها برای من مفید است! و برخی می‌تواند مفید باشد!)

بنابراین در قرن ۲۱ با مشکل مواجه هستیم. ما اطلاعات زیادی داریم و همچنان در حال تولید مقدار بیشتر از آن هستیم. بشر ماشین‌های کوچک که داده‌ها را ۲۴/۷ جمع‌آوری کنند، ساخته است و کار ما این است که همه آن‌ها را درک کنیم. سن داده را وارد کنید. این سن زمانی است که ماشین‌ها را با اجداد قرن نوزدهم و داده‌های ایجادشده توسط همتایان قرن بیست و یکم و ایجاد فهم و منابع دانش که هر انسان بر روی زمین می‌تواند از آن بهره‌مند شوند، در نظر می‌گیریم. ایالات‌متحده نقش کلیدی جدیدی در دولت برای متخصصان پیشرو در داده ایجاد کرد. شرکت‌های فناوری مانند Reddit، که تاکنون هیچ متخصص داده‌ای در تیم خود نداشتند، اکنون آن‌ها را به‌سوی خود می‌برند. این مزیت کاملا آشکار است - استفاده از داده‌ها برای پیش‌بینی دقیق و شبیه‌سازی، دیدی از دنیا به ما می‌دهد که پیش از این هرگز نبوده است.

به نظر عالی می‌رسد، اما دست آورد چیست؟

در این فصل، اصطلاحات و واژگان مرتبط با علوم داده مدرن را بررسی خواهیم کرد. کلمات و عبارات کلیدی را که در بحث ما در مورد علوم داده در این کتاب ضروری است، خواهیم دید. همچنین خواهیم دید که چرا از علم داده‌ها استفاده می‌کنیم و سه حوزه کلیدی علم داده‌ها از پیش تعیین‌شده در کد پایتون که زبان‌اصلی مورداستفاده در این کتاب است را شروع می‌کنیم:

- اصطلاحات پایه علم داده
- سه حوزه علم داده
- دستورالعمل پایه پایتون

علم داده چیست؟

پیش از هر چیز، به برخی از تعاریف اساسی که در سراسر این کتاب استفاده می‌کنیم خواهیم پرداخت. یک مسئله در این فیلد[1] این است که این فیلد به‌قدری جوان است که تعاریف آن می‌تواند در کتاب درسی تا روزنامه، تا متن مقاله متفاوت باشند.

اصطلاحات پایه

به‌طورکلی تعاریف زیر به‌قدری عمومی هستند که در گفتگوهای روزانه مورداستفاده قرار می‌گیرند و در جهت هدف کتاب مقدمه‌ای بر اصول علم داده است.

[1] Field

با این تعریف شروع می‌کنیم که داده چیست. در ابتدا ممکن است به‌عنوان یک تعریف احمقانه به نظر برسد، اما بسیار مهم است. هرگاه ما از کلمه Data یا "داده" استفاده می‌کنیم، درواقع، به مجموعه‌ای از اطلاعات در فرمت سازمان‌یافته یا غیر سازمان‌دهی شده اشاره می‌کنیم:

- داده‌های سازمان‌یافته: به داده‌هایی اشاره دارد که در ساختار ردیف / ستون مرتب‌شده‌اند. به‌طوری‌که هر سطر مشاهدات واحد را نشان می‌دهند و ستون‌ها ویژگی‌های آن مشاهدات را نشان می‌دهند.
- داده‌های سازمان‌دهی نشده: این نوع داده‌ها در فرم آزاد هستند، معمولا متن یا صوت / سیگنال خام است که باید برای سازمان‌دهی بیشتر تجزیه شود.
- زمانی که اکسل (یا هر برنامه دیگری از صفحات گسترده) را باز کنید، یک ساختار سطر / ستون خالی در انتظار اطلاعات سازمان‌یافته را می‌بینید. این برنامه‌ها در رابطه با داده‌های سازمان‌دهی نشده به‌خوبی عمل نمی‌کنند. در بیشتر موارد، ما با داده‌های سازمان‌دهی شده به‌عنوان ساده‌ترین راه برای درک و فهم سروکار داریم، اما از پرداختن به داده‌هایی به‌صورت متن خام و روش‌های پردازش داده‌های سازمان‌دهی نشده نیز خسته نخواهیم شد.

علم داده‌ها، هنر و کسب دانش از طریق داده‌ها است. این تعریف کوچکی برای چنین عنوان بزرگی است. علم داده‌ها بسیاری از موارد را پوشش می‌دهد که صفحات آن را گرفته و همگی را لیست می‌کند.

علم داده در مورد چگونگی گرفتن داده‌ها و استفاده از آن برای کسب دانش است، سپس از این دانش برای انجام موارد زیر استفاده می‌کنیم:

- تصمیم‌گیری
- پیش‌بینی آینده
- درک گذشته / حال
- ایجاد صنایع / محصولات جدید

در این کتاب همه‌چیز راجع به روش‌های علوم داده است، ازجمله نحوه پردازش داده‌ها، جمع‌آوری مفاهیم و استفاده از این مفاهیم برای تصمیم‌گیری و پیش‌بینی‌های آگاهانه.

علم داده در مورداستفاده از داده‌ها به‌منظور به دست آوردن بینش جدید است که در غیر این صورت از دست می‌رفت.

به‌عنوان‌مثال، تصور کنید که شما در کنار یک میز با سه نفر دیگر نشسته‌اید. چهار نفر از شما باید بر اساس برخی داده‌ها تصمیم‌گیری کنند. چهار نظر برای بررسی وجود دارد. شما می‌توانید از علم داده‌ها برای آوردن افکار پنجم، ششم و حتی هفتم به جدول استفاده کنید.

به همین دلیل است که علم داده جایگزینی برای مغز انسان نیست، بلکه مکمل آن است، در کنار آن کار می‌کند. علم داده‌ها نباید به‌عنوان یک راه حل نهایی برای نگرانی‌های داده‌ای ما در نظر گرفته شود؛ این فقط یک نظر است، یک نظر بسیار آگاهانه است، اما به‌هرحال یک نظر است.

چرا علم داده؟

در عصر داده‌ها، واضح است که انبوهی از داده‌ها را داریم. اما ضرورت دانستن یک مجموعه جدید از واژگان چیست؟ مگر تجزیه‌وتحلیل پیشین ما چه ایرادی داشت؟ اولا، حجم مطلق داده‌ها باعث می‌شود که به معنای واقعی کلمه برای یک انسان غیرممکن باشد که آن را در زمان قابل‌قبول تحلیل کند. داده‌ها در اشکال مختلف و از منابع مختلف و اغلب به‌صورت غیر سازمان‌دهی شده جمع‌آوری می‌شوند.

داده‌ها می‌توانند ازدست‌رفته، ناقص و یا فقط ظاهرا اشتباه باشند. اغلب داده‌هایی در مقیاس بسیار متفاوت داریم که کار مقایسه آن‌ها را نیز مشکل می‌کند. داده‌های مربوط به ماشین‌های مورداستفاده در قیمت‌گذاری را در نظر بگیرید. یکی از مشخصه‌های یک خودرو سال ساخت آن و دیگری ممکن است تعداد مسافت پیموده شده (مایل[1]) آن ماشین باشد. هنگامی‌که داده‌ها را تمیز می‌کنیم (که ما زمانی معادل خواندن این کتاب را صرف می‌کنیم)، روابط بین داده‌ها بیشتر آشکار می‌شود و دانشی که در عمق میلیون‌ها ردیف از داده‌ها محو شده بود، به‌سادگی آشکار می‌شود. یکی از اهداف اصلی علوم داده‌ها، ایجاد روش‌های صریح برای کشف و به‌کارگیری این روابط در داده‌ها است.

پیش از این، به علم داده با دیدگاه تاریخی نگاه کردیم، حال با استفاده از یک مثال ساده، نقش علم داده را در کسب‌وکار امروزی موردبحث قرار خواهیم داد.

مثال – تکنولوژی‌های Sigma

Ben Runkle، مدیرعامل، تکنولوژی‌های سیگما[2]، در حال تلاش برای حل یک مشکل بزرگ است. این شرکت به‌طور مداوم مشتریان همیشگی خود را از دست می‌دهد. وی نمی‌داند چرا آن‌ها می‌روند، اما

[1] Mile
[2] Sigma

باید سریع‌تر کاری انجام دهد. او متقاعد شده است که باید محصولات و ویژگی‌های جدیدی ایجاد کند و فن‌آوری‌های موجود را یکپارچه کند. او برای اطمینان، به دکتر جسی هوگان[1]، متخصص ارشد داده خود مشورت می‌کند. بااین‌حال، جسی متقاعد نشده است که محصولات و ویژگی‌های جدید به‌تنهایی این شرکت را نجات دهد. در عوض، Jessie به رونوشت‌های اخیر بلیط های خدمات مشتری رو می‌آورد. Jessie به Runkle آخرین رونوشت‌ها را نشان می‌دهد و چیزی شگفت‌آور پیدا می‌کند:

- "... مطمئن نیستید چگونه این صادرات را انجام دهید؛ همین‌طور است؟"
- "دکمه‌ای که یک لیست جدید ایجاد می‌کند کجاست؟"
- "صبر کنید، آیا شما می‌دانید کجا لغزنده است؟"
- "اگر امروز نمی‌توانم این را تصور کنم، این‌یک مشکل واقعی است ..."

واضح است که مشتریان با UI / UX موجود مشکل داشتند و به دلیل کمبود امکانات ناراحت نشدند. Runkle و Hughan یک پیاده‌سازی UI / UX انبوه را سازمان‌دهی کردند و فروش آن‌ها هرگز بهتر نشده بود.

البته، علم مورداستفاده در آخرین مثال، حداقل بود، اما یک نکته دارد که ما تمایل داریم افرادی مانند Runkle، را یک محرک بنامیم. امروزه رایج است مدیرعامل[2] شرکت سهامی عام می‌خواهد همه تصمیمات را به‌سرعت بگیرد و راه‌حل‌ها را تا زمانی که آن چیز کار کند تکرار کند. دکتر هوگان بسیار تحلیل‌گر است. او می‌خواهد مشکل را بیشتر ازآنچه Runkle می‌خواهد، حل کند، او به‌جای احساس خستگی، برای پاسخ، به داده‌های تولیدشده کاربر رو آورد. علم داده در مورد به‌کارگیری مهارت‌های ذهنی تحلیلی و استفاده از آن‌ها به‌عنوان یک محرک است.

هر دو این ذهنیت‌ها در شرکت‌های امروز جای دارند. بااین‌حال، این روش تفکر هوگان است که بر ایده‌های علوم داده حکم‌فرماست یعنی – استفاده از داده‌های تولیدشده یک شرکت به‌عنوان منبع اطلاعات به‌جای این‌که فقط یک راه حل را برگزیند و با آن کنار بیایند.

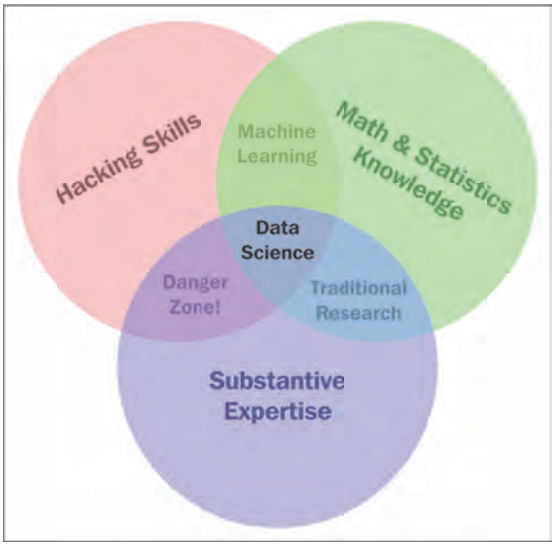
---

[1] Dr. Jessie Hughan
[2] CEO : Chief Executive Officer

نمودار ون علم داده[1]

یک تصور غلط رایج این است که گفته می‌شود تنها افراد نابغه و یا فارغ‌التحصیلان دکتری می‌توانند ریاضی به‌کاررفته در علوم داده را درک کنند. این مطلب کاملا نادرست است. درک علوم داده با سه حوزه اساسی زیر همراه است:

- ریاضی / آمار: این حوزه به استفاده از معادلات و فرمول‌ها برای انجام تجزیه‌وتحلیل می‌پردازد.
- برنامه‌نویسی کامپیوتر: در این حوزه به توانایی استفاده از کد برای دستیابی به نتایج در کامپیوتر پرداخته می‌شود.
- دانش دامنه: در این حوزه به درک دامنه مشکل (موردمطالعه) اشاره می‌شود. (پزشکی، امور مالی، علوم اجتماعی و غیره)
- نمودار ون زیر یک نمایش دیداری را از چگونگی ارتباط این سه حوزه نشان می‌دهد:



شکل۲ نمودار ون علم داده

افراد دارای مهارت‌های هک کردن[2] می‌توانند الگوریتم‌های پیچیده را با استفاده از زبان‌های کامپیوتری درک کنند. داشتن دانش در زمینه ریاضی و آمار، امکان فرضیه‌سازی و ارزیابی

[1] The data science Venn diagram
[2] Hacking

الگوریتم‌ها را فراهم نموده و روش‌های موجود با موقعیت‌های خاص را تنظیم می‌کند. با داشتن تخصص اصلی (تخصص دامنه) امکان به‌کارگیری مفاهیم و نتایج به روشی مؤثر و معنی‌دار فراهم می‌شود.

برای این‌که در این حوزه، از مهارت‌های کافی بهره‌مند شوید (در این حوزه باهوش باشید) داشتن تنها دو مورد از تخصص‌های ذکرشده لازم است اما کافی نیست و سبب ایجاد شکاف در این زمینه از علم خواهد شد. فرض کنید شما فرد ماهری در کدنویسی (برنامه‌نویسی) هستید ضمن این‌که در زمینه تجارت روزانه نیز از آموزش رسمی برخوردار شده باشید. در این صورت می‌توانید یک سیستم خودکار برای تجارت در محل خود ایجاد کنید. اما ازآنجاکه مهارت‌های ریاضی برای ارزیابی الگوریتم‌های خود ندارید بنابراین درنهایت، در بلندمدت پول خود را از دست می‌دهید. تنها زمانی می‌توانید علم داده را به‌درستی انجام دهید که در سه حوزه برنامه‌نویسی، ریاضی و دانش حوزه موردنظر، از مهارت‌های کافی برخوردار باشید.

دانش حوزه[1]، موردی است که ممکن است برای شما تعجب‌برانگیز باشد. در حقیقت دانش حوزه، همان دانش خاص حوزه‌ای است که در آن کار می‌کنید. اگر یک تحلیل‌گر مالی بخواهد داده‌های مربوط به حملات قلبی را تجزیه‌وتحلیل کند، آنگاه برای درک تعدادی زیادی از اعداد به کمک یک متخصص قلب نیاز خواهد داشت.

علم داده اشتراک سه حوزه اصلی است که پیش‌تر ذکر شد. برای دستیابی به دانش از روی داده، باید از برنامه‌نویسی کامپیوتر برای دسترسی به داده استفاده نمود. از ریاضی برای فهم مدل‌هایی که ایجاد می‌شوند بهره برد و از همه مهم‌تر این‌که موقعیت‌های تجزیه‌وتحلیل را در حوزه دانشی که هستیم درک کنیم. این شامل ارائه داده‌ها می‌شود. اگر یک مدل برای پیش‌بینی حملات قلبی در بیماران ایجاد کنیم، آنگاه بهتر است برنامه یا پی‌دی‌افی از اطلاعات تهیه نمود به‌طوری‌که بتوان اعداد (شماره‌ها) را در آن تایپ نمود و پیش‌بینی سریع را دریافت کرد. همه این تصمیمات باید توسط یک متخصص علوم داده گرفته شود.

همچنین توجه داشته باشید که اشتراک دو حوزه ریاضی و برنامه‌نویسی، یادگیری ماشین است. در این کتاب با جزئیات بیشتری به یادگیری ماشین خواهیم پرداخت. اما توجه به این مطلب مهم است که الگوریتم‌های یادگیری ماشین دارای قابلیت‌های واضحی برای تعمیم مدل یا نتایج در هر حوزه‌ای به شمار می‌روند که به‌راحتی بر روی کامپیوتر قرار می‌گیرند. شما ممکن است بهترین الگوریتم را برای پیش‌بینی سرطان داشته باشید. می‌توانید بیماری سرطان را با صحت 99% بر اساس داده‌های گذشته

[1] Domain knowledge

بیماران مبتلابه سرطان پیش‌بینی کنید اما اگر توانایی به‌کارگیری این مدل را از لحاظ عملی نداشته باشید، نظیر این‌که پزشکان و پرستاران بتوانند به‌راحتی از این مدل استفاده کنند، آنگاه این مدل بی‌فایده است.

هر دو حوزه برنامه‌نویسی کامپیوتر و ریاضی به‌طور گسترده در این کتاب پوشش داده شده است. دانش حوزه با هر دو جنبه عملی علم داده و خواندن نمونه‌هایی از تحلیل‌های دیگران همراه است.

ریاضی

بیشتر مردم وقتی فردی در مورد کلمه math صحبت می‌کند، گوش دادن را متوقف می‌کنند و تلاش می‌کنند خود را بی‌اطلاع از موضوع نشان دهند.

در سراسر این کتاب به ریاضیات موردنیاز برای علوم داده، به‌ویژه دو مبحث آمار و احتمال می‌پردازیم. هم‌چنین با استفاده از این زیر حوزه از ریاضی به ایجاد مدل‌ها خواهیم پرداخت.

یک مدل داده[1] به یک ارتباط رسمی و سازمان‌یافته بین عناصر داده اشاره دارد که معمولا باهدف شبیه‌سازی یک پدیده در دنیای واقعی ایجاد می‌شود.

به‌طور اساسی، از ریاضی برای فرمول سازی روابط بین متغیرها استفاده خواهیم نمود. به‌عنوان یک ریاضیدان (محض) سابق و معلم ریاضی فعلی، می‌دانم که چقدر این مرحله می‌تواند دشوار باشد. تا حد امکان این مراحل را با بهترین توضیح به‌صورت واضح انجام خواهم داد. در بین سه حوزه علوم داده، ریاضی قابلیت حرکت از یک حوزه به حوزه‌ای دیگر را فراهم می‌کند. درک این نظریه امکان به‌کارگیری مدل ساخته‌شده در صنعت مد را برای ایجاد یک مدل مالی نیز فراهم می‌کند.

مباحث ریاضی که در این کتاب پوشش داده می‌شود شامل جبر پایه تا مدل‌سازی‌های پیشرفته آماری و احتمالی هست. این فصل‌ها را از قلم نیندازید حتی اگر مطالبی را در مورد آن‌ها از قبل می‌دانید. هر مفهوم ریاضی را با دقت معرفی نموده و مثال‌ها و اهداف آن را توضیح خواهم داد. ریاضیات ارائه‌شده در این کتاب برای متخصصان علوم داده ضروری است.

مثال ـ مدل‌های spawner-recruit

در زیست‌شناسی، در میان بسیاری از مدل‌ها، از مدل شناخته‌شده spawner-recruit استفاده می‌کنیم که این مدل برای تشخیص سلامت بیولوژیکی یک‌گونه جانوری به کار می‌رود. در این مدل به ارتباط

[1] Data model

اساسی بین تعداد واحدهای سالم والدین یک‌گونه و تعداد واحدهای جدید در یک گروه از جانوران پرداخته می‌شود. در مجموعه داده‌های عمومی تعداد salmon spawners و recruits، نمودار زیر برای تجسم ارتباط بین این دو گونه ترسیم شده است. به‌طورقطع نوعی ارتباط مثبت وجود دارد. (به دنبال افزایش یکی، دیگری نیز افزایش می‌یابد.) اما چگونه می‌توان این رابطه را فرمول سازی کرد؟ برای مثال، اگر تعداد spawnerها در یک جمعیت معلوم باشد، آیا می‌توان تعداد recruitهای به‌دست‌آمده در یک گروه را پیش‌بینی نمود و برعکس؟

به‌طور اساسی، با به‌کارگیری مدل‌ها می‌توان یک متغیر را به متغیری دیگر مرتبط کرد. مثال زیر را در نظر بگیرید:

$$Recruits = 0.5 * Spawners + 60$$

در این مثال، فرض کنید تعداد salmonها 1.15 (در هزار) در spawnerها باشد. در این صورت معادلات زیر را خواهیم داشت:

$$Recruits = 0.5 * 1.15 + 60$$

$$Recruits = 60.575 \left(in\, thousands \right)$$

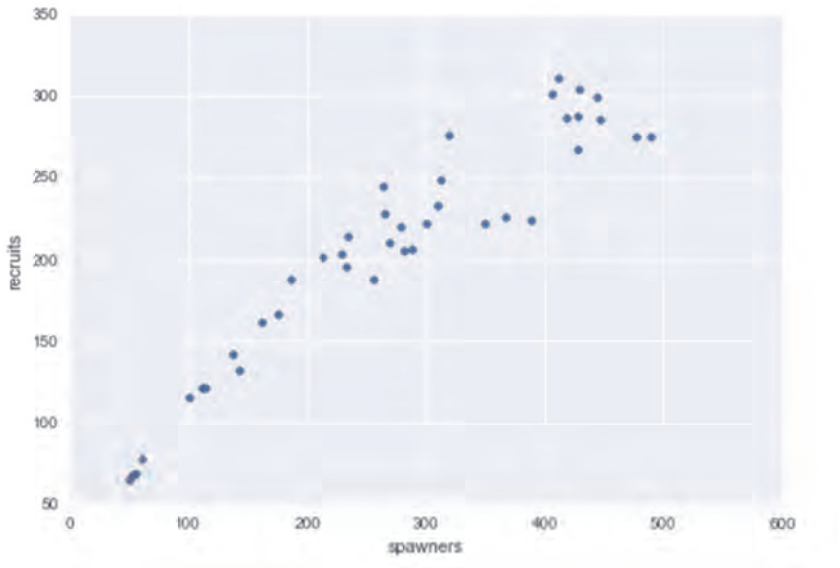
این نتیجه برای برآورد این‌که چه طور سلامت یک جامعه تغییر می‌کند، می‌تواند بسیار مفید باشد. اگر این مدل را ایجاد کنیم، آنگاه می‌توان به‌صورت عینی مشاهده کرد که چه طور ارتباط بین این دو متغیر می‌تواند تغییر کند.

انواع زیادی از مدل‌های داده‌ای شامل مدل‌های آماری و احتمالی وجود دارند. هر دو این‌ها زیرمجموعه‌ای از یک نمونه بزرگ‌تر هستند که یادگیری ماشین[1] نامیده می‌شود. ایده اصلی در این سه موضوع این است که از داده‌ها برای ارائه بهترین مدل ممکن استفاده کنیم. دیگر به غرایز انسانی وابسته نیستیم بلکه به داده‌ها متکی هستیم.

هدف از این مثال نشان دادن این است که چگونه می‌توان ارتباط بین عناصر داده‌ای را با استفاده از معادلات ریاضی تعریف کرد. این واقعیت که مثلا از داده‌های سلامت salmon استفاده شد، بی‌اهمیت است! در سراسر این کتاب، ارتباط‌هایی شامل بازاریابی دلار، تمایل داده‌ها (احساس)، بازدیدهای رستوران و بسیاری موارد دیگر را خواهیم دید. هدف اصلی این است که خوانندگان تا حد امکان در معرض هر دامنه‌ای قرارگرفته و با آن حوزه خاص آشنا شوند.

---

[1] Machine learning



The spawner-recruit model

ریاضی و برنامه‌نویسی ابزارهایی هستند که این امکان را برای متخصصان علوم داده فراهم می‌کنند که به عقب بازگشته و مهارت‌های خود را تقریبا در هرکجا اعمال کنند.

برنامه‌نویسی کامپیوتر

بیایید باهم صادق باشیم. احتمالا فکر می‌کنید علوم کامپیوتر جالب‌تر از ریاضی باشد. بسیار خوب، شما را سرزنش نمی‌کنم. محتوای اخبار شامل خبرهایی از ریاضی نیست، بلکه حاوی خبرهایی از فن‌آوری است. تلویزیون را روشن نمی‌کنید که خبری را در مورد قضیه جدیدی از اعداد اول ببینید، بلکه به دنبال گزارش‌های تحقیقاتی هستید که چگونه آخرین گوشی‌های هوشمند می‌توانند باکیفیت بهتر، از گربه‌ها یا هر چیز دیگری عکس‌برداری کنند. زبان‌های کامپیوتری چگونگی برقراری ارتباط با ماشین را نشان داده و پیشنهادها را برای انجام به ماشین می‌دهند. یک کامپیوتر مانند یک کتاب می‌تواند به زبان‌های زیادی صحبت نموده و همین‌طور به زبان‌های زیادی نوشته شود. به همین ترتیب (به‌طور مشابه) علم داده نیز می‌تواند توسط زبان‌های بسیاری انجام شود. Python ، Julia و R ازجمله زبان‌های در دسترس هستند. این کتاب منحصرا در استفاده از پایتون تمرکز خواهد کرد.

چرا Python ؟

به دلایل مختلف زیر از پایتون استفاده خواهیم کرد:

- پایتون یک‌زبان بسیار ساده برای خواندن و نوشتن است، حتی اگر پیش‌تر هرگز کدنویسی نکرده باشید، مثال‌های بعدی را حتی پس از اینکه شما این کتاب را خواندید آسان می‌سازد.
- یکی از رایج‌ترین زبان‌هایی است که هم در تولید و هم در محیط دانشگاهی (یکی از سریع‌ترین پیشرفت‌ها، به‌عنوان یک واقعیت) مورداستفاده قرار می‌گیرد.
- انجمن آنلاین این زبان به‌صورت گسترده و دوستانه است، به این معنی که یک جست‌وجوی سریع در گوگل باید شامل نتایج متعدد از افرادی باشد که آن‌ها نیز با چنین شرایط مشابه (نه به‌طور کاملا مشابه) مواجه شده و (مشکل آن‌ها) حل‌شده باشد.
- پایتون دارای ماژول‌های[1] داده‌ای از پیش ساخته‌شده است که هر دو گروه افراد مبتدی و خبره در حوزه علوم داده می‌توانند از آن بهره‌مند شوند.

آخرین مورد احتمالا بزرگ‌ترین دلیلی است که بر روی پایتون تمرکز خواهیم نمود. این ماژول‌های از پیش ساخته‌شده نه‌تنها قدرتمند هستند، بلکه به‌سادگی ایجاد می‌شوند. در پایان چند فصل اول، از این ماژول‌ها به‌راحتی استفاده خواهید نمود. برخی از این ماژول‌ها عبارت‌اند از:

- pandas
- sci-kit learn
- seaborn
- numpy/scipy
- requests (to mine data from the Web)
- BeautifulSoup (for the Web-HTML parsing)

پایتون در عمل

پیش از این‌که ادامه دهیم، مهم است که به بسیاری از مهارت‌های برنامه‌نویسی موردنیاز در پایتون اشاره کنیم.

در پایتون، متغیرهایی داریم که متغیرهایی برای اشیاء هستند. در ابتدا تنها روی مواردی اندک از این نوع اشیاء اساسی تمرکز خواهیم کرد:

- **int (an integer)**
  - **Examples: 3, 6, 99, -34, 34, 11111111**
- **float (a decimal):**
  - **Examples: 3.14159, 2.71, -0.34567**

[1] Modules