Finding Ghosts in Your Data Anomaly Detection Techniques with Examples in Python

پیدا کردن ارواح در داده های شما تکنیک های تشخیص ناهنجاری با مثال در پایتون :

معرفی

به این کتاب در مورد تشخیص ناهنجاری خوش آمدید! در طول این کتاب، ما برای ساخت موتور تشخیص ناهنجاری در پایتون می رویم. برای انجام این کار، ابتدا باید به این سوال پاسخ دهید که "ناهنجاری چیست؟" چنین سوالی پاسخ ساده ای دارد اما در با ارائه پاسخ ساده، در را به روی سؤالات بیشتری باز می کنیم که پاسخ آنها باز است هنوز درهای بیشتر این شادی و نفرین دنیای آکادمیک است: ما همیشه می توانیم کمی پیش برویم کمی پایین تر از سوراخ خرگوش. با این حال، قبل از شروع غواصی در سوراخ های خرگوش، اجازه دهید سطح انتظارات را تعیین کنیم. همه کدهای این کتاب در پایتون خواهد بود. این مطمئنا تنها زبانی نیست که می توانید برای این منظور استفاده کنید - من و داور فنی محترم، همکار دیگری نوشتیم یک موتور تشخیص ناهنجاری با استفاده از ترکیبی از C# و R، بنابراین هیچ چیزی به آن نیاز ندارد ما از پایتون استفاده می کنیم. ما زبان جلد و دیگر انتخاب‌های طراحی را در کتاب انجام می‌دهیم، بنابراین صرفه‌جویی می‌کنم شما بقیه اینجا تا آنجا که سطح راحتی شما با پایتون پیش می رود، هدف این کتاب است این نیست که به شما زبان یاد بدهم، بنابراین احتمال می‌دهم با زبان آشنایی داشته باشم. البته من زمینه کدی را که می نویسیم ارائه می کنم و زمان بیشتری را صرف آن خواهم کرد مفاهیمی که کمتر شهودی هستند. علاوه بر این، تمام کدهایی که در کتاب استفاده خواهیم کرد در مخزن GitHub همراه در دسترس است هدف من در این کتاب فقط نوشتن یک موتور تشخیص ناهنجاری نیست، بلکه راه رفتن است خط بین دنیای دانشگاهی و توسعه ادبیات غنی در اطراف وجود دارد تشخیص ناهنجاری، اما بسیاری از ادبیات متراکم و آغشته به منطق رسمی است. من می‌خواهم برخی از بهترین بینش‌ها را از آن ادبیات آکادمیک برای شما بیاورم، اما در معرض دید قرار دهم آن را به گونه ای که برای اکثریت بزرگ توسعه دهندگان منطقی است. به همین دلیل، هر یک بخشی از کتاب حداقل یک فصل به نظریه اختصاص داده شده است. علاوه بر این، بسیاری از فصل‌های کدنویسی نیز با تئوری شروع می‌شوند، زیرا صرفاً کافی نیست چند دستور را تایپ کنید یا readme پروژه را برای فراخوانی روش نمونه بررسی کنید. من می خواهم برای کمک به درک اینکه چرا چیزی مهم است، زمانی که یک رویکرد می تواند کار کند، و زمانی که رویکرد ممکن است شکست بخورد. علاوه بر این، اگر بخواهید خودتان به آن شیرجه بزنید ادبیات، کتابشناسی در پایان کتاب شامل انواع آکادمیک است

منابع

قبل از اینکه امضا کنم و وارد کتاب شویم، می‌خواهم تشکر ویژه‌ای از خودم داشته باشم

همکار و ویراستار فنی، تینگ چو. من برای مهارت های تینگ نهایت احترام را قائل هستم، بنابراین

تا حدی که سعی کردم او را وادار کنم تا کتاب را با من بنویسد! او برای حفظ من کارهای زیادی انجام داد

در مسیر درست قرار گرفته و شکل نهایی این کتاب را به شدت تحت تاثیر قرار داده است

انتخاب الگوریتم ها و بخش هایی از پشته فناوری که استفاده خواهیم کرد. گفت، هر گونه خطا

البته تنها مال من و من هستند. متاسفانه

اگر نظری در مورد کتاب یا تشخیص ناهنجاری دارید، خوشحال می‌شوم از آن بشنوم

شما. ساده ترین راه ارتباط از طریق ایمیل است

در

در ضمن امیدوارم از کتاب لذت ببرید

اهمیت ناهنجاری ها و تشخیص ناهنجاری ها:

قبل از اینکه شروع به ساخت یک آشکارساز ناهنجاری کنیم، مهم است که بدانیم به طور خاص، ناهنجاری چیست. بخش اول این فصل یک مقدمه را در درک ناهنجاری ها اختیار ما قرار می دهد. سپس در قسمت دوم فصل به کاربرد آن برای تشخیص ناهنجاری در صنایع مختلف می پردازیم .. در قسمت سوم فصل، ما سه کلاس تشخیص ناهنجاری را برمی‌شماریم که به آنها باز خواهیم گشت این کتاب، به دست آوردن درک سطح بالایی از هر یک در آماده سازی برای غواصی عمیق تر بعد. در نهایت، این فصل را با چند نکته در مورد اینکه چه زمانی باید ساختن یک آشکارساز ناهنجاری خودمان در نظر بگیریم به پایان می‌رسانیم .

تعریف ناهنجاری ها :

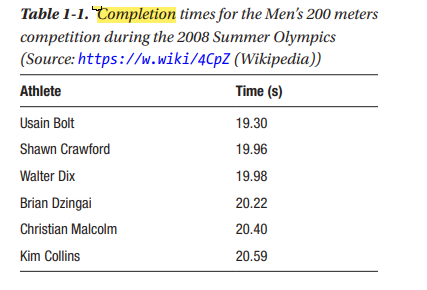
قبل از اینکه بتوانیم شروع به ساختن یک آشکارساز ناهنجاری کنیم، باید بفهمیم که ناهنجاری چیست ؟ به طور مشابه، قبل از اینکه بتوانیم مفهوم ناهنجاری را تعریف کنیم، نیاز داریم که موارد پرت را درک کنید .

داده های خارج از محدوده

نقطه پرت یک نقطه داده است که به طور قابل توجهی با مقادیر عادی متفاوت است. مثلا،در بازی‌های المپیک تابستانی ۲۰۰۸، یوسین بولت یک رکورد جهانی را در ۲۰۰ متر مردان با زمان 19.30 ثانیه ثبت کرد. بیش از شش دهم ثانیه سریعتر نسبت به مقام دوم شاون کرافورد به پایان رساند .. پنج رقیب باقی مانده (به جز دو دوندگانی که محروم شدند) زمان هایی بین 19.96 ثانیه تا 20.59 ثانیه داشتند.

به این معنی که تفاوت بین مکان دوم و آخرین کمی بیشتر بود از تفاوت بین مقام اول و دوم. درنتیجه نتایج برای رتبه دوم تا رتبه ششم، مقادیر درونی یا مقادیری هستند که تفاوت قابل توجهی از هنجار(معمول) ندارند .

جدول 1-1 زمان حضور ورزشکاران در فینال را نشان می دهد.



بر اساس این اطلاعات، واضح است که زمان بولت دور از ذهن بوده (داده پرت) است. سوال بعدی آیا این یک ناهنجاری بود؟

# نویز در مقابل ناهنجاری ها (Noise vs. Anomalies)

در دسته بندی های پرت، دو زیر مجموعه را تعریف می کنیم: نویز و ناهنجاری ها. تفاوت بین این دو اصطلاح در نهایت به نفع انسان است. نویز موارد پرتی است که برای ما جالب نیست، در حالی که ناهنجاری ها موارد پرتی هستند که جالب هستند.

این تعریف از ناهنجاری ها با تعریف آگاروال (1-4) مطابقت دارد، اما این تعریف کلی و عمومی نیست - برای مثال، Mehrotra و همکاران. از عبارات "داده پرت " و "ناهنجاری" را بصورت مترادف استفاده می کند (مهروترا و همکاران، 4).

یکی دیگر از راه‌های جداسازی نویز از ناهنجاری‌ها، فکر کردن بر اساس مدل‌ها با شرایط خطا است.

فرض کنید تابعی داریم که مجموعه ای از ورودی ها را می گیرد و تخمین اینکه چند ثانیه طول می کشد تا آن شخص 200 متر را در رقابت (مسابقه) بدود برمی گرداند. . اگر ورودی های مختلفی را وارد کنیم، به طور بالقوه می توانیم شکل آن تابع را ترسیم کنیم ، درک این موضوع که بین برآوردی که تولید می کنیم و و زمان واقعی ،مقداری واریانس وجود دارد را ما ، این را با اضافه کردن یک عبارت خطا نشان می دهیم.

TAthlete = f(A, B, . . .) +ᵋ

گاهی اوقات، با توجه به مجموعه ای از مقادیر ورودی و زمان واقعی ما ممکن است یک ربع ثانیه تفاوت با زمان مورد انتظارمان ببینیم . اگر باور کنیم (اعتقاد داشته باشیم)که این اجرا (دویدن) خاص همچنان از همان تابع اساسی پیروی می کند و کل تفاوت به دلیل عبارت خطا است ، پس تفاوت بین تخمین ما و زمان واقعی ، به درستی از میزان نویز صحبت نمودیم (اگر اینطوری باور داشته باشیم این اختلاف را، که زیاد است را بعنوان نویز در نظر میگیریم نه یک رفتار ناهنجار). همان میزان خطا یا واریانس را نویز در نطر گرفتیم این رنگها را من نوشتم و ترجمه کتاب نیست .در مقابل ، ، فرض کنید برای ورزشکاری افسانه ای مانند بولت عملکرد جداگانه ای وجود دارد.

Tbolt = f(A, B, . . .) +ᵋ

این تابع تخمین قابل توجهی کمتری از زمان را نسبت به زمان اجرا ورزشکار المپیکی (اگر چنین چیزی امکان پذیر باشد!) ارائه می دهد. اگر از این تابع استفاده کنیم و یک چهارم ثانیه سریعتر از نقاط داده با استفاده از مدل معمولی را نتیجه بگیریم (بدست بیاوریم) ، نتیجه بسیار منطقی است اگر ما انتظار خروجی معمولی ورزشکار را داشتیم و در نهایت به نتیجه می رسیدیم

خروجی بولت، با این حال، نتیجه یک ناهنجاری خواهد بود: نتیجه ای که با آن متفاوت است

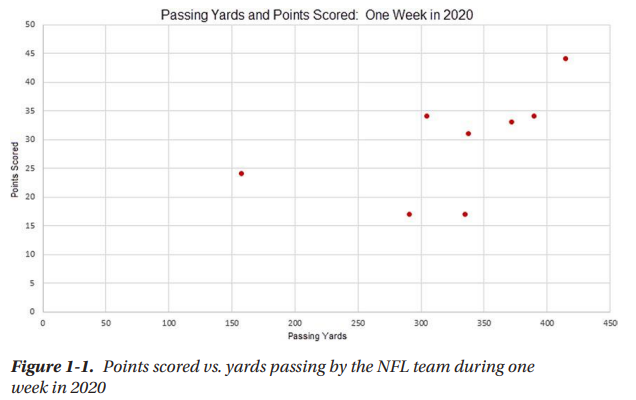
هنجار و به دلیل تفاوت در عملکرد زیربنایی مورد علاقه ما است

نتیجه را ایجاد کند.

در عمل، این درک از یک ناهنجاری احتمال بیشتری نسبت به واقعیت دارد. ما لزوماً تابع خاصی که هر نقطه داده خاصی را ایجاد کرده است را نمی دانیم ، و این امر جداسازی نویز از ناهنجاری ها را دشوار می کند. در مورد بولت، ما می توانیم انتظار داشته باشیم که این یک ناهنجاری باشد تا نویز، نه تنها به این دلیل که بولت رکورد جهانی جدیدی را برای این رویداد رقم زد بلکه به این دلیل که زمان او بسیار کمتر از زمان های دیگر ورزشکاران در سطح جهانی بود ، به این ایده که او یک کلاس متفاوت است اعتبار می بخشد.

# تشخیص یک مثال (Diagnosing an Example)

با مشاهده یک مثال، اجازه دهید یک مثال گرافیکی را مرور کنیم. شکل 1-1 الف نمودار پراکنده امتیازهای کسب شده در مقابل یاردهای پاس در یک هفته لیگ ملی فوتبال بازی های فصل 2020 را نشان می دهد.



ما به عنوان انسان، توانایی های ذاتی و آموزش دیده خاصی برای تجزیه و تحلیل این اطلاعات و نتیجه گیری از آن موضوع را داریم ، که ما با جزئیات بیشتر به آنها در فصل 2 خواهیم پرداخت . در حال حاضر، کافی است بگوییم که به نظر می رسد یک همبستگی مثبت بین تعداد یاردهای عبور و امتیازهای کسب شده در این نمودار وجود دارد.

با توجه به اطلاعات موجود، می‌توانیم باند نسبتاً فشرده‌ای را تصور کنیم که تعداد امتیازهای به دست آمده برای تعداد یارد عبور کرده معینی را تخمین می‌زند، رسم یک خط ضخیم که نزدیک یا به طور کامل اکثر نقاط داده را پوشش می دهد. انجام این کار ما را به این نتیجه رساند که تیمی که کوارتربکش(مدافعانش) 300 یارد پاس می دهد ممکن است تقریباً 28 تا 35 امتیاز در یک روز عادی امتیاز بدست اورد.

پس از طرح این مدل نظری، سپس می‌توانیم به دو حالت داده پرت بپردازیم که در آن‌ها تیم ها علیرغم پرتاب تقریباً به ترتیب 290 و 335 یارد توسط مدافعان خود، تنها 17 امتیاز کسب کردند. این منجر به دو سؤال اساسی می شود که حداقل به طور موقت - به منظور پیشبرد روند تشخیص ناهنجاری باید به آنها پاسخ دهیم.

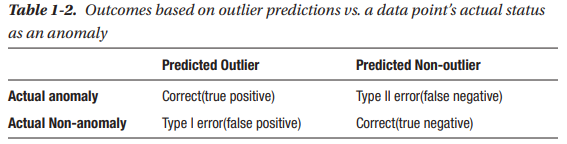
اولین سوال این است که آیا مدل ما به اندازه کافی خوب است که درونی ها را از پرت جدا کند؟ اگر پاسخ به این "بله" است، سؤال دوم این است که چقدر محتمل است که نقاط داده پرت از فرآیندی کاملاً متفاوت از موارد درونی را دنبال می کنند؟

در طول کتاب، به ایده‌ها و تکنیک‌هایی می‌پردازیم که به شما کمک می‌کنند تا به این دو سوال پاسخ دهیم با این حال، برای پاسخ دادن به سوال دوم به طور قاطع ممکن است بسیار دشوار باشد، زیرا این امر به درک ما از عملکرد داخلی فرآیندی که این نقاط داده را تولید می کند بستگی دارد.

این منجر به توجه(نکته) مهم دیگری می شود: دانش دامنه (دانش آن حوزه) برای تشخیص داده پرت ضروری نیست ، اما برای تشخیص ناهنجاری لازم است. بازگشت به مثال در شکل 1-1، فردی که هیچ دانشی از فوتبال آمریکایی ندارد، می تواند رابطه در داده ها و همچنین دو نقطه پرت در مجموعه داده آن را درک کند . با این حال، آنچه که شخص نمی تواندانجام دهد چیست نتیجه گیری کند که آیا این نقاط داده صرفاً نویزهستند یا خیر و یا اگر دلالت بر یک الگوی رفتاری جداگانه دارند این را نمیتواند فرد تشخیص دهدکه نویز هست یا رفتار جدید. ساختن این ساختار، اگر آن تیم ها فقط 17 امتیاز به دلیل بدشانسی، گل های از دست رفته در زمین، یا مواجهه با منطقه قرمز برتری در دفاع کسب کرد، این عوامل به نویز بودن نقاط پرت اشاره می کنند. اگر از طرف دیگر، این تیم ها استراتژی های مربیگری ناکارآمدی دارند که توپ را به سمت پایین زمین حرکت می دهد اما به طور مداوم در تبدیل امتیاز آن به حرکت تهاجمی شکست می خوریم، در این صورت ممکن است به درستی این نقاط را ناهنجاری نامید. یک متخصص آن حوزه با اطلاعات کافی می تواند به ما کمک کند که دو نتیجه را از هم جدا کنید همانطور که بعداً در این فصل خواهیم دید، قصد ما از ساختن یک آشکارساز ناهنجاری ارائه اطلاعات کافی برای چنین متخصص حوزه ای است تا بتواند آن را تجزیه و تحلیل و تفسیر کند و سپس بر اساس سیگنال هایی که ما ارسال می کنیم عمل کنید.

# اگر اشتباه کنیم چه؟ (What If We’re Wrong?)

هنگام تصمیم گیری در مورد اینکه آیا یک نقطه یک ناهنجاری است، می توانیم عمل طبقه بندی (classification) را با تقسیم نتایج به دو یا چند کلاس بر اساس برخی ویژگی های اساسی انجام دهیم. در مورد تشخیص ناهنجاری، ما معمولاً با دو کلاس مواجه می‌شویم: این چیزیک ناهنجاری است ، یا این چیز یک ناهنجاری نیست. اگر دو کلاس ممکن داشته باشیم (این یک outlier است یا این یک پرت نیست) و ما می توانیم هر یک از دو کلاس را پیش بینی کنیم، در نهایت با نتیجه ای مانند جدول 1-2به پایان می رسیم ، که نتایج ممکن را با توجه به یک کلاس پیش بینی شده و کلاس واقعی آن به ما می دهد .



با نگاهی به این جدول، ما دو بعد را می بینیم: آیا یک نقطه داده معین یک ناهنجاری واقعاً بوده است یا خیر و اینکه آیا پیش بینی ما در آن نقطه داده مثبت بود یا خیر (که بله، این یک نقطه پرت است) یا منفی (خیر، این یک پرت نیست). جدول دارای دو بعد است آیا یک داده ناهنجار هست یا نه این بعد اول است و اینکه آیا ما درست پیش بینی کردهایم نیز بعد دوم است .زمانی که ما به درستی تعیین کنیم که یک نقطه داده پرت در واقع یک ناهنجاری است یا یک داده غیر پرت در واقع یک ناهنجاری نیست، توجه داشته باشیم که طبقه بندی ما درست عمل کرده و می توان ادامه داد. با این حال، دوتا از راه هایی که در آنها ممکن است اشتباه کنیم (در دو صورت اشتباه پیش میآید). اولین موردی که ممکن است اشتباه کنیم این است که پیش بینی کنیم که یک نقطه داده یک نقطه پرت خواهد بود، اما معلوم شود که یک ناهنجاری نیست- وفقط نویز بوده است این به عنوان خطای نوع I یا مثبت کاذب نیز شناخته می شود. برعکس این ،در صورتی است که برخی از داده‌ها علیرغم واقعی بودن، ما آنها را یک نقطه داده پرت ناهنجاری تعیین کنیم . خرابی (شکست) اخیر به عنوان خطای نوع II یا منفی کاذب شناخته می شود. تا حد امکان، ما می خواهیم از هر دو نوع خطا جلوگیری کنیم، اما یک سوال جالب ناشی از این واقعیت که ما این خطاها را جداگانه طبقه بندی می کنیم: کدام یک از این دو بدتر است؟

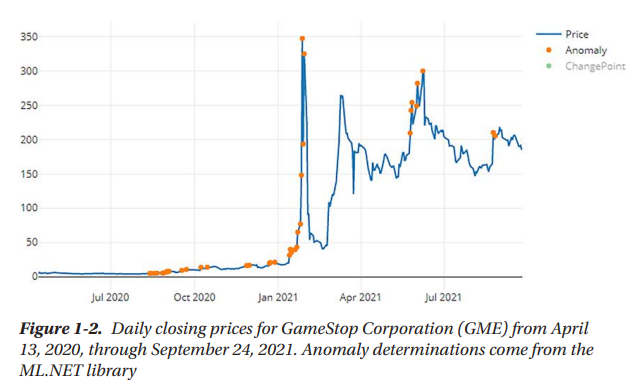
پاسخ، آماده باش مشاور معروف است: بستگی دارد. در این مورد، بررسی و شرایط کلیدی شامل1- پیامدهای منفی از دست دادن حداقل یک ناهنجاری واقعی و عدم اقدام در مورد آن در مقابل دریافت هشدار یا انجام تحقیق در مورد نقاط داده که کاملاً عادی بود اما توسط سرویس تشخیص ما اشتباه طبقه بندی شد.2- اینکه چه مقدار از موارد مثبت کاذب را مایل به پذیرش آنها هستید تا احتمال منفی کاذب را محدود کنید به چند عامل بستگی دارد، 2-الف)مانند فراوانی فعالیت مثبت کاذب، چه سطحی از تلاش مورد نیاز هست برای تعیین اینکه آیا یک مورد که پیش‌بینی شده واقعاً یک ناهنجاری است یا خیر، و2-ب)هزینه یک نتیجه منفی کاذب چقدر است ؟عواملی از این قبیل به شما در تصمیم گیری کمک می کند آشکارساز بیرونی شما چقدر باید سختگیر باشد، اگرچه ما یک رویکرد جایگزین را خواهیم دید از فصل 4 شروع می شود که می تواند به حل این مشکل کمک کند.

# ناهنجاری در طبیعت (Anomalies in the Wild)

در این بخش، به طور خلاصه به چند صنعت که تشخیص ناهنجاری در تصمیم گیری آن ها نقش مهمی در بازی می کند، می پردازیم . تجزیه و تحلیل دقیق هر یک از این صنایع بسیار فراتر از محدوده این کار است ، اما حتی یک خلاصه مختصر باید در طرح بندی کافی این تز که تشخیص ناهنجاری یک نگرانی بین صنعتی باشد موثر است .

دارایی، مالیه، سرمایه گذاری (Finance)

تشخیص ناهنجاری اهداف مختلفی را در بخش مالی انجام می دهد. یکی از مهمترین سناریوهای واضح برای این امر به قیمت گذاری دارایی ها مربوط می شود. نماد و سمبل قیمت گذاری دارایی غیرعادی باید شرکت GameStop (GME) باشد، که قیمت سهام روزانه آن بسته شدن (در اخر روز قیمت سهام را ارائه میکند)انجام می شودشکل 1-2.



سهام نسبتا کمی آشفتگی را تجربه می‌کند که با GameStop شاهد دوره چند ماهه آن بودیم ، اما بر اساس اطلاعاتی که در زمان نگارش این کار در دسترس داریم ، بسیاری از نقاط داده پرت به کمپین‌های رسانه‌ای مرتبط است (خواه آنها انبوه یا اجتماعی باشد)، موعد مقرر برای گزینه‌ها، و تحرک‌کنندگان بازار در یکی از بزرگترین رویدادهای بازار این نسل آنها طرف می‌گیرند (تاثیر میگذارند) .قیمت گذاری دارایی تنها جایی نیست که تشخیص ناهنجاری در آن صنعت مالی منطقی باشد. موقعیت دیگری که در آن تشخیص ناهنجاری می تواند مهم باشد حسابداری قانونی است، یعنی بررسی شیوه های حسابداری و مالی اسنادی برای تعیین اینکه آیا اعضای یک سازمان اقدامات کلاهبرداری کرده اند یا خیر. حساب های پزشکی قانونی بسیار بیشتر از منفذ وروزنه بر روی ترازنامه و ورودی‌ها مجله انجام می دهند.، اما وقتی به اسناد مالی نگاه می‌کنند، به دنبال الگوها و ناهنجاری ها هستند

یکی از این الگوها، قانون بنفورد است که بیان می‌کند که رقم سرب به اندازه کافی

مجموعه بزرگی از اعداد اغلب از یک الگوی خاص پیروی می کنند که در آن اعداد کوچکتر هستند

بیشتر از اعداد بزرگتر ظاهر می شود. شکل 1-3 نشان می دهد که توزیع در

فرم گرافیکی همه مجموعه‌های داده از این الگو پیروی نمی‌کنند، اما یک رویکرد سریع "قاعده سرانگشتی" این است

اگر یک مجموعه داده شامل نقاط داده ای باشد که در حداقل سه (یا ترجیحاً بیشتر) سفارش پخش شده اند

از بزرگی، پس احتمالاً از قانون بنفورد پیروی می کند. این قانون سرانگشتی پابرجاست زیرا

بسیاری از انواع توزیع هایی که به قانون بنفورد پایبند هستند - مانند نمایی

یا توزیع های log-normal—معمولاً شامل مقادیر مختلفی است که در چندین مورد پخش شده اند

مرتبه های بزرگی، در حالی که توزیع هایی که به قانون بنفورد پایبند نیستند-

توزیع های یکنواخت و گاوسی رایج ترین هستند - اغلب اینطور نیست. برای یک

فرو رفتن دقیق در ریاضیات پشت قانون بنفورد، من هال (2019) را توصیه می کنم.

با این حال، برای اهداف ما، درک این نکته کافی است که قانون بنفورد می تواند

ارائه یک حسابدار قانونی با انتظارات از نحوه توزیع برخی از

مجموعه داده ها باید نگاه کنند، به ویژه مجموعه داده هایی مانند مجموعه ورودی های مجله برای a

شرکت یا دفتر دولتی انحراف قابل توجهی از قانون بنفورد وجود ندارد

به خودی خود اثبات تخلف است، اما این یک پرچم قرمز برای حسابرس است و معمولاً شایستگی بیشتری دارد.

تحقیق برای درک اینکه چرا این مجموعه داده خاص معمولی نیست.

دارو

تشخیص ناهنجاری در زمینه پزشکی بسیار مهم است. بدن عظیمی از وجود دارد

تحقیق در مورد اقدامات بدن سالم در مقابل ناسالم. نمونه هایی از این موارد

اقدامات شامل فشار خون، دمای بدن، اشباع اکسیژن خون و

داده های نوار قلب (ECG) می توان گفت که یک فرد در حال حاضر در شرایط ناسالم است

وضعیت بر اساس تغییرات قابل توجه در اندازه گیری های آن شخص در مقابل هنجار. توجه داشته باشید

که این هنجار می تواند در نتیجه جمع آوری و میانگین گیری داده ها از یک بزرگ باشد

جمعیت افراد یا می تواند از مواردی باشد که فرد در یک حالت قرار دارد

از سلامت خوب شناخته شده

با الکترونیک مدرن، ما می توانیم داده ها را به روش های قبلی جمع آوری و مقایسه کنیم

خارج از اتاق بیمارستان غیرممکن است یک نمونه از یک محصول مصرفی که این را پر می کند

جایگاه اپل واچ است که قادر است داده های ضربان قلب را ردیابی و نمایش دهد

از طریق یک برنامه نوار قلب این دستگاه می تواند در مورد ضربان قلب غیر طبیعی بالا یا پایین نیز اطلاع دهد

به عنوان ریتم نامنظم قلب مهروترا و همکاران مثال‌های دیگری از ناهنجاری‌ها ارائه کنید

تشخیص می تواند در زمینه پزشکی ارزشمند باشد (11-12).

تجزیه و تحلیل ورزشی

ورزش یک تجارت بزرگ در سرتاسر جهان است و فوربس 12 فرنچایز ورزشی را رتبه بندی کرده است

ارزش حداقل 3.5 میلیارد دلار تا سال 2020. علاوه بر این، بسته به ورزش، یک تیم ممکن است سرمایه گذاری کند.

200 میلیون دلار یا بیشتر در هر فصل حقوق بازیکنان. به منظور به حداکثر رساندن این بازده

در سرمایه گذاری، تیم ها زمان و هزینه قابل توجهی را در زمینه ورزش صرف کرده اند

تجزیه و تحلیل. با این تأکید بر تجزیه و تحلیل، تجزیه و تحلیل و درک مهم است

برای اینکه فریب نخوریم که موفقیت آماری در یک فصل به این نتیجه برسد

لزوما به بعدی منتقل شود.

یک اشتباه 23 میلیون دلاری

کریس جانسون اولین بازیکن پایه سوم برای آتلانتا بریوز بود که وارد میدان شد

فصل 2013. او آن فصل را با میانگین ضربات 0.321 به پایان رساند که برای او سودآور بود

قرارداد 23 میلیون دلاری در چهار سال آینده. در حال آمدن به فصل 2014، شجاعان

انتظار داشت که او یکی از عوامل اصلی حمله باشد. در عوض، میانگین ضربه زدن او به پایین آمد

.263، و در اواخر سال 2015، جان هارت، مدیر کل آتلانتا، کریس جانسون را به عنوان بخشی از

تبادل قراردادهای بد با سرخپوستان کلیولند. در زمان عقد قراردادش، آنجا

دلیل خوبی برای این باور بود که جانسون به آن عمل نمی کند، و شکل 1-4 نشان می دهد چرا، با استفاده از آماری به نام میانگین ضربه زدن به توپ در بازی (BABIP). BABIP به چگونگی نگاه می کند

یک بازیکن وقتی توپ را وارد بازی می کند، به خوبی این کار را انجام می دهد، یعنی نادیده گرفتن هوم ران ها نیز همینطور

سناریوهایی مانند ضربه زدن یا راه رفتن در جایی که توپ هرگز در بازی نیست. یک مقدار عادی برای

BABIP معمولاً در محدوده 0.270-.330 قرار دارد، اما BABIP کریس جانسون با 0.394 او را ساخته است.

شماره 1 از 383 بازیکن واجد شرایط با حداقل 150 خفاش در طول فصل 2013. که جانسون لیگ را رهبری کرد و BABIP حداقل 60 امتیاز داشت (یعنی 0.060)

بالاتر از هنجار مهم است زیرا شواهد قابل توجهی وجود دارد که دفاع ها هستند

مسئول اصلی تعیین اینکه آیا یک توپ در بازی تبدیل به اوت می شود یا نه

خمیر خود را برای این اصل کلی استثنائاتی وجود دارد، اما مفهوم این است

که به طور معمول، بازیکنی که BABIP بسیار بالایی در یک فصل دارد، بعید است که تکرار کند

این در فصل بعد در مورد جانسون، او از 0.394 در سال 2013 به 0.345 در سال 2014 کاهش یافت.

جانسون که استعداد قابل توجهی در ضربه زدن به قدرت یا گرفتن پایه نداشت، بیشتر بود

نسبت به اکثر بازیکنان به میانگین ضربه زدن برای خروجی تهاجمی خود وابسته است، و این

بازگشت به مقداری نزدیکتر به میانگین، عامل فروپاشی تهاجمی او بود. داشته است

شجاعان فهمیدند که جانسون چقدر در این معیار مهم فاصله دارد،

شاید بعد از فصل 2013 به او پیشنهاد بدی ندادند و به دیگران اجازه دادند

تیم در عوض این اشتباه را انجام می دهد.

یک ناهنجاری پایدار

گرگ مادوکس یکی از بهترین پرتاب کننده های تمام دوران است. او به طور مداوم از خمیرها جلوگیری می کرد

برای 23 فصل در لیگ های بزرگ، به لطف کنترل باورنکردنی او بر زمین ها و

توانایی تقریباً بی‌نظیر تغییر سرعت در زمین‌ها بدون تسلیم شدن به ضربات.

مادوکس همچنین افتخار این را دارد که پارچ با بیشترین تفاوت بین آنها باشد رتبه‌بندی میانگین اجرا (ERA) و فیلدینگ مستقل Pitching (FIP) او برای

بازیکنان بین سال‌های 1988 و 2007، محدوده‌ای که نقطه‌ای را که مدوکس در آن چرخید را پوشش می‌دهد

گوشه به عنوان یک شروع کننده جوان به سال قبل از آویزان کردن گیره ها برای همیشه.

FIP به‌عنوان یک اندازه‌گیری در تحقیقاتی در مورد درک عواملی که a

pitcher قادر است در مقابل دفاع خود را کنترل کند. نتیجه این تحقیق این است که پارچ ها هستند

در درجه اول مسئول سه چیز است: پیاده روی، اعتصاب و دویدن در خانه. در غیر این صورت،

اکثر سهم به توپ های زده شده در بازی دقیقاً در دفاع قرار می گیرد - درست مثل

آنچه در بخش قبل با BABIP دیدیم.

یک نتیجه گیری از FIP این است که تمایل بیشتری به پیش بینی ERA یک پارچ دارد

فصل بعد نسبت به دوران پارچ در فصل جاری، چرا که دفاع ها به سمت

منظور داشتن. این به این معنی است که اگر پارتی را ببینید که ERA آن به طور قابل توجهی کمتر از FIP او است،

می‌توانیم انتظار عقب‌نشینی در سال آینده را داشته باشیم. یعنی مگر اینکه گرگ مادوکس باشد. شکل 1-5 نشان می دهد

1703 فصل را از پارچ ها خارج می کند و تفاوت بین ERA و FIP آنها را نشان می دهد،

که در آن یک مقدار منفی به این معنی است که یک پارچ از FIP و یک مقدار مثبت بهتر عمل کرده است

یعنی کم کار کرده گرگ مادوکس مسئول دو تفاوت بزرگ بین ERA و FIP است.

و آنها در سال های متوالی آمدند به بوت، به این معنی که او به عقب نشینی

آنطور که انتظار دارید معنی بدهید مدوکس مسئول چهار فصل از 30 فصل برتر بازیکن است

در این بازه زمانی و به طور مداوم از FIP خود بهتر عمل کرد. شکل 1-6 جمع را نشان می دهد

تفاوت بین ERA و FIP در طول زمان برای همه پارچ ها بین سال های 1988 و 2007،

و مادوکس در صدر فهرست قرار دارد و پس از آن پدرو مارتینز، خوزه ریخو و دنیس قرار دارند.

مارتینز. در این مورد، ما یک بار دیگر شاهد یک حالت پرت هستیم: مادوکس به طور مداوم در عملکرد خود بهتر عمل می کند.

ERA در مقایسه با انتظارات FIP، چیزی است که تعداد کمی از پارچ ها می توانند ادعا کنند. بر اساس

مجموعه کارهایی که او انجام داد، این موردی است که در آن وضعیت پرت مادوکس یک ناهنجاری است: او

الگوهای رفتاری به اندازه کافی با سایر پارچ ها متفاوت بود که منصفانه است بگوییم

او عملکرد متفاوتی را نسبت به سایر پرتابگرها دنبال کرد.

تجزیه و تحلیل وب

تجزیه و تحلیل وب سایت به دو دلیل یک زمینه عالی برای تشخیص ناهنجاری است. اول، وب سایت ها

تمایل به ایجاد ترافیک کافی دارند که یک انسان نتواند از طریق داده ها در هیچ کدام تجزیه و تحلیل کند

زمان معقول این امر مستلزم استفاده از ابزارهای تجمیع گزارش و تجزیه و تحلیل است

کار را ساده کنید، اما حتی با این ابزارها، زمان مورد نیاز برای تجزیه و تحلیل کامل است

ممکن است دلهره آور باشد دلیل دوم حتی مهمتر است: یک ثابت وجود دارد

جریان پنهانی از خطاها در هر سایت ابزارهای خودکار به طور مداوم وب سایت ها را اسکن می کنند و

آدرس های IP عمومی برای آسیب پذیری های شناخته شده و رایج. این می تواند منجر به 404 شود (پرونده نیست

یافت شد) خطاهای درخواست برای فایلی مانند index.php در یک سایت میزبان NET. اینها نیستند

درخواست های مشروع و به اندازه کافی مکرر هستند که در هر کدام از آنها بیش از حد نویز می شود

سیستم ساده ای که با فرض اینکه شما این پیام ها را فیلتر نکنید، هشدارها را پخش می کند.

ممکن است فیلتر کردن خطاهای نامشروع کافی باشد، اما حداقل دو خطای دیگر وجود دارد

کلاس های خطای غیر جالب: پیوندهای قدیمی و خطاهای تصادفی. برای مثال

از اولی، فرض کنید من یک ایمیل با پیوند به عکسی که آنلاین قرار داده ام ارسال می کنم. چند سال بعد، من عکس را از وب سایت حذف کردم، اما شما هنوز

لینک را در ایمیل خود داشته باشید اگر آن لینک را باز کنید، با خطای 404 (فایل یافت نشد) مواجه خواهید شد.

چون منبع دیگر وجود ندارد. برای مثال دومی، ممکن است سعی کنم به a دسترسی داشته باشم

URL از حافظه و، در طول فرآیند، به طور تصادفی کلید اشتباه، منجر شود

به اشتباه تایپی این می تواند منجر به خطای 404 نیز شود. به طور مشابه، ممکن است به طور تصادفی روی آن کلیک کنم

یک پیوند محافظت شده، از شما خواسته می شود برای اعتبارنامه ها، و سپس لغو را انتخاب کنید، که منجر به 403 می شود.

خطا (مجاز نیست). اینها خطاهای قانونی هستند، اما از نقطه نظر یک سیستم

مدیر، این یک رفتار طبیعی و قابل انتظار است. چیزی برای سیستم ها وجود ندارد

مدیر به "رفع" زیرا همه چیز طبق طراحی کار می کند.

جایی که تشخیص ناهنجاری می تواند جالب باشد، یافتن خطاهای بالاتر و فراتر از آن است

«تابش پس‌زمینه اینترنت» (با اعتبار استیو گیبسون برای قیاس)

و اشتباهات رایج فرض کنید یک محصول جدید منتشر می شود و توسعه می یابد

تیم پرچم ویژگی را فعال کرده است که 1٪ از ترافیک را به یک صفحه جدید ارسال می کند و 99٪ را باقی می گذارد.

برای ماندن در صفحه موجود فرض کنید در صفحه جدید مشکلی وجود دارد،

به طوری که درصدی از افراد خطاهای 404 مربوط به منابع از دست رفته را دریافت می کنند

در صفحه. اکنون ما یک الگوی رفتاری متفاوت داریم - دیگر یک یا دو نیست

درخواست های ناموفق برای یک منبع معین، و نه نویز مبتنی بر گسترده است. تشخیص ناهنجاری

سیستمی که می تواند این نوع خطاها را پیدا کند و در عین حال که بیشتر خطاها را فیلتر می کند، هشدار دهد

نویز قبلی می تواند مدیران سیستم و تیم های توسعه را به میزان قابل توجهی نجات دهد

در حالی که هنوز مشکلات انتشار کد را قبل از اینکه مشتری یک بلیط مشکل صادر کند، پیدا کنید.

و خیلی بیشتر

در این مرحله، باید روشن شود که تشخیص ناهنجاری به یک صنعت محدود نمی شود

یا رشته تحصیلی اما در عوض به طور گسترده قابل استفاده است. علاوه بر این، معلوم می شود که وجود دارد

مجموعه ای متداول از تکنیک های تشخیص ناهنجاری که مستقل از زمینه یا صنعت هستند،

به این معنی که تکنیک هایی که در این کتاب یاد می گیریم صرف نظر از شما قابل اجرا خواهد بود

صنعت.

اکنون که بررسی مختصری از ناهنجاری ها در طبیعت را تکمیل کرده ایم، مرحله بعدی این است

برای بررسی سه کلاس تشخیص ناهنجاری.

صفحه 32 یا همان 16 خود کتاب:

**کلاس های تشخیص ناهنجاری:**

سه کلاس گسترده از روش های تشخیص ناهنجاری در دسترس ما وجود دارد. که به هر کدام به نوبت خود بر اساس ترتیب ظاهر شدن آنها در فصل های بعدی خواهیم پرداخت ،همه سه تکنیک در حوزه های خود ارزشمند هستند و درک آن را مولفه های کلیدی هر یک به منظور اعمال بهترین رویکرد برای یک مشکل معین ارزشمند می کند .

تشخیص ناهنجاری آماری:

اولین کلاس از تکنیک های تشخیص ناهنجاری، تکنیک تشخیص ناهنجاری آماری است ، با این تکنیک‌ها، فرض می‌کنیم که داده‌ها از برخی توزیعهای اساسی خاص پیروی می کنند(در این تکنیک ها ما فرض میکنیم داده ها از توزیع های اساسی مثل توزیع مثلا گوسی تبعیت میکنند) - در حال حاضر، فکر کنید که داده ها دارای شکل هستند – نسبت به سایر توزیع از داده‌ها،{ در حال حاضر، به این فکر کنید که غیر از توزیع داده ها ما ، شکل یک مجموعه داده را داریم} ما هیچ فرضی در مورد اینکه این داده‌ها «باید» چگونه باشند، نداریم . به عبارت دیگر، ما مدلی نداریم با توجه به اینکه مجموعه ای از ورودی ها آن را داشته باشیم تا ، برخی از خروجی های پیش بینی شده خاص را دریافت کنید.

برای تکنیک های تشخیص ناهنجاری آماری، موارد پرت مواردی هستند که به طور قابل توجهی از مرکز توزیع ما دور هستند - یعنی به اندازه کافی دور از مرکزی که ما می توانیم انتظار داشته باشیم هستند که آنها واقعاً بخشی از یک توزیع نیستند. این تعاریف "به طور قابل توجهی متفاوت" و "به اندازه کافی دور" لزوماً آزاد هستند (loose منظوزر این است که تعاریف دقیق که در همه جا ثابت باشند نیستند و بسته به حوزه ای که داریم کار انجام می دهیم متفاوت هستند) آنها ممکن است توسط کاربر قابل تنظیم باشند و احتمالاً به دامنه بستگی دارند.

فصول 6 تا 8 بر گنجاندن تکنیک های آماری در یک سرویس تشخیص ناهنجاری کلی(تعمیم یافته) متمرکز است .

تشخیص ناهنجاری خوشه ای :

دسته دوم تکنیک های تشخیص ناهنجاری بر خوشه بندی متمرکز است. مفهوم خوشه بندی این است که نقاط را بر اساس فاصله آنها از یکدیگر گروه بندی می کنیم.وقتی نقاط به اندازه کافی به یکدیگر نزدیک باشند، آن گروه بندی ها را خوشه می نامیم. ما هیچ فرضی در مورد روابط زیربنایی (ارتباطاتی که داده ها با هم دارند را ما نمی دانیم فقط یکسری داده هستند) در داده ها نداشته باشیم - یعنی چگونه نقاط داده منفرد در نهایت با هم خوشه می شوند. ما فرض نمی کنیم که داده ها ازهیچ توزیع خاصی پیروی نمایند.

و ما تلاشی برای ایجاد مدلی از ورودی های مربوط به آن به خروجی ها نداریم . در عوض، ما فقط به خروجی ها و خوشه هایی که آنها تشکیل می دهند توجه می کنیم.هنگام انجام تشخیص پرت با استفاده از تکنیک های خوشه بندی، یک داده را اگر از هر خوشه ای به اندازه کافی دور باشد، نقطه پرت در نظر می گیریم .این منطقی است که تعریف ما از نقطه پرت به عنوان نقطه ای که به اندازه کافی از هنجار فاصله داشته باشد اما مستلزم آن است که ما درک خود را از "هنجار" از یک نقطه واحد گسترش دهیم - مانند مرکز یک توزیع - به مجموعه ای از نقاط، که هر یک ممکن است مستقل از نقاط دیگر باشد.(یعنی بهتر است که نقطه ای که میخواهیم فاصله سایر نقاط از این نقطه برای تعیین ناهنجار وعادی در نظر بگیریم بهتر است که آن نقطه بعنوان مرکز توزیع در نظر بگیریم).

فصل‌های 9 تا 11 بر ترکیب تکنیک‌های خوشه‌بندی در سرویس تشخیص ناهنجاری تعمیم یافته تمرکز خواهند داشت ، و فصل 12 یک نوع جالب را در خوشه بندی معرفی می کند.

تشخیص ناهنجاری مبتنی بر مدل:

آخرین کلاس تکنیک هایی که در این کتاب به آنها خواهیم پرداخت، تکنیک های مدل سازی است. در این مورد ،ما نوع مدل رفتار داده مورد انتظار را آموزش می دهیم(مدلی که انتخاب می کنیم با رفتار مطابقت دارد)، به طور معمول، در این مدل ها با توجه به برخی ورودی ها یا مجموعه ای از ورودی ها مسیر مشابهی را دنبال میکنیم ، ما برای هر ورودی وزن بر اساس ارزش آن و سپس ترکیب آن ورودی های وزنی با هم در برخی راهها برای تولید خروجی ما تعیین می کنیم . این تکنیک به رگرسیون معروف است و فوق العاده در دنیای علم داده رایج است. فرمول ما برای پیش بینی زمان دونده سرعت یک نمونه ای از رگرسیون است:



در این مورد، ما دقیقاً مشخص نمی کنیم که وزن ها چیست و چگونه متغیرها را ترکیب می کنیم ، زیرا جزییات پیاده سازی به ماهیت مسئله خاص آن بستگی دارند، به عنوان مثال، ممکن است فرض کنیم که پیش بینی زمان یک ورزشکار المپیکی بر اساس یک مدل خطی است که در آن هر متغیر وزنی با هم جمع می شود.برای رسیدن به یک تخمین از زمان، مانند آن :



نقاط پرت در این مورد مواردی هستند که در آن نقطه داده محاسبه شده درواقعیت تفاوت قابل ملاحظه ای با نقطه داده پیش بینی شده دارد. برگردیم به پرونده یوسین بولت، ممکن است مجموعه ورودی های وزنی خود را قرار دهیم و (به عنوان مثال) زمان 20.10 ثانیه برای مسابقات 200 متر مردان را دریافت کنیم. وقتی او آن را در 19.30 ثانیه اجرا می کند، پیش‌بینی ما به اندازه‌ای دور است که زمان بولت به یک داده پرت تبدیل شود. سپس، ما می توانیم جهش در شهود بسازیم (تغییر دهیم نظرمان را) و آن این است که بولت از مدل معمولی ما پیروی نمی کند، بلکه از یک مدل مختلف پیروی می کند . مدلی که ممکن است وزن های متفاوتی روی متغیرهای ورودی داشته باشد یا ممکن است حتی شامل متغیرهایی باشد که با سایر متغیرها ارتباط نداشته باشد (مربوط نیستند). هر چه که باشد، یک بار فرض کنیم که بولت از حالت دیگری پیروی می کند ، اکنون می توانیم ادعا کنیم که آن نقاط پرت در واقع ناهنجاری ها هستند . آنها از فرآیندی متفاوت از فرآیندی که اکثر داده های ما را ایجاد می کند ایجاد شده اند."""در مثال اضافه شدن یک وسیله جدید دیدیم که آن داده پرت در یک دوره زمانی مشخص دارد تکرار میشود پس این مورد باعث میشود که آن نقطه را نویز ندانیم .این مثال را من نوشتم و داخل کتاب نبوده است."""

انجام این جهش(تغییر دیدگاه که منظور همان درنظر گرفتن داده پرت برای یک نقطه داده است) مستلزم یک فرض کلیدی است:

مدلی که ما انتخاب کردیم رفتاری مناسب در پدیده اساسی زمینه ای در اکثر موارد به اندازه کافی خوب است. اگر مدل ما در پیش‌بینی نتایج خوب نیست، در نهایت با تصویرهای نادرست از آن نقاط پرت مواجه خواهیم شد ، زیرا مدل به اندازه کافی در تعیین چیزی که رفتار نرمال است خوب نیست. اگر چنین است، نباید به مدل اعتماد کنیم که به ما بگوید چه چیزی یک داده خارج از محدوده است.(اوت لایر) .

فصل های 13 تا 17 بر روی یک نوع خاص جالب از تشخیص ناهنجاری مبتنی بر مدل تمرکز دارند : تجزیه و تحلیل سری های زمانی

# ساخت ردیاب (آشکار ساز ) ناهنجاری (Building an Anomaly Detector)

اکنون که ما یک ایده سطح بالا در مورد چیستی ناهنجاری و چه تکنیک هایی برای یافتن ناهنجاری ها داریم ممکن است از آن استفاده کنیم، زمان آن فرا رسیده است که فرآیند ساختن یک سرویس تشخیص ناهنجاری را آغاز کنیم ، چیزی که در طول این کتاب به سمت آن کار خواهیم رفت. با این حال ، قبل ازایتکه ما یک خط کد بنویسیم یا در مقالات تحقیقاتی غوطه ور شویم، ، باید درک بهتری از چگونگی استفاده مشتریان مان از خدمات تشخیص ناهنجاری به دست آوریم .در حال حاضر، ما به سه مفهوم کلیدی طراحی نگاه خواهیم کرد.

اهداف کلیدی (Key Goals) :

اولین چیزی که باید بدانیم این است که هدف از سرویس تشخیص ناهنجاری ما چیست ؟ دلایل مختلفی وجود دارد که چرا یک فرد ممکن است بخواهد از آن سرویس تشخیص ناهنجاری استفاده کند واولین دلیل هشدار دادن به انسان در صورت وجود یک مشکل است . به عنوان مثال، با بازگشت به پرونده تجزیه و تحلیل وب، می خواهیم سیستم به مدیران و توسعه دهندگان ها زمانی که مشکلی وجود دارد که باعث می شود مردم منابع مربوطه وبشان را نمیبینند هشدار دهیم . در عین حال، ما نمی‌خواهیم کارکنان (در اینجا استفاده کنندگان معنی کنید یا اعضا) را نسبت به مواردی از این دست آگاه کنیم. ربات‌های خودکار که به دنبال آسیب‌پذیری‌های رایج هستند(و ناتوانی در یافتن آنها) و یا افراد تایپ اشتباه URL ها را انجام دادند. دلیل دوم، کاملا جدا از هشدار، آزمایش و بررسی یک حدس و گمان است. جدول 1-1 به ما نشان داد که یوسین بولت در المپیک 2008 چقدر سریع بود و ما می توانستیم ببینیم که تفاوت زمان بین بولت و بقیه رقبا بسیار زیاد بود. با این حال، ممکن است بخواهیم تحلیل عینی تری از داده ها را ببینیم، چیزی که ممکن است این احتمال را تعیین کند که یک ورزشکار معمولی المپیک بتواند 200 متررا مانند بولت در 19.30 ثانیه بدود. این تحلیل عقب مانده {با نگاهی به گذشته}(backward-looking analysis) لزوماً منجر به برخی الزام هشدار بعدی نمی شود ، علامت گذاری آن به عنوان یک مورد استفاده جداگانه.

دلیل سومی که ممکن است بخواهیم از سرویس تشخیص ناهنجاری استفاده کنیم، از سناریوی هشدار در اولین مورد استفاده(بعنوان اولین دلیل ما گفتیم که هشدار دهد)، اما یک قدم فراتر می رود که در ادامه آمده است. : ما می توانیم از یک سرویس تشخیص ناهنجاری استفاده کنیم تا به برنامه اجازه دهد بدون دخالت انسان مشکلی را برطرف کند.به عنوان مثال، فرض کنید مانیتورهای کامپیوتری رده بالا را می فروشید با میانگین ارزش خرده فروشی 1500 دلار آمریکا. یک تعطیلات بزرگ در راه است و شما می خواهید هزینه مانیتور خود را 150 دلار کاهش دهید و در مجموع آن را به قیمت 1350 دلار آمریکا بفروشید. متأسفانه در تلاش برای اعمال تخفیف، قیمت را 150 دلار آمریکا تعیین کرده اید و یک تعداد زیادی سفارش دریافت می شود زیرا مردم متوجه می شوند که این نمایشگر با 90٪ تخفیف در قیمت خرده فروشی آن مشخص شده است. این امر شرکت را در موقعیت بدی قرار می دهد و نیاز به اجازه دادن به سفارشات و جذب ضرر عمده دارد یا لغو سفارشات و به طور بالقوه افزایش خشم مشتریان و همچنین پلت فرم خرید به طور بالقوه به خطر می اندازد - در مورد دوم را در پی خواهد داشت،

ارتباط مشتری با پلتفرم تا جایی که مشتری به آن برسد از فروش در آن پلتفرم منع شده است. یک سرویس هشدار ممکن است نشان دهنده این قیمت باشد محدوده این مانیتور هرگز از 1300 تا 1500 دلار منحرف نشده است و این قیمت جدید کاملا متفاوت است ما می توانیم سیستمی را تصور کنیم که از تغییر قیمت جلوگیری می کند قبل از اینکه انسان تایید کند که این در واقع قیمت مورد نظر است.

در نهایت، ما حتی ممکن است ندانیم «عادی» چگونه به نظر می رسد و بخواهیم این را کشف کنیم.برای مثال، فرض کنید می‌خواهیم بفهمیم که چه نوع شخصی بیشترین احتمال خرید یک مانیتور 1500 دلاری با توجه به جزئیات جمعیت شناختی را دارد، ممکن است متوجه شویم که خوشه های خاصی وجود دارد از افرادی که تمایل به خرید مانیتورهای ما دارند و ممکن است ما به سمت آن مشتریان بازاریابی کنیم نقاط پرت از این فرآیند نیز جالب خواهد بود زیرا ممکن است نشان دهنده گروه های بالقوه مشتریانی است که مانیتور ما را خریداری نکرده اند اگر ما در مورد خواسته های آنها بیشتر بدانیم، ممکن است تمایل داشته باشیم. اطلاعاتی در مورد در شرایط دورافتاده، تحلیلگران ما ممکن است تحقیقات بازار را انجام دهند و تعیین کنند که آیا بخش هایی از بازار موجود وجود دارد یا خیر و اگر چنین است، چگونه می توان به این افراد دسترسی پیدا کرد.

# چگونه انسان ها با ناهنجاری ها برخورد می کنند؟ (How Do Humans Handle Anomalies?)

نکته بعدی این است که از خود بپرسید، افرادی که از سرویس تشخیص ناهنجاری شما استفاده می کنند چگونه عمل می کننددر موردگزارش‌های داده پرت را بررسی می‌کند و تعیین می‌کند که آیا یک نقطه پرت داده شده است در واقع یک ناهنجاری است یا خیر یا اگر صرفاً نویز هست؟ به عنوان مثال، آیا فرآیندی تحقیق خاصی وجود دارد که یک انسان باید دنبال کند؟ در یک سناریوی تولید، ممکن است تصور کنیم تشخیص ناهنجاری در سنسورها مانند دمای دستگاه یا میزان فشار یک دستگاه مشخص برای فشار دادن و تا کردن یک جعبه تراوش می کند. فرض کنید که یک انسان دمای ماشین را مشاهده می کند و اگر دما از حد قابل قبول بالاتر رفت، انسان دکمه ای را فشار می دهد تا ماده خنک کننده را پخش کند. اگر آن عمل دما را کاهش نمی دهد تا به حد قابل قبولی برساند ، سپس انسان دکمه دیگری را فشار می دهد تا عملیات کند شود. در نهایت، اگر دما همچنان خارج از حد پارامترهای نرمال باقی بماند ، انسان دکمه دیگری را فشار می دهد تا خط را تا زمانی که دستگاه خنک شود یا یک تکنسین بررسی دستگاه را کامل نماید متوقف کند، دلیل فهمیدن اینکه مردم چگونه از سرویس تشخیص ناهنجاری شما استفاده می کنند مهم است آن این است که ممکن است بتوانید به آنها کمک کنید تا فرآیندهای خاصی را خودکار کنند. در سناریوی قبلی، ما به راحتی می‌توانیم سیستمی را تصور کنیم که داده‌های دما را به سرویس تشخیص ناهنجاری ما تغذیه می‌کند.(اینکه دما را به سیستم ما بعنوان ورودی بدهد فرآیندی آسان است). و پس از دریافت تعداد معینی از موارد داده پرت در یک بازه زمانی معین، به طور خودکار هر یک از سه مرحله را انجام دهید. ارائه راه حل ظریف بیشتر ، شاید این سرویس توزیع(پخش) عامل خنک کننده را در حین نوشتن اخطار به گزارش خودکار کند. سپس، اگر سیستم سرعت عملیات را کاهش دهد، ممکن است به کارکنان تکنسین های هشدار دهد تا آنها را از مشکل آگاه کنند و گزینه خاموش شدن کامل برای انسان از آنجا خارج شوند (یعنی کارها اتومات شود و این گزینه را که باید توسط انسان ماشین را خاموش کند را حذف نماید) .

در صورتی که نتوانیم این نوع فعالیت را خودکار کنیم، کاربران همچنان ممکن است ارزش داشتن یک سرویس تشخیص ناهنجاری را پیدا کنند (اگر نتواند سیستم را خودکار نماییم ممکن است سیستم ناهنجاری را همچنان ارزشمند دانست)زیرا می تواند مبنایی را برای یک تحقیق فراهم کند. بازگشت به مثال تجزیه و تحلیل وب سایت، یک تعداد غیر طبیعی از خطاها – در کل چه تعداد خطاها به طور غیرعادی زیاد باشد یا یک تعداد غیر عادی یک کلاس خاص از خطا مانند 404 (فایل یافت نشد) یا 403 (نه مجاز) - می تواند به مدیران سیستم اطلاعات کافی برای بررسی بیشتر و ردیابی آنچه ممکن است باعث تغییر رفتار شده باشد در گزارش ها بدهد ، خواه استقرار کد باشد(ناشی از پیاده سازی کد باشد)، تغییر زیرساخت، یا خدمات ناموفق.

در بین این دو سطح تعامل، ما همچنین می‌توانیم یک سرویس پیش‌بینی را که بستگی به سرویس تشخیص ناهنجاری دارد و پس از دریافت اطلاعات پرت، صف‌های چیزی برای بررسی دستی دارد یک نمونه از این می تواند بررسی دستی تغییرات قیمت خودکار محصولات در اکثر مواقع باشد

، ممکن است دستی نخواهیم

بررسی تغییرات زیرا تعداد محصولات تغییر یافته می تواند بسیار زیاد باشد. یک

اتفاقی که در آن قیمت یک محصول به طور چشمگیری با استاندارد آن متفاوت است، با این حال،

یک مورد عالی برای بررسی دستی قبل از تغییر قیمت واقعی خواهد بود. این

ممکن است یک شرکت را به میزان قابل توجهی نجات دهد.

هنگامی که سعی می کنید اهداف خاص تماس گیرندگان خدمات خود را تشخیص دهید، بیشتر از همه مهمترین سوال این است که وقتی هشدار یک ناهنجاری بالقوه به آنها می دهید؟ چه دانشی در انسان لازم است ، یک ناهنجاری هست با استفاده از تعریف پذیرفته شده ما، لزوماً جالب است . انسان چه اطلاعاتی دارد که این چیز را جالب می کند و نه فقط یک حادثه دیگر از سر و صدا؟ علاوه بر این، چه مقدار از آن اطلاعات می توانیم اتوماسیون یا منتقل به ماشین آلات کنیم؟ همه چیز قابل اتوماسیون نیست، چه به دلیل کمبود تجهیزات مناسب و یا به دلیل مقدار مختلف و ناهمگون . دانش در ذهن افراد مختلف از راه حل خودکار جلوگیری می کند قابل دوام بودن گفته می شود، دلیل معمولی وجود یک سرویس تشخیص ناهنجاری است

ارائه اطلاعات در مورد ناهنجاری های احتمالی برای بررسی های بعدی و اقدام بالقوه.

# ناشناخته های شناخته شده (Known Unknowns)

همانطور که ما این بحث در مورد ناهنجاری ها را به پایان می رسانیم، یک نکته نهایی (پایانی) داریم: چه مقدار اطلاعات در حال حاضر در مورد ناهنجاری های موجود داریم؟ فرض کنید یک تاریخچه ای از تمام دما های ثبت شده برای یک ماشین معین را داریم . علاوه بر این، یک متخصص از طریق هر یک از آن ثبت ها و تعیین اینکه آیا دستگاه به اشتباه رفتار می کند یا خیررا مرتب کرده است ، به ما اجازه می دهد دمای بالا را با محصولاتی که با ماشینکاری نادرست برای تعمیر یا اسقاط نیاز داشتیم را مرتبط کنیم (یعنی اگر دما بالا بود نیاز به تعمیر یا اسقاط است). ما با این اطلاعات مسلح شدیم، ما می توانیم یادگیری نظارت شده انجام دهیم ، زیر مجموعه ای از یادگیری ماشینی که از اطلاعات خوب شناخته شده برای نتیجه گیری در مورد اطلاعات ناشناخته استفاده می کند . یادگیری تحت نظارت به ویژه کلاس محبوبی از یادگیری ماشین در دنیای تجارت است - اکثر مسائل یادگیری ماشینی تمایل دارند که بوسیله سناریوهای یادگیری نظارت شده تنظیم شوند و تعداد زیادی از ابزار و تحقیق در مورد تکنیک های یادگیری تحت نظارت وجود دارد . متأسفانه برای ما، تشخیص ناهنجاری در دنیای یادگیری تحت نظارت قرار نمی گیرد ، زیرا ما به ندرت این تجمل که متخصصان مجموعه داده های جامعی را برای ما بسازند را داریم. اغلب، ما مجبوریم یادگیری بدون نظارت، کلاس دیگری از یادگیری ماشین را انجام دهیم. که در آن ما پاسخ های خوبی در مورد اینکه آیا یک نقطه داده خاص یک ناهنجاری بود را نمی دانیم.(برچسب نداریم داده حاشیه نویسی شده نداریم) در عوض، این به ما بستگی دارد که کشف کنیم که آیا آن نقطه داده ارائه شده غیر عادی بود یا خیر؟ همانطور که در سراسر این کتاب خواهیم دید، تکنیک های زیادی وجود دارد که می توانیم برای یادگیری بدون نظارت استفاده کنیم، اما همه آنها با این هشدار همراه هستند که ما به راحتی نمی توانیم تعیین کنید که آشکارساز ناهنجاری ما چقدر درست (یا نادرست) است، زیرا اطلاعاتی خوب در مورد اینکه چه چیزی یک ناهنجاری در مقابل یک غیرعادی است را نداریم. گاهی اوقات، ممکن است از برخی سناریوهای ناهنجاری قبلی آگاهی داشته باشیم. این یک حد وسط بین دو نوع است که به عنوان یادگیری نیمه نظارت شناخته می شود. ایده در اینجا این است که ما ممکن است مقدار کمی از داده های خوب شناخته شده، مانند تاریخچه ثبت شده از برخی ناهنجاری ها، داشته باشیم ، و از این اطلاعات برای آموزش یک مدل استفاده کنید، همانطور که ما در مورد سناریو. نظارت شده انجام می دهیم از این نمونه کوچک، ما از تکنیک های خوشه بندی برای تعیین اینکه کدام نقاط داده به اندازه کافی شبیه به ناهنجاری های شناخته شده ما هستند استفاده می کنیم و اقدام به برچسب گذاری داده ها مشابه به عنوان ناهنجاری ها اشاره می کنند می کنیم. ما می‌توانیم آن نقاطی را که به‌عنوان نقاط پرت مشخص شده‌اند به انسان نشان دهیم و از آنها بخواهید به ما بگویند که آیا پیش‌بینی‌های ما درست بوده یا نه، در نتیجه ما داده های ورودی شناخته شده را بدون استفاده از زمان و انرژی زیاد برای بررسی مقادیر زیادی از داده ها گسترش می‌دهد. . یادگیری نیمه نظارتی معمولاً دقیق تر از یادگیری بدون نظارت است. با فرض اینکه داده های معتبر کافی برای ایجاد یک مدل اولیه دارید.

# نتیجه (Conclusion)

در طول این فصل اول، منجر به ماهیت نقاط پرت، نویز و ناهنجاری ها شد. مقادیر پرت مقادیری هستند که به اندازه کافی با هنجار (داده عادی) متفاوت هستند. نویز ها و ناهنجاری ها پارتیشن هایی(بخشهایی) از مجموعه نقاط پرت هستند. نقاط پرت که مورد توجه نیستند نویز هستند، در حالی که نقاط پرت جالب ناهنجاری هستند. در هر کجا تجزیه و تحلیل آماری از مسائل وجود دارد تشخیص ناهنجاری مهم است ، به این معنی که تشخیص ناهنجاری دارای یک خانه در یک طیف گسترده ای از صنایع و زمینه ها است . تشخیص ناهنجاری دارای یک جایگاه در صنایع و زمینههای زیادی است .

سه کلاس از تکنیک های تشخیص ناهنجاری وجود دارد که ما در سراسر کتاب روی آنها تمرکز خواهیم کرد : تکنیک های آماری، خوشه بندی، و مبتنی بر مدل. هر کلاس از تکنیک ها، مفروضات خاص خود را از داده ها و سناریوها به ارمغان می آورد و هر کدام از نظر خودش ارزشمند هستند. خواهيم ديد كه انواع معيني از مشكلات براي يكي از کلاس ها (تکنیک ها) سازگارتر هستند کلاس تکنیک نسبت به سایرین، بنابراین داشتن درک خوب از هر سه کلاس مهم است. در نهایت، ما به عنصر انسانی تشخیص ناهنجاری نگاه کردیم. مهم نیست چقدر خوب سرویس شما، ناهنجاری‌ها را تشخیص می‌دهد، از آنجایی که برای انسان‌ها ارزش دارد، مفید است. در این حد، درک نیازهای کسانی که از خدمات شما استفاده می کنند می تواند به شما در ساختن چیزی با ارزش کمک کند اگر چه ما بر روی جنبه های فنی شناسایی داده پرت و توسعه نرم افزار تمرکز خواهیم کرد ، مطمئن شوید که کاربر نهایی خود را در ذهن داشته باشید تا زمان توسعه چیز واقعی

بعبارتی میگوید که نیاز کاربران نهایی را در هنگام توسعه سیستم در نظر داشته باشید.