

یادگیری ماشین قابل تفسیر

راهنمای ساخت مدل‌های جعبه سیاه قابل توضیح

راهنمای ایجاد قابلیت توضیح برای مدل‌های جعبه سیاه

راهنمای قابل توضیح کردن مدل‌های جعبه سیاه

فهرست مطالب

Error! Bookmark not defined.	خلاصه
۱۰	فصل ۱ پیشگفتار نویسنده
۱۳	فصل ۲ مقدمه
۱۵	۲.۱ زمان داستان
۲۲	۲.۲ یادگیری ماشین چیست؟
۲۵	۲.۳ اصطلاحات
۲۹	فصل ۳ تفسیر پذیری
۳۰	۳.۱ اهمیت تفسیرپذیری
۳۷	۳.۲ طبقه‌بندی روش‌های تفسیرپذیری
۳۸	۳.۳ دامنه تفسیرپذیری
۳۸	۳.۳.۱ شفافیت الگوریتم
۳۹	۳.۳.۲ تفسیرپذیری مدل کل نگر
۳۹	۳.۳.۳ تفسیرپذیری مدل جهانی در سطح مدولار
۴۰	۳.۳.۴ تفسیر محلی برای یک پیش‌بینی واحد
۴۰	۳.۳.۵ تفسیر محلی برای گروهی از پیش‌بینی‌ها
۴۰	۳.۴ ارزیابی تفسیرپذیری
۴۱	۳.۵ خواص توضیحات
۴۴	۳.۶ توضیحات انسان پسند
۴۴	۳.۶.۱ توضیح چیست؟
۴۵	۳.۶.۲ یک توضیح خوب چیست؟
۵۰	فصل ۴ مجموعه داده‌ها
۵۰	۴.۱ اجاره دوچرخه (بازگشت)
۵۱	۴.۲ نظرات هرزنامه YouTube طبقه‌بندی متن (
۵۲	۴.۳ عوامل خطر برای سرطان دهانه رحم (طبقه‌بندی)
۵۴	فصل ۵ مدل‌های قابل تفسیر
۵۵	۵.۱ رگرسیون خطی
۵۷	۵.۱.۱ تفسیر

۵۹	۵.۱.۲ مثال
۶۱	۵.۱.۳ تفسیر بصری
۶۳	۵.۱.۴ پیش‌بینی‌های فردی را توضیح دهد.
۶۵	۵.۱.۵ رمزگذاری ویژگی‌های دسته بندی
۶۷	۵.۱.۶ آیا مدل‌های خطی توضیحات خوبی ایجاد می‌کنند؟
۶۷	۵.۱.۷ مدل‌های خطی پراکنده
۷۱	۵.۱.۸ مزایا
۷۲	۵.۱.۹ معایب
۷۲	۵.۲ رگرسیون لجستیک
۷۲	۵.۲.۱ رگرسیون خطی برای طبقه‌بندی چه اشکالی دارد؟
۷۴	۵.۲.۲ نظریه
۷۵	۵.۲.۳ تفسیر
۷۷	۵.۲.۴ مثال
۷۸	۵.۲.۵ مزایا و معایب
۷۹	۵.۲.۶ نرم افزار
۷۹	۵.۳ GAM و موارد دیگر GLM
۸۱	۵.۳.۱ نتایج غیر گاووسی - GLMs
۸۷	۵.۳.۲ فعل و انفعالات
۹۲	۵.۳.۳ جلوه‌های غیر خطی - GAM
۹۷	۵.۳.۴ مزایا
۹۸	۵.۳.۵ معایب
۹۸	۵.۳.۶ نرم افزار
۹۹	۵.۳.۷ برنامه‌های افروزنی بیشتر
۱۰۰	۵.۴ درخت تصمیم
۱۰۲	۵.۴.۱ تفسیر
۱۰۳	۵.۴.۲ مثال
۱۰۵	۵.۴.۳ مزایا
۱۰۶	۵.۴.۴ معایب

۱۰۶	۵.۴.۵ نرم افزار.....
۱۰۷	۵.۵ قوانین تصمیم گیری.....
۱۰۹	۵.۵.۱ یادگیری قوانین از یک ویژگی واحد(OneR)
۱۱۴	۵.۵.۲ پوشش متوالی.....
۱۱۸	۵.۵.۳ فهرست قوانین بیزی.....
۱۲۴	۵.۵.۴ مزايا
۱۲۵	۵.۵.۵ معایب
۱۲۶	۵.۵.۶ نرم افزار و جایگزین
۱۲۶	۵.۶ RuleFit.....
۱۲۸	۵.۶.۱ تفسیر و مثال.....
۱۳۰	۵.۶.۲ نظریه
۱۳۰	۵.۶.۳ مزايا
۱۳۴	۵.۶.۴ معایب
۱۳۵	۵.۷ سایر مدل‌های قابل تفسیر
۱۳۵	۵.۷.۱ طبقه‌بندی کننده ساده بیز
۱۳۵	۵.۷.۲ K-نzdیکترین همسایه‌ها
۱۳۷	فصل عروش‌های مدل-آگنوستیک
۱۴۰	فصل ۷ توضیحات مبتنی بر مثال
۱۴۲	فصل ۸ مدل جهانی-روشهای آگنوستیک
۱۴۳	۸.۱ طرح وابستگی جزئی(PDP)
۱۴۴	۸.۱.۱ اهمیت ویژگی مبتنی بر PDP
۱۴۵	۸.۱.۲ مثال
۱۴۸	۸.۱.۳ مزايا
۱۴۹	۸.۱.۴ معایب
۱۵۰	۸.۱.۵ نرم افزار و جایگزین
۱۵۱	۸.۲ طرح جلوه‌های محلی انباسته(ALE)
۱۵۱	۸.۲.۱ انگیزه و شهود
۱۵۵	۸.۲.۲ نظریه

۱۵۶	برآورد ۸.۲.۳
۱۶۰	مثالها ۸.۲.۴
۱۶۹	مزایا ۸.۲.۵
۱۷۰	معایب ۸.۲.۶
۱۷۲	اجرا و جایگزین ۸.۲.۷
۱۷۳	تعامل با ویژگی‌ها ۸.۳
۱۷۳	تعامل ویژگی؟ ۸.۳.۱
۱۷۴	نظریه: آماره H فریدمن ۸.۳.۲
۱۷۶	مثالها ۸.۳.۳
۱۷۸	مزایا ۸.۳.۴
۱۷۸	معایب ۸.۳.۵
۱۸۰	پیاده سازی‌ها ۸.۳.۶
۱۸۰	گزینه‌های جایگزین ۸.۳.۷
۱۸۱	تجزیه عملکردی ۸.۴
۱۸۳	چگونه کامپوننت‌ها را محاسبه نکنیم I ۸.۴.۱
۱۸۴	تجزیه عملکردی ۸.۴.۲
۱۸۵	چگونه کامپوننت‌ها را محاسبه نکنیم II ۸.۴.۳
۱۸۵	ANOVA ۸.۴.۴ عملکردی
۱۸۷	ANOVA ۸.۴.۵ عملکردی تعمیم یافته برای ویژگی‌های وابسته
۱۸۸	نمودارهای اثر محلی انباسته شده ۸.۴.۶
۱۸۹	مدل‌های رگرسیون آماری ۸.۴.۷
۱۹۰	پاداش: طرح وابستگی جزئی ۸.۴.۸
۱۹۰	مزایا ۸.۴.۹
۱۹۱	معایب ۸.۴.۱۰
۱۹۲	اهمیت ویژگی جایگشت ۸.۵
۱۹۲	نظریه ۸.۵.۱
۱۹۳	آیا باید اهمیت داده‌های آموزش یا آزمون را محاسبه کنم؟ ۸.۵.۲
۱۹۶	مثال و تفسیر ۸.۵.۳

۱۹۸	۸.۵.۴ مزایا
۱۹۹	۸.۵.۵ معایب
۲۰۱	۸.۵.۶ گزینه‌های جایگزین
۲۰۱	۸.۵.۷ نرم افزار
۲۰۲	۸.۶ جانشین جهانی
۲۰۲	۸.۶.۱ نظریه
۲۰۴	۸.۶.۲ مثال
۲۰۵	۸.۶.۳ مزایا
۲۰۶	۸.۶.۴ معایب
۲۰۶	۸.۶.۵ نرم افزار
۲۰۷	۸.۷ نمونه‌های اولیه و انتقادات
۲۰۸	۸.۷.۱ نظریه
۲۱۴	۸.۷.۲ مثالها
۲۱۴	۸.۷.۳ مزایا
۲۱۵	۸.۷.۴ معایب
۲۱۶	۸.۷.۵ کد و جایگزین
۲۱۷	فصل ۹ مدل محلی-روش‌های آگنوستیک
۲۱۸	۹.۱ انتظار شرطی فردی (ICE)
۲۱۸	۹.۱.۱ مثال‌ها
۲۲۳	۹.۱.۲ مزایا
۲۲۳	۹.۱.۳ معایب
۲۲۳	۹.۱.۴ نرم افزار و جایگزین
۲۲۵	۹.۲.۱ LIME برای داده‌های جدولی
۲۲۸	۹.۲.۱.۱ مثال
۲۲۹	۹.۲.۲ LIME برای متن
۲۳۱	۹.۲.۳ LIME برای تصاویر
۲۳۲	۹.۲.۴ مزایا
۲۳۳	۹.۲.۵ معایب

۲۳۵	توضیحات خلاف واقع	۹.۳
۲۳۸	ایجاد توضیحات خلاف واقع	۹.۳.۱
۲۴۳	مثال	۹.۳.۲
۲۴۴	مزایا	۹.۳.۳
۲۴۵	معایب	۹.۳.۴
۲۴۵	نرم افزار و جایگزین	۹.۳.۵
۲۴۷	قوانين محدوده (لنگرها)	۹.۴
۲۴۹	یافتن لنگرها	۹.۴.۱
۲۵۲	پیچیدگی و زمان اجرا	۹.۴.۲
۲۵۲	مثال داده‌های جدولی	۹.۴.۳
۲۵۶	مزایا	۹.۴.۴
۲۵۷	معایب	۹.۴.۵
۲۵۷	نرم افزار و جایگزین	۹.۴.۶
۲۵۸	ارزش‌های شپلی	۹.۵
۲۵۸	ایده کلی	۹.۵.۱
۲۶۱	مثال‌ها و تفسیر	۹.۵.۲
۲۶۳	ارزش Shapley در جزئیات	۹.۵.۳
۲۶۴	ارزش Shapley	۹.۵.۳.۱
۲۶۷	مزایا	۹.۵.۴
۲۶۸	معایب	۹.۵.۵
۲۶۹	نرم افزار و جایگزین	۹.۵.۶
۲۷۰	SHAP توضیحات افزودنی(SHapley)	۹.۶
۲۷۰	تعریف	۹.۶.۱
۲۷۲	KernelSHAP	۹.۶.۲
۲۷۵	TreeSHAP	۹.۶.۳
۲۷۷	مثالها	۹.۶.۴
۲۷۸	Aهمیت ویژگی SHAP	۹.۶.۵
۲۷۹	طرح خلاصه SHAP	۹.۶.۶

۲۸۱	۹.۶.۷ طرح وابستگی SHAP
۲۸۱	۹.۶.۸ ارزش‌های تعامل SHAP
۲۸۲	۹.۶.۹ خوشبندی مقادیر Shapley
۲۸۳	۹.۶.۱۰ مزایا
۲۸۴	۹.۶.۱۱ معایب
۲۸۴	۹.۶.۱۲ نرم افزار
۲۸۶	فصل ۱۱ انگاهی به توب کریستالی
۲۸۸	فصل ۱۳ با استناد به این کتاب
۲۸۹	فصل ۱۴ ترجمه‌ها
۲۹۰	فصل ۱۵ سپاسگزاریها

خلاصه

یادگیری ماشین پتانسیل زیادی برای بهبود محصولات، فرایندها و تحقیقات دارد. اما رایانه‌ها معمولاً پیش‌بینی‌های خود را توضیح نمی‌دهند که مانع برای پذیرش یادگیری ماشین است. این کتاب درباره تفسیر مدل‌های یادگیری ماشین و تصمیمات آن‌هاست.

پس از بررسی مفاهیم تفسیرپذیری، با مدل‌های ساده و قابل تفسیری مانند درخت تصمیم، قوانین تصمیم‌گیری و رگرسیون خطی آشنا خواهید شد. تمرکز کتاب بر روی روش‌های مدل-آگنوستیک برای تفسیر مدل‌های جعبه سیاه مانند اهمیت ویژگی و اثرات محلی انباسته شده و توضیح پیش‌بینی‌های فردی با مقادیر LIME و Shapley است.

همه روش‌های تفسیر به طور عمیق توضیح داده شده و به صورت انتقادی مورد بحث قرار می‌گیرند. چگونه زیر روپوش کار می‌کنند؟ نقاط قوت و ضعف آنها در چیست؟ چگونه می‌توان خروجی‌های آنها را تفسیر کرد؟ این کتاب شما را قادر می‌سازد تا روش تفسیری را که برای پروژه یادگیری ماشین شما مناسب‌تر است، انتخاب و به درستی اعمال کنید. خواندن این کتاب برای یادگیران یادگیری ماشین، دانشمندان داده، آماردانان و هر کسی که علاقه‌مند به تفسیرپذیر ساختن مدل‌های یادگیری ماشین است، توصیه می‌شود.

درباره من: نام من Christoph Molnar است، من یک آماردان و یک متخصص یادگیری ماشین هستم. هدف من این است که یادگیری ماشین را قابل تفسیر کنم.

@ChristophMolnar
من را در توییتر دنبال کنید!
جلد توسط @YvonneDoinel

همچنین کتاب دوم من Modeling Mindsets را مشاهده کنید.
این کتاب تحت مجوز Creative Commons Attribution-NonCommercial-ShareAlike 4.0 بین‌المللی مجوز دارد.

فصل ۱ پیشگفتار نویسنده

این کتاب زمانی که من به عنوان آمارگیر در تحقیقات بالینی کار می‌کردم به عنوان یک پروژه جانبی شروع شد. چهار روز در هفته کار می‌کردم و در روزهای بیکاری روی پروژه‌های جانبی کار می‌کردم. در نهایت، یادگیری ماشین قابل تفسیر به یکی از پروژه‌های جانبی من تبدیل شد. در ابتدا قصد نوشتن کتاب نداشتم. فقط علاقه‌مند به یافتن اطلاعات بیشتر در مورد یادگیری ماشین قابل تفسیر بودم و به دنبال منابع خوبی برای آموختن بودم. با توجه به موقیت یادگیری ماشین و اهمیت تفسیرپذیری، من انتظار داشتم که تعداد زیادی کتاب و آموزش در مورد این موضوع وجود داشته باشد. اما من فقط چند مقاله تحقیقاتی و چند پست وبلاگ پراکنده در سراسر اینترنت را پیدا کردم و هیچ منبع جامعی پیدا نکردم. نه کتاب، نه آموزش، نه مقاله مروری، نه هیچ چیز دیگری. این خلاً باعث شد من شروع به نوشتن این کتاب کنم. در نهایت شروع به نوشتن کتابی کردم که آرزو داشتم زمانی که مطالعه خود را در مورد یادگیری ماشین قابل تفسیر شروع کردم، وجود داشته باشد. قصد من از این کتاب دو چیز بود: برای خودم یاد بگیرم و این دانش جدید را با دیگران به اشتراک بگذارم.

من مدرک لیسانس و فوق لیسانس خود را در رشته آمار در LMU مونیخ آلمان دریافت کردم. بیشتر دانش من در مورد یادگیری ماشین به صورت خودآموز و شرکت در دوره‌های آنلاین، مسابقات، پروژه‌های جانبی و فعالیت‌های حرفه‌ای است. پیشینه آماری من مهارت بسیار خوبی برای ورود به یادگیری ماشین و بهویژه برای تفسیرپذیری بود. در آمار، تمرکز عمدۀ بر ساخت مدل‌های رگرسیون قابل تفسیر است. بعد از اینکه فوق لیسانس آمار را تمام کردم تصمیم گرفتم به مقطع دکتری نرم‌ware، چون از نوشتن پایان‌نامه فوق لیسانس لذت نبردم. نوشتن خیلی به من استرس وارد می‌کرد؛ بنابراین به عنوان دانشمند داده در استارت‌آپ Fintech و به عنوان آماردان در تحقیقات بالینی مشغول به کار شدم. بعد از سه سال کار در صنعت، نوشتن این کتاب را شروع کردم و چند ماه بعد، دکترای خود را در زمینه یادگیری ماشین تفسیرپذیر شروع کردم. در حین کار بر روی این کتاب، لذت نوشتن را دوباره کشف کردم و به من کمک کرد تا اشتیاقم به تحقیق را زیادتر کنم.

این کتاب بسیاری از تکنیک‌های یادگیری ماشین قابل تفسیر را پوشش می‌دهد. در فصل اول، مفهوم تفسیرپذیری را معرفی می‌کنم و انگیزه لازم را برای تفسیرپذیری بیان می‌کنم. چند داستان کوتاه برای درک بهتر این موضوع آورده شده است! این کتاب در مورد ویژگی‌های مختلف توضیحات و آنچه که انسان فکر می‌کند توضیح خوبی است، بحث می‌کند. سپس مدل‌های یادگیری ماشین که ذاتاً قابل تفسیر هستند، مانند مدل‌های رگرسیون و درخت‌های تصمیم موردبخت قرار می‌دهیم. تمرکز اصلی این کتاب بر روی روش‌های تفسیرپذیری مدل-آگنوستیک است. مدل-آگنوستیک به این معنی است که این روش‌ها را می‌توان برای هر مدل یادگیری ماشین اعمال کرد و پس از آموزش مدل اعمال می‌شود. این استقلال از مدل، روش‌های مدل-آگنوستیک را بسیار انعطاف‌پذیر و قدرتمند می‌کند. برخی از تکنیک‌ها چگونگی پیش‌بینی‌های فردی را توضیح می‌دهند، مانند

توضیحات مدل-آگنوستیک محلی قابل تفسیر (LIME) و مقادیر Shapley. سایر تکنیک‌ها میانگین رفتار مدل را در یک مجموعه‌داده توصیف می‌کنند. در اینجا با نمودار وابستگی جزئی، اثرات محلی انباشته شده، اهمیت ویژگی جای گشت و بسیاری از روش‌های دیگر آشنا می‌شویم. یک دسته خاص، روش‌های مبتنی بر مثال است که نقاط داده را به عنوان توضیحات تولید می‌کند. توضیحات خلاف واقع، نمونه‌های اولیه، نمونه‌های تأثیرگذار و مثال‌های مختصراً روش‌های مبتنی بر مثال هستند که در این کتاب موربuth قرار گرفته‌اند. این کتاب با برخی تأملات در مورد آینده یادگیری ماشین قابل تفسیر به پایان می‌رسد.

شما مجبور نیستید کتاب را از ابتدا تا انتهای بخوانید، می‌توانید به جلو و عقب بروید و روی تکنیک‌هایی تمرکز کنید که بیشتر مورد علاقه شما هستند. من فقط توصیه می‌کنم که از مقدمه و فصل تفسیرپذیری شروع کنید. اکثر بخش‌ها از ساختار مشابهی پیروی می‌کنند و بر یک روش تفسیری تمرکز می‌کنند. پاراگراف اول روش را خلاصه می‌کند. سپس سعی می‌کنم بدون اتكا به فرمول‌های ریاضی، روش را به صورت شهودی توضیح دهم. سپس به تئوری روش می‌پردازم تا درک عمیقی از نحوه عملکرد آن به دست آوریم. این قسمت حاوی فرمول‌هایی خواهد بود. من معتقدم که یک روش جدید با استفاده از مثال‌ها به بهترین وجه قابل درک است. بنابراین، هر روش برای داده‌های واقعی اعمال می‌شود. برخی افراد می‌گویند که آماردانان افراد بسیار منتقدی هستند. این موضوع برای من صدق می‌کند، زیرا هر فصل شامل بحث‌های انتقادی در مورد مزايا و معایب روش تفسیر مربوطه است. این کتاب تبلیغی برای روش‌ها نیست، اما باید به شما کمک کند تصمیم بگیرید که آیا این روش برای تحقیق شما خوب است یا خیر. در بخش آخر هر بخش، نرم افزارهای پیاده سازی موجود آورده شده است.

یادگیری ماشین مورد توجه بسیاری از افراد در تحقیقات و صنعت قرار گرفته است. گاهی اوقات یادگیری ماشین بیش از حد در رسانه‌ها مطرح می‌شود، در حالیکه کاربردی واقعی و تأثیرگذار معینی وجود دارد. یادگیری ماشین یک فناوری قدرتمند برای محصولات، تحقیقات و اتماسیون است. به عنوان مثال، امروزه از یادگیری ماشین در موارد زیر استفاده می‌شود: برای شناسایی تراکنش‌های مالی تقلیبی، توصیه فیلم‌ها و طبقه‌بندی تصاویر. اغلب مهم است که مدل‌های یادگیری ماشین قابل تفسیر باشند. تفسیرپذیری به توسعه دهنده‌گان در رفع اشکال و بهبودها کمک می‌کند، اعتماد به مدل ایجاد می‌کند، پیش‌بینی‌های مدل را توجیه می‌کند و به بینش‌های جدید منجر می‌شود. افزایش نیاز به تفسیرپذیری یادگیری ماشین نتیجه طبیعی افزایش استفاده از یادگیری ماشین است. این کتاب منبعی ارزشمند برای بسیاری از افراد می‌تواند باشد. مربیان آموزشی می‌توانند از این کتاب برای معرفی دانش آموزان خود با مفاهیم یادگیری ماشین قابل تفسیر استفاده می‌کنند. من از چندین دانشجوی کارشناسی ارشد و دکتری ایمیل دریافت کرده‌ام. دانشجویانی که به من گفتند این کتاب نقطه شروع و مهم‌ترین مرجع پایان نامه‌های آنها بوده است. این کتاب به محققان کاربردی در زمینه‌های بوم‌شناسی، مالی، روانشناسی و غیره که از یادگیری ماشین برای درک داده‌های خود استفاده می‌کنند کمک کرده است. دانشمندان داده که در صنعت کار

می‌کنند به من گفتند که از کتاب "یادگیری ماشین قابل تفسیر" برای کار خود استفاده می‌کنند و آن را به همکاران خود توصیه می‌کنند. خوشحالم که افراد زیادی از این کتاب بهره برده‌اند و در تفسیر مدل متخصص شدند. من این کتاب را به علاقمندانی توصیه می‌کنم که می‌خواهند مرواری بر تکنیک‌های تفسیرپذیر تر کردن مدل‌های یادگیری ماشین خود داشته باشند. همچنین برای دانشجویان و محققین (و هر کس دیگری) که به موضوع علاقه‌مند است، مفید خواهد بود. برای استفاده حداکثری از این کتاب، باید درک اولیه‌ای از یادگیری ماشین داشته باشید. همچنین باید درک درستی از ریاضیات پایه دانشگاهی داشته باشید تا بتوانید تئوری و فرمول‌های این کتاب را دنبال کنید. با این حال، درک توصیف شهودی روش در ابتدای هر فصل بدون ریاضیات نیز باید امکان‌پذیر باشد.

امیدوارم از کتاب لذت ببرید!

فصل ۲ مقدمه

این کتاب به شما توضیح می‌دهد که چگونه می‌توانید مدل‌های یادگیری ماشین (با نظارت) را قابل تفسیر کنید. بخش‌ها حاوی برخی فرمول‌های ریاضی هستند، اما شما باید بتوانید ایده‌های پشت روش‌ها را حتی بدون فرمول‌ها درک کنید. این کتاب برای افرادی نیست که سعی می‌کنند یادگیری ماشین را از ابتدا یاد بگیرند. اگر در یادگیری ماشین تازه‌کار هستید، کتاب‌ها و منابع دیگری برای یادگیری اصول اولیه وجود دارد. من کتاب «عناصر یادگیری آماری» اثر Andrew Ng (۲۰۰۹) و دوره آنلاین «یادگیری ماشین» coursera.com در پلتفرم یادگیری آنلاین^۱ را برای شروع با یادگیری ماشین توصیه می‌کنم. هم کتاب و هم دوره رایگان در دسترس هستند! روش‌های جدید برای تفسیر مدل‌های یادگیری ماشین با سرعتی سراسام‌آور منتشر می‌شوند. همگام‌شدن با هر آنچه منتشر می‌شود غیرممکن است. به همین دلیل است که در این کتاب جدیدترین و فانتزی‌ترین روش‌ها را پیدا نمی‌کنید، بلکه روش‌های ثبت شده و مفاهیم اساسی تفسیرپذیری یادگیری ماشین را پیدا خواهید کرد. این اصول شما را برای ساختن مدل‌های یادگیری ماشین قابل تفسیر آماده می‌کند. درک مفاهیم اساسی به شما این امکان را می‌دهد که هر مقاله جدیدی در مورد تفسیرپذیری منتشر شده در arxiv.org در ۵ دقیقه گذشته از زمان شروع خواندن این کتاب را بهتر درک و ارزیابی کنید (ممکن است در میزان انتشار اغراق کنم).

این کتاب با چند داستان کوتاه (کابوس وار) شروع می‌شود که برای درک کتاب مورد نیاز نیست، اما امیدوارم شما را سرگرم کند و به فکر فروبرد. سپس این کتاب مفاهیم تفسیرپذیری یادگیری ماشین را بررسی می‌کند. ما در مورد اینکه تفسیرپذیری مهم است و انواع مختلف توضیحاتی که وجود دارد بحث خواهیم کرد. اصطلاحات استفاده شده در سراسر کتاب را می‌توان در بخش اصطلاحات جستجو کرد. بیشتر مدل‌ها و روش‌های توضیح داده شده، با استفاده از نمونه‌های داده واقعی ارائه شده‌اند که در فصل داده‌ها شرح داده شده است. یکی از راه‌های قابل تفسیر کردن یادگیری ماشین، استفاده از مدل‌های قابل تفسیر، مانند مدل‌های خطی یا درخت‌های تصمیم‌گیری است. گزینه دیگر استفاده از ابزارهای تفسیر مدل-آگنوستیک که می‌توانند برای هر مدل یادگیری ماشین نظرات شده‌ای اعمال شوند. روش‌های مدل-آگنوستیک را می‌توان به روش‌های کلی که رفتار میانگین مدل را توصیف می‌کند و روش‌های محلی که پیش‌بینی‌های فردی را توضیح می‌دهند، تقسیم کرد. فصل روش‌های مدل-آگنوستیک به روش‌هایی مانند نمودارهای واپستگی جزئی^۲ و اهمیت ویژگی^۳ می‌پردازد. روش‌های مدل-آگنوستیک با تغییر ورودی مدل یادگیری ماشین و اندازه‌گیری تغییرات در خروجی کار می‌کنند. این کتاب با یک چشم‌انداز خوش‌بینانه در مورد آینده یادگیری ماشین قابل تفسیر به پایان می‌رسد.

می‌توانید کتاب را از ابتدا تا انتهای بخوانید یا مستقیماً به روش‌های مورد علاقه خود بروید.

¹ <https://www.coursera.org/learn/machine-learning>

² Partial dependence plots

³ Feature importance

امیدوارم از خواندن لذت ببرید!

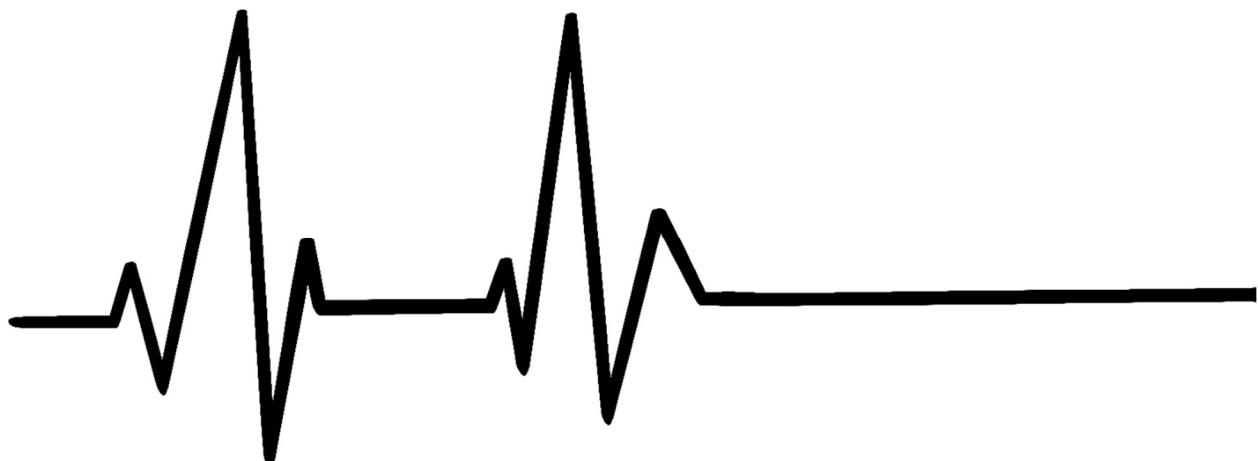
۲.۱ زمان داستان

با چند داستان کوتاه شروع می‌کنیم. هر داستان یک فراخوان اغراق‌آمیز برای یادگیری ماشین قابل تفسیر است. اگر عجله دارید، می‌توانید از داستان‌ها صرف‌نظر کنید. اگر می‌خواهید سرگرم شوید و انگیزه ندارید، ادامه مطلب را بخوانید!

این قالب از داستان‌های فناوری Import AI Newsletter در خبرنامه Jack Clark او الهام گرفته شده است. اگر این نوع داستان‌ها را دوست دارید یا اگر به هوش مصنوعی علاقه‌مند هستید، توصیه می‌کنم در این خبرنامه ثبت‌نام کنید.

رعدوبرق هرگز دو بار نمی‌زند

۲۰۳۰ : یک آزمایشگاه پزشکی در سوئیس



تام گفت: "قطعًا این بدترین راه برای مردن نیست!" و سعی کرد چیز مثبتی در تراژدی پیدا کند. او پمپ را از قطب داخل وریدی خارج کرد.

لنا افزود: "او فقط به دلایل اشتباه مرد."

تام در حالی که پیچ پشتی پمپ را باز می‌کرد، شکوه کنان گفت: "و مطمئنًا با پمپ مرفین خراب! فقط کار برای ما زیاد کردا!" بعد از برداشتن تمام پیچ‌ها، صفحه را بلند کرد و کنار گذاشت. او یک کابل را به درگاه تشخیص وصل کرد.

لنا لبخند تمسخرآمیزی به او زد و گفت: "شما فقط از داشتن شغل شکایت نکردید، نه؟"

تام با لحنی کنایه‌آمیز داد زد: "البته که نه. هرگزا" و کامپیوتر پمپ را بوت کرد.

لنا سر دیگر کابل را به تبلتش وصل کرد و گفت: "بسیار خوب، تشخیص در حال انجام است. من واقعاً کنجکاو هستم که ببینم چه اشتباهی رخداده است".

تام گفت: "مطمئناً از طرف ناشناسی به مریض شلیک شده است. غلظت بالای این مواد مورفین. این اولین بار است، درسته؟ معمولاً یک پمپ شکسته اصلاً موادی نمی‌دهد یا مقدار بسیار کمی را تولید می‌کند. اما هرگز، این مقدار زیاد تولید نمی‌کند."

لنا تبلتش را بالا آورد و گفت: "این را نگاه کن. این پیک را اینجا می‌بینی؟ این قدرت ترکیب مسکن‌ها است. نگاه کن این خط سطح مرجع را نشان می‌دهد. بیچاره مخلوطی از مسکن در سیستم خونش داشت که می‌توانست ۱۷ بار او را بکشد. توسط پمپ ما در اینجا تزریق می‌شود. و در اینجا... او با انگشت خود تند تند گفت: "در اینجا می‌توانید لحظه مرگ بیمار را ببینید".

تام از سرپرستش پرسید: "پس، آیا می‌دانید چه اتفاقی افتاده است، رئیس؟"

لنا گفت: "حسگرها به نظر سالم هستند. ضربان قلب، سطح اکسیژن، گلوکز، و ... داده‌ها همان‌طور که انتظار می‌رفت جمع‌آوری شد. برخی از مقادیر ازدست‌رفته در داده‌های اکسیژن خون وجود دارند، اما این غیرعادی نیست. اینجا را نگاه کن. سنسورها همچنین کاهش ضربان قلب بیمار و سطوح بسیار پایین کورتیزول ناشی از مشتقات مورفین و سایر عوامل مسدود‌کننده درد را تشخیص داده‌اند." او همچنان به مرور گزارش تشخیصی ادامه داد.

تام مجذوب صفحه‌نمایش شده بود. این اولین تحقیق او در مورد خرابی واقعی دستگاه بود.

لنا به تام گفت: "خوب، اینجا اولین قطعه از پازل ماست. این سیستم در ارسال اخطار به کanal ارتباطی بیمارستان ناموفق بود. هشدار ایجاد شد، اما در سطح پروتکل رد شد. ممکن است تقصیر ما باشد، اما ممکن است تقصیر بیمارستان نیز باشد".

تام درحالی که چشمانت همچنان به صفحه‌نمایش خیره شده بود سر تکان داد.

لنا ادامه داد: "عجب است. این هشدار همچنین باید باعث خاموش شدن پمپ شود. اما مشخص است که موفق به انجام این کار نشده است. این باید یک اشکال باشد. چیزی که تیم با کیفیت از دست داد. یه چیز واقعاً بد. شاید به مشکل پروتکل مربوط باشد".

تام با تعجب گفت: "بنابراین، سیستم اورژانسی پمپ به نوعی خراب شد، اما چرا پمپ پر شد و این همه مسکن به جان دو تزریق کرد؟"

لنا توضیح داد: "سؤال خوبی بود. حق با شمامست. جدا از خرابی اضطراری پروتکل، پمپ اصلاً نباید آن مقدار دارو را تجویز می‌کرد. با توجه به سطح پایین کورتیزول و سایر علائم هشدار، الگوریتم باید خیلی زودتر به خودی خود متوقف می‌شد."

تام پرسید: "شاید یک بدشานسی، مانند یک در میلیون، مانند برخورد با رعدوبرق؟"

لنا توضیح داد: "نه، تام. اگر اسنادی را که برایتان فرستادم خوانده بودید، می‌دانستید که پمپ ابتدا در آزمایش‌های حیوانی و سپس روی انسان‌ها آموزش داده شد تا بر اساس ورودی حسی، مقدار مناسبی از مسکن‌ها را تزریق کند. الگوریتم پمپ ممکن است مبهم و پیچیده باشد، اما تصادفی نیست. این بدان معناست که در شرایط مشابه، پمپ دوباره دقیقاً به همان روش عمل می‌کند. بیمار ما دوباره می‌میرد. ترکیبی یا اثر متقابل نامطلوب ورودی‌های حسی باید باعث رفتار اشتباه پمپ شده باشد. به همین دلیل است که ما باید عمیق‌تر بگردیم و بفهمیم اینجا چه اتفاقی افتاده است."

تام که در فکر فرو رفته بود پاسخ داد: "می‌بینم. آیا به هر حال بیمار به زودی نخواهد مرد؟ به خاطر سرطان یا مريضى مشابه آن؟"

لنا درحالی که گزارش تحلیل را می‌خواند سر تکان داد.

تام بلند شد و به سمت پنجره رفت. به بیرون نگاه کرد، چشمانش به نقطه‌ای در دوردست دوخته شد. "شاید دستگاه به او لطفی کرده است که او را از درد رهایی بخشد. دیگر رنجی نیست. شاید کار درست را انجام داده است. مثل یک رعدوبرق، اما، می‌دانید، یک رعدوبرق خوب. منظورم مثل قرعه کشی است، اما نه تصادفی. اما به دلیلی دیگر، اگر من جای پمپ بودم، همین کار را می‌کردم".

بالاخره لنا سرش را بلند کرد و به او نگاه کرد.

تام مدام به چیزی بیرون نگاه می‌کرد.

هر دو برای چند لحظه سکوت کردند.

لنا دوباره سرش را پایین انداخت و به تحلیل ادامه داد. "نه، تام. این یک اشکال است... فقط یک باگ لعنی".

به افتادن اعتماد کن

۲۰۵۰: یک ایستگاه مترو در سنگاپور



با عجله به سمت ایستگاه متروی بیشان رفت. با افکارش از قبل سر کار بود. آزمایشات برای معماری عصبی جدید باید تا الان کامل شده باشد. او بازطراحی «سیستم پیش‌بینی وابستگی مالیاتی برای اشخاص حقیقی» را مدیریت می‌کرد که پیش‌بینی می‌کند آیا شخص پول را از اداره مالیات پنهان می‌کند یا خیر. تیم او یک قطعه مهندسی ظریف را ارائه کرده است. در صورت موفقیت، این سیستم نه تنها به اداره مالیات خدمت می‌کند، بلکه به سیستم‌های دیگر مانند سیستم هشدار ضد تروریسم و ثبت تجاری نیز وارد می‌شود. یک روز، دولت حتی می‌تواند پیش‌بینی‌ها را در امتیاز اعتماد مدنی ادغام کند. امتیاز اعتماد مدنی تخمین می‌زند که یک فرد چقدر قابل اعتماد است. این تخمین بر هر بخش از زندگی روزمره شما تأثیر می‌گذارد، مانند دریافت وام یا مدت زمانی که باید برای پاسپورت جدید صبر کنید. وقتی از پله برقی پایین می‌آمد، او تصور کرد که ادغام سیستم تیمش، در سیستم امتیاز اعتماد مدنی چگونه خواهد بود.

او به طور معمول دست خود را روی دستگاه RFID خوان بدون کاهش سرعت راه رفتنش کشید. ذهن او درگیر بود، اما ناهمانگی انتظارات حسی و واقعیت زنگ خطر را در مغزش به صدا درآورد. خیلی دیر.

دماغ ابتدا وارد دروازه ورودی مترو شد و با پشت به زمین افتاد. قرار بود در باید باز می‌شد، اما باز نشد. مات و مبهوت از جایش بلند شد و به صفحه‌نمایش کنار ورودی نگاه کرد. یک شکلک دوستانه روی صفحه پیشنهاد کرد: «لطفاً یک بار دیگر امتحان کنید». شخصی از آنجا گذشت و بی توجه به او دستش را از روی صفحه گذراند. در باز شد و او رفت. در دوباره بسته شد. بینی اش را پاک کرد. درد داشت ولی حداقل خونریزی نداشت. سعی کرد در را باز کند، اما دوباره در باز نشد. عجیب بود. شاید حساب حمل و نقل عمومی او توکن کافی نداشته باشد. او برای بررسی موجودی حساب به ساعت هوشمند خود نگاه کرد.

ساعتش به او اعلام کرد: "ورود رد شد. لطفاً با دفتر مشاوره شهروندان خود تماس بگیرید!".

احساس تهوع مثل مشت به شکمش خورد. او مشکوک بود که چه اتفاقی افتاده است. برای تایید نظریه خود، او بازی موبایل "Sniper Guild" را شروع کرد که یک مسابقه تیراندازی بود. برنامه به طور خودکار بسته شد، که نظریه او را تایید کرد. گیج شد و دوباره روی زمین نشست.

تنها یک توضیح ممکن وجود داشت: امتیاز اعتماد مدنی او بطور قابل ملاحظه‌ای کاهش یافته بود. یک افت کوچک به معنای محرومیت‌های جزئی بود، مانند عدم دریافت پروازهای درجه یک یا نیاز به کمی بیشتر صبر کردن برای استناد رسمی. نمره اعتماد پایین نادر بود و به این معنی بود که شما به عنوان یک تهدید برای جامعه طبقه‌بندی می‌شوید. یکی از اقدامات در برخورد با این افراد دور نگه داشتن آنها از مکان‌های عمومی مانند مترو بود. دولت تراکنش‌های مالی افراد دارای امتیاز اعتماد مدنی پایین را محدود کرد. آنها همچنین شروع به نظارت فعالانه بر رفتار شما در رسانه‌های اجتماعی کردند و حتی تا آنجا پیش رفتند که محتوای خاصی مانند بازی‌های

خشونت آمیز را محدود کردند. افزایش امتیاز اعتماد مدنی هر چه کمتر بود به طور تصاعدی دشوارتر می‌شد. امتیاز افراد با نمره بسیار پایین معمولاً هرگز بهبود نمی‌یابند.

او نمی‌توانست به هیچ دلیلی فکر کند که چرا نمره او باید پایین می‌آمد. امتیاز بر اساس یادگیری ماشین بود. سیستم امتیاز اعتماد مدنی مانند موتور روغن کاری شده‌ای عمل می‌کرد که جامعه را اداره می‌کرد. عملکرد سیستم امتیاز اعتماد همیشه به دقت نظارت می‌شد. یادگیری ماشین از ابتدای قرن بسیار بهتر شده بود. آنقدر کارآمد شده بود که تصمیمات اتخاذ شده توسط سیستم امتیاز اعتماد دیگر قابل بحث نبود. یک نظام خطاپذیر. او با نالمیدی خندهید. نظام معصوم. این سیستم به ندرت شکست خورده است. اما شکست خورد. او باید یکی از آن موارد خاص باشد. خطای سیستم؛ از این به بعد یک طرد شده. هیچ کس جرات نداشت سیستم را زیر سوال ببرد. آنقدر در دولت، در خود جامعه ادغام شده بود که نمی‌توان آن را زیر سوال برد. در معدود کشورهای دموکراتیک باقیمانده، تشکیل جنبش‌های ضد دموکراتیک ممنوع بود، نه به این دلیل که ذاتاً بدخواهانه بودند، بلکه به این دلیل که سیستم فعلی را بی ثبات می‌کردند. همین منطق در مورد الگوکراسی‌های رایج‌تر هم اعمال می‌شود. نقد در الگوریتم‌ها به دلیل خطر برای وضع موجود ممنوع بود.

اعتماد الگوریتمی تار و پود نظم اجتماعی بود. برای منافع عمومی، اشتباهات نادر به طور ضمنی پذیرفته شد. صدها سیستم پیش‌بینی و پایگاه داده دیگر به امتیاز وارد شده‌اند و نمی‌توان مشخص کرد چه چیزی باعث افت امتیاز او شده است. او احساس می‌کرد که یک سوراخ تاریک بزرگ در ذهن او باز شده است. با وحشت به فضای خالی نگاه کرد.

سیستم وابستگی مالیاتی او در نهایت در سیستم امتیاز اعتماد مدنی ادغام شد، اما او هرگز با آن را نشناخت.

گیره‌های فرمی

۶۱۲ سال پس از استقرار مریخ: موزه‌ای در مریخ



زولا با دوستش زمزمه کرد: «تاریخ کسل کننده است». زولا، دختری با موهای آبی، با تنبلی یکی از پهپادهای پروژکتوری را که در اتاق زمزمه می‌کرد، با دست چپش تعقیب می‌کرد. معلم با صدایی ناراحت گفت: "تاریخ مهم است." زولا سرخ شد. او انتظار نداشت معلمش او را بشنود.

معلم از او پرسید: "زولا، چه چیزی یاد گرفتی؟" او با دقت گفت: "اینکه مردم باستان تمام منابع سیاره زمین را مصرف کردند و سپس مردند؟". دختر دیگری به نام لین افزواد: "نه. کسانی که آب و هوا را گرم کردند، مردم نبودند. کامپیوتر و ماشین بودند. و این سیاره زمین است، نه سیاره زمین". زولا سری به تایید تکان داد. معلم با احساس غرور لبخندی زد و سری تکان داد. «حق با هر دوی شماست. میدونی چرا اینطوری شد؟" زولا پرسید: "چون مردم کوته فکر و حریص بودند؟" لین گفت: "مردم نمی‌توانند ماشین‌های خود را متوقف کنند!".

معلم گفت: «باز هم، هر دوی شما درست می‌گویید، اما موضوع بسیار پیچیده‌تر از این است. بیشتر مردم در آن زمان از آنچه در حال رخ دادن بود آگاه نبودند. برخی تغییرات شدید را دیدند، اما نتوانستند آنها را معکوس کنند. مشهورترین قطعه از این دوره شعری از نویسنده ناشناس است که به بهترین شکل آنچه را که در آن زمان اتفاق افتاد به تصویر می‌کشد. با دقت گوش کنید!»

معلم شعر را شروع کرد. تعداد زیادی از پهپادهای کوچک در مقابل کودکان قرار گرفتند و شروع به پخش ویدیویی در چشمان آنها کردند. ویدیو، فردی را با کت و شلوار نشان می‌داد که در جنگلی ایستاده بود که در آن فقط کنده‌های درخت باقی مانده بود. شروع کرد به صحبت کردن:

ماشین‌ها محاسبه می‌کنند. ماشین‌ها پیش‌بینی می‌کنند
جوری که بخشی از آن هستیم آن‌ها را دنبال می‌کنیم.
ما به عنوان یک آموزش‌دیده به دنبال یک بهینه هستیم.
بهینه یک بعدی، محلی و بدون محدودیت است.

سیلیکون و گوشت، در تعقیب تصاعدی.
رشد ذهنیت ماست.

وقتی همه جوایز جمع‌آوری شد،
و عوارض جانبی نادیده گرفته شد.
وقتی تمام سکه‌ها استخراج می‌شوند،
و طبیعت عقب افتاده است.

دچار مشکل خواهیم شد،
به هر حال، رشد تصاعدی یک حباب است.
تراژدی عوام آشکار می‌شود،

منفجر شدن،
جلوی چشمان ما
محاسبات سرد و طمع سرد،
زمین را از گرما پر کنید.
همه چیز در حال مرگ است،
و ما رعایت می کنیم.

ما مانند اسبهایی با چشم بند در مسابقه خلقت خود مسابقه می دهیم،
به سوی فیلتر بزرگ تمدن.

و بنابراین ما بی امان تعقیب می کنیم.
همان طور که ما بخشی از ماشین هستیم.
آنتروپی را در بر می گیرد.

علم برای شکستن سکوت اتاق گفت: "یک خاطره تاریک. در کتابخانه شما آپلود خواهد شد. تکلیف شما این است که آن را تا هفته آینده آن را حفظ کنید." زولا آهی کشید. او موفق شد یکی از پهپادهای کوچک را بگیرد. پهپاد از CPU و موتورها گرم بود. زولا دوست داشت چون دستان او را گرم می کرد.

۲.۲ یادگیری ماشین چیست؟

یادگیری ماشین مجموعه‌ای از روش‌هایی است که رایانه‌ها برای انجام و بهبود پیش‌بینی‌ها یا رفتارها، بر اساس داده‌ها، از آن‌ها استفاده می‌کنند.

برای مثال، برای پیش‌بینی ارزش یک خانه، کامپیوتر، الگوهایی از فروش خانه‌های قبلی، یاد می‌گیرد. این کتاب بر یادگیری ماشین نظارت شده^۱ تمرکز دارد، که همه مسائل پیش‌بینی را پوشش می‌دهد. در این نوع، مجموعه داده‌ای داریم که در آن نتیجه^۲ مورد نظر را می‌دانیم (مثلاً قیمت‌های قبلی خانه) و می‌خواهیم یاد بگیریم که چگونه نتیجه را برای داده‌های جدید پیش‌بینی کنیم. وظایف^۳ خوبه‌بندی^۴ (= یادگیری بدون نظارت^۵) که در آن ما یک نتیجه خاص نداریم، اما می‌خواهیم خوبه‌هایی از نقاط داده را پیدا کنیم، جزء یادگیری تحت نظارت قرار نمی‌گیرد. همچنین مواردی مانند یادگیری تقویتی^۶، که در آن یک عامل^۷ یاد می‌گیرد با انجام دادن یک عمل در محیط پاداش خاصی را بهینه کند (مثلاً رایانه‌ای که Tetris بازی می‌کند)، مستثنی می‌شوند. هدف یادگیری تحت نظارت، یادگیری یک مدل پیش‌بینی است که ویژگی‌های داده‌ها (به عنوان مثال اندازه خانه، مکان، نوع طبقه، و ...) را به یک خروجی (مثلاً قیمت خانه) نگاشت می‌کند. اگر خروجی طبقه‌ای^۸ باشد، کار را طبقه‌بندی و اگر عددی باشد، رگرسیون نامیده می‌شود. الگوریتم یادگیری ماشین یک مدل را با تخمین^۹ پارامترها (مانند وزن‌ها) یا ساختارهای یادگیری (مانند درختان) یاد می‌گیرد. الگوریتم توسط یک امتیاز یا تابع خطا^{۱۰} کنترل می‌شود تا بتواند آن را به حداقل برساند. در مثال ارزش خانه، یادگیری ماشین، تفاوت بین قیمت تخمینی خانه و قیمت پیش‌بینی شده را به حداقل می‌رساند. سپس می‌توان از یک مدل یادگیری ماشین کاملاً آموزش دیده برای پیش‌بینی نمونه‌های^{۱۱} جدید استفاده کرد.

تخمین قیمت خانه، پیشنهادهای محصول، تشخیص تابلوهای خیابان، پیش‌بینی اولیه اعتبار و تشخیص تقلب: همه این مثال‌ها، وجه اشتراکی دارند. این وجه مشترک آن است که می‌توان آن‌ها را با یادگیری ماشین حل کرد. وظایف متفاوت است، اما رویکرد یکسان است:

¹ supervised machine learning / supervised machine learning

² Outcome

³ tasks

⁴ Clustering

⁵ unsupervised learning

⁶ reinforcement learning

⁷ agent

⁸ categorical

⁹ estimating

¹⁰ loss function

¹¹ instances

مرحله ۱: جمع‌آوری داده‌ها^۱. هرچقدر بیشتر بهتر. داده‌ها باید حاوی نتیجه‌ای باشد که می‌خواهید پیش‌بینی کنید و اطلاعات اضافی که با استفاده از آن می‌توان پیش‌بینی کرد. برای آشکارساز علائم خیابان ("آیا تابلوی خیابان در تصویر وجود دارد؟")، تصاویر خیابان را جمع‌آوری می‌کنید و برچسب می‌زنید که آیا تابلوی خیابان قابل مشاهده است یا خیر. برای یک پیش‌بینی‌کننده اولیه اعتبار، به داده‌های گذشته در مورد وام‌های واقعی، اطلاعاتی در مورد اینکه آیا مشتریان وام‌های خود را نکول کرده‌اند، و داده‌هایی که به شما در انجام پیش‌بینی‌ها کمک می‌کنند، مانند درآمد، اعتبارات گذشته اولیه و غیره نیاز دارد. برای یک برنامه تخمين زن خودکار ارزش خانه، می‌توانید داده‌ها را از فروش‌های قبلی خانه و اطلاعات مربوط به املاک مانند اندازه، مکان و غیره جمع‌آوری کنید.

مرحله ۲: این اطلاعات را در یک الگوریتم یادگیری ماشین وارد کنید که یک مدل آشکارساز علامت، یک مدل رتبه‌بندی اعتبار یا یک تخمين گر ارزش خانه ایجاد می‌کند.

مرحله ۳: از مدل با داده‌های جدید استفاده کنید. مدل را در یک محصول یا فرآیند ادغام کنید، مانند ماشین خودران، فرآیند درخواست اعتبار یا وب سایت بازار املاک.

ماشین‌ها در بسیاری از کارها، مانند بازی شطرنج (یا اخیراً Go) یا پیش‌بینی آب و هوا از انسان‌ها پیشی می‌گیرند. حتی اگر ماشین به خوبی یک انسان باشد یا در یک وظیفه کمی بدتر باشد، مزایای زیادی از نظر سرعت، تکرارپذیری و مقیاس‌پذیری وجود دارد. یک مدل یادگیری ماشین پیاده‌سازی شده، می‌تواند یک کار را بسیار سریع‌تر از انسان‌ها انجام دهد، نتایج قابل اعتمادی را ارائه می‌دهد و می‌تواند بی‌نهایت بار کپی شود. تکرار یک مدل یادگیری ماشین در ماشین دیگر سریع و ارزان است. آموزش یک انسان برای انجام یک وظیفه می‌تواند چندین دهه طول بکشد (مخصوصاً در جوانی) و بسیار پرهزینه است. یک عیب عمده استفاده از یادگیری ماشین این است که بینش در مورد داده‌ها و وظیفه‌ای که ماشین حل می‌کند در مدل‌های پیچیده، پنهان است. برای توصیف یک شبکه عصبی عمیق^۲ به میلیون‌ها عدد نیاز دارد، و هیچ راهی برای درک کامل مدل وجود ندارد. مدل‌های دیگر، مانند جنگل تصادفی^۳، از صدھا درخت تصمیم تشکیل شده‌اند که به پیش‌بینی‌ها رأی^۴ می‌دهند. برای درک چگونگی تصمیم گیری، باید به آرا و ساختار هر یک از صدھا درخت نگاه کنید. این کار، صرفنظر از اینکه چقدر باهوش هستید یا حافظه شما چقدر خوب کار می‌کند، ممکن نیست. بهترین مدل‌ها اغلب ترکیبی از چندین مدل (ممولاً ترکیبی^۵ نامیده می‌شوند) هستند که قابل تفسیر نیستند، حتی اگر هر مدل منفرد قابل تفسیر باشد. اگر فقط بر روی بهبود عملکرد تمرکز کنید، به طور خودکار مدل‌های غیرشفاف‌تری^۶ خواهید داشت. مدل‌های برنده در

¹ Data collection

² Deep neural network

³ Random forest

⁴ Vote

⁵ Ensemble

⁶ Opaque

مسابقات یادگیری ماشین اغلب ترکیبی از مدل‌های بسیار پیچیده مانند درختان تقویت شده^۱ یا شبکه‌های عصبی عمیق هستند.

¹ Boosted trees

۲.۳ اصطلاحات

برای جلوگیری از سردرگمی به دلیل ابهام، در اینجا چند تعریف از اصطلاحات استفاده شده در این کتاب آورده شده است:

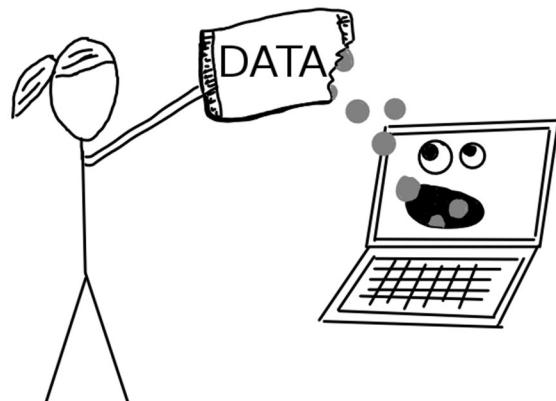
الگوریتم^۱ مجموعه‌ای از قوانین است که یک ماشین برای رسیدن به یک منظور خاص از آنها پیروی می‌کند (*Definition of Algorithm.*, 2017). یک الگوریتم را می‌توان به عنوان دستور العملی در نظر گرفت که ورودی‌ها، خروجی‌ها و تمام مراحل موردنیاز برای رسیدن از ورودی‌ها به خروجی را تعریف می‌کند. دستور العمل‌های آشپزی، الگوریتم‌هایی هستند در آن‌ها مواد اولیه آشپزی ورودی، غذای پخته شده خروجی و مراحل آماده سازی و پخت دستور العمل‌های الگوریتم هستند.

یادگیری ماشین^۲ مجموعه‌ای از روش‌هایی است که به رایانه‌ها اجازه می‌دهد از داده‌ها برای انجام و بهبود پیش‌بینی‌ها (مثلًاً تشخیص سرطان، فروش هفتگی، اعتبار اولیه) یاد بگیرند. یادگیری ماشین یک تغییر پارادایم از «برنامه‌نویسی عادی» است که در آن تمام دستور العمل‌ها باید به صراحت به رایانه داده شود، به «برنامه‌نویسی غیرمستقیم» که داده‌ها به رایانه ارائه می‌شود.

Without Machine Learning



With Machine Learning



یادگیرنده^۳ یا الگوریتم یادگیری ماشین^۴ برنامه‌ای است که برای یادگیری مدل یادگیری ماشین از داده‌ها استفاده می‌شود. نام دیگر "القاگر"^۵ است (به عنوان مثال "القاگر درخت").

¹ Algorithm

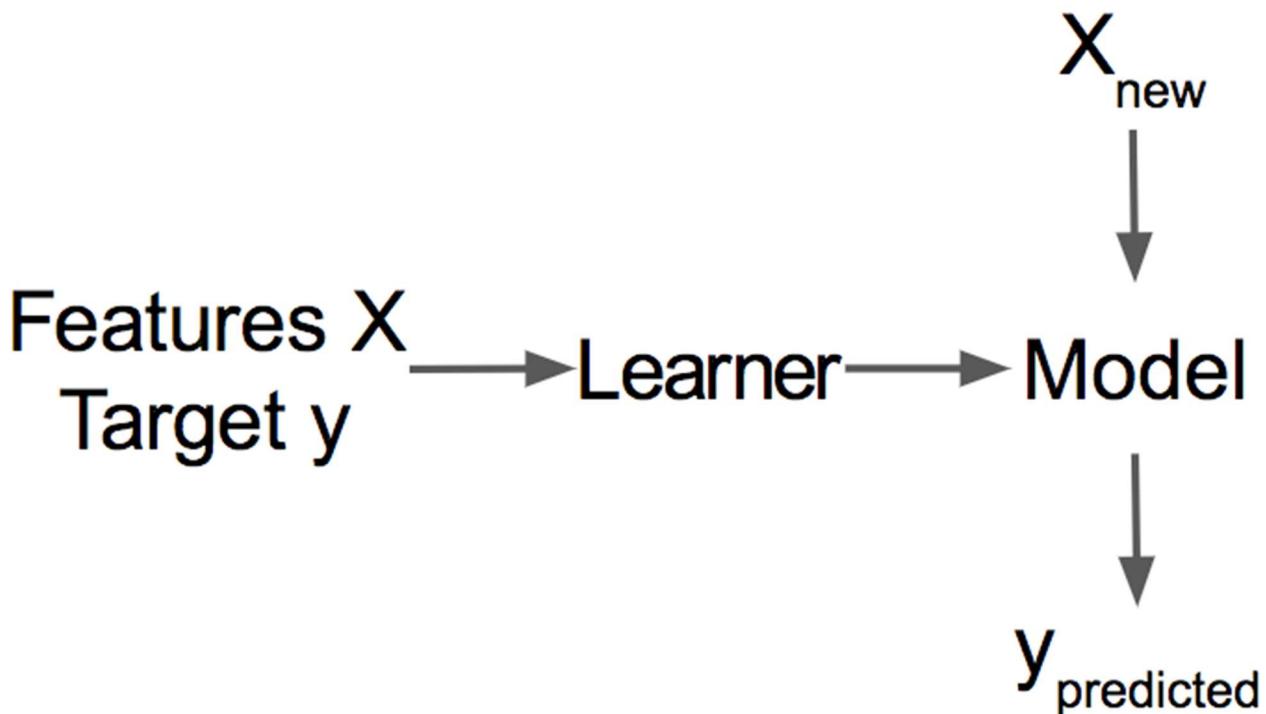
² Machine Learning

³ Learner

⁴ Machine Learning Algorithm

⁵ Inducer

مدل یادگیری ماشین^۱ برنامه‌ای است که ورودی‌ها را به پیش‌بینی‌ها نگاشت می‌کند. این مدل می‌تواند مجموعه‌ای از وزن‌ها برای یک مدل خطی یا شبکه عصبی باشد. نام‌های دیگر کلمه "مدل"، "پیش‌بینی‌کننده"^۲ یا - بسته به کار - "طبقه‌بند"^۳ یا "مدل رگرسیون"^۴ است. در فرمول‌ها، مدل یادگیری ماشین آموزش‌دیده \hat{f} یا $\hat{f}(x)$ نامیده می‌شود.



شکل ۱: یک یادگیرنده، مدلی را از داده‌های آموزشی برچسب گذاری شده یاد می‌گیرد. این مدل برای پیش‌بینی استفاده می‌شود.

مدل جعبه سیاه^۵ سیستمی است که مکانیسم‌های داخلی خود را آشکار نمی‌کند. در یادگیری ماشین، "جعبه سیاه" به مدل‌هایی اطلاق می‌شود که نمی‌توان با نگاه‌کردن به پارامترهای آنها، درکشان کرد (مثلاً یک شبکه عصبی). نقطه مقابل جعبه سیاه گاهی اوقات به عنوان جعبه سفید^۶ شناخته می‌شود که در این کتاب به عنوان مدل قابل تفسیر از آن یاد می‌شود. روش‌های مدل‌آگنوستیک برای تفسیرپذیری، مدل‌های یادگیری ماشین را به عنوان جعبه‌های سیاه در نظر می‌گیرند، حتی اگر چنین نباشند.

¹ Machine Learning Model

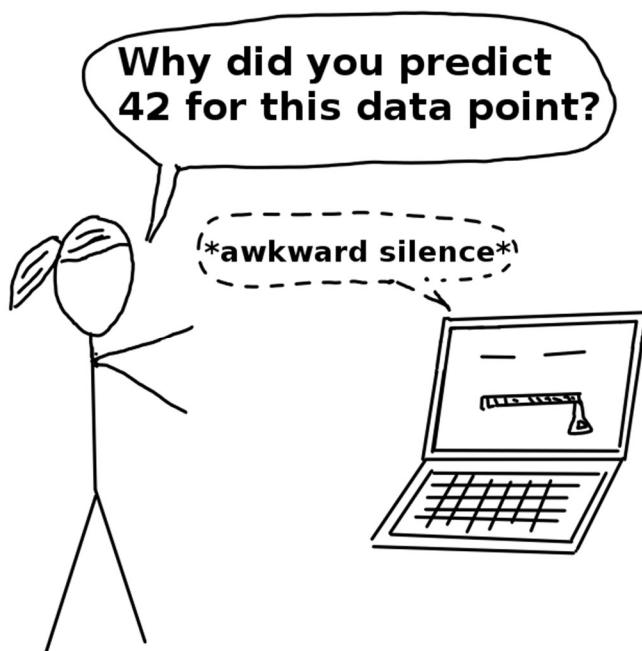
² Predictor

³ Classifier

⁴ Regression model

⁵ Black box model

⁶ White box



یادگیری ماشین قابل تفسیر^۱ به روش‌ها و مدل‌های اشاره دارد که رفتار و پیش‌بینی‌های سیستم‌های یادگیری ماشین را برای انسان قابل درک می‌کنند.

مجموعه‌داده^۲، جدولی از داده‌هاست که ماشین از آن یاد می‌گیرد. مجموعه‌داده شامل ویژگی‌ها و هدف پیش‌بینی است. مجموعه‌داده، هنگامی که برای آموزش یک مدل استفاده می‌شود، داده آموزشی^۳ نامیده می‌شود. نمونه^۴، یک ردیف در مجموعه‌داده است. نامهای دیگر «نمونه» عبارت‌اند از: (داده) نقطه^۵، مثال^۶، مشاهده^۷. یک نمونه از مقادیر ویژگی⁽ⁱ⁾ x و مقدار معلوم y_i تشکیل شده است.

ویژگی‌ها^۸، ورودی‌هایی هستند که برای پیش‌بینی یا طبقه‌بندی استفاده می‌شوند. یک ویژگی یک ستون در مجموعه‌داده است. در سرتاسر کتاب، ویژگی‌ها قابل تفسیر فرض می‌شوند، به این معنی که درک معنای آنها آسان است، مانند دمای یک روز معین یا قد یک فرد. تفسیرپذیری ویژگی‌ها یک فرض بزرگ است. اگر درک ویژگی‌های ورودی سخت باشد، درک اینکه مدل چه کاری انجام می‌دهد بسیار سخت‌تر است. ماتریس تمام ویژگی‌ها X نامیده می‌شود و $x^{(i)}$ برای یک نمونه. بردار یک ویژگی واحد برای همه نمونه‌ها x_j است و مقدار ویژگی⁽ⁱ⁾ x_j برای نمونه i می‌باشد.

¹ Interpretable Machine Learning

² Dataset

³ Training

⁴ Instance

⁵ Point

⁶ Example

⁷ Observation

⁸ Features

هدف^۱، اطلاعاتی است که ماشین یاد می‌گیرد تا آن را پیش‌بینی کند. در فرمول‌های ریاضی معمولاً هدف y یا y نامیده می‌شود.

وظیفه یادگیری ماشین^۲ ترکیب مجموعه ویژگی‌ها با یک هدف است. بسته به نوع هدف، وظیفه می‌تواند به عنوان مثال طبقه‌بندی، رگرسیون، تجزیه و تحلیل بقا^۳، خوش‌بندی^۴، یا تشخیص داده پرت^۵ باشد. پیش‌بینی^۶ مقدار هدفی است که مدل یادگیری ماشین بر اساس ویژگی‌های داده شده، «حدس می‌زند^۷». در این کتاب، پیش‌بینی مدل با $\hat{f}(x^{(i)})$ یا \hat{y} نشان داده شده است.

¹ Target

² Machine Learning Task

³ survival analysis

⁴ clustering

⁵ outlier detection

⁶ Prediction

⁷ guesses

فصل ۳ تفسیرپذیری

تعریف تفسیرپذیری (از نظر ریاضی) دشوار است. من تعریف (غیر ریاضی) تفسیرپذیری که توسط Miller (2019) ارائه شده است را دوست دارم. این تعریف عبارت این‌گونه است: تفسیرپذیری درجه‌ای است که یک انسان می‌تواند علت یک تصمیم را درک کند. یکی تعریف دیگر این است: تفسیرپذیری درجه‌ای است که یک انسان می‌تواند به طور مداوم نتیجه مدل را پیش‌بینی کند (Kim et al., 2016). هرچه قابلیت تفسیر یک مدل یادگیری ماشین بالاتر باشد، درک اینکه چرا تصمیم‌ها یا پیش‌بینی‌های خاصی گرفته شده‌اند، برای فردی آسان‌تر است. یک مدل از مدل دیگر قابل تفسیرتر است اگر درک تصمیمات آن برای انسان آسان‌تر از تصمیمات مدل دیگر باشد. من از هر دو اصطلاح قابل تفسیر^۱ و قابل توضیح^۲ به جای یکدیگر استفاده خواهم کرد. مانند Miller (2019)، من فکر می‌کنم منطقی است که بین اصطلاحات تفسیرپذیری / توضیح پذیری و توضیح تفاوت قائل شویم. من از "توضیح"^۳ برای توضیح پیش‌بینی‌های فردی استفاده خواهم کرد. برای یادگیری آنچه که ما انسان‌ها به عنوان یک توضیح خوب می‌دانیم، بخش توضیحات انسان پسند را مطالعه فرمایید.

یادگیری ماشین قابل تفسیر یک اصطلاح مفید است که "استخراج دانش از یک مدل یادگیری ماشین درباره روابط موجود در داده‌ها یا یادگیری شده توسط مدل" را نشان می‌دهد (Murdoch et al., 2019).

¹ Interpretable

² Explainable

³ Explanation

۳. اهمیت تفسیرپذیری

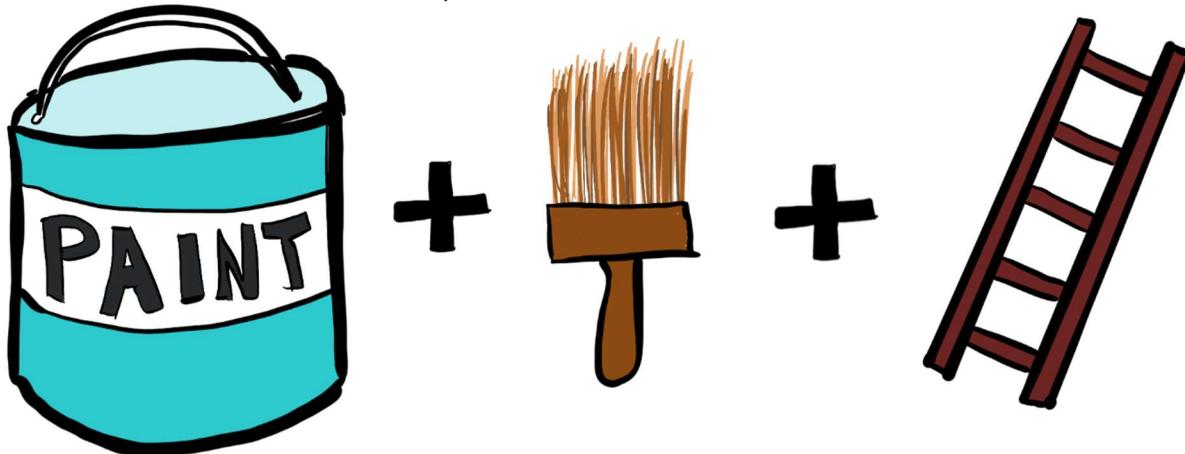
اگر یک مدل یادگیری ماشین عملکرد خوبی دارد، چرا فقط به مدل اعتماد نمی‌کنیم و از چرایی تصمیم خاصی صرفنظر نمی‌کنیم؟ مشکل این است که یک معیار واحد، مانند دقت طبقه‌بندی، توصیف ناقصی از اکثر وظایف دنیای واقعی است (Doshi-Velez & Kim, 2017).

اجازه دهید به دلایلی که چرا تفسیرپذیری بسیار مهم است، عمیق‌تر بپردازیم. وقتی نوبت مدل‌سازی پیش‌بینی فرا می‌رسد، باید یک مبادله انجام دهید: آیا فقط می‌خواهید بدانید چه چیزی پیش‌بینی می‌شود؟ به عنوان مثال، احتمال اینکه یک مشتری سرگردان شود یا اینکه برخی از داروها چقدر برای بیمار موثر است. یا می‌خواهید بدانید چرا پیش‌بینی انجام شد و احتمالاً برای تفسیرپذیری بهتر، چقدر کاهش عملکرد پیش‌بینی‌کننده را تحمل می‌کنید؟ در برخی موارد، برای شما مهم نیست که چرا تصمیم گرفته شده است، کافی است بدانید که عملکرد پیش‌بینی‌کننده روی مجموعه‌داده آزمایشی خوب بوده است. اما در موارد دیگر، دانستن «چرایی» می‌تواند به شما کمک کند تا درباره مساله، داده‌ها و دلیل شکست یک مدل بیشتر بدانید. برخی از مدل‌ها ممکن است نیازی به توضیح نداشته باشند زیرا در یک محیط کم خطر استفاده می‌شوند، به این معنی که یک اشتباه عاقب جدی در پی نخواهد داشت (مثلاً سیستم پیشنهاد دهنده فیلم) یا روشی که قبلاً به طور گسترده مورد مطالعه و ارزیابی قرار گرفته است (مثلاً تشخیص کاراکتر نوری). نیاز به تفسیرپذیری از نقصان در فرمول بندی مسئله ناشی می‌شود (Doshi-Velez & Kim, 2017)، به این معنی که برای مشکلات یا وظایف خاص، پیش‌بینی (چه چیزی) کافی نیست. مدل علاوه بر پیش‌بینی، باید توضیح دهد که چگونه به پیش‌بینی رسید (چرا)، زیرا یک پیش‌بینی صحیح فقط تا حدی مشکل اصلی شما را حل می‌کند. دلایل زیر باعث نیاز برای تفسیرپذیری و توضیح می‌شود (Doshi-Velez & Kim, 2017).

کنجکاوی و یادگیری انسان: انسان‌ها یک مدل ذهنی از محیط خود دارند که زمانی که اتفاق غیرمنتظره ای رخ می‌دهد به روز می‌شود. این به روز رسانی با یافتن توضیحی برای رویداد غیرمنتظره انجام می‌شود. به عنوان مثال، یک انسان به طور غیرمنتظره ای احساس بیماری می‌کند و می‌پرسد: "چرا اینقدر احساس بیماری می‌کنم؟". او یاد می‌گیرد که هر بار که توت قرمز را می‌خورد بیمار می‌شود. او مدل ذهنی خود را به روز می‌کند و به این نتیجه می‌رسد که توت‌ها باعث بیماری شده‌اند و بنابراین باید از آن‌ها اجتناب شود. هنگامی که از مدل‌های یادگیری ماشین غیرشفاف در تحقیقات استفاده می‌شود، اگر مدل فقط پیش‌بینی‌هایی را بدون توضیح ارائه دهد، یافته‌های علمی کاملاً پنهان می‌مانند. برای تسهیل یادگیری و ارضای کنجکاوی در مورد اینکه چرا برخی پیش‌بینی‌ها یا رفتارها توسط ماشین‌ها ایجاد می‌شوند، تفسیرپذیری و توضیحات بسیار مهم است. البته انسان‌ها برای هر اتفاقی که می‌افتد نیازی به توضیح ندارند. برای اکثر مردم اشکالی ندارد که ندانند کامپیوتر چگونه کار می‌کند. اتفاقات غیرمنتظره ما را کنجکاو می‌کند. به عنوان مثال: چرا کامپیوتر من به طور غیرمنتظره ای خاموش می‌شود؟

میل انسان به یافتن معنا در جهان ارتباط نزدیک با یادگیری دارد. ما می‌خواهیم تضادها یا ناسازگاری‌ها را با عناصر ساختارهای دانش خود هماهنگ کنیم. "چرا سگم مرا گاز گرفت با توجه به اینکه قبلًاً این کار را نکرده بود؟" ممکن است یک انسان بپرسد بین آگاهی از رفتار گذشته سگ و تجربه ناخوشایند گاز تازه گرفته شده تناقض وجود دارد. توضیحات دامپزشک تضاد صاحب سگ را برطرف می‌کند: "سگ تحت استرس بود و گاز گرفته بود." هر چه تصمیم یک ماشین بیشتر بر زندگی یک فرد تأثیر بگذارد، توضیح رفتار ماشین اهمیت بیشتری دارد. فرض کنید یک مدل یادگیری ماشین به صورت غیرمنتظره درخواست وام متقاضیان را رد کند. این ناسازگاری بین انتظار متقاضیان و واقعیت را فقط با نوعی توضیح می‌توان آشتبانی داد. در عمل، توضیحات نباید شرایط را به طور کامل توضیح دهند، بلکه باید به یک علت اصلی بپردازنند. مثال دیگر پیشنهاد محصول الگوریتمی است. من شخصاً همیشه به این فکر می‌کنم که چرا محصولات یا فیلم‌های خاصی به صورت الگوریتمی به من توصیه شده‌اند. اغلب کاملاً واضح است: تبلیغات، من را در اینترنت دنبال می‌کنند. چون اخیراً یک ماشین لباسشویی خریدم، می‌دانم که در روزهای آینده تبلیغات ماشین لباسشویی دریافت خواهم کرد. بله، اگر در خرید قبل کلاه زمستانی در سبد خریدم وجود داشته باشد، پیشنهاد دستکش منطقی است. الگوریتم این فیلم را پیشنهاد می‌کند، زیرا کاربرانی که فیلمی که من دوست داشتم، را دوست داشته‌اند، از فیلم پیشنهادی لذت برده‌اند. شرکت‌های اینترنتی به طور فزاینده‌ای توضیحاتی را به توصیه‌های خود اضافه می‌کنند. یک مثال خوب، توصیه‌های محصول بر اساس ترکیبات محصولاتی است که اغلب با یکدیگر خریداری می‌شوند:

Frequently Bought Together



شکل ۳.۱: محصولات پیشنهادی که اغلب با هم خریداری می‌شوند.
در بسیاری از رشته‌های علمی تغییری از روش‌های کیفی به کمی (به عنوان مثال جامعه شناسی، روانشناسی)، و همچنین به سمت یادگیری ماشین (زیست شناسی، ژنومیک) وجود دارد. هدف علم کسب دانش است، اما

بسیاری از مشکلات با مجموعه داده‌های بزرگ و مدل‌های یادگیری ماشین جعبه سیاه حل می‌شوند. گاهی خود مدل به جای داده، به منبع دانش تبدیل می‌شود. تفسیرپذیری امکان استخراج این دانش اضافی، که توسط مدل ایجاد شده است را ممکن می‌سازد.

مدل‌های یادگیری ماشین وظایف دنیای واقعی را انجام می‌دهند که به اقدامات ایمنی و تست نیاز دارند. تصور کنید یک ماشین خودران به طور خودکار دوچرخه سواران را بر اساس یک سیستم یادگیری عمیق، شناسایی می‌کند. شما می‌خواهید ۱۰۰٪ مطمئن باشید که انتزاعی که سیستم یادگرفته است بدون خطا است، زیرا زیزگرفتن دوچرخه سواران بسیار ناگوار است. یک توضیح ممکن است نشان دهد که مهم‌ترین ویژگی یادگرفته شده، تشخیص دو چرخ دوچرخه است، و این توضیح به شما کمک می‌کند در مورد، شرایطی که این چرخ‌ها را پوشش می‌دهند مانند دوچرخه با کیسه‌های جانبی که تا حدی چرخ‌ها را می‌پوشانند فکر کنید.

به‌طور پیش‌فرض، مدل‌های یادگیری ماشین، سوگیری‌ها را از داده‌های آموزشی دریافت می‌کنند. این می‌تواند مدل‌های یادگیری ماشین شما را به ماشین سوگیری تبدیل کند که علیه گروه‌های دارای نمایندگی کمتر، تبعیض قائل شوند. تفسیرپذیری یک ابزار اشکال زدایی مفید برای تشخیص سوگیری است. در مدل‌های یادگیری ماشین ممکن است این اتفاق بیفتد که مدل یادگیری ماشین که برای تأیید یا رد خودکار درخواست‌های اعتباری آموزش داده‌اید، علیه اقلیتی که از لحاظ تاریخی از حق امتیاز محروم شده‌اند تبعیض قائل شود. هدف اصلی شما اعطای وام فقط به افرادی است که در نهایت آن وام‌ها را بازپرداخت خواهند کرد. ناقص بودن فرمول مساله در این مورد، در این واقعیت نهفته است که شما نه تنها می‌خواهید نکول وام را به حداقل برسانید، بلکه موظف هستید بر اساس جمعیت‌شناسی خاص تبعیض قائل نشوید. این یک محدودیت اضافی است که بخشی از فرمول مساله شما است (اعطای وام به روشی کم خطر و قابل قبول) که توسط تابع خطایی که مدل یادگیری ماشین برای آن بهینه سازی شده است، در نظر گرفته نمی‌شود.

فرآیند ادغام ماشین‌ها و الگوریتم‌ها در زندگی روزمره ما نیازمند تفسیرپذیری برای افزایش پذیرش اجتماعی است. مردم، باورها، امیال، نیات و غیره را به اشیا نسبت می‌دهند. در یک آزمایش معروف، Heider and Simmel (1944) به شرکت کنندگان در آزمایش، فیلم‌هایی از اشکال نشان داد که در این فیلم‌ها، یک دایره، یک "در" را برای ورود به یک "اتاق" (که یک مستطیل ساده بود) باز می‌کرد. شرکت کنندگان اعمال شکل‌ها را همانند یک عامل انسانی توصیف می‌کردند. نیات و حتی احساسات و ویژگی‌های شخصیتی، به اشکال نسبت می‌دادند. ربات‌ها مثال خوبی هستند. من نام جاروبرقی ام را Doge گذاشته ام. اگر Doge گیر کند، فکر می‌کنم: "Doge" می‌خواهد به تمیز کردن ادامه دهد، اما از من کمک می‌خواهد چون گیر کرده است." وقتی Doge تمیز کردن را تمام می‌کند و پریز خانه را برای شارژ مجدد جستجو می‌کند، فکر می‌کنم: "Doge" می‌لیل به شارژ مجدد دارد و قصد دارد پریز خانه را پیدا کند." همچنین ویژگی‌های شخصیتی را به Doge نسبت می‌دهم: "Doge" کمی خنگ است، اما

زیباست." اینها افکار من است، به خصوص وقتی متوجه می‌شوم که Doge در حالی که خانه را با جاروبرقی، جارو می‌کشد، گیاهی را له کرده است. ماشین یا الگوریتمی که پیش‌بینی‌های خود را توضیح می‌دهد، مقبولیت بیشتری پیدا می‌کند. در بخش توضیحات انسان پسند را گفته می‌شود، که استدلال می‌کند که توضیحات، یک فرآیند اجتماعی هستند.

از توضیحات برای مدیریت تعاملات اجتماعی استفاده می‌شود. توضیح دهنده با ایجاد معنای مشترک از چیزی، بر اعمال، احساسات و باورهای گیرنده توضیح تأثیر می‌گذارد. برای اینکه یک ماشین با ما تعامل داشته باشد، ممکن است نیاز داشته باشد که احساسات و باورهای ما را شکل دهد. ماشین‌ها باید ما را مقاعده کنند تا بتوانند به هدف مورد نظر خود برسند. اگر ربات جاروبرقی، خود را تا حدی توضیح نمی‌داد، به طور کامل نمی‌پذیرفتش. جاروبرقی با توضیح اینکه گیر کرده است، به عنوان مثال، یک «حادثه» (مانند گیر کردن دوباره روی فرش حمام...) به جای توقف کار بدون هیچ توضیحی، معنای مشترکی ایجاد می‌کند. جالب اینجاست که ممکن است بین هدف ماشین توضیح دهنده (ایجاد اعتماد) و هدف گیرنده (درک پیش‌بینی یا رفتار) ناهماهنگی وجود داشته باشد. شاید توضیح کامل برای اینکه چرا Doge گیر کرده می‌تواند این باشد که شارژ باتری، بسیار کم است، یکی از چرخ‌ها به درستی کار نمی‌کند و یک اشکال وجود دارد که باعث می‌شود ربات، بارها و بارها به نقطه‌ای برود که مانع وجود دارد. این دلایل (و چند مورد دیگر) باعث شد ربات گیر کند، اما فقط توضیح داد که چیزی مانع است و همین برای من کافی بود تا به رفتار آن اعتماد کنم و معنای مشترک آن تصادف را دریافت کنم. به هر حال، Doge دوباره در حمام گیر کرد. هربار که قصدارم Doge را در حالت جاروبرقی بگذارم باید فرش را بردارم. اما فقط توضیح داد که مانع در راه است و همین برای من کافی بود تا به رفتار آن اعتماد کنم و معنای مشترکی از آن حادثه پیدا کنم.



شکل ۳.۲: Doge، جاروبرقی ما، گیر کرده است. به عنوان توضیحی برای تصادف، Doge به ما گفت که باید روی سطح صاف باشد.

مدل‌های یادگیری ماشین را فقط زمانی می‌توان اشکال زدایی و بازرسی کرد، که بتوان آنها را تفسیر نمود. حتی در محیط‌های کم خطر، مانند توصیه‌های فیلم، توانایی تفسیر، هم در مرحله تحقیق و توسعه و هم پس از توسعه، ارزشمند است. وقتی از یک مدل در یک محصول استفاده می‌شود، ممکن است همه چیز اشتباه شود. تفسیر یک پیش‌بینی اشتباه، به درک علت خطا کمک می‌کند. این موضوع یک جهت گیری برای نحوه تعمیر سیستم ارائه می‌دهد. به عنوان مثال، یک طبقه‌بندی کننده هاسکی از گرگ را در نظر بگیرید که برخی هاسکی‌ها را به اشتباه به عنوان گرگ طبقه‌بندی می‌کند. با استفاده از روش‌های یادگیری ماشین قابل تفسیر، متوجه می‌شوید که طبقه‌بندی اشتباه به دلیل برف در تصویر است. طبقه‌بندی کننده یاد گرفته است که از برف به عنوان ویژگی برای یافتن تصاویر گرگ استفاده کند. استفاده از این ویژگی ممکن است برای تفکیک گرگ‌ها از هاسکی در مجموعه داده‌های آموزشی منطقی باشد، اما در دنیای واقعی این چنین نیست.

اگر می‌توانید اطمینان حاصل کنید که مدل یادگیری ماشین می‌تواند تصمیماتش را توضیح دهد، می‌توانید موارد زیر را نیز راحت‌تر بررسی کنید (Doshi-Velez & Kim, 2017):

انصارف^۱: حصول اطمینان از اینکه پیش‌بینی‌ها بی‌طرفانه هستند و به طور ضمنی یا صریح علیه گروه‌های مهجور تبعیض قائل نمی‌شوند. یک مدل قابل تفسیر می‌تواند به شما بگوید که چرا تصمیم گرفته است که

^۱ Fairness

یک فرد خاص نباید وام دریافت کند، و قضاوت در مورد اینکه آیا این تصمیم بر اساس یک سوگیری جمعیت شناختی (مثلاً نژادی) است یا خیر، برای یک انسان آسان‌تر می‌شود.

حریم خصوصی^۱: اطمینان از این که از اطلاعات حساس در داده‌ها محافظت می‌شود.
قابلیت اطمینان^۲ یا استحکام^۳: اطمینان از اینکه تغییرات کوچک در ورودی منجر به تغییرات بزرگ در پیش‌بینی نمی‌شود.

علیت^۴: بررسی این موضوع که فقط روابط علی انتخاب شده باشند.
اعتماد^۵: در مقایسه با جعبه سیاه، اعتماد به سیستمی که تصمیمات خود را توضیح می‌دهد برای انسان آسان‌تر است.

زمانی که نیازی به تفسیرپذیری نداریم.

سناریوهای زیر نشان می‌دهند که چه زمانی نیازی به تفسیرپذیری مدل‌های یادگیری ماشین نداریم یا حتی نمی‌خواهیم.

اگر مدل تأثیر قابل توجهی نداشته باشد، نیازی به تفسیرپذیری نیست. تصور کنید شخصی به نام مایک روی یک پروژه جانبی یادگیری ماشین کار می‌کند تا بر اساس داده‌های فیس بوک پیش‌بینی کند که دوستانش برای تعطیلات بعدی خود کجا خواهند رفت. مایک فقط دوست دارد دوستانش را با حدسهای صحیح که آنها در تعطیلات کجا خواهند رفت، غافلگیر کند. اگر مدل اشتباه باشد مشکلی جدی به وجود نمی‌آید (در بدترین حالت فقط کمی باعث خجالت مایک است) و اگر مایک نتواند خروجی مدل خود را توضیح دهد نیز مشکلی وجود ندارد. اگر مایک شروع به ایجاد یک کسب و کار در مورد پیش‌بینی‌های مقصد تعطیلات کند، وضعیت تغییر می‌کند. اگر مدل اشتباه باشد، کسبوکار ممکن است ضرر کند، یا این که مدل ممکن است به دلیل تعصبات نژادی یادگیری‌شده برای برخی افراد بدتر عمل کند. به محض اینکه مدل تأثیر قابل توجهی، خواه مالی یا اجتماعی داشته باشد، قابلیت تفسیر، مهم می‌شود.

وقتی مسئله به خوبی مطالعه شده باشد، نیازی به تفسیرپذیری نیست. برخی از کاربردها به اندازه کافی مورد مطالعه قرار گرفته‌اند، به طوری که تجربه عملی خوبی از مدل وجود دارد و مشکلات مدل در طول زمان حل شده است. یک مثال خوب یک مدل یادگیری ماشین برای تشخیص کاراکترهای نوری است که تصاویر را از پاکتهای

¹ Privacy

² Reliability

³ Robustness

⁴ Causality

⁵ Trust

نامه پردازش می کند و آدرس ها را استخراج می کند. سال ها تجربه با این سیستمها وجود دارد و مشخص است که خوب کار می کنند. علاوه بر این، ما واقعاً علاقه ای به کسب بینش اضافی نداریم.

تفسیرپذیری ممکن است افراد یا برنامه ها را قادر به دستکاری سیستم کنند. مشکلات با کاربرانی که یک سیستم را فریب می دهند ناشی از عدم تطابق بین اهداف سازنده و کاربر یک مدل است. امتیازدهی اعتباری چنین سیستمی است زیرا بانک ها می خواهند اطمینان حاصل کنند که وامها فقط به متقارضیانی داده می شود که احتمالاً آن ها را پس می دهند و متقارضیان قصد دارند وام را دریافت کنند حتی اگر بانک نخواهد به آنها وام بدهد. این عدم تطابق بین اهداف، انگیزه هایی را برای متقارضیان ایجاد می کند تا با سیستم بازی کنند تا شанс خود را برای دریافت وام افزایش دهند. اگر متقارضی بدانند که داشتن بیش از دو کارت اعتباری بر امتیاز او تأثیر منفی می گذارد، به سادگی سومین کارت اعتباری خود را برای بهبود امتیاز خود باطل می کند و پس از تأیید وام، کارت جدیدی اخذ می کند. در حالی که امتیاز او بهبود یافت، احتمال واقعی بازپرداخت وام بدون تغییر باقی ماند. سیستم را تنها در صورتی می توان بازی داد که ورودی ها یک ویژگی علی، وکالتی^۱ باشند و این ویژگی وکالتی تاثیر علی واقعی ندارد. در صورت امکان، باید از ویژگی های وکالتی اجتناب شود، زیرا آنها مدل ها را قابل بازی می کنند. به عنوان مثال، گوگل را سیستمی به نام Google Flu Trends برای پیش بینی شیوع آنفولانزا ایجاد کرد. این سیستم جستجوهای گوگل را با شیوع آنفولانزا همبسته می کرد و عملکرد ضعیفی داشته است. توزیع پرس و جوهای جستجو تغییر کرد و Google Flu Trends بسیاری از شیوع آنفولانزا را از دست داد. جستجوی گوگل باعث آنفولانزا نمی شود. هنگامی که افراد علائمی مانند "تب" را جستجو می کنند، صرفاً این جستجو یک همبستگی با شیوع واقعی آنفولانزا دارد. به صورت ایده آل، مدل فقط باید از ویژگی های علی استفاده کند که قابلیت بازی کردن با آن ها وجود نداشته باشد.

¹ Proxy

۳.۲ طبقه‌بندی روش‌های تفسیرپذیری

روش‌های تفسیرپذیری یادگیری ماشین را می‌توان بر اساس معیارهای مختلف طبقه‌بندی کرد. ذاتی یا تعقیبی؟ این معیار تشخیص می‌دهد که آیا تفسیرپذیری با محدود کردن پیچیدگی مدل یادگیری ماشین (Intrinsic) یا با استفاده از روش‌هایی که مدل را پس از آموزش تجزیه و تحلیل می‌کند (post hoc) به دست می‌آید. تفسیرپذیری ذاتی به مدل‌های یادگیری ماشین اشاره دارد که به دلیل ساختار ساده‌شان قابل تفسیر در نظر گرفته می‌شوند، مانند درخت‌های تصمیم کوتاه یا مدل‌های خطی پراکنده. تفسیرپذیری تعقیبی به کاربرد روش‌های تفسیری پس از آموزش مدل اشاره دارد. اهمیت ویژگی جایگشت، به عنوان مثال، یک روش تفسیر تعقیبی است. روش‌های تعقیبی نیز می‌توانند برای مدل‌های قابل تفسیر ذاتی اعمال شوند. به عنوان مثال، اهمیت ویژگی جایگشت را می‌توان برای درختان تصمیم محاسبه کرد. سازماندهی فصول در این کتاب با تمایز بین مدل‌های ذاتاً قابل تفسیر و روش‌های تفسیر تعقیبی (و مدل-آگنوستیک) تعیین می‌شود.

نتیجه روش تفسیر: روش‌های مختلف تفسیر را می‌توان به طور تقریبی با توجه به نتایج آنها متمایز کرد.

- **آمار خلاصه ویژگی:** بسیاری از روش‌های تفسیر، آمار خلاصه‌ای را برای هر ویژگی ارائه می‌دهند. برخی از روش‌ها یک عدد واحد را برای هر ویژگی برمی‌گردانند، مانند اهمیت ویژگی، یا یک نتیجه پیچیده‌تر، مانند نقاط قوت تعامل ویژگی‌های زوجی، که شامل یک عدد برای هر جفت ویژگی است.
- **تجسم خلاصه ویژگی:** بیشتر آمار خلاصه ویژگی‌ها قابل ترسیم نیز هستند. برخی از خلاصه‌های ویژگی‌ها در واقع تنها زمانی معنادار هستند که ترسیم شوند و انتخاب یک جدول برای ارائه، اشتباه می‌باشد. وابستگی جزئی یک ویژگی، چنین موردی است. نمودارهای وابستگی جزئی منحنی‌هایی هستند که یک ویژگی و میانگین نتیجه پیش‌بینی شده را نشان می‌دهند. بهترین راه برای ارائه وابستگی‌های جزئی، رسم منحنی به جای چاپ مختصات است.

- **محتویات داخلی مدل (مثلاً وزن‌های آموزش داده شده):** تفسیر مدل‌های قابل تفسیر ذاتی در این دسته قرار می‌گیرد. به عنوان مثال وزن در مدل‌های خطی یا ساختار درختی آموزش داده شده درخت‌های تصمیم (ویژگی‌ها و آستانه‌های مورد استفاده برای تقسیم‌ها) هستند. مرز بین محتویات داخلی مدل و آمار خلاصه ویژگی، به عنوان مثال، در مدل‌های خطی از بین می‌رود، زیرا وزن‌ها هم زمان محتویات داخلی مدل و هم آمار خلاصه ویژگی‌ها هستند. روش دیگری که مدل‌های داخلی را خروجی می‌دهد، تجسم آشکارسازهای ویژگی است که در شبکه‌های عصبی کانولوشن آموخته شده‌اند. روش‌های تفسیرپذیری که داخلی‌های مدل خروجی را به دست می‌آورند، طبق تعریف، مختص مدل هستند (معیار بعدی را ببینید).
- **نقطه داده:** این دسته شامل تمام روش‌هایی است که نقاط داده (از قبل موجود یا تازه ایجاد شده) را برای قابل تفسیر کردن یک مدل برمی‌گرداند. یکی از روش‌ها توضیحات خلاف واقع نامیده می‌شود. برای

توضیح پیش‌بینی یک نمونه داده، این روش با تغییر برخی از ویژگی‌هایی که نتیجه پیش‌بینی شده به روشنی مرتبط تغییر می‌کند، یک نقطه داده مشابه پیدا می‌کند (مثلاً یک تلنگر در کلاس پیش‌بینی شده). مثال دیگر شناسایی نمونه‌های اولیه کلاس‌های پیش‌بینی شده است. برای مفید بودن، روش‌های تفسیری که نقاط داده جدید را تولید می‌کنند، مستلزم آن هستند که خود نقاط داده قابل تفسیر باشند. این برای تصاویر و متون به خوبی کار می‌کند، اما برای داده‌های جدولی با صدها ویژگی کمتر مفید است.

- **مدل قابل تفسیر ذاتی:** یک راه حل برای تفسیر مدل‌های جعبه سیاه، تقریب آنها (به صورت جهانی یا محلی) با یک مدل قابل تفسیر است. خود مدل قابل تفسیر با نگاه کردن به پارامترهای مدل داخلی یا آمار خلاصه ویژگی تفسیر می‌شود.

مدل خاص یا مدل آگنوستیک؟ ابزارهای تفسیر مدل خاص به کلاس‌های مدل خاص محدود می‌شوند. تفسیر وزن‌های رگرسیون در یک مدل خطی یک تفسیر مدل خاص است، زیرا - طبق تعریف - تفسیر مدل‌های ذاتی قابل تفسیر همیشه مختص مدل است. ابزارهایی که فقط برای تفسیر شبکه‌های عصبی کار می‌کنند، مختص مدل هستند. ابزارهای مدل-آگنوستیک را می‌توان در هر مدل یادگیری ماشین استفاده کرد و پس از آموزش مدل (post hoc) استفاده می‌شود. این روش‌های آگنوستیک معمولاً با تجزیه و تحلیل جفت‌های ورودی و خروجی ویژگی کار می‌کنند. طبق تعریف، این روش‌ها نمی‌توانند به اجزای داخلی مدل مانند وزن یا اطلاعات ساختاری دسترسی داشته باشند.

محلي یا جهاني؟ آيا روش تفسير يك پيش‌بیني فردی يا کل رفتار مدل را توضیح می‌دهد؟ يا حوزه مابین این دو حالت است؟ برای اطلاعات بیشتر در مورد معیار حوزه تفسیر بخش بعدی را مطالعه کنید.

۳.۳ حوزه تفسیرپذیری

یک الگوریتم مدلی را آموزش می‌دهد که پیش‌بینی‌ها را تولید می‌کند. هر مرحله را می‌توان از نظر شفافیت یا تفسیر پذیری ارزیابی کرد.

۳.۳.۱ شفافیت الگوریتم

الگوریتم چگونه مدل را ایجاد می‌کند؟

شفافیت الگوریتم در مورد چگونگی یادگیری الگوریتم از داده‌ها و نوع روابطی است که می‌تواند یاد بگیرد، بحث می‌کند. اگر از شبکه‌های عصبی کانولوشن برای طبقه‌بندی تصاویر استفاده می‌کنید، می‌توانید توضیح دهید که الگوریتم آشکارسازهای لبه و فیلترها را در پایین‌ترین لایه‌ها یاد می‌گیرد. این درک نحوه عملکرد الگوریتم است. اما این توضیح برای مدل خاصی که در پایان آموزش داده می‌شود و برای چگونگی پیش‌بینی‌های فردی نیست. شفافیت الگوریتم فقط به دانش الگوریتم نیاز دارد و کاری به داده‌ها یا مدل‌های آموزش داده شده ندارد. این کتاب بر تفسیرپذیری مدل تمرکز دارد و نه شفافیت الگوریتم. الگوریتم‌هایی مانند روش حداقل مربعات برای مدل‌های

خطی به خوبی مطالعه و درک شده‌اند. این الگوریتم‌ها به عنوان شفافیت بالا شناخته می‌شوند. رویکردهای یادگیری عمیق (اعمال کردن یک گرادیان از طریق شبکه‌ای با میلیون‌ها وزن) کمتر درک شده‌اند و تحقیق در مورد کارکردهای درونی در مرکز توجه محققان می‌باشد. این موارد به عنوان ناشفافیت کم در نظر گرفته می‌شوند.

۳.۳.۲ تفسیرپذیری مدل کل نگر

مدل آموزش‌دیده چگونه پیش‌بینی می‌کند؟

اگر بتوانید کل مدل را یکجا درک کنید، می‌توانید یک مدل را قابل تفسیر توصیف کنید (Lipton, 2018). برای توضیح خروجی مدل جهانی، به مدل آموزش‌دیده، دانش الگوریتم و داده‌ها نیاز دارید. این سطح از تفسیرپذیری در مورد درک چگونگی تصمیم‌گیری مدل، بر اساس یک دیدگاه جامع از ویژگی‌های آن و هر یک از اجزای آموخته شده مانند وزن‌ها، پارامترهای دیگر، و ساختار است. کدام ویژگی‌ها مهم هستند و چه نوع تعاملاتی بین آنها وجود دارد؟ تفسیرپذیری مدل جهانی به درک توزیع نتیجه هدف شما بر اساس ویژگی‌ها کمک می‌کند. دستیابی به تفسیرپذیری مدل جهانی در عمل بسیار دشوار است. هر مدلی که بیش از تعداد انگشت شماری پارامتر یا وزن داشته باشد، بعيد است که در حافظه کوتاه مدت یک انسان معمولی جای بگیرد. من استدلال می‌کنم که شما واقعاً نمی‌توانید یک مدل خطی با ۵ ویژگی را تصور کنید، زیرا این به معنای ترسیم ابر صفحه تخمینی به صورت ذهنی در یک فضای ۵ بعدی است. هر فضای ویژگی با بیش از ۳ بعد به سادگی برای انسان قابل تصور نیست. معمولاً هنگامی که افراد سعی در درک یک مدل دارند، فقط بخش‌هایی از آن را در نظر می‌گیرند، مانند وزن‌ها در مدل‌های خطی.

۳.۳.۳ تفسیرپذیری مدل جهانی در سطح مدولار

چگونه بخش‌هایی از مدل بر پیش‌بینی‌ها تأثیر می‌گذارد؟

یک مدل ساده بیز با صدھا ویژگی برای من و شما بزرگ‌تر از آن است که بتوانیم آن را در حافظه کاری خود نگه داریم. و حتی اگر بتوانیم تمام وزن‌ها را به خاطر بسپاریم، نمی‌توانیم به سرعت برای نقاط داده جدید پیش‌بینی کنیم. علاوه بر این، شما باید توزیع مشترک همه ویژگی‌ها را در ذهن خود داشته باشید تا اهمیت هر ویژگی و اینکه ویژگی‌ها به طور متوسط چگونه بر پیش‌بینی‌ها تأثیر می‌گذارند، تخمین بزنید. این یک کار غیر ممکن است. اما شما به راحتی می‌توانید یک وزن را درک کنید. درحالی که تفسیرپذیری مدل جهانی معمولاً دور از دسترس است، شанс خوبی برای درک حداقل برخی از مدل‌ها در سطح مدولار وجود دارد. همه مدل‌ها در سطح پارامتر قابل تفسیر نیستند. برای مدل‌های خطی، بخش‌های قابل تفسیر وزن‌ها هستند، برای درخت‌ها تقسیم‌ها (ویژگی‌های انتخابی به اضافه نقاط برش) و پیش‌بینی‌های گره برگ است. به عنوان مثال، به نظر می‌رسد که مدل‌های خطی را می‌توان به طور کامل در یک سطح مدولار تفسیر کرد، اما تفسیر یک وزن منفرد، با تمام وزن‌های دیگر در هم تنیده است. تفسیر وزن منفرد همیشه با این پاورقی همراه می‌شود که سایر ویژگی‌های ورودی در همان مقدار

باقی می‌مانند، که در مورد بسیاری از شرایط واقعی صدق نمی‌کند. یک مدل خطی که ارزش یک خانه را پیش‌بینی می‌کند، که هم اندازه خانه و هم تعداد اتاق‌ها را در نظر می‌گیرد، می‌تواند وزن منفی برای ویژگی اتاق داشته باشد. دلیل این امر این است که بین این ویژگی و ویژگی اندازه خانه همبستگی شدید وجود دارد. در بازاری که مردم اتاق‌های بزرگ‌تر را ترجیح می‌دهند، یک خانه با اتاق‌های کمتر می‌تواند ارزش بیشتری نسبت به خانه‌هایی با اتاق‌های بیشتر داشته باشد، اگر هر دو خانه دارای اندازه یکسان باشند. وزن‌ها فقط در چارچوب سایر ویژگی‌های مدل معنا می‌یابند.

۳.۳.۴ تفسیر محلی برای یک پیش‌بینی واحد

چرا مدل برای یک نمونه، پیش‌بینی خاصی انجام داد؟

می‌توانید روی یک نمونه تمرکز کنید و آنچه را که مدل برای این ورودی پیش‌بینی می‌کند بررسی کنید و توضیح دهید که چرا. اگر به یک پیش‌بینی خاص نگاه کنید، رفتار مدل پیچیده‌تر ممکن است قابل فهم تر باشد. به طور محلی، پیش‌بینی ممکن است فقط به صورت خطی یا یکنواخت به برخی ویژگی‌ها بستگی داشته باشد، نه اینکه وابستگی پیچیده‌ای به آنها داشته باشد. به عنوان مثال، ارزش یک خانه ممکن است به طور غیرخطی به اندازه آن بستگی داشته باشد. اما اگر فقط به یک خانه ۱۰۰ متر مربعی خاص نگاه می‌کنید، این احتمال وجود دارد که برای آن زیر مجموعه‌داده، پیش‌بینی مدل شما به صورت خطی به اندازه بستگی دارد. شما می‌توانید با شبیه سازی نحوه تغییر قیمت پیش‌بینی شده با افزایش یا کاهش اندازه ۱۰ متر مربع به این موضوع پی ببرید. بنابراین توضیحات محلی می‌توانند دقیق‌تر از توضیحات جهانی باشند. این کتاب روش‌هایی را ارائه می‌کند که می‌توانند پیش‌بینی‌های فردی را در مدل‌های آگنوستیک قابل تفسیر تر کنند.

۳.۳.۵ تفسیر محلی برای گروهی از پیش‌بینی‌ها

چرا مدل پیش‌بینی‌های خاصی را برای گروهی از نمونه‌ها انجام داد؟

پیش‌بینی‌های مدل برای نمونه‌های متعدد را می‌توان با روش‌های تفسیر مدل جهانی (در سطح مدولار) یا با توضیح نمونه‌های جداگانه توضیح داد. روش‌های سراسری را می‌توان این گونه اعمال کرد: در نظر گرفتن قسمتی از نمونه‌ها، رفتار با آن‌ها به عنوان کل مجموعه‌داده و استفاده از روش‌های جهانی برای این زیرمجموعه. روش‌های توضیح فردی را می‌توان در هر نمونه استفاده کرد و سپس برای کل گروه فهرست یا تجمعی کرد.

۳.۴ ارزیابی تفسیرپذیری

هیچ اتفاق نظر مشخصی در مورد اینکه تفسیرپذیری در یادگیری ماشین چیست، وجود ندارد. همچنین نحوه اندازه گیری آن مشخص نیست. اما برخی تحقیقات اولیه در این مورد و تلاشی برای تدوین برخی رویکردها برای ارزیابی، همان‌طور که در بخش بعدی توضیح داده شده است، وجود دارد.

Doshi-Velez and Kim (2017) سه سطح اصلی را برای ارزیابی تفسیرپذیری پیشنهاد می‌کنند:

ارزیابی سطح کاربردی (وظیفه واقعی): توضیحات را در محصول قرار دهید و آن را توسط کاربر نهایی آزمایش کنید. نرم افزار تشخیص شکستگی را با یک قطعه یادگیری ماشین تصور کنید که شکستگی‌ها را با استفاده از اشعه ایکس مکان یابی و علامت گذاری می‌کند. در سطح کاربردی، رادیولوژیست‌ها نرم افزار تشخیص شکستگی را مستقیماً برای ارزیابی مدل آزمایش می‌کنند. این نیاز به یک چیدمان تجربی خوب و درک چگونگی ارزیابی کیفیت دارد. یک مبنای خوب برای این موضوع این است که یک انسان چقدر در توضیح همان تصمیم خوب است. **ارزیابی سطح انسانی (وظیفه ساده):** یک ارزیابی سطح کاربردی ساده شده، است. تفاوت این است که این آزمایش‌ها با متخصصان حوزه انجام نمی‌شود، بلکه با افراد عادی انجام می‌شود. این امر آزمایش‌ها را ارزان‌تر می‌کند (مخصوصاً اگر متخصصان حوزه رادیولوژیست باشند) و یافتن آزمایش‌کنندگان بیشتر، آسان‌تر می‌شود. به عنوان مثال، توضیحات متفاوتی به کاربر نشان داده می‌شود و کاربر بهترین را انتخاب می‌کند.

ارزیابی سطح عملکرد (وظیفه جایگزین): به انسان نیاز ندارد. این ارزیابی وقتی بهترین حالت را دارد که کلاس مدل مورد استفاده، قبلًا توسط شخص دیگری در ارزیابی سطح انسانی، ارزیابی شده باشد. به عنوان مثال، ممکن است مشخص شود که کاربران نهایی درخت تصمیم را درک می‌کنند. در این مورد، یک جایگزین برای کیفیت توضیح ممکن است عمق درخت باشد. درختان کوتاه‌تر نمره توضیح پذیری بهتری را دریافت می‌کنند. البته اضافه کردن این محدودیت زمانی منطقی است که عملکرد پیش‌بینی درخت خوب باقی بماند و در مقایسه با درخت بزرگ‌تر خیلی کاهش نیابد.

فصل بعدی بر ارزیابی توضیحات برای پیش‌بینی‌های فردی در سطح عملکرد مرکز دارد. مشخصات مرتبط با توضیحاتی که برای ارزیابی آنها در نظر می‌گیریم، چیست؟

۳.۵ خواص توضیحات

ما می‌خواهیم پیش‌بینی‌های یک مدل یادگیری ماشین را توضیح دهیم. برای رسیدن به این هدف، ما به برخی از روش‌های توضیحی، که الگوریتمی هستند که توضیحات را تولید می‌کند، تکیه می‌کنیم. یک توضیح معمولاً مقادیر ویژگی یک نمونه را به روشنی قابل درک برای انسان به پیش‌بینی مدل آن مرتبط می‌کند. انواع دیگر توضیحات شامل مجموعه‌ای از نمونه‌های داده است (مثلاً در مورد مدل k نزدیک‌ترین همسایه). برای مثال، می‌توانیم خطر سرطان را با استفاده از یک ماشین بردار پشتیبان پیش‌بینی کنیم و پیش‌بینی‌ها را با استفاده از روش جایگزین محلی (local surrogate method) توضیح دهیم، که درخت‌های تصمیم را به عنوان توضیح، تولید می‌کند. یا می‌توانیم به جای ماشین بردار پشتیبان از مدل رگرسیون خطی استفاده کنیم. مدل رگرسیون خطی، مجهز به یک روش توضیحی (تفسیر وزن‌ها) می‌باشد.

ما نگاهی دقیق‌تر به خواص روش‌های توضیحات و توضیحات می‌اندازیم (Robnik-Šikonja & Bohanec, 2018) از این خواص می‌توان برای قضاوت در مورد خوب بودن روش توضیح یا توضیح استفاده کرد. البته مشخص نیست که این ویژگی‌ها چگونه به درستی اندازه گیری می‌شوند و در نتیجه یکی از چالش‌ها نحوه محاسبه آنها است.

خواص روش‌های توضیح

- قدرت توضیح، «زبان» یا ساختار توضیحاتی است که روش قادر به ایجاد آن است. یک روش توضیحی می‌تواند قوانین IF-THEN، درخت‌های تصمیم، جمع وزنی، زبان طبیعی یا چیز دیگری را ایجاد کند.
- توضیح می‌دهد که روش توضیح تا چه حد به بررسی مدل یادگیری ماشین، مانند Translucency پارامترهای آن، متکی است. برای مثال، روش‌های تبیین متکی بر مدل‌های قابل تفسیر ذاتی مانند مدل رگرسیون خطی (مخصوص مدل) بسیار شفاف هستند. روش‌هایی که تنها به دستکاری ورودی‌ها و مشاهده پیش‌بینی‌ها تکیه می‌کنند، شفافیت صفر دارند. بسته به سناریو، سطوح مختلف شفافیت ممکن است مطلوب باشد. مزیت شفافیت بالا این است که روش می‌تواند به اطلاعات بیشتری برای تولید توضیحات تکیه کند. مزیت شفافیت کم این است که روش توضیح قابل حمل تر است.
- قابلیت حمل طیفی از مدل‌های یادگیری ماشین را توصیف می‌کند که می‌توان با آن از روش توضیحی استفاده کرد. روش‌هایی با شفافیت پایین، قابلیت حمل بالاتری دارند، زیرا با مدل یادگیری ماشین مانند یک جعبه سیاه رفتار می‌کنند. مدل‌های جایگزین ممکن است روش توضیحی با بالاترین قابلیت حمل باشند. روش‌هایی که فقط برای شبکه‌های عصبی مکرر کار می‌کنند، قابلیت حمل کمی دارند.
- پیچیدگی الگوریتمی پیچیدگی محاسباتی روشی را که توضیح را ایجاد می‌کند، توصیف می‌کند. زمانی که زمان محاسبه یک گلوگاه در تولید توضیحات است، این ویژگی مهم است که در نظر گرفته شود.

خواص تبیین‌های فردی

- دقت: یک توضیح چقدر داده‌های دیده نشده را پیش‌بینی می‌کند؟ دقت بالا به ویژه در صورتی مهم است که توضیح برای پیش‌بینی‌ها به جای مدل یادگیری ماشین استفاده شود. دقت پایین می‌تواند خوب باشد اگر دقت مدل یادگیری ماشین نیز پایین باشد، و اگر هدف توضیح این باشد که مدل جعبه سیاه چه کاری انجام می‌دهد. در این مورد فقط وفاداری مهم است.
- وفاداری: توضیح چقدر به پیش‌بینی مدل جعبه سیاه نزدیک است؟ وفاداری بالا یکی از مهم‌ترین ویژگی‌های توضیح است، زیرا توضیح با وفاداری پایین برای توضیح مدل یادگیری ماشین بی فایده است. دقت و وفاداری ارتباط نزدیکی با هم دارند. اگر مدل جعبه سیاه دقت بالایی داشته باشد و توضیحات دارای وفاداری بالا باشد، توضیحات نیز از دقت بالایی برخوردار است. برخی از توضیحات فقط وفاداری محلی را ارائه می‌دهند، به این معنی که توضیح فقط به خوبی به پیش‌بینی مدل برای زیرمجموعه‌ای از

داده‌ها (مثلًاً مدل‌های جایگزین محلی) یا حتی برای یک نمونه داده منفرد (مثلًاً مقادیر Shapley) تقریب دارد.

سازگاری: یک توضیح بین مدل‌هایی که برای یک کار آموزش‌دهد اند و پیش‌بینی‌های مشابهی تولید می‌کنند چقدر تفاوت دارد؟ برای مثال، من یک ماشین بردار پشتیبان و یک مدل رگرسیون خطی را برای یک کار آموزش می‌دهم و هر دو پیش‌بینی‌های بسیار مشابهی را تولید می‌کنند. من توضیحات را با استفاده از روشی که انتخاب می‌کنم محاسبه می‌کنم و تفاوت‌های توضیحات را تجزیه و تحلیل می‌کنم. اگر توضیحات بسیار شبیه به هم باشند، توضیحات بسیار سازگار هستند. من این ویژگی را تا حدودی دشوار می‌دانم، زیرا این دو مدل می‌توانند از ویژگی‌های متفاوتی استفاده کنند، اما پیش‌بینی‌های مشابهی دریافت می‌کنند (همچنین «اثر راشومون» نامیده می‌شود). در این مورد سازگاری بالا مطلوب نیست زیرا توضیحات باید بسیار متفاوت باشند. اگر مدل‌ها واقعًا بر روابط مشابه متکی باشند، سازگاری بالا مطلوب است.

پایداری: توضیحات برای نمونه‌های مشابه چقدر شبیه است؟ در حالی که سازگاری توضیحات بین مدل‌ها را مقایسه می‌کند، ثبات توضیحات بین نمونه‌های مشابه را برای یک مدل ثابت مقایسه می‌کند. پایداری بالا به این معنی است که تغییرات جزئی در ویژگی‌های یک نمونه، توضیح اساسی را تغییر نمی‌دهد (مگر اینکه این تغییرات جزئی نیز پیش‌بینی را به شدت تغییر دهد). عدم ثبات می‌تواند نتیجه واریانس زیاد روش تبیین باشد. به عبارت دیگر، روش توضیح به شدت تحت تأثیر تغییرات جزئی مقادیر ویژگی نمونه مورد توضیح قرار می‌گیرد. فقدان ثبات همچنین می‌تواند ناشی از مؤلفه‌های غیر قطعی روش توضیح باشد، مانند مرحله نمونه‌گیری داده‌ها، مانند روش جایگزین محلی. پایداری بالا همیشه مطلوب است.

قابل‌درک بودن: انسان‌ها چقدر توضیحات را درک می‌کنند؟ این دقیقاً مانند یک ملک دیگر در میان بسیاری به نظر می‌رسد، اما فیل در اتاق است. تعریف و اندازه گیری دشوار است، اما درست کردن آن بسیار مهم است. بسیاری از مردم قبول دارند که قابل‌درک بودن به مخاطب بستگی دارد. ایده‌هایی برای اندازه گیری در کپذیری شامل اندازه گیری اندازه توضیح (تعداد ویژگی‌ها با وزن غیر صفر در یک مدل خطی، تعداد قوانین تصمیم‌گیری، و ...) یا آزمایش اینکه افراد چقدر می‌توانند رفتار مدل یادگیری ماشین را از توضیحات پیش‌بینی کنند، می‌شود. قابل‌درک بودن ویژگی‌های استفاده شده در توضیح نیز باید در نظر گرفته شود. تغییر پیچیده ویژگی‌ها ممکن است کمتر از ویژگی‌های اصلی قابل‌درک باشد.

قطعیت: آیا توضیح، قطعیت مدل یادگیری ماشین را منعکس می‌کند؟ بسیاری از مدل‌های یادگیری ماشین فقط پیش‌بینی می‌کنند بدون اینکه بیانیه‌ای در مورد مدل‌ها وجود داشته باشد، اطمینان دارند که پیش‌بینی درست است. اگر مدل یک احتمال 4 درصدی سرطان را برای یک بیمار پیش‌بینی کند، آیا

به اندازه احتمال ۴ درصدی است که بیمار دیگر با مقادیر ویژگی‌های متفاوت دریافت کرده است؟ توضیحی که شامل قطعیت مدل باشد بسیار مفید است.

درجه اهمیت: توضیح چقدر اهمیت ویژگی‌ها یا بخش‌هایی از توضیح را منعکس می‌کند؟ به عنوان مثال، اگر یک قاعده تصمیم به عنوان توضیحی برای یک پیش‌بینی فردی ایجاد شود، آیا مشخص است که کدام یک از شرایط قانون مهم‌ترین بوده است؟

تازگی: آیا توضیح نشان می‌دهد که آیا نمونه داده ای که باید توضیح داده شود از یک منطقه "جدید" به دور از توزیع داده‌های آموزشی آمده است؟ در چنین مواردی، مدل ممکن است نادرست باشد و توضیح ممکن است بی فایده باشد. مفهوم تازگی با مفهوم یقین مرتبط است. هر چه نوآوری بالاتر باشد، احتمال اینکه مدل به دلیل کمبود داده از اطمینان پایینی برخوردار باشد بیشتر است.

نمایندگی: توضیح چند مورد را پوشش می‌دهد؟ توضیحات می‌توانند کل مدل را پوشش دهند (مثلاً تفسیر وزن‌ها در مدل رگرسیون خطی) یا فقط یک پیش‌بینی فردی را نشان دهند (مثلاً مقادیر Shapley).

۳.۶ توضیحات انسان پسند

بیایید عمیق‌تر آنچه را که ما انسان‌ها به عنوان توضیحات «خوب» می‌پسندیم، بررسی کنیم و پیامدهای آن برای یادگیری ماشین قابل تفسیر پیدا کنیم. تحقیقات علوم انسانی می‌تواند به ما در یافتن این موضوع کمک کند. (Miller, 2019) تحقیقات مفصلی درباره توضیحات انجام داده است و این بخش، خلاصه‌ای بر اساس این تحقیقات است.

در این بخش، می‌خواهم شما را به موارد زیر آشنا کنم: به عنوان توضیحی برای یک رویداد، انسان‌ها توضیحات کوتاه (فقط ۱ یا ۲ علت) را ترجیح می‌دهند که موقعیت فعلی را با موقعیتی که در آن رویداد رخ نمی‌داد، مقایسه کند. به خصوص آوردن علل غیرعادی، توضیحات خوبی محسوب می‌شوند. توضیحات، تعاملات اجتماعی بین توضیح دهنده و گیرنده توضیح هستند و بنابراین زمینه اجتماعی تأثیر زیادی بر محتوای واقعی توضیح دارد. وقتی برای یک پیش‌بینی یا رفتار خاص به توضیحاتی با همه عوامل نیاز دارید، توضیحی انسان‌پسند نمی‌خواهد، بلکه یک توصیف علی کامل می‌خواهد. اگر از نظر قانونی ملزم به تعیین همه ویژگی‌های تأثیرگذار هستید یا اگر مدل یادگیری ماشین را اشکال‌زدایی می‌کنید، احتمالاً می‌خواهید یک توصیف علی داشته باشید. در این صورت به نکات زیر توجه نکنید. در سایر موارد که افراد غیرمتخصص یا افراد با کم وقت، دریافت کننده توضیحات هستند، بخش‌های زیر باید برای شما جالب باشد.

۳.۶.۱ توضیح چیست؟

توضیح پاسخ به یک سوال چرایی است (Miller, 2019).

چرا درمان روی بیمار جواب نداد؟

چرا وام من رد شد؟

چرا هنوز زندگی فرازمنی با ما تماس نگرفته است؟

دو سؤال اول را می‌توان با توضیح «روزمره» پاسخ داد، درحالی که سؤال سوم از دسته «پدیده‌های کلی‌تر علمی و سؤال‌های فلسفی» می‌باشد. ما روی توضیحات نوع «روزانه» تمرکز می‌کنیم، زیرا این توضیحات مربوط به یادگیری ماشین قابل تفسیر است. سؤالاتی که با «چگونه» شروع می‌شوند معمولاً می‌توانند به عنوان سؤالات «چرا» بازنویسی شوند: «چگونه وام من رد شد؟» را می‌توان به «چرا وام من رد شد؟» تبدیل کرد.

در ادامه، اصطلاح «توضیح» آورده شد به فرآیند اجتماعی و شناختی توضیحات و همچنین محصول این فرآیندها اشاره دارد. توضیح دهنده می‌تواند یک انسان یا یک ماشین باشد.

۳.۶.۲ یک توضیح خوب چیست؟

این بخش، مطالب ارائه شده توسط Miller (2019) را در مورد توضیحات "خوب" بیشتر فشرده می‌کند و مفاهیم ملموسی را برای یادگیری ماشین قابل تفسیر اضافه می‌کند.

۱- توضیحات مقایسه ای هستند (Lipton, 1990). انسان‌ها معمولاً نمی‌پرسند که چرا یک پیش‌بینی خاص انجام شده است، بلکه می‌پرسند چرا این پیش‌بینی به جای پیش‌بینی دیگری انجام شده است. ما تمایل داریم در موارد خلاف واقع فکر کنیم، به عنوان مثال "اگر ورودی X متفاوت بود، پیش‌بینی چگونه بود؟". برای پیش‌بینی قیمت مسکن، صاحب خانه ممکن است علاقه‌مند باشد که چرا قیمت پیش‌بینی شده در مقایسه با قیمتی که انتظار داشته است، بالاتر است. اگر درخواست وام من رد شود، اهمیتی برای شنیدن همه عواملی که باعث رد شدن وام شدند، ندارم. من مشتاق به شنیدن عواملی هستم که برای دریافت وام باید تغییر کنند. من می‌خواهم تفاوت بین درخواست من و نسخه مورد پذیرش درخواستم را بدانم. تشخیص اینکه توضیحات مقایسه ای اهمیت دارند، یافته مهمی برای یادگیری ماشین قابل توضیح، است. از اکثر مدل‌های قابل تفسیر، می‌توانید توضیحی را استخراج کنید که به طور ضمنی پیش‌بینی یک نمونه را با پیش‌بینی یک نمونه داده مصنوعی یا میانگین نمونه‌ها مقایسه کند. پژوهشکار ممکن است بپرسند: "چرا دارو برای بیمار من جواب نمی‌دهد؟" و ممکن است توضیحی بخواهند که بیمارشان را با بیماری که دارو برای او موثر بوده و مشابه بیمار بدون پاسخ است، مقایسه کند. در ک توضیحات متضاد آسانتر از توضیحات کامل است. توضیح کاملی در مورد سوال پژوهش که چرا دارو کار نمی‌کند ممکن است شامل موارد زیر باشد: بیمار به مدت ۱۰ سال به این بیماری مبتلا بوده است، ۱۱ ژن over-expressed هستند، the patients body is very quick ... توضیح متضاد ممکن است بسیار ساده تر باشد: بدن بیمار در تجزیه دارو به مواد شیمیایی بی اثر بسیار سریع عمل می‌کند in breaking the drug down into ineffective chemicals برخلاف بیمار پاسخ دهنده به دارو، بیمار بدون پاسخ دارای ترکیب خاصی از ژن‌ها است که اثربخشی دارو را کاهش

می‌دهد. بهترین توضیح، توضیحی است که بیشترین تفاوت را بین شرایط مورد نظر و شرایط مرجع را برجسته کند.

معنی آن برای یادگیری ماشین قابل تفسیر: انسان‌ها توضیح کاملی برای یک پیش‌بینی نمی‌خواهند، اما می‌خواهند تفاوت‌ها را با پیش‌بینی نمونه دیگری (که می‌تواند مصنوعی باشد) مقایسه کنند. ایجاد توضیحات مقایسه‌ای وابسته به کاربرد است زیرا به نقطه مرجع برای مقایسه نیاز دارد. این توضیح ممکن است به نقطه داده ای که باید توضیح داده شود و به کاربر دریافت کننده توضیح، بستگی داشته باشد. یک کاربر وب‌سایت پیش‌بینی قیمت خانه، ممکن است بخواهد توضیحی در مورد پیش‌بینی قیمت خانه در مقایسه با خانه خود یا شاید خانه دیگری در وب‌سایت یا شاید با یک خانه متوسط در همسایگی اش داشته باشد. راه حل برای ایجاد خودکار توضیحات مقایسه‌ای، ممکن است شامل یافتن نمونه‌های اولیه (prototypes) یا کهن الگوها (archetypes) در داده‌ها باشد.

۲- توضیحات انتخاب شده است. مردم انتظار توضیحی ندارند که فهرست واقعی و کاملی از علی یک رویداد را ارائه دهد. ما عادت کرده ایم که یک یا دو علت را از میان انواع علل احتمالی به عنوان توضیح انتخاب کنیم. به عنوان مدرک، اخبار تلویزیون را روشن کنید: "کاهش قیمت سهام به دلیل واکنش فراینده علیه محصول شرکت به دلیل مشکلات مربوط به آخرین به روز رسانی نرم افزار است".

سوپاسا و تیمش به دلیل دفاع ضعیف بازی را باختند: آنها به حریفان خود فضای زیادی برای اجرای استراتژی خود دادند.

بی اعتمادی فراینده به نهادهای مستقر و دولت ما، عوامل اصلی کاهش مشارکت رای دهنگان است. این واقعیت که یک رویداد را می‌توان با علل مختلف توضیح داد، اثر را شومون نامیده می‌شود. را شومون یک فیلم ژاپنی است که داستان‌ها (توضیحات) مقایسه‌ای و جایگزین درباره مرگ یک سامورایی را روایت می‌کند. برای مدل‌های یادگیری ماشین مطلوبست، اگر بتوان یک پیش‌بینی خوب با استفاده از ویژگی‌های مختلف انجام داد. روش‌های Ensemble که چندین مدل را با ویژگی‌های مختلف (توضیحات مختلف) ترکیب می‌کنند عموماً عملکرد خوبی دارند، زیرا میانگین‌گیری زیادی از آن «داستان‌ها» پیش‌بینی‌ها را قوی‌تر و دقیق‌تر می‌کند. اما همچنین به این معنی است که بیش از یک توضیح انتخابی وجود دارد که چرا یک پیش‌بینی خاص انجام شده است.

معنی آن برای یادگیری ماشین قابل تفسیر: توضیح را خیلی کوتاه بیان کنید، فقط ۱ تا ۳ دلیل بیاورید، حتی اگر دنیا پیچیده‌تر باشد. روش LIME کارکرد خوبی در این زمینه دارد.

۳- توضیحات اجتماعی هستند. آنها بخشی از مکالمه یا تعامل بین توضیح دهنده و گیرنده توضیح هستند. بافت اجتماعی، محتوا و ماهیت توضیحات را تعیین می‌کند. اگر بخواهم به یک فرد فنی، توضیح دهم که چرا ارزهای دیجیتال اینقدر ارزش دارند، مواردی از این قبیل می‌گوییم: «دفتر غیرمت مرکز، توزیع شده، مبتنی بر بلاک

چین، که توسط یک نهاد مرکزی قابل کنترل نیست، افرادی که می خواهند امنیت داشته باشند را به خرید تشویق می کند و در نتیجه قیمت بالا می رود". اما به مادربزرگم می گفتم: "بین، مادربزرگ: ارزهای دیجیتال کمی شبیه طلای رایانه‌ای هستند. مردم طلا را دوست دارند و برای آن پول زیادی می پردازند و جوانان نیز طلای کامپیوتری را دوست دارند و هزینه زیادی برای آن می پردازند".

معنی آن برای یادگیری ماشین قابل تفسیر چیست: به محیط اجتماعی برنامه یادگیری ماشین خود و مخاطبان هدف توجه کنید. دریافت درست بخش اجتماعی مدل یادگیری ماشین کاملاً به برنامه خاص شما بستگی دارد. متخصصانی از علوم انسانی (به عنوان مثال روانشناسان و جامعه شناسان) را پیدا کنید تا به شما کمک کنند.

۴- توضیحات بر موارد غیرعادی تمرکز دارند. مردم برای توضیح رویدادها بیشتر بر علل غیرعادی تمرکز می کنند (Kahneman & Tversky, 1981). این موارد، علی هستند که احتمال کمی داشتند اما با این وجود اتفاق افتادند. حذف این علل غیرطبیعی نتیجه را تا حد زیادی تغییر می داد (توضیح خلاف واقع counterfactual explanation). انسان‌ها این نوع علل «غیرعادی» را به عنوان توضیحات خوبی در نظر می‌گیرند. مثالی از Štrumbelj and Kononenko (2011) است: فرض کنید مجموعه داده ای از موقعیت‌های آزمون بین معلمان و دانش آموزان داریم. دانش آموزان در یک دوره شرکت می‌کنند و پس از ارائه موفقیت آمیز دوره را مستقیماً می‌گذرانند. معلم این گزینه را دارد که علاوه بر آن از دانش آموز سوالاتی بپرسد تا دانش آنها را محک بزند. دانش آموزانی که نتوانند به این سوالات پاسخ دهنند در دوره مردود خواهند بود. دانش آموزان می‌توانند سطوح آمادگی متفاوتی داشته باشند، که به احتمالات متفاوتی برای پاسخ صحیح به سوالات معلم ترجمه می‌شود (اگر تصمیم به امتحان دانش آموز داشته باشد). ما می‌خواهیم پیش‌بینی کنیم که آیا یک دانش آموز این دوره را سپری می‌کند و پیش‌بینی خود را توضیح دهد. در صورتی که استاد هیچ سوال اضافی نپرسد، شانس قبولی ۱۰۰٪ است، در غیر این صورت احتمال قبولی بستگی به سطح آمادگی دانش آموز و احتمال پاسخگویی صحیح در نتیجه به سوالات دارد.

سناریوی ۱: معلم معمولاً از دانش آموزان سوالات اضافی می‌پرسد (مثلاً ۹۵ از ۱۰۰ بار). دانش آموزی که درس نخوانده است (۱۰٪ شانس قبولی در بخش سوال) جزو افراد خوش شانس نبوده و سوالات اضافی دریافت می‌کند که نمی‌تواند به درستی پاسخ دهد. چرا دانش آموز در درس مردود شد؟ می‌گوییم تقصیر دانشجو بود که درس نخواند.

سناریوی ۲: معلم به ندرت سوالات اضافی می‌پرسد (مثلاً ۲ از ۱۰۰ بار). برای دانش آموزی که برای سوالات مطالعه نکرده است، احتمال گذراندن دوره را زیاد پیش‌بینی می‌کنیم، زیرا سوالات بعید است. البته یکی از دانش آموزان برای سوالات آماده نشد که ۱۰ درصد شانس قبولی در سوالات را به او می‌دهد. او بدشانس است و معلم سوالات اضافی می‌پرسد که دانش آموز نمی‌تواند به آنها پاسخ دهد و در درس مردود می‌شود. دلیل رد شدن چیست؟ من

استدلال می کنم که اکنون، توضیح بهتر این است که "چون معلم دانش آموز را امتحان کرد". بعید بود که معلم امتحان بگیرد، بنابراین معلم رفتار غیرعادی داشت.

معنی آن برای یادگیری ماشین قابل تفسیر چیست: اگر یکی از ویژگی‌های ورودی برای یک پیش‌بینی به هر معنا غیرعادی بود (مانند یک دسته نادر از یک ویژگی طبقه‌بندی شده) و این ویژگی بر پیش‌بینی تأثیر گذاشت، باید در توضیح گنجانده شود، حتی اگر سایر ویژگی‌های «عادی» مشابه باشند. تأثیر بر پیش‌بینی به عنوان یک غیرعادی است. یک ویژگی غیرعادی در مثال پیش‌بینی قیمت خانه ما ممکن است این باشد که یک خانه نسبتاً گران دو بالکن دارد. حتی اگر برخی از روش‌های انتساب نشان دهند که دو بالکن به اندازه اندازه خانه متوسط، همسایگی خوب یا بازسازی اخیر در تفاوت قیمت نقش دارند، ویژگی غیرعادی "دو بالکن" ممکن است بهترین توضیح برای این که چرا خانه چنین گران است.

۵- توضیحات درست است. توضیحات خوب در واقعیت (یعنی در موقعیت‌های دیگر) صادق هستند. اما به طرز نگران‌کننده‌ای، این مهم‌ترین عامل برای توضیح «خوب» نیست. به عنوان مثال، به نظر می‌رسد انتخابی مهم‌تر از صداقت است. توضیحی که فقط یک یا دو علت احتمالی را انتخاب می‌کند، به ندرت کل فهرست علل مرتبط را پوشش می‌دهد. انتخاب بخشی از حقیقت را حذف می‌کند. این درست نیست که مثلاً فقط یک یا دو عامل باعث سقوط بورس شده است. در عوض حقیقت این است که میلیون‌ها علت وجود دارد که میلیون‌ها نفر را تحت تأثیر قرار می‌دهد تا به گونه‌ای عمل کنند که در نهایت باعث سقوط بورس شود.

معنی آن برای یادگیری ماشین قابل تفسیر: توضیح باید رویداد را تا حد امکان صادقانه پیش‌بینی کند، که در یادگیری ماشین گاهی اوقات به آن وفاداری (fidelity) می‌گویند. بنابراین اگر بگوییم که بالکن دوم قیمت یک خانه را افزایش می‌دهد، باید برای خانه‌های دیگر (یا حداقل برای خانه‌های مشابه) نیز صدق کند. برای انسان‌ها، وفاداری یک توضیح به اندازه انتخابی بودنش، مقایسه‌ای بودنش و جنبه اجتماعی آن مهم نیست.

۶- توضیحات خوب با باورهای قبلی توضیح دهنده مطابقت دارد. انسان‌ها تمایل دارند اطلاعاتی را نادیده بگیرند که با باورهای قبلی آنها همخوانی ندارد. این اثر سوگیری تایید (confirmation bias) نامیده می‌شود (Nickerson, 1998). توضیحات از این نوع سوگیری در امان نیستند. مردم تمایل دارند توضیحاتی را که با عقاید آنها همخوانی ندارد بی ارزش بدانند یا نادیده بگیرند. مجموعه باورها از فردی به فرد دیگر متفاوت است، اما باورهای گروهی مانند جهان بینی سیاسی نیز وجود دارد.

معنی آن برای یادگیری ماشین قابل تفسیر چیست: توضیحات خوب با باورهای قبلی سازگار است. ادغام این با یادگیری ماشین دشوار است و احتمالاً عملکرد پیش‌بینی را به شدت به خطر می‌اندازد. اعتقاد قبلی ما برای تأثیر اندازه خانه بر قیمت پیش‌بینی شده این است که هر چه خانه بزرگ‌تر باشد، قیمت بالاتر است. اجازه دهید فرض کنیم که یک مدل همچنین اثر منفی اندازه خانه را بر قیمت پیش‌بینی شده برای چند خانه نشان می‌دهد.

مدل این را یاد گرفته است زیرا عملکرد پیش‌بینی را بهبود می‌بخشد (به دلیل برخی از تعاملات پیچیده)، اما این رفتار به شدت با باورهای قبلی ما در تضاد است. می‌توانید محدودیت‌های یکنواختی (monotonicity) را اعمال کنید (یک ویژگی فقط می‌تواند در یک جهت بر پیش‌بینی تأثیر بگذارد) یا از چیزی مانند یک مدل خطی استفاده کنید که این خاصیت را دارد.

۷- توضیحات خوب کلی و محتمل است. علتی که می‌تواند بسیاری از رویدادها را توضیح دهد بسیار کلی است و می‌تواند توضیح خوبی در نظر گرفته شود. توجه داشته باشید که این با این ادعا که علل غیرعادی توضیحات خوبی ارائه می‌دهند، تناقض دارد. همان‌طور که می‌بینم، علل غیرعادی بر علل عمومی‌غلبه می‌کنند. علل غیرعادی بنا به تعریف در سناریوی داده شده نادر هستند. در صورت عدم وجود یک رویداد غیرعادی، یک توضیح کلی توضیح خوبی در نظر گرفته می‌شود. همچنین به یاد داشته باشید که مردم تمایل دارند احتمالات رویدادهای مشترک را اشتباه ارزیابی کنند. (جو یک کتابدار است. آیا او یک فرد خجالتی است یا یک فرد خجالتی که دوست دارد کتاب بخواند؟) یک مثال خوب این است که "خانه گران است چون بزرگ است" که توضیح بسیار کلی و خوب برای گرانی یا ارزانی خانه‌هاست.

معنی آن برای یادگیری ماشین قابل تفسیر چیست: به طور کلی می‌توان به راحتی با پشتیبانی ویژگی اندازه‌گیری کرد، که تعداد نمونه‌هایی است که توضیح برای آن‌ها اعمال می‌شود تقسیم بر تعداد کل نمونه‌ها.

فصل ۴ مجموعه داده‌ها

در سراسر کتاب، تمام مدل‌ها و تکنیک‌ها بر روی مجموعه داده‌های واقعی که به‌طور رایگان و به‌صورت آنلاین در دسترس هستند، اعمال می‌شوند. ما از مجموعه داده‌های مختلف برای وظایف مختلف استفاده خواهیم کرد: طبقه‌بندی، رگرسیون و طبقه‌بندی متن.

۴.۱ اجاره دوچرخه (رگرسیون)

این مجموعه‌داده شامل تعداد روزانه دوچرخه‌های کرایه شده از شرکت اجاره دوچرخه Capital-Bikeshare در واشنگتن دی سی به همراه آب و هوا و اطلاعات فصلی است. داده‌ها با سخاوت توسط Capital-Bikeshare در دسترس قرار گرفت. (Fanaee-T and Gama 2014) داده‌های آب و هوا و اطلاعات فصل را اضافه کردند. هدف این است که پیش‌بینی کنید بسته به آب و هوا و روز چند دوچرخه اجاره می‌شود. داده‌ها را می‌توان از مخزن یادگیری ماشین UCI دانلود کرد.

ویژگی‌های جدیدی به مجموعه‌داده اضافه شد و همه ویژگی‌های اصلی برای مثال‌های این کتاب استفاده نشده است. در اینجا لیستی از ویژگی‌هایی مورد استفاده، آورده شده است:

تعداد دوچرخه‌های اجاره داده شده به کاربران عادی و ثبت‌نام شده. تعداد دوچرخه‌های اجاره داده شده، به عنوان هدف در کار رگرسیون استفاده می‌شود.
فصل، شامل بهار، تابستان، پاییز یا زمستان.
نشان می‌دهد که آیا روز تعطیل بود یا نه.
سال، ۲۰۱۱ یا ۲۰۱۲.

تعداد روزهای پس از ۱۱.۱۰.۱۰ (اولین روز در مجموعه‌داده). این ویژگی برای در نظر گرفتن روند در طول زمان معرفی شد.

نشان می‌دهد که آیا روز یک روز کاری یا آخر هفته بوده است.

- وضعیت آب و هوا در آن روز. یکی از:
 - صف، کمی‌ابر، نیمه ابری، ابری
 - مه + ابر، مه + ابرهای شکسته، مه + چند ابر، مه
 - برف خفیف، باران خفیف + رعدوبرق + ابرهای پراکنده، باران خفیف + ابرهای پراکنده
 - باران شدید + پالت‌های یخ + رعدوبرق + مه، برف + غبار
- دما بر حسب درجه سانتیگراد.

رطوبت نسبی بر حسب درصد (۰ تا ۱۰۰).

سرعت باد بر حسب کیلومتر در ساعت.

برای مثال‌های این کتاب، داده‌ها کمی پردازش شده است. می‌توانید R-script پردازشی را در مخزن GitHub کتاب به همراه فایل RData نهایی پیدا کنید.

۴.۲ نظرات هرزنامه YouTube (طبقه‌بندی متن)

به عنوان نمونه ای برای طبقه‌بندی متن، ما با ۱۹۵۶ نظر از ۵ ویدیوی مختلف YouTube کار می‌کنیم. خوشبختانه، نویسنده‌گانی که از این مجموعه داده در مقاله ای در مورد طبقه‌بندی هرزنامه استفاده کردند، داده‌ها را به صورت رایگان در دسترس قرار دادند (Alberto et al., 2015).

نظرات از طریق API YouTube از پنج ویدیو از ده ویدیوی پربازدید YouTube در نیمه اول سال ۲۰۱۵ جمع‌آوری شد. هر ۵ مورد، ویدیو موزیک هستند. یکی از آنها Gangnam Style اثر هنرمند کره‌ای Psy است. هنرمندان دیگر Shakira و Eminem، LMFAO، Katy Perry بودند.

برخی از نظرات را بررسی کنید. نظرات به صورت دستی به عنوان هرزنامه یا قانونی برچسب گذاری شدند. هرزنامه با "۱" و نظرات قانونی با "۰" کدگذاری شد.

جدول ۴.۱: نمونه نظرات از مجموعه داده‌های هرزنامه YouTube

محتوا	کلاس
Huh, anyway check out this you [tube] channel: kobyoshi02	۱
Hey guys check out my new channel and our first vid THIS IS US THE MONKEYS!!! I'm the monkey in the white shirt,please leave a like comment and please subscribe!!!!	۱
just for test I have to say murdev.com	۱
me shaking on my channel enjoy ^_^	۱
watch?v=vtaRGgvGtWQ Check this out .	۱
Hey, check out my new website!! This site is about kids stuff. kidsmediausa . com	۱
Subscribe to my channel	۱
i turned it on mute as soon as i came on i just wanted to check the views...	۰
You should check my channel for Funny VIDEOS!!	۱
and u should.d check my channel and tell me what I should do next!	۱

همچنین می‌توانید به یوتیوب بروید و به بخش نظرات نگاهی بیندازید. اما لطفاً در جهنم یوتیوب گرفتار نشوید و در نهایت به تماشی ویدئوهایی از میمون‌ها که در حال دزدیدن و نوشیدن کوکتل از گردشگران در ساحل هستند، بنشینید. آشکارساز هرزنامه گوگل نیز احتمالاً از سال ۲۰۱۵ تغییرات زیادی کرده است. ویدیوی رکورددشکنی "Gangnam Style" را در اینجا تماشا کنید.

اگر می خواهید با داده ها بازی کنید، می توانید فایل RData را به همراه اسکریپت R با برخی عملکردهای راحت در مخزن GitHub کتاب پیدا کنید.

۴.۳ عوامل خطر برای سرطان دهانه رحم (طبقه بندی)

مجموعه داده های سرطان دهانه رحم شامل شاخص ها و عوامل خطر برای پیش بینی اینکه آیا یک زن به سرطان دهانه رحم مبتلا می شود یا خیر. این ویژگی ها شامل داده های جمعیت شناختی (مانند سن)، سبک زندگی و سابقه پزشکی است. داده ها را می توان از مخزن یادگیری ماشین UCI دانلود کرد و توسط Fernandes et al (۲۰۱۷) ارائه شده است.

زیرمجموعه ویژگی های داده استفاده شده در مثال های کتاب عبارت اند از:

سن بر حسب سال

تعداد شرکای جنسی

اولین رابطه جنسی (سن در سال)

تعداد حاملگی ها

سیگار کشیدن بله یا خیر

سیگار کشیدن (در سال)

داروهای ضد بارداری هورمونی بله یا خیر

داروهای ضد بارداری هورمونی (در سال)

دستگاه داخل رحمی بله یا خیر (IUD)

تعداد سالهای استفاده از دستگاه داخل رحمی (IUD)

آیا بیمار تا به حال بیماری مقاربی (STD) داشته است بله یا خیر

تعداد تشخیص های STD

زمان از اولین تشخیص STD

زمان از آخرین تشخیص STD

نتیجه نمونه برداری "سالم" یا "سرطان" است. خروجی هدف.

نمونه برداری به عنوان استاندارد طلایی برای تشخیص سرطان دهانه رحم عمل می کند. برای مثال های این کتاب، نتیجه نمونه برداری به عنوان هدف مورد استفاده قرار گرفت. مقادیر گمشده برای هر ستون با حالت (متداول ترین مقدار) نسبت داده می شود، که احتمالاً راه حل بدی است، زیرا پاسخ واقعی می تواند با احتمال گم شدن یک مقدار مرتبط باشد. احتمالاً سوگیری وجود دارد زیرا سؤالات ماهیت بسیار خصوصی دارند. اما این کتابی در مورد انتساب داده های از دست رفته نیست، بنابراین انتساب حالت باید برای مثال ها کافی باشد.

برای باز تولید نمونه های این کتاب با این مجموعه داده، پیش پردازش R-script و فایل RData نهایی را در مخزن GitHub کتاب پیدا کنید.

فصل ۵ مدل‌های قابل تفسیر

ساده‌ترین راه برای دستیابی به تفسیرپذیری استفاده از زیر مجموعه‌ای از الگوریتم‌هایی است که مدل‌های قابل تفسیر را ایجاد می‌کنند. رگرسیون خطی، رگرسیون لجستیک و درخت تصمیم معمولاً از مدل‌های قابل تفسیر استفاده می‌شوند.

در فصل‌های بعدی در مورد این مدل‌ها صحبت خواهیم کرد. نه در جزئیات، فقط اصول اولیه، زیرا در حال حاضر تعداد زیادی کتاب، فیلم، آموزش، مقالات و مطالب بیشتری در دسترس است. ما بر نحوه تفسیر مدل‌ها تمرکز خواهیم کرد. این کتاب رگرسیون خطی، رگرسیون لجستیک، دیگر پسوندهای رگرسیون خطی، درختان تصمیم، قوانین تصمیم گیری و الگوریتم RuleFit را با جزئیات بیشتری مورد بحث قرار می‌دهد. همچنین سایر مدل‌های قابل تفسیر را فهرست می‌کند.

تمام مدل‌های تفسیرپذیر توضیح داده شده در این کتاب در سطح مدولار قابل تفسیر هستند، به استثنای روش k-nearest-همساخه. جدول زیر یک نمای کلی از انواع مدل‌های قابل تفسیر و ویژگی‌های آنها ارائه می‌دهد. یک مدل خطی است اگر ارتباط بین ویژگی‌ها و هدف به صورت خطی مدل شود. یک مدل با محدودیت‌های یکنواختی تضمین می‌کند که رابطه بین یک ویژگی و نتیجه هدف همیشه در یک جهت در کل محدوده ویژگی پیش می‌رود؛ افزایش در مقدار ویژگی یا همیشه منجر به افزایش یا همیشه به کاهش هدف می‌شود. نتیجه یکنواختی برای تفسیر یک مدل مفید است زیرا درک یک رابطه را آسان‌تر می‌کند. برخی از مدل‌ها می‌توانند به طور خودکار تعامل بین ویژگی‌ها را برای پیش‌بینی نتیجه هدف داشته باشند. می‌توانید با ایجاد دستی ویژگی‌های تعامل، تعاملات را در هر مدلی بگنجانید. فعل و انفعالات می‌توانند عملکرد پیش‌بینی را بهبود بخشنند، اما تعاملات زیاد یا بسیار پیچیده می‌تواند به تفسیرپذیری آسیب برساند. برخی از مدل‌ها فقط رگرسیون، برخی فقط طبقه‌بندی و برخی دیگر هر دو را مدیریت می‌کنند.

از این جدول، می‌توانید یک مدل قابل تفسیر مناسب برای کار خود انتخاب کنید، رگرسیون (regr) یا طبقه‌بندی (کلاس):

وظیفه	کلاس، رگر	کلاس	کلاس، رگر	regr
آثر متقابل	آره	آره	آره	خیر
یکنواخت	آره	آره	آره	خیر
خطی	آره	آره	آره	آره
الگوریتم	رگرسیون خطی	رگرسیون لجستیک	درختان تصمیم	RuleFit

وظیفه	اثر متقابل	یکنواخت	خطی	الگوریتم
کلاس	خیر	آره	خیر	بیز ساده لوح
کلاس، رگر	خیر	خیر	خیر	-کانزدیکترین همسایگان

شما می‌توانید استدلال کنید که هم رگرسیون لجستیک و هم ساده لوحانه، توضیحات خطی را مجاز می‌دانند. با این حال، این فقط برای لگاریتم هدف صادق است: افزایش یک ویژگی به اندازه یک نقطه، لگاریتم احتمال هدف را به میزان معینی افزایش می‌دهد، با فرض ثابت ماندن همه ویژگی‌های دیگر.

۱.۵. رگرسیون خطی

یک مدل رگرسیون خطی، هدف را با استفاده از مجموع وزنی ویژگی‌های ورودی پیش‌بینی می‌کند. خطی بودن رابطه آموزش داده شده، تفسیر را آسان می‌کند. مدل‌های رگرسیون خطی مدت‌هاست که توسط آماردانان، دانشمندان کامپیوتر و سایر افرادی که به مسائل کمی رسیدگی می‌کنند، استفاده می‌شود.

مدل‌های خطی می‌توانند برای مدل سازی وابستگی یک هدف رگرسیونی y به برخی از ویژگی‌های x استفاده شوند. روابط آموزش داده شده خطی هستند و می‌توان آنها را برای نمونه \hat{y} به صورت زیر نوشت:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \epsilon$$

نتیجه پیش‌بینی شده یک نمونه، مجموع وزنی از ویژگی‌های p آن است. بتاهای (β_i) وزن یا ضرایب ویژگی‌های آموزش داده شده را نشان می‌دهد. وزن اول در مجموع (β_0) عرض از مبدا نامیده می‌شود و با ویژگی ضرب نمی‌شود. اپسیلون (ϵ) خطایی است که پیش‌بینی ما دارد، یعنی تفاوت بین پیش‌بینی و نتیجه واقعی. فرض بر این است که این خطاهای از یک توزیع گاووسی پیروی می‌کنند، به این معنی که ما در هر دو جهت منفی و مثبت خطایی کنیم و بسیاری از خطاهای کوچک و تعداد کمی از خطاهای بزرگ را انجام می‌دهیم.

برای تخمین وزن بهینه می‌توان از روش‌های مختلفی استفاده کرد. عمدها از روش حداقل مربعات معمولی، برای یافتن وزن‌هایی استفاده می‌شود، که اختلاف محدود بین نتایج واقعی و برآورد شده را به حداقل می‌رسانند:

$$\beta = \arg \min_{\beta_0 \dots \beta_p} \sum_{i=1}^n \left(y^{(i)} - \left(\beta_0 + \sum_{j=1}^p \beta_j x_j^{(i)} \right) \right)^2$$

ما در مورد چگونگی یافتن وزن‌های بهینه به تفصیل بحث نخواهیم کرد، اما اگر علاقه‌مند هستید، می‌توانید فصل ۳.۲ کتاب "عناصر یادگیری آماری" (Hastie, 2009) یا یکی دیگر از کتاب‌های آنلاین در مورد مدل‌های رگرسیون خطی را مطالعه کنید.

بزرگ‌ترین مزیت مدل‌های رگرسیون خطی، خطی بودن است: این روش تخمین را ساده می‌کند و مهم‌تر از همه، این معادلات خطی تفسیر آسانی در سطح مدولار (یعنی وزن‌ها) دارند. این یکی از دلایل اصلی گسترش مدل خطی و همه مدل‌های مشابه در زمینه‌های دانشگاهی مانند پزشکی، جامعه‌شناسی، روان‌شناسی و بسیاری دیگر از زمینه‌های تحقیقاتی کمی است. به عنوان مثال، در زمینه پزشکی، نه تنها پیش‌بینی نتیجه بالینی یک بیمار مهم است، بلکه تعیین کمیت تأثیر دارو و در عین حال در نظر گرفتن جنسیت، سن و سایر ویژگی‌ها به روشنی قابل تفسیر است نیز اهمیت دارد.

وزن‌های تخمینی با فواصل اطمینان همراه است. فاصله اطمینان محدوده‌ای برای تخمین وزن است که وزن «واقعی» را با اطمینان خاصی پوشش می‌دهد. به عنوان مثال، فاصله اطمینان ۹۵٪ برای وزن ۲ می‌تواند از ۱ تا ۳ متغیر باشد. تفسیر این فاصله به این صورت خواهد بود: اگر تخمین را ۱۰۰ بار با داده‌های نمونه گیری جدید تکرار کنیم، فاصله اطمینان در ۹۵ مورد از ۱۰۰ مورد شامل وزن واقعی می‌شود، با فرض اینکه مدل رگرسیون خطی، مدل درست داده‌ها باشد.

اینکه آیا مدل، مدل «درست» است بستگی به این دارد که آیا روابط موجود در داده‌ها مفروضات خاصی را برآورده می‌کنند که این مفروضات عبارت‌اند از خطی بودن، نرمال بودن، همسانی، استقلال، ویژگی‌های ثابت و عدم وجود چند خطی.

خطی بودن

مدل رگرسیون خطی، پیش‌بینی را مجبور می‌کند که ترکیبی خطی از ویژگی‌ها باشد، که هم بزرگ‌ترین مزیت و هم بزرگ‌ترین محدودیت آن است. خطی بودن منجر به مدل‌های قابل تفسیر می‌شود. کمی کردن و توصیف اثرات خطی آسان است. آنها افزودنی هستند، بنابراین به راحتی می‌توان اثرات را از هم تفکیک کرد. اگر اثرات متقابل ویژگی یا ارتباط غیرخطی یک ویژگی با مقدار هدف مشکوک هستید، می‌توانید عبارات تعامل را اضافه کنید یا از اسپیلاین‌های رگرسیون استفاده کنید.

نرمال بودن

فرض بر این است که هدف، ویژگی‌هایی مشخص، از توزیع نرمال پیروی می‌کند. اگر این فرض نقض شود، فواصل اطمینان تخمینی، وزن ویژگی‌ها، نامعتبر است.

همسانی (واریانس ثابت)

واریانس عبارات خطی، در کل فضای ویژگی ثابت فرض می‌شود. فرض کنید می‌خواهید ارزش یک خانه را با توجه به مساحت نشیمن بر حسب متر مربع پیش‌بینی کنید. شما یک مدل خطی را تخمین می‌زنید که فرض می‌کند، صرف نظر از اندازه خانه، خطای در اطراف پاسخ پیش‌بینی شده واریانس یکسانی دارد. این فرض اغلب در واقعیت نقض می‌شود. در مثال خانه، قابل قبول است که واریانس شرایط خطای در اطراف قیمت پیش‌بینی شده برای

خانه‌های بزرگ‌تر بیشتر باشد، زیرا قیمت‌ها بالاتر هستند و فضای بیشتری برای نوسانات قیمت وجود دارد. فرض کنید میانگین خطأ (تفاوت بین قیمت پیش‌بینی شده و واقعی) در مدل رگرسیون خطی شما ۵۰۰۰۰ یورو باشد. اگر همسان بودن را فرض کنید، فرض می‌کنید که میانگین خطأ ۵۰۰۰۰ برای خانه‌هایی که ۱ میلیون قیمت دارند و برای خانه‌هایی که فقط ۴۰۰۰۰ قیمت دارند یکسان است.

مستقل بودن

فرض بر این است که هر نمونه مستقل از هر نمونه دیگری است. اگر اندازه‌گیری‌های مکرر را انجام دهید، مانند آزمایش‌های خون متعدد برای هر بیمار، نقاط داده مستقل نیستند. برای داده‌های وابسته به مدل‌های رگرسیون خطی خاص، مانند مدل‌های اثر مختلط یا GEE نیاز دارید. اگر از مدل رگرسیون خطی "عادی" استفاده می‌کنید، ممکن است نتیجه گیری اشتباهی از مدل بگیرید.

ویژگی‌های ثابت

ویژگی‌های ورودی "ثابت" در نظر گرفته می‌شوند. ثابت به این معنی است که آنها به عنوان "ثابت داده شده" و نه به عنوان متغیرهای آماری در نظر گرفته می‌شوند. این بدان معناست که آنها قادر خطاها را اندازه‌گیری هستند. این یک فرض نسبتاً غیر واقعی است. با این حال، بدون این فرض، شما باید مدل‌های خطای اندازه‌گیری بسیار پیچیده‌ای را که خطاهای اندازه‌گیری ویژگی‌های ورودی شما را محاسبه می‌کنند، برآش دهید. و معمولاً شما نمی‌خواهید این کار را انجام دهید.

فقدان چند خطی

شما ویژگی‌های قوی همبسته را نمی‌خواهید، زیرا این تخمین وزن‌ها را به هم می‌زنند. در شرایطی که دو ویژگی به شدت همبستگی دارند، تخمین وزن‌ها مشکل ساز می‌شود، زیرا اثرات ویژگی افزایشی هستند و غیرقابل تعیین می‌شود که به کدام یک از ویژگی‌های همبسته نسبت داده شود.

۵.۱.۱ تفسیر

تفسیر وزن در مدل رگرسیون خطی به نوع ویژگی مربوطه بستگی دارد.

ویژگی عددی: افزایش ویژگی عددی به اندازه یک واحد، نتیجه تخمینی را به اندازه وزن آن ویژگی تغییر می‌دهد. یک مثال از یک ویژگی عددی اندازه یک خانه است.

ویژگی باینری: ویژگی که یکی از دو مقدار ممکن را برای هر نمونه می‌گیرد. به عنوان مثال ویژگی "خانه همراه با یک باغ" است. یکی از مقادیر به عنوان دسته مرجع (در برخی از زبان‌های برنامه نویسی که با کدگذاری شده‌اند) به حساب می‌آید، مانند "بدون باغ". تغییر ویژگی از دسته مرجع به دسته دیگر، نتیجه تخمینی را بر اساس وزن ویژگی تغییر می‌دهد.

ویژگی طبقه‌بندی با دسته‌های متعدد: ویژگی با تعداد ثابتی از مقادیر ممکن. به عنوان مثال ویژگی «نوع کف» با دسته‌های احتمالی «فرش»، «لمینت» و «پارکت» است. یک راه حل برای مقابله با بسیاری از دسته‌ها، رمزگذاری یک گرم است، به این معنی که هر دسته دارای ستون باینری خاص خود است. برای یک ویژگی طبقه‌بندی با دسته‌های L ، شما فقط به ستون‌های $1-L$ نیاز دارید، زیرا ستون $-L$ امین اطلاعات اضافی دارد (به عنوان مثال وقتی ستون‌های 1 تا $-L$ همه دارای مقدار 0 برای یک مثال هستند، می‌دانیم که ویژگی طبقه‌بندی این نمونه در رد L قرار می‌گیرد. سپس تفسیر برای هر دسته مانند تفسیر ویژگی‌های باینری است. برخی از زبان‌ها، مانند R، به شما امکان می‌دهند تا ویژگی‌های دسته‌بندی را به روش‌های مختلف رمزگذاری کنید، همان‌طور که در ادامه این فصل توضیح داده شد.

عرض از مبدا β_0 : عرض از مبدا، وزن ویژگی برای "ویژگی ثابت" است که همیشه برای همه موارد ۱ است. اکثر بسته‌های نرم افزاری به طور خودکار این ویژگی "۱" را برای تخمین عرض از مبدا اضافه می‌کنند. تفسیر این است: برای مثال، با تمام مقادیر ویژگی‌های عددی در صفر و مقادیر ویژگی‌های طبقه‌بندی شده در دسته‌های مرجع، پیش‌بینی مدل وزن عرض از مبدا است. تفسیر عرض از مبدا معمولاً مطرح نیست، زیرا نمونه‌هایی با مقادیر همه ویژگی‌ها در صفر اغلب معنی ندارند. تفسیر تنها زمانی معنادار است که ویژگی‌ها استاندارد شده باشند (میانگین صفر، انحراف معیار یک). در این حالت، عرض از مبدا، نتیجه پیش‌بینی شده نمونه‌ای را منعکس می‌کند که در آن همه ویژگی‌ها در مقدار میانگین خود هستند. تفسیر ویژگی‌ها در مدل رگرسیون خطی را می‌توان با استفاده از الگوهای متنی زیر خودکار کرد.

تفسیر یک ویژگی عددی

با افزایش ویژگی x_k به اندازه یک واحد، پیش‌بینی y را β_k واحد افزایش می‌دهد، زمانی که تمام مقادیر ویژگی‌های دیگر ثابت باقی می‌ماند.

تفسیر یک ویژگی طبقه‌بندی شده

تغییر ویژگی x_k از دسته مرجع به دسته دیگر، پیش‌بینی y را β_k واحد افزایش می‌دهد زمانی که تمام ویژگی‌های دیگر ثابت می‌مانند.

اندازه گیری مهم دیگر برای تفسیر مدل‌های خطی، اندازه گیری R-squared است. R-squared به شما می‌گوید که چه مقدار از واریانس کل نتیجه هدف شما توسط مدل توضیح داده شده است. هرچه R-squared بالاتر باشد، مدل شما داده‌ها را بهتر توضیح می‌دهد. فرمول محاسبه R-squared به صورت زیر است:

$$R^2 = 1 - SSE/SST$$

SSE مجموع مجذور عبارات خطأ است:

$$SSE = \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2$$

SST مجموع مجذور واریانس داده است:

$$SSE = \sum_{i=1}^n (y^{(i)} - \hat{y})^2$$

SSE به شما می‌گوید که پس از برآش مدل خطی، چقدر واریانس باقی می‌ماند، که با اختلاف مجذور بین مقادیر هدف پیش‌بینی شده و واقعی اندازه گیری می‌شود SST. واریانس کل نتیجه هدف است R-squared. به شما می‌گوید که چه مقدار از واریانس شما را می‌توان با مدل خطی توضیح داد R-squared. معمولاً بین ۰ برای مدل‌هایی که مدل اصلًا داده‌ها را توضیح نمی‌دهد و ۱ برای مدل‌هایی که تمام واریانس داده‌های شما را توضیح می‌دهند، متغیر است. همچنین ممکن است R-squared بدون نقض قوانین ریاضی یک مقدار منفی به خود بگیرد. این زمانی اتفاق می‌افتد که SSE بزرگ‌تر از SST باشد، به این معنی که یک مدل روند داده‌ها را نمی‌گیرد و بدتر از استفاده از میانگین هدف به عنوان پیش‌بینی با داده‌ها مطابقت دارد.

یک نکته وجود دارد، زیرا R-squared با تعداد ویژگی‌های مدل افزایش می‌یابد، حتی اگر اصلًا حاوی اطلاعاتی در مورد مقدار هدف نباشند. بنابراین، بهتر است از R-squared تنظیم شده استفاده کنید که تعداد ویژگی‌های استفاده شده در مدل را به حساب می‌آورد. محاسبه آن این است:

$$R^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1}$$

که در آن p تعداد ویژگی‌ها و n تعداد نمونه‌ها است.

تفسیر یک مدل با R-squared (تعدیل شده) بسیار کم معنی دار نیست، زیرا چنین مدلی اساساً واریانس زیادی را توضیح نمی‌دهد. هر گونه تفسیری از اوزان معنادار نخواهد بود.

اهمیت ویژگی

اهمیت یک ویژگی در مدل رگرسیون خطی را می‌توان با قدر مطلق آماره t اندازه گیری کرد. آماره t وزن تخمین زده شده با خطای استاندارد آن است.

$$t_{\beta_j} = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

اجازه دهید بررسی کنیم که این فرمول به ما چه می‌گوید: اهمیت یک ویژگی با افزایش وزن افزایش می‌یابد. این منطقی است. هر چه وزن تخمینی واریانس بیشتری داشته باشد (= هر چه نسبت به مقدار صحیح اطمینان کمتری داشته باشیم)، اهمیت ویژگی کمتر است. این نیز منطقی است.

۵.۱.۲ مثال

در این مثال، ما از مدل رگرسیون خطی برای پیش‌بینی تعداد دوچرخه‌های اجاره‌ای در یک روز خاص، با توجه به اطلاعات آب و هوا و تقویم استفاده می‌کنیم. برای تفسیر، وزن‌های رگرسیون برآورده شده را بررسی می‌کنیم.

ویژگی‌های عددی و طبقه‌ای تشکیل شده است. برای هر ویژگی، جدول وزن تخمینی، خطای استاندارد تخمین (SE) و قدر مطلق آماره ($|t|$) را نشان می‌دهد.

	Weight	SE	$ t $
(Intercept)	2399.4	238.3	10.1
seasonSPRING	899.3	122.3	7.4
seasonSUMMER	138.2	161.7	0.9
seasonFALL	425.6	110.8	3.8
holidayHOLIDAY	-686.1	203.3	3.4
workingdayWORKING DAY	124.9	73.3	1.7
weathersitMISTY	-379.4	87.6	4.3
weathersitRAIN/SNOW/STORM	-1901.5	223.6	8.5
temp	110.7	7	15.7
hum	-17.4	3.2	5.5
windspeed	-42.5	6.9	6.2
days_since_2011	4.9	0.2	28.5

تفسیر یک ویژگی عددی (دما): افزایش دما به میزان ۱ درجه سانتیگراد، تعداد پیش‌بینی‌شده دوچرخه‌ها را تا ۱۱۰.۷ افزایش می‌دهد، زمانی که سایر ویژگی‌ها ثابت می‌مانند.

تفسیر یک ویژگی طبقه‌بندی شده ("Weathersit") تعداد تخمینی دوچرخه‌ها در هنگام باران، برف یا طوفان - ۱۹۰.۵ کمتر است، در مقایسه با آب و هوای خوب - دوباره با فرض اینکه همه ویژگی‌های دیگر تغییر نمی‌کنند. وقتی هوا مه آلود است، با توجه به ثابت ماندن سایر ویژگی‌ها، تعداد دوچرخه‌های پیش‌بینی شده منفی ۳۷۹.۴ در مقایسه با هوای خوب کمتر است.

همه تفاسیر همیشه با ذکر این نکته همراه می‌شوند که "همه ویژگی‌های دیگر ثابت می‌مانند". این به دلیل ماهیت مدل‌های رگرسیون خطی است. هدف پیش‌بینی شده ترکیبی خطی از ویژگی‌های وزنی است. معادله خطی برآورده شده یک ابر صفحه در فضای ویژگی/هدف است (یک خط ساده در مورد یک ویژگی واحد). وزن‌ها شبیه (گرادیان) ابر صفحه را در هر جهت مشخص می‌کنند. جنبه خوب این است که تفسیر یک اثر افزودنی یک ویژگی بخصوص از سایر ویژگی‌های جداست. این امکان‌پذیر است زیرا تمام تاثیرات ویژگی (= وزن ضربدر مقدار ویژگی) در معادله با یک مثبت ترکیب می‌شوند. در جنبه بد، تفسیر توزیع مشترک ویژگی‌ها را نادیده می‌گیرد. افزایش یک ویژگی،

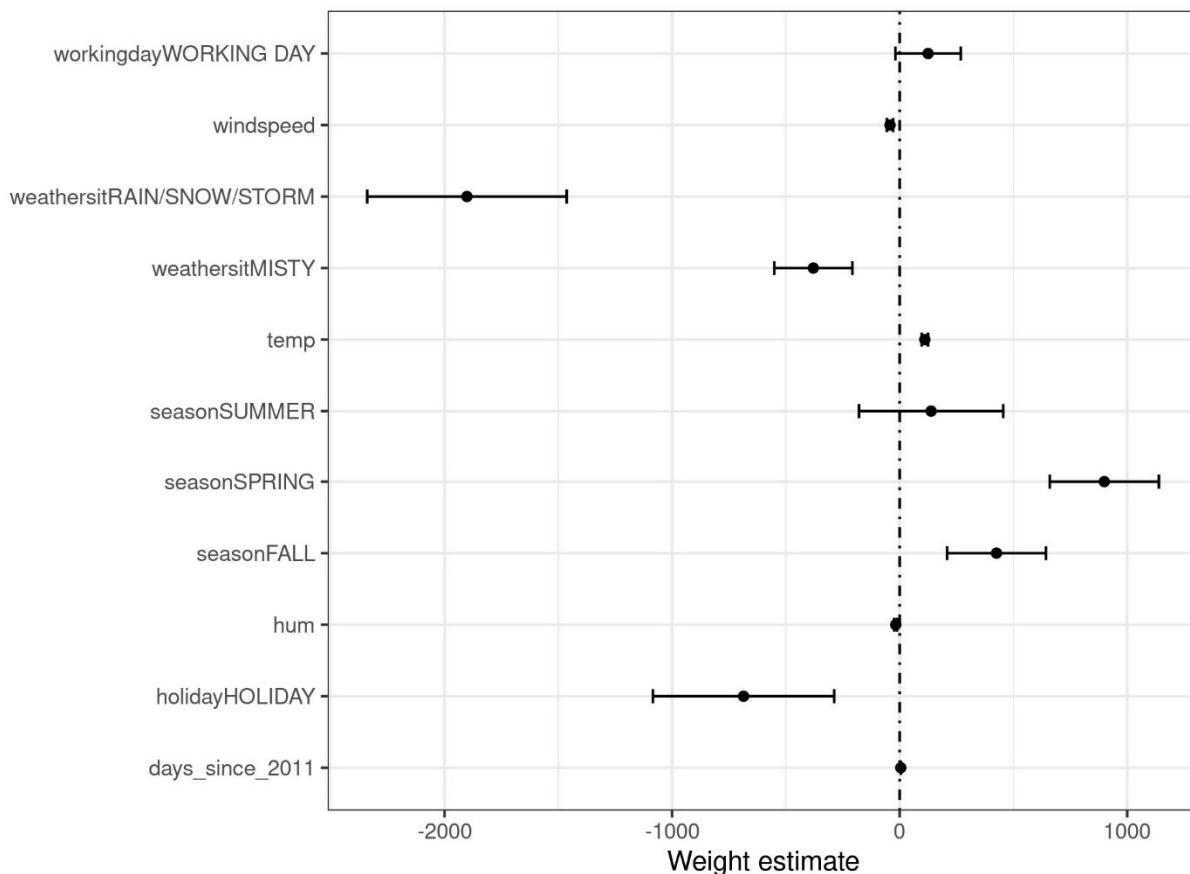
اما عدم تغییر ویژگی دیگر، می‌تواند به نقاط داده غیر واقعی یا حداقل بعيد منجر شود. به عنوان مثال افزایش تعداد اتاق‌ها بدون افزایش اندازه خانه ممکن است غیرواقعی باشد.

۵.۱.۳ تفسیر بصری

گرافیک‌های مختلف، مدل رگرسیون خطی را برای انسان آسان و سریع درک می‌کند.

۵.۱.۳.۱ نمودار وزن

اطلاعات جدول وزنی (تخمین وزن و واریانس) را می‌توان در نمودار وزنی مشاهده کرد. نمودار زیر نتایج حاصل از مدل رگرسیون خطی قبلی را نشان می‌دهد.



شکل ۱.۵: وزن‌ها به صورت نقاط و فاصله‌های اطمینان ۹۵ درصد به صورت خطوط نمایش داده می‌شوند. نمودار وزن نشان می‌دهد که هوای بارانی/برفی/اطوفانی تأثیر منفی قوی بر تعداد پیش‌بینی شده دوچرخه دارد. وزن ویژگی روز کاری نزدیک به صفر است و صفر در بازه ۹۵٪ لحاظ شده است که به این معنی است که اثر از نظر آماری معنی دار نیست. برخی از فواصل اطمینان بسیار کوتاه و برآوردها نزدیک به صفر هستند، با این حال اثرات ویژگی از نظر آماری معنی دار بود. دما یکی از این نامزدها است. مشکل نمودار وزن این است که ویژگی‌ها در مقیاس‌های مختلف اندازه گیری می‌شوند. در حالی که برای آب و هوا، وزن تخمینی تفاوت بین هوای خوب و

بارانی/طوفانی/برفی را نشان می‌دهد، برای دما فقط افزایش ۱ درجه سانتی‌گراد را نشان می‌دهد. قبل از برازش مدل خطی، می‌توانید وزن‌های تخمینی را با مقیاس بندی ویژگی‌ها (میانگین صفر و انحراف استاندارد یک) قابل مقایسه تر کنید.

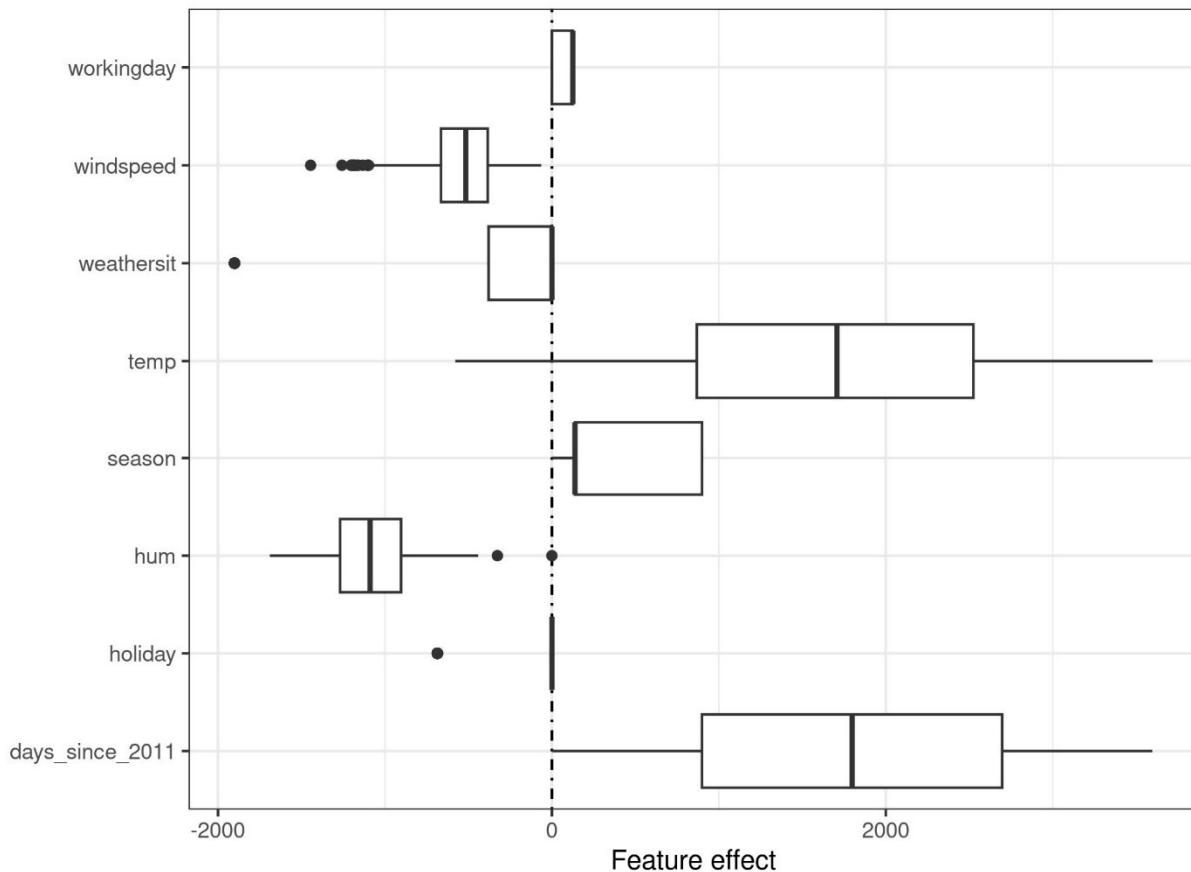
۵.۱.۳.۲ نمودار اثر

وزن‌های مدل رگرسیون خطی زمانی که در مقادیر واقعی ضرب شوند می‌توانند به طور معنی داری تحلیل شوند. وزن‌ها به مقیاس ویژگی‌ها بستگی دارد و اگر ویژگی‌ای داشته باشد که مثلاً قد یک فرد را اندازه‌گیری می‌کند و از متر به سانتی‌متر تغییر می‌دهید، متفاوت خواهد بود. وزن تغییر خواهد کرد، اما اثرات واقعی در داده‌های شما تغییر نخواهد کرد. همچنانی دانستن توزیع ویژگی خود در داده‌ها مهم است، زیرا اگر واریانس بسیار پایینی دارید، به این معنی است که تقریباً همه نمونه‌ها سهم مشابهی از این ویژگی دارند. نمودار اثر می‌تواند به شما کمک کند تا بفهمید که ترکیب وزن و ویژگی چقدر به پیش‌بینی‌های داده‌های شما کمک می‌کند. با محاسبه اثرات شروع کنید، که وزن هر ویژگی ضربدر مقدار ویژگی یک نمونه است:

$$effect_j^{(i)} = w_j x_j^{(i)}$$

اثرات را می‌توان با نمودارهای باکس پلات مستطیلی تجسم کرد. مستطیل در یک باکس پلات شامل محدوده اثر برای نیمی از داده‌ها (۲۵٪ تا ۷۵٪ چندک اثر). خط عمودی در کادر، اثر میانه است، یعنی ۵۰ درصد از نمونه‌ها تأثیر کمتر و نیمی‌دیگر تأثیر بیشتری بر پیش‌بینی دارند. نقاط پرت، نقاطی که بیش از ۱.۵ برابر دامنه بین چارکی (یعنی تفاوت بین ربع اول و سوم) بالای ربع سوم، یا کمتر از ۱.۵ برابر بین چارکی زیر چارک اول تعريف می‌شوند. دو خط افقی که whiskers پایینی و بالایی نامیده می‌شوند، نقاط زیر چارک اول و بالایی چارک سوم را که پرت نیستند به هم متصل می‌کنند. اگر نقاط پرت وجود نداشته باشد، whiskers به مقادیر حداقل و حداکثر گسترش می‌یابند.

اثرات طبقه‌بندی ویژگی را می‌توان در یک باکس پلات خلاصه کرد، در مقایسه با نمودار وزن، که در آن هر دسته ردیف خاص خود را دارد.



شکل ۵.۲: نمودار اثر ویژگی، توزیع اثرات (= ارزش ویژگی ضربدر وزن ویژگی) را در بین داده‌ها در هر ویژگی نشان می‌دهد.

بیشترین سهم در تعداد مورد انتظار دوچرخه‌های اجاره‌ای مربوط به ویژگی دما و ویژگی روز است که روند اجره دوچرخه را در طول زمان نشان می‌دهد. دما دامنه وسیعی از مشارکت در پیش‌بینی دارد. ویژگی روند روز از صفر به مقادیر مثبت بزرگ افزایش است، زیرا اولین روز در مجموعه‌داده (۱۰۱.۲۰۱۱) تأثیر روند بسیار کمی دارد و وزن تخمینی برای این ویژگی مثبت است (۴.۹۳). این به این معنی است که اثر با هر روز افزایش می‌یابد و برای آخرین روز در مجموعه‌داده (۳۱.۱۲.۲۰۱۲) دارای بالاترین میزان است. توجه داشته باشید که برای اثراتی با وزن منفی، نمونه‌هایی با اثر مثبت آنهاست هستند که دارای ارزش ویژگی منفی هستند. به عنوان مثال، روزهایی که سرعت باد دارای اثر منفی زیاد است، روزهایی هستند که سرعت باد زیاد است.

۵.۱.۴ پیش‌بینی‌های فردی را توضیح دهید

هر یک از ویژگی‌های یک نمونه چقدر در پیش‌بینی کمک کرده است؟ این را می‌توان با محاسبه اثرات برای این مثال پاسخ داد. تفسیر اثرات خاص نمونه فقط در مقایسه با توزیع اثر برای هر ویژگی منطقی است. ما می‌خواهیم

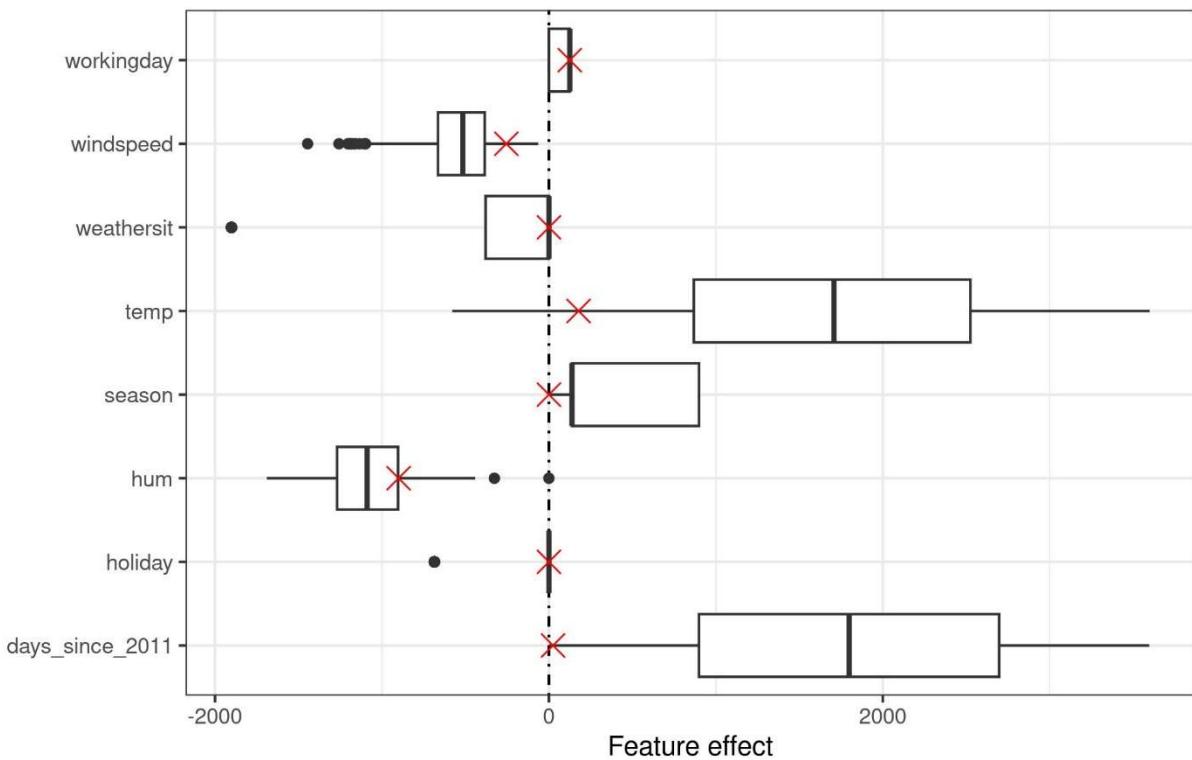
پیش‌بینی مدل خطی را برای نمونه ششم از مجموعه‌داده دوچرخه توضیح دهیم. نمونه دارای مقادیر ویژگی زیر است.

جدول ۵.۱: مقادیر ویژگی برای نمونه ۶

Feature	Value
season	WINTER
yr	2011
mnth	JAN
holiday	NO HOLIDAY
weekday	THU
workingday	WORKING DAY
weathersit	GOOD
temp	1.604356
hum	51.8261
windspeed	6.000868
cnt	1606
days_since_2011	5

برای به دست آوردن اثرات ویژگی این نمونه، باید مقادیر ویژگی آن را در وزن‌های آن ویژگی در مدل رگرسیون خطی ضرب کنیم. برای مقدار "روز کاری" ویژگی "روز کاری"، اثر ۱۲۴.۹ است. برای دمای ۱.۶ درجه سانتیگراد، اثر ۱۷۷.۶ است. ما این اثرات را به صورت علامت ضربدر به نمودار اثرات اضافه می‌کنیم، که توزیع اثرات در داده‌ها را به ما نشان می‌دهد. این کار به ما اجازه می‌دهد تا اثرات فردی را با توزیع اثرات در داده‌ها مقایسه کنیم.

Predicted value for instance: 1571
 Average predicted value: 4504
 Actual value: 1606



شکل ۵.۳: نمودار اثر برای یک نمونه، توزیع اثر و اثرات نمونه مورد نظر را نشان می‌دهد.

اگر ما از پیش‌بینی نمونه‌های داده‌های آموزشی میانگین بگیریم، عدد ۴۵۰۴ به دست می‌آید. در مقایسه با این میانگین، پیش‌بینی نمونه ششم کوچک است، زیرا تنها ۱۵۷۱ کرایه دوچرخه پیش‌بینی شده است. نمودار اثر دلیل آن را نشان می‌دهد. باکس پلات‌ها، اثرات را برای همه نمونه‌های مجموعه‌داده نشان می‌دهند، علامت‌های ضربدر اثرات را برای نمونه ۶ نشان می‌دهند. نمونه ششم تأثیر دمای پایینی دارد زیرا در این روز دما ۲ درجه بود که در مقایسه با اکثر روزهای دیگر پایین است (و به یاد داشته باشید که وزن ویژگی دما مثبت است). همچنین تأثیر ویژگی روند «days_since_2011» در مقایسه با سایر نمونه‌های داده کم است، زیرا این نمونه مربوط به اوایل سال ۲۰۱۱ (۵ روز) است و ویژگی روند نیز وزن مثبتی دارد.

۵.۱.۵ رمزگذاری ویژگی‌های دسته بندی
 راههای مختلفی برای رمزگذاری یک ویژگی طبقه‌بندی وجود دارد و انتخاب هر حالت، بر تفسیر وزن‌ها تأثیر می‌گذارد.

استاندارد در مدل‌های رگرسیون خطی، کدگذاری تیمار است که در اکثر موارد مناسب است. استفاده از رمزگذاری‌های مختلف منجر به ایجاد ماتریس‌های مختلف (طراحی) از یک ستون واحد با ویژگی طبقه‌بندی

می‌شود. این بخش سه کدگذاری مختلف را ارائه می‌کند، اما تعداد بیشتری وجود دارد. مثال مورد استفاده دارای شش نمونه و یک ویژگی طبقه‌بندی شده با سه دسته است. برای دو نمونه اول، ویژگی دسته A را می‌گیرد، برای نمونه سه و چهار، دسته B، و برای دو مورد آخر، دسته C.

کدگذاری تیمار

در کدگذاری تیمار، وزن هر دسته، تفاوت برآورد شده در پیش‌بینی بین دسته مربوطه و دسته مرجع است. عرض از مبدا مدل خطی، میانگین دسته مرجع است (زمانی که سایر ویژگی‌ها ثابت می‌مانند). ستون اول ماتریس طراحی، عرض از مبدا است که همیشه ۱ است. ستون دو نشان می‌دهد که آیا نمونه در رد B قرار دارد یا خیر، ستون سه نشان می‌دهد که آیا در دسته C قرار دارد یا خیر. برای دسته A نیازی به ستون نیست، زیرا پس از آن معادله خطی بیش از حد مشخص می‌شود و هیچ راه حل منحصر به فردی برای وزن‌ها نمی‌توان یافت. کافی است بدانیم که یک نمونه در رد B یا C نیست.

ماتریس ویژگی:

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

کدگذاری اثر

وزن هر دسته، تفاوت تخمینی y از دسته مربوطه به میانگین کلی است (باتوجه به اینکه همه ویژگی‌های دیگر صفر هستند یا دسته مرجع). ستون اول برای تخمین فاصله استفاده می‌شود. وزن β مرتبط با رهگیری نشان دهنده میانگین کلی و β ، وزن ستون دو، تفاوت بین میانگین کلی و دسته B است. اثر کل دسته B است β تفسیر دسته C معادل است. برای رد مرجع A، تفاوت به میانگین کلی است و β اثر کلی

ماتریس ویژگی:

$$\begin{pmatrix} 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

کدنویسی مصنوعی

این β در هر دسته، میانگین تخمینی مقدار y برای هر دسته است (باتوجه به اینکه تمام مقادیر ویژگی‌های دیگر صفر هستند یا دسته مرجع). توجه داشته باشید که وقفه در اینجا حذف شده است تا بتوان یک راه حل منحصر

به فرد برای وزن‌های مدل خطی پیدا کرد. راه دیگر برای کاهش این مشکل چند خطی، کنار گذاشتن یکی از دسته‌بندی‌ها است.

ماتریس ویژگی:

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

اگر می‌خواهید کمی عمیق‌تر در رمزگذاری‌های مختلف ویژگی‌های طبقه‌بندی شده مطالعه کنید، این صفحه و ب (<https://stats.oarc.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/>) و این پست وبلاگ (<http://heidiseibold.github.io/page7/>) را بررسی کنید.

۵.۱.۶ آیا مدل‌های خطی توضیحات خوبی ایجاد می‌کنند؟

با قضاوت بر اساس شزايطی که توضیح خوب دارد، همان‌طور که در فصل توضیحات انسان پسند ارائه شده است، مدل‌های خطی بهترین توضیحات را ایجاد نمی‌کنند. مدل‌های خطی مقایسه‌ای هستند، اما نمونه مرجع یک نقطه داده است که در آن همه ویژگی‌های عددی صفر هستند و ویژگی‌های طبقه‌بندی در دسته‌های مرجع خود قرار دارند. این معمولاً یک نمونه مصنوعی و بی معنی است که بعید است در داده‌های شما یا واقعیت رخ دهد. یک استثنا وجود دارد: اگر همه ویژگی‌های عددی در مرکز میانگین باشند (ویژگی منهای میانگین ویژگی) و همه ویژگی‌های طبقه‌بندی با کدگذاری اثر آورده شده باشند، نمونه مرجع نقطه داده‌ای است که در آن همه ویژگی‌ها مقدار میانگین ویژگی را می‌گیرند. در این حالت نیز ممکن است یک نقطه داده وجود نداشته باشد، اما حداقل ممکن است محتمل‌تر یا معنادار‌تر باشد. در این مورد، وزن‌ها ضربدر مقادیر ویژگی (اثرات ویژگی) سهم ویژگی را در نتیجه پیش‌بینی‌شده در مقایسه با «میانگین نمونه» توضیح می‌دهند. یکی دیگر از جنبه‌های توضیح خوب، انتخاب پذیری است که در مدل‌های خطی با استفاده از ویژگی‌های کمتر یا با آموزش مدل‌های خطی پراکنده می‌توان به آن دست یافت. اما به طور پیش‌فرض، مدل‌های خطی توضیحات انتخابی ایجاد نمی‌کنند. مدل‌های خطی توضیحات درستی ایجاد می‌کنند، تا زمانی که معادله خطی مدل مناسبی برای رابطه بین ویژگی‌ها و نتیجه باشد. هر چه رفتارهای غیر خطی و تعاملات بیشتر باشد، دقت مدل خطی کمتر خواهد بود و توضیحات کمتر صادق می‌باشند. خطی بودن توضیحات را کلی‌تر و ساده‌تر می‌کند. من معتقدم ماهیت خطی مدل، عامل اصلی استفاده از مدل‌های خطی برای توضیح روابط است.

۵.۱.۷ مدل‌های خطی محدود

نمونه‌های مدل‌های خطی که من انتخاب کرده‌ام همگی زیبا و مرتب هستند، اینطور نیست؟ اما در واقعیت ممکن است شما فقط چند ویژگی نداشته باشید، بلکه صدها یا هزاران ویژگی را داشته باشید. و همچنین مدل‌های

رگرسیون خطی شما؟ تفسیرپذیری به سراسیبی می‌رود. حتی ممکن است در موقعیتی قرار بگیرید که ویژگی‌های بیشتری نسبت به نمونه‌ها وجود دارد و اصلاً نمی‌توانید یک مدل خطی استاندارد را برآش دهید. خبر خوب این است که راههایی برای معرفی محدودیت (= چند ویژگی) در مدل‌های خطی وجود دارد.

۵.۱.۷.۱ لاسو

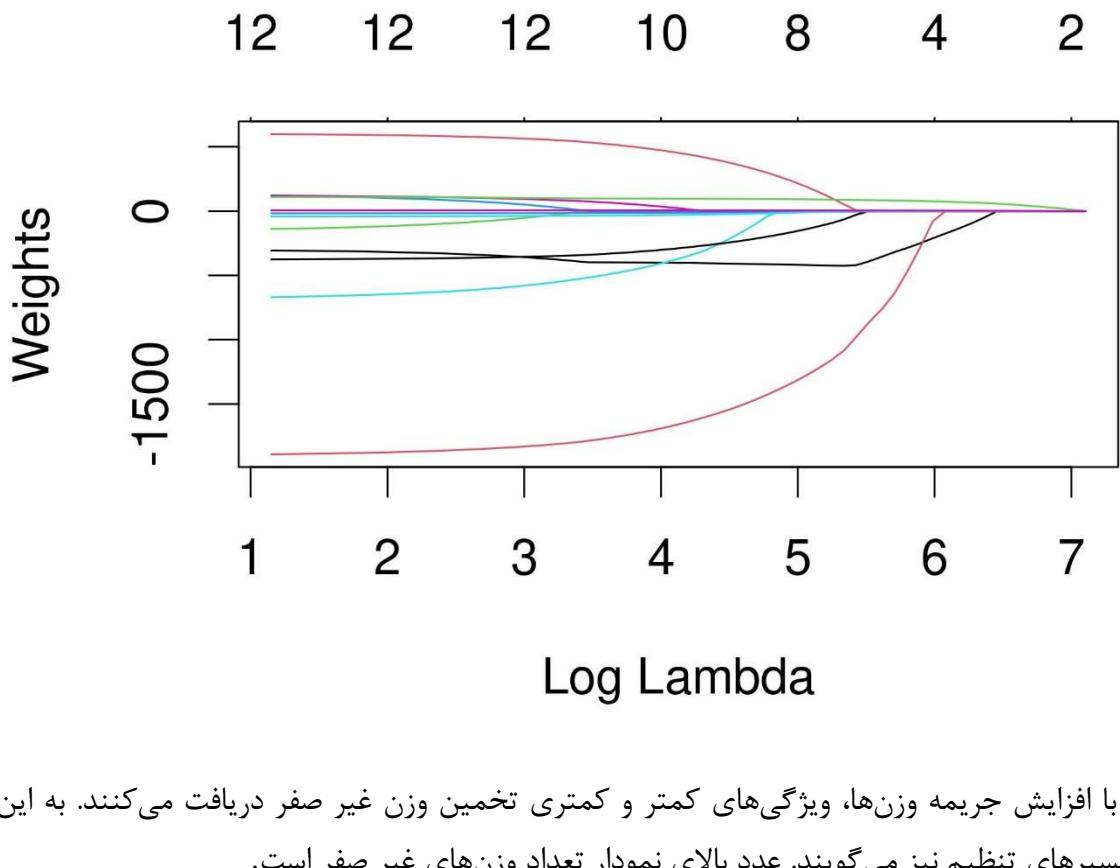
لاسو یک راه خودکار و راحت برای اعمال محدودیت به مدل رگرسیون خطی است. Lasso مخفف حداقل انقباض مطلق و عملگر انتخاب است و هنگامی که در یک مدل رگرسیون خطی اعمال می‌شود، وظیفه انتخاب ویژگی و تنظیم وزن ویژگی انتخاب شده را انجام می‌دهد. اجازه دهید مساله کمینه‌سازی، که وزن‌ها را بهینه می‌کنند در نظر بگیریم:

$$\min_{\beta} \left(\frac{1}{n} \sum_{i=1}^n (y^{(i)} - x_i^T \beta)^2 \right)$$

یک عبارت به این مسئله بهینه سازی اضافه می‌کند.

$$\min_{\beta} \left(\frac{1}{n} \sum_{i=1}^n (y^{(i)} - x_i^T \beta)^2 + \lambda \|\beta\|_1 \right)$$

عبارت β بردار ویژگی است و منجر به جریمه وزن‌های بزرگ می‌شود. از آنجایی که از L1-norm استفاده می‌شود، بسیاری از وزن‌ها تخمینی ۰ دریافت می‌کنند و بقیه کوچک می‌شوند. پارامتر لامبدا (λ) قدرت تنظیم را کنترل می‌کند و معمولاً با اعتبارسنجی متقطع تنظیم می‌شود. به خصوص هنگامی که لامبدا بزرگ است، بسیاری از وزن‌ها به صفر تبدیل می‌شوند. وزن هر ویژگی با یک منحنی در شکل زیر نشان داده شده است.



شکل ۵.۴: با افزایش جریمه وزن‌ها، ویژگی‌های کمتر و کمتری تخمین وزن غیر صفر دریافت می‌کنند. به این منحنی‌ها مسیرهای تنظیم نیز می‌گویند. عدد بالای نمودار تعداد وزن‌های غیر صفر است. چه مقداری را برای لامبدا انتخاب کنیم؟ اگر عبارت جریمه را به عنوان یک پارامتر تنظیم می‌بینید، می‌توانید لامبدا را پیدا کنید که خطای مدل را با اعتبارسنجی متقاطع به حداقل می‌رساند. همچنان می‌توانید لامبدا را به عنوان پارامتری برای کنترل تفسیرپذیری مدل در نظر بگیرید. هر چه جریمه بزرگ‌تر باشد، ویژگی‌های کمتری در مدل وجود دارد (زیرا وزن آنها صفر است) و بهتر می‌توان مدل را تفسیر کرد.

مثال با لاسو

اجاره دورچرخه را با استفاده از لاسو پیش‌بینی می‌کنیم. تعداد ویژگی‌هایی که می‌خواهیم در مدل داشته باشیم را از قبل تعیین می‌کنیم. اجازه دهید ابتدا عدد را روی ۲ ویژگی تنظیم کنیم:

Weight
seasonWINTER
seasonSPRING
seasonSUMMER

season	FALL	0
holiday	HOLIDAY	0
workingday	WORKING DAY	0
weathersit	MISTY	0
weathersit	RAIN/SNOW/STORM	0
temp		52.33
hum		0
windspeed		0
days_since_2011		2.15

دو ویژگی اول با وزن‌های غیرصفر در مسیر لاسو دما ("temp") و روند زمانی ("days_since_2011") هستند.
اکنون، اجازه دهید ۵ ویژگی را انتخاب کنیم:

	Weight
season	WINTER -389.99
season	SPRING 0
season	SUMMER 0
season	FALL 0
holiday	HOLIDAY 0
workingday	WORKING DAY 0
weathersit	MISTY 0
weathersit	RAIN/SNOW/STORM -862.27
temp	85.58
hum	-3.04
windspeed	0
days_since_2011	3.82

توجه داشته باشید که وزن‌های "temp" و "days_since_2011" با مدل با دو ویژگی متفاوت است. دلیل این امر این است که با کاهش لامبدا، حتی ویژگی‌هایی که قبلاً در مدل هستند، کمتر جریمه می‌شوند و ممکن است وزن مطلق بیشتری به دست آورند. تفسیر وزن‌های لاسو با تفسیر وزن‌ها در مدل رگرسیون خطی مطابقت دارد. فقط باید به استاندارد بودن یا نبودن ویژگی‌ها توجه کنید، زیرا این روی وزن‌ها تأثیر می‌گذارد. در این مثال، ویژگی‌ها توسط نرمافزار استاندارد شده بودند، اما وزن‌ها به طور خودکار برای ما تغییر شکل دادند تا با مقیاس‌های ویژگی اصلی مطابقت داشته باشند.

روش‌های دیگر برای پراکندگی در مدل‌های خطی

طیف گسترده‌ای از روش‌ها را می‌توان برای کاهش تعداد ویژگی‌ها در یک مدل خطی استفاده کرد.

روش‌های پیش‌پردازش:

- ویژگی‌های انتخاب شده دستی: همیشه می‌توانید از دانش متخصصان، برای انتخاب یا حذف برخی از ویژگی‌ها استفاده کنید. اشکال بزرگ این است که نمی‌توان آن را خودکار کرد و شما باید به کسی دسترسی داشته باشید که داده‌ها را درک کند.
- انتخاب تک متغیره: یک مثال ضریب همبستگی است. شما فقط ویژگی‌هایی را در نظر می‌گیرید که از آستانه مشخصی از همبستگی بین ویژگی و هدف فراتر می‌روند. نقطه ضعف روش این است که ویژگی‌ها را فقط به صورت جداگانه در نظر می‌گیرد. برخی از ویژگی‌ها ممکن است تا زمانی که مدل خطی برخی ویژگی‌های دیگر را در نظر نگرفته باشد، همبستگی نشان ندهند. این ویژگی‌ها را با روش‌های انتخاب تک متغیره از دست خواهید داد.

روش‌های گام به گام:

- انتخاب رو به جلو: مدل خطی را با یک ویژگی برازش دهید. این کار را با سایر ویژگی‌ها نیز انجام دهید. مدلی را انتخاب کنید که بهترین عملکرد را دارد (مثلاً بالاترین R-squared). اکنون دوباره، ویژگی‌های باقی‌مانده را یک به یک به مدل تک ویژگی قبلی اضافه کنید و بهترین ویژگی بعدی را بیابید. این کار را تا رسیدن به معیاری مانند حداکثر تعداد ویژگی‌های مدل ادامه دهید.
 - انتخاب رو به عقب: مشابه انتخاب رو به جلو می‌باشد. اما به جای افزودن ویژگی‌ها، با مدلی شروع کنید که شامل همه ویژگی‌ها است و سعی کنید یک ویژگی را حذف کنید در حالیکه بالاترین افزایش عملکرد را داشته باشید. این کار را تا رسیدن به معیار توقف تکرار کنید.
- توصیه می‌کنم از Lasso استفاده کنید، زیرا می‌تواند خودکار باشد، همه ویژگی‌ها را به طور همزمان در نظر بگیرد و از طریق لامبدا قابل کنترل باشد. همچنین برای مدل رگرسیون لجستیک (طبقه‌بندی) نیز کار می‌کند.

۵.۱.۸ مزایا

مدل‌سازی پیش‌بینی‌ها به عنوان یک جمع وزنی، نحوه تولید پیش‌بینی‌ها را شفاف می‌کند. و با لاسو می‌توانیم اطمینان حاصل کنیم که تعداد ویژگی‌های مورد استفاده کم باقی می‌ماند. بسیاری از افراد از مدل‌های رگرسیون خطی استفاده می‌کنند. این بدان معناست که در بسیاری از جاها برای مدل سازی پیش‌بینی و انجام استنتاج پذیرفته شده است. سطح بالایی از تجربه و تخصص جمعی، از جمله مطالب آموزشی در مورد مدل‌های رگرسیون خطی و پیاده سازی در نرم افزار وجود دارد. توابع رگرسیون خطی را می‌توان در R، Python، Java، Julia، Scala، Javascript و ... یافت.

از نظر ریاضی، تخمین وزن‌ها ساده است و شما تضمینی برای یافتن وزن‌های بهینه دارید (باتوجهه به اینکه تمام مفروضات مدل رگرسیون خطی توسط داده‌ها برآورده می‌شوند).

همراه با وزن‌ها، فواصل اطمینان، آزمون‌ها و تئوری آماری را دریافت می‌کنید. همچنین توسعه‌های زیادی برای مدل رگرسیون خطی وجود دارد (به فصل GLM، GAM و موارد دیگر مراجعه کنید).

۵.۱.۹ معایب

مدل‌های رگرسیون خطی فقط می‌توانند روابط خطی را مدل کنند، یعنی مجموع وزنی ویژگی‌های ورودی. **هرگونه رابطه غیرخطی** یا تعامل باید به صورت دستی اعمال شود و به عنوان یک ویژگی ورودی به صورت صریح به مدل وارد شود.

همچنین اغلب عملکرد پیش‌بینی مدل‌های خطی پیش‌بینی کننده خوب نیست، زیرا روابطی را که می‌توانند بیاموزند، بسیار محدود است و معمولاً پیچیدگی واقعیت را بیش از حد ساده می‌کنند.

تفسیر یک وزن می‌تواند غیر شهودی باشد زیرا وزن آن ویژگی به تمام ویژگی‌های دیگر بستگی دارد. دو ویژگی با همبستگی مثبت بالا در پیش‌بینی خروجی y ، یکی ممکن است وزن مثبت در مدل خطی داشته باشد و دیگری وزن منفی. دلیل وزن منفی ویژگی دوم این است که، با توجه به ویژگی همبسته دیگر، در فضای با ابعاد بالا با پیش‌بینی y همبستگی پیدا می‌کند. ویژگی‌های کاملاً همبسته حتی یافتن یک راه حل منحصر به فرد برای معادله خطی را غیرممکن می‌کند. یک مثال: شما یک مدل برای پیش‌بینی ارزش یک خانه دارید و دارای ویژگی‌هایی مانند تعداد اتاق‌ها و اندازه خانه هستید. اندازه خانه و تعداد اتاق‌ها به شدت مرتبط هستند: هر چه خانه بزرگ‌تر باشد، اتاق‌های بیشتری دارد. اگر هر دو ویژگی را در یک مدل خطی قرار دهید، ممکن است این اتفاق بیفتاد که اندازه خانه پیش‌بینی کننده بهتری باشد و وزن مثبت زیادی دریافت کند. تعداد اتاق‌ها ممکن است وزن منفی داشته باشد، زیرا با توجه به اینکه یک خانه دارای اندازه یکسانی است، افزایش تعداد اتاق‌ها می‌تواند ارزش آن را کاهش دهد. زمانی که همبستگی خیلی قوی باشد، پایداری معادله خطی کم می‌شود.

۵.۲ رگرسیون لجستیک

رگرسیون لجستیک احتمالات مسائل طبقه‌بندی را با دو نتیجه ممکن، مدل‌سازی می‌کند. این مدل، توسعه مدل رگرسیون خطی برای مسائل طبقه‌بندی است.

۵.۲.۱ رگرسیون خطی برای طبقه‌بندی چه اشکالی دارد؟

مدل رگرسیون خطی می‌تواند برای رگرسیون خوب کار کند، اما برای طبقه‌بندی شکست می‌خورد. چرا اینطور است؟ در صورت وجود دو کلاس، می‌توانید یکی از کلاس‌ها را با 0 و دیگری را با 1 برچسب گذاری کنید و از رگرسیون خطی استفاده کنید. از نظر فنی کار می‌کند و اکثر برنامه‌های مدل خطی وزن‌ها را برای شما محاسبه بیرون می‌کنند. اما این روش چند مشکل دارد:

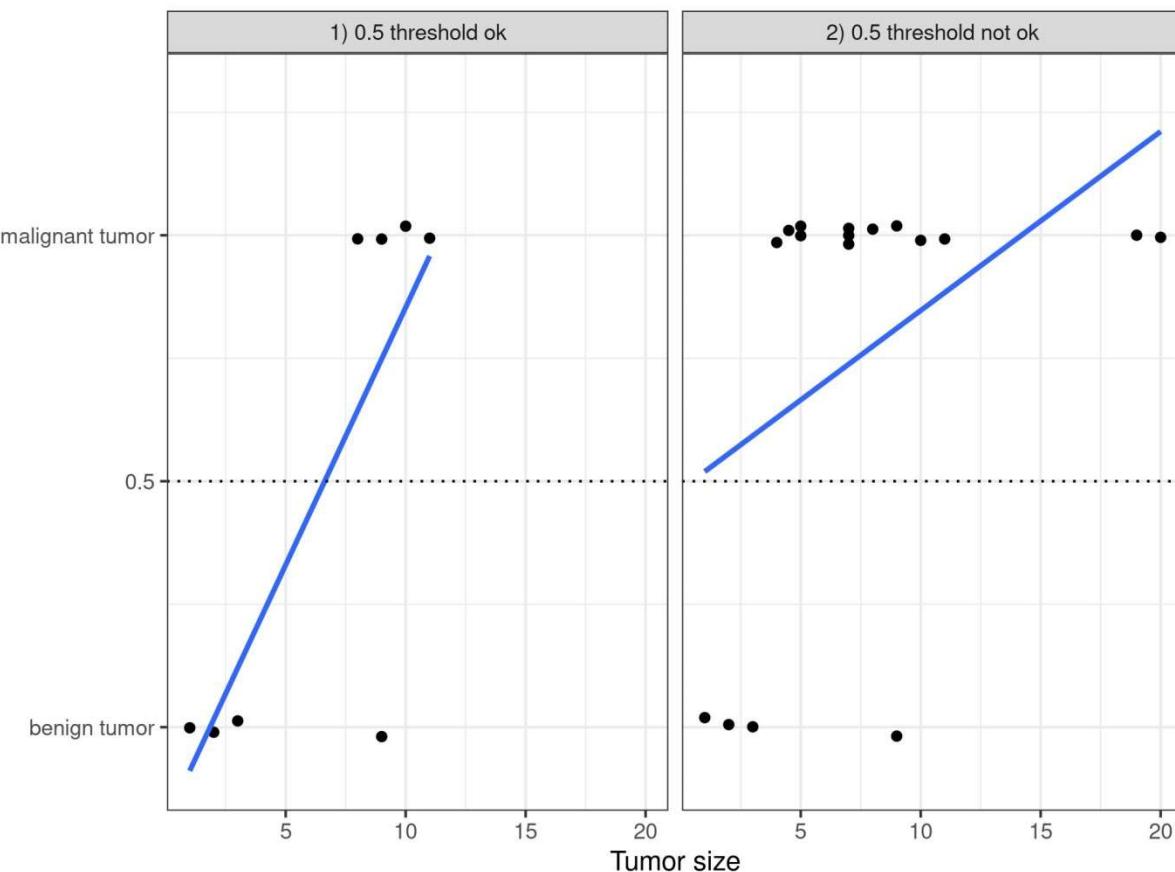
یک مدل خطی، احتمالات را به عنوان خروجی محاسبه نمی‌کند. این مدل، کلاً سهای را به عنوان اعداد $(0 \text{ و } 1)$ در نظر می‌گیرد و با بهترین ابر صفحه (که برای تک ویژگی، یک خط است)، فاصله بین نقاط و ابر صفحه را به حداقل

می‌رساند. بنابراین به سادگی بین نقاط درون یابی می‌شود و شما نمی‌توانید آن خروجی را به عنوان احتمال تفسیر کنید.

یک مدل خطی برون یابی نیز می‌کند و مقادیر زیر صفر و بالای یک را به شما خروجی می‌دهد. این نشانه خوبی است که ممکن است رویکرد هوشمندانه‌تری برای طبقه‌بندی وجود داشته باشد.

از آنجایی که نتیجه پیش‌بینی شده یک احتمال نیست، بلکه یک درونیابی خطی بین نقاط است، هیچ آستانه معنی‌داری وجود ندارد که در آن بتوانید یک کلاس را از کلاس دیگر تشخیص دهید. تصویر خوبی از این موضوع Stackoverflow(<https://stats.stackexchange.com/questions/22381/why-not-approach-classification-through-regression>) در

مدل‌های خطی به مسائل طبقه‌بندی با کلاس‌های متعدد تعمیم نمی‌یابند. شما باید شروع به برچسب زدن کلاس بعدی با ۲، سپس ۳ و غیره کنید. کلاس‌ها ممکن است ترتیب معنی‌داری نداشته باشند، اما مدل خطی، ساختار عجیبی را در ارتباط بین ویژگی‌ها و پیش‌بینی‌های کلاس شما ایجاد می‌کند. هر چه ارزش یک ویژگی با وزن مثبت بیشتر باشد، بیشتر به پیش‌بینی کلاسی با عدد بالاتر کمک می‌کند، حتی اگر کلاس‌هایی که اتفاقاً عدد مشابهی به دست می‌آورند از کلاس‌های دیگر نزدیک‌تر نباشند.



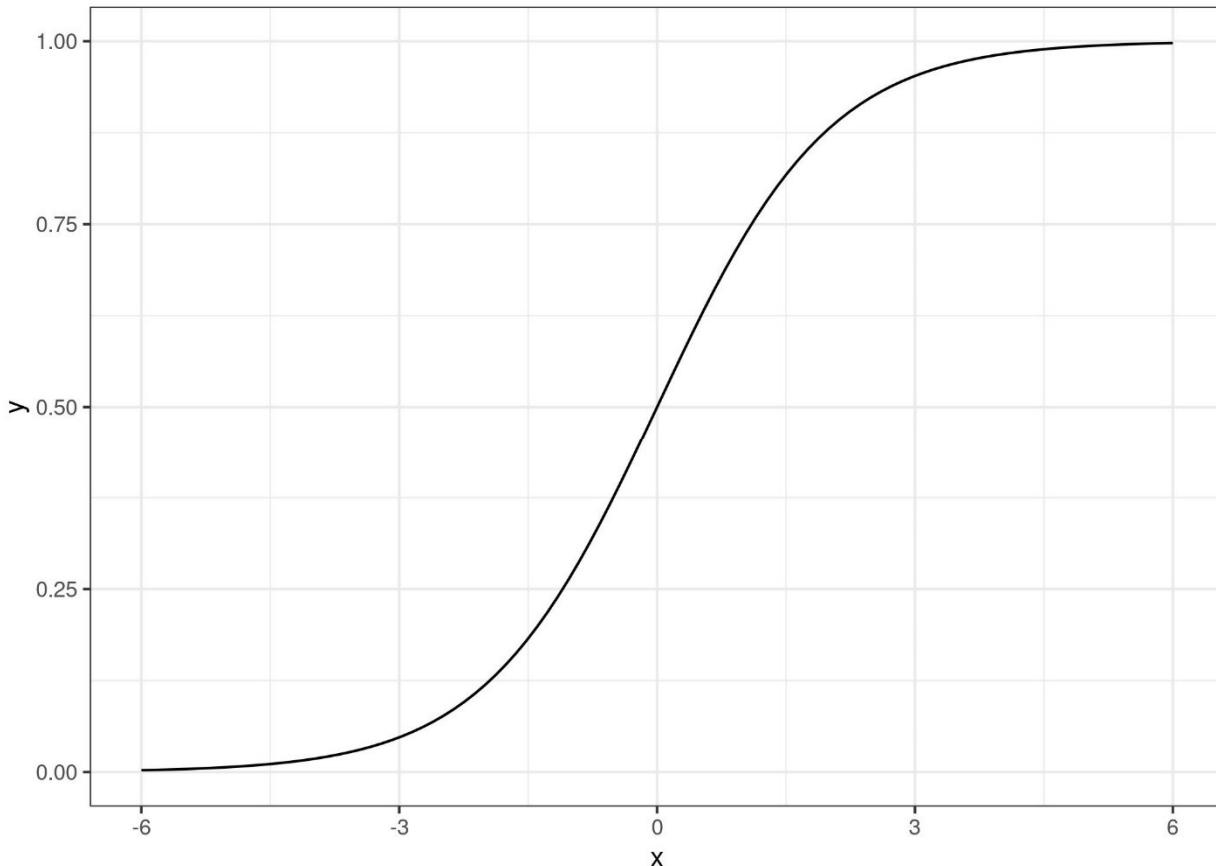
شکل ۵.۵: یک مدل خطی تومورها را با توجه به اندازه آنها به عنوان بدخیم (۱) یا خوش خیم (۰) طبقه‌بندی می‌کند. خطوط پیش‌بینی مدل خطی را نشان می‌دهد. برای داده‌های سمت چپ، می‌توانیم از ۰.۵ به عنوان آستانه طبقه‌بندی استفاده کنیم. پس از معرفی چند مورد تومور بدخیم دیگر، خط رگرسیون تغییر می‌کند و آستانه ۰.۵ دیگر کلاس‌ها را از هم جدا نمی‌کند. برای کاهش ترسیم بیش از حد، نقاط کمی‌تکان می‌خورند.

۵.۲.۲ نظریه

یک راه برای طبقه‌بندی رگرسیون لجستیک است. مدل رگرسیون لجستیک به جای برازش یک خط مستقیم یا ابر صفحه، از تابع لجستیک برای فشرده کردن خروجی یک معادله خطی بین ۰ و ۱ استفاده می‌کند. تابع لجستیک به صورت زیر تعریف می‌شود:

$$\text{logistic}(\eta) = \frac{1}{1 + \exp(-\eta)}$$

و بدین شکل است:



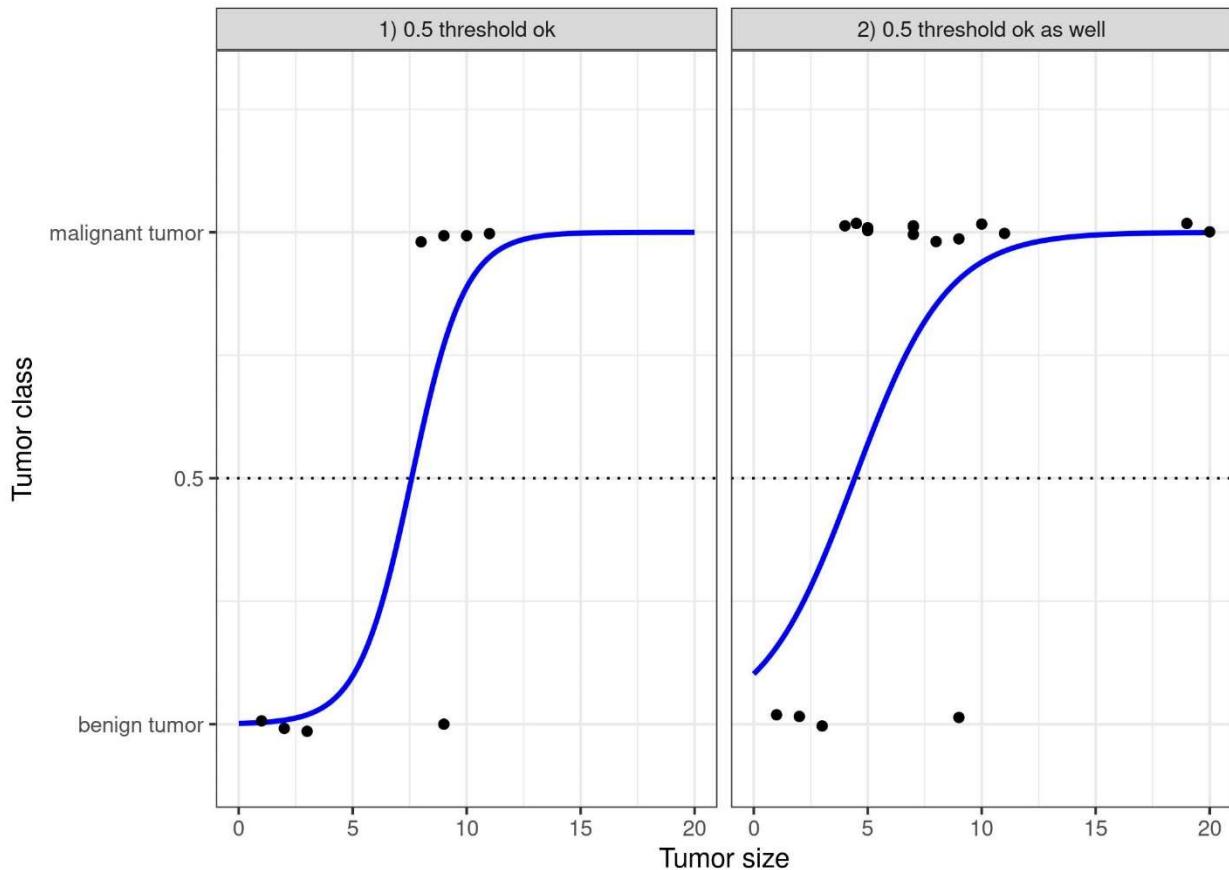
شکل ۵.۶: تابع لجستیک. خروجی اعداد بین ۰ و ۱ در ورودی ۰، خروجی ۰.۵ است. رسیدن از رگرسیون خطی به رگرسیون لجستیک ساده است. در مدل رگرسیون خطی، ما رابطه بین نتیجه و ویژگی‌ها را با یک معادله خطی مدل کرده ایم:

$$\hat{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}$$

برای طبقه‌بندی، احتمالات بین ۰ و ۱ را ترجیح می‌دهیم، بنابراین سمت راست معادله را در تابع لجستیک قرار می‌دهیم. این امر خروجی را مجبور می‌کند که فقط مقادیر بین ۰ و ۱ را در نظر بگیرد.

$$P(y^{(i)} = 1) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}))}$$

اجازه دهدید دوباره مثال اندازه تومور را بررسی کنیم. اما به جای مدل رگرسیون خطی، از مدل رگرسیون لجستیک استفاده کنیم:



شکل ۵.۷: مدل رگرسیون لجستیک مرز تصمیم گیری صحیح بین بدخیم و خوش خیم را بسته به اندازه تومور پیدا می‌کند. مرز، تابع لجستیکی است که برای برآذش داده‌ها، جابجا و فشرده می‌شود. با رگرسیون لجستیک بهتر می‌توان طبقه‌بندی کرد و در هر دو مورد می‌توانیم از ۰.۵ به عنوان آستانه استفاده کنیم. اضافه کردن نقاط جدید تاثیر زیادی بر منحنی برآذش شده نمی‌گذارد.

۵.۲.۳ تفسیر

تفسیر وزن‌ها در رگرسیون لجستیک با تفسیر وزن‌ها در رگرسیون خطی متفاوت است، زیرا نتیجه در رگرسیون لجستیک احتمالی بین ۰ و ۱ است. وزن‌ها دیگر روی احتمال به صورت خطی تأثیر نمی‌گذارند. مجموع وزنی

توسط تابع لجستیک به یک احتمال تبدیل می‌شود. بنابراین ما باید معادله را برای تفسیر دوباره فرمول بندی کنیم تا فقط عبارت خطی در سمت راست فرمول باشد.

$$\ln\left(\frac{P(y=1)}{1-P(y=1)}\right) = \log\left(\frac{P(y=1)}{P(y=0)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

ما عبارت موجود در تابع "ln" را "شانس" می‌نامیم (احتمال رویداد تقسیم بر احتمال عدم رویداد) و وقتی لگاریتم آن محاسبه می‌شود به آن لگاریتم احتمالات (log odds) گفته می‌شود.

این فرمول نشان می‌دهد که مدل رگرسیون لجستیک یک مدل خطی از لگاریتم احتمالات است. عالی! این مفید به نظر نمی‌رسد! با کمی تحلیل، می‌توانید بفهمید که هنگامی یکی از ویژگی‌ها x_j به اندازه یک واحد تغییر می‌کند، پیش‌بینی چقدر تغییر می‌کند. برای انجام این کار، ابتدا می‌توانیم تابع $\exp()$ را در دو طرف معادله اعمال کنیم:

$$\frac{P(y=1)}{1-P(y=1)} = odds = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$$

بنابراین ما آنچه را پس از تغییر یک واحد در ویژگی اتفاق می‌افتد، بررسی می‌نماییم. اما به جای بررسی تفاوت، به نسبت دو پیش‌بینی نگاه می‌کنیم:

$$\frac{odds_{x_j+1}}{odds} = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_j(x_j + 1) + \cdots + \beta_p x_p)}{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j + \cdots + \beta_p x_p)}$$

ما قانون زیر را اعمال می‌کنیم:

$$\frac{\exp(a)}{\exp(b)} = \exp(a - b)$$

و بسیاری از عبارات را حذف می‌کنیم:

$$\frac{odds_{x_j+1}}{odds} = \exp(\beta_j(x_j + 1) - \beta_j x_j) = \exp(\beta_j)$$

در پایان، ما عبارتی به سادگی $\exp()$ از وزن ویژگی داریم. تغییر در یک ویژگی به اندازه یک واحد، نسبت شانس (ضریب $\exp(\beta_j)$) را با ضریب $\exp(\beta_j)$ تغییر می‌دهد. همچنین می‌توانیم آن موضوع را این‌گونه تفسیر کنیم: تغییر در x_j به اندازه یک واحد لگاریتم احتمالات را به اندازه مقدار وزن مربوطه افزایش می‌دهد. اکثر مردم نسبت شانس را تفسیر می‌کنند زیرا فکر کردن در مورد $\ln()$ یک عبارت، برای مغز سخت است. تفسیر نسبت احتمالات از قبل نیاز به تمرین دارد. به عنوان مثال، اگر شما احتمال ۲ دارید، به این معنی است که احتمال $y=1$ دو برابر $y=0$ است. اگر وزن (نسبت لگاریتم احتمالات) 0.7 دارید، آنگاه افزایش ویژگی مربوطه در یک واحد، احتمال را در $\exp(0.7)$ ضرب می‌کند (قریباً ۱. ۳). احتمال به ۴ تغییر می‌کند. اما معمولاً شما با احتمال کار نمی‌کنید و وزن‌ها را فقط به عنوان نسبت‌های احتمالات تفسیر می‌کنید. زیرا برای محاسبه واقعی احتمال باید یک مقدار برای هر ویژگی تعیین کنید، که تنها زمانی منطقی است که بخواهید یک نمونه خاص از مجموعه داده را در نظر بگیرید.

در اینجا تفاسیر مدل رگرسیون لجستیک با انواع ویژگی‌های مختلف آورده شده است:

- ویژگی عددی: اگر ارزش ویژگی x را یک واحد افزایش دهید، احتمال تخمین زده شده با ضریب β تغییر می‌کند.
- ویژگی دسته‌ای باینری: یکی از دو مقدار ویژگی، دسته مرجع است (در برخی زبان‌ها، مقداری که با L کدگذاری می‌شود). تغییر ویژگی x از دسته مرجع به دسته دیگر، احتمال تخمین زده شده، با ضریب β تغییر می‌کند.
- ویژگی دسته‌ای با بیش از دو دسته: یک راه حل برای با برخورد با مسائل چندین دسته، one-hot encoding است. بدین صورت که هر دسته ستون خاص خود را دارد. شما فقط به $L-1$ ستون برای یک L ویژگی دسته‌ای نیاز دارید، در غیر این صورت بیش از حد پارامتر تعریف شده است. دسته L ام، دسته مرجع است. شما می‌توانید از هر رمزگذاری دیگری که می‌تواند در رگرسیون خطی به کار برد شود، استفاده کنید. بعد از این کار، تفسیر برای هر دسته معادل تفسیر ویژگی‌های باینری است.
- عرض از مبدا β_0 : وقتی همه ویژگی‌های عددی صفر هستند و ویژگی‌های طبقه‌بندی در دسته مرجع قرار دارند، احتمال تخمین زده شده (β_0) $\exp(\beta_0)$ می‌باشد. تفسیر عرض از مبدا معمولاً انجام نمی‌شود.

۵.۲.۴ مثال

ما از مدل رگرسیون لجستیک برای پیش‌بینی سرطان دهانه رحم بر اساس برخی عوامل خطر استفاده می‌کنیم. جدول زیر وزن‌های تخمینی، نسبت‌های احتمال مربوطه و خطای استاندارد تخمین‌ها را نشان می‌دهد.

جدول ۵.۲: نتایج برآش یک مدل رگرسیون لجستیک بر روی مجموعه‌داده سرطان دهانه رحم. ویژگی‌های مورد استفاده در مدل، وزن‌های تخمینی و نسبت‌های احتمال مربوطه و خطاهای استاندارد وزن‌های تخمینی آورده شده است.

	Weight	Odds ratio	Std. Error
Intercept	-2.91	0.05	0.32
Hormonal contraceptives y/n	-0.12	0.89	0.30
Smokes y/n	0.26	1.30	0.37
Num. of pregnancies	0.04	1.04	0.10
Num. of diagnosed STDs	0.82	2.27	0.33
Intrauterine device y/n	0.62	1.86	0.40

تفسیر یک ویژگی عددی (Num. of diagnosed STDs): افزایش تعداد STD های تشخیص داده شده (بیماری‌های مقاربتی) شانس ابتلا به سرطان در مقایسه با عدم وجود سرطان را با ضریب ۲.۲۷ تغییر می‌دهد (افزایش می‌دهد) در حالی که همه ویژگی‌های دیگر همان باقی می‌مانند. به خاطر داشته باشد که همبستگی به معنای علیت نیست.

تفسیر یک ویژگی طبقه‌بندی شده ("داروهای ضد بارداری هورمونی بله یا خیر): برای زنانی که از داروهای ضد بارداری هورمونی استفاده می‌کنند نسبت به زنان بدون ضد بارداری هورمونی، شانس ابتلا به سرطان در مقایسه با بدون سرطان ۰.۸۹ کمتر است، در صورتی که سایر ویژگی‌ها یکسان باقی بمانند.
مانند مدل خطی، تفاسیر همیشه با این بند آمده است که "همه ویژگی‌های دیگر ثابت می‌مانند".

۵.۲.۵ مزایا و معایب

بسیاری از مزایا و معایب مدل رگرسیون خطی در مورد مدل رگرسیون لجستیک نیز صدق می‌کند. رگرسیون لجستیک به طور گسترده توسط افراد مختلف مورد استفاده قرار گرفته است، اما با توجه به عبارات محدود خود (مثلًاً تعاملات باید به صورت دستی اضافه شوند) مشکلاتی دارد و مدل‌های دیگر ممکن است عملکرد پیش‌بینی بهتری داشته باشند.

یکی دیگر از معایب مدل رگرسیون لجستیک این است که تفسیر دشوارتری دارد، زیرا تفسیر اوزان ضربی (additive) است و افزایشی (multiplicative) نیست.

رگرسیون لجستیک در موارد جداولی کامل مشکل دارد. اگر ویژگی ای وجود داشته باشد که این دو کلاس را کاملاً از هم جدا کند، مدل رگرسیون لجستیک دیگر قابل آموزش نیست. این به این دلیل است که وزن آن ویژگی همگرا (converge) نمی‌شود، زیرا وزن بهینه بی نهایت خواهد بود. این واقعًا کمی تا سف بار است، زیرا چنین ویژگی واقعًا مفید است. اما اگر قانون ساده ای دارید که هر دو کلاس را از هم جدا می‌کند، نیازی به یادگیری ماشین ندارید. مشکل جداسازی کامل را می‌توان با معرفی جریمه وزن‌ها یا تعریف یک توزیع احتمال اولیه از وزن‌ها (prior probability distribution of weights) حل کرد.

از طرفی، مدل رگرسیون لجستیک نه تنها یک مدل طبقه‌بندی است، بلکه احتمالاتی را نیز به شما می‌دهد. این یک مزیت بزرگ نسبت به مدل‌هایی است که فقط می‌توانند طبقه‌بندی نهایی را ارائه دهند. دانستن اینکه یک نمونه برای یک کلاس ۹۹ درصد احتمال دارد در مقایسه با ۵۱ درصد، تفاوت بزرگی ایجاد می‌کند.

رگرسیون لجستیک همچنین می‌تواند از طبقه‌بندی باینتری به طبقه‌بندی چند طبقه گسترش یابد. در این حالت به آن رگرسیون چند جمله‌ای می‌گویند.

۵.۲۶ نرم افزار

من ازتابع `glm` در R برای همه مثال‌ها استفاده کردم. شما می‌توانید رگرسیون لجستیک را در هر زبان برنامه نویسی که می‌تواند برای انجام تجزیه و تحلیل داده‌ها استفاده شود، پیدا کنید، مانند پایتون، جاوا، استاتا، متلب و GLM5.۳، GAM و موارد دیگر

بزرگ‌ترین نقطه قوت و همچنین بزرگ‌ترین ضعف مدل رگرسیون خطی این است که پیش‌بینی به عنوان مجموع وزنی ویژگی‌ها، مدل می‌شود. علاوه بر این، مدل خطی با سیاری از مفروضات دیگر همراه است. خبر بد این است (خوب، واقعاً خبری نیست) این است که همه این فرضیات اغلب در واقعیت نقض می‌شوند: نتیجه با توجه به ویژگی‌ها ممکن است توزیع غیر گاووسی داشته باشد، ویژگی‌ها ممکن است برهم کنش داشته باشند و رابطه بین ویژگی‌ها و خروجی ممکن است غیرخطی باشد. خبر خوب این است که جامعه آمار تغییرات مختلفی را ایجاد کرده است که مدل رگرسیون خطی را از یک تیغه ساده به یک چاقوی سوئیسی تبدیل می‌کند.

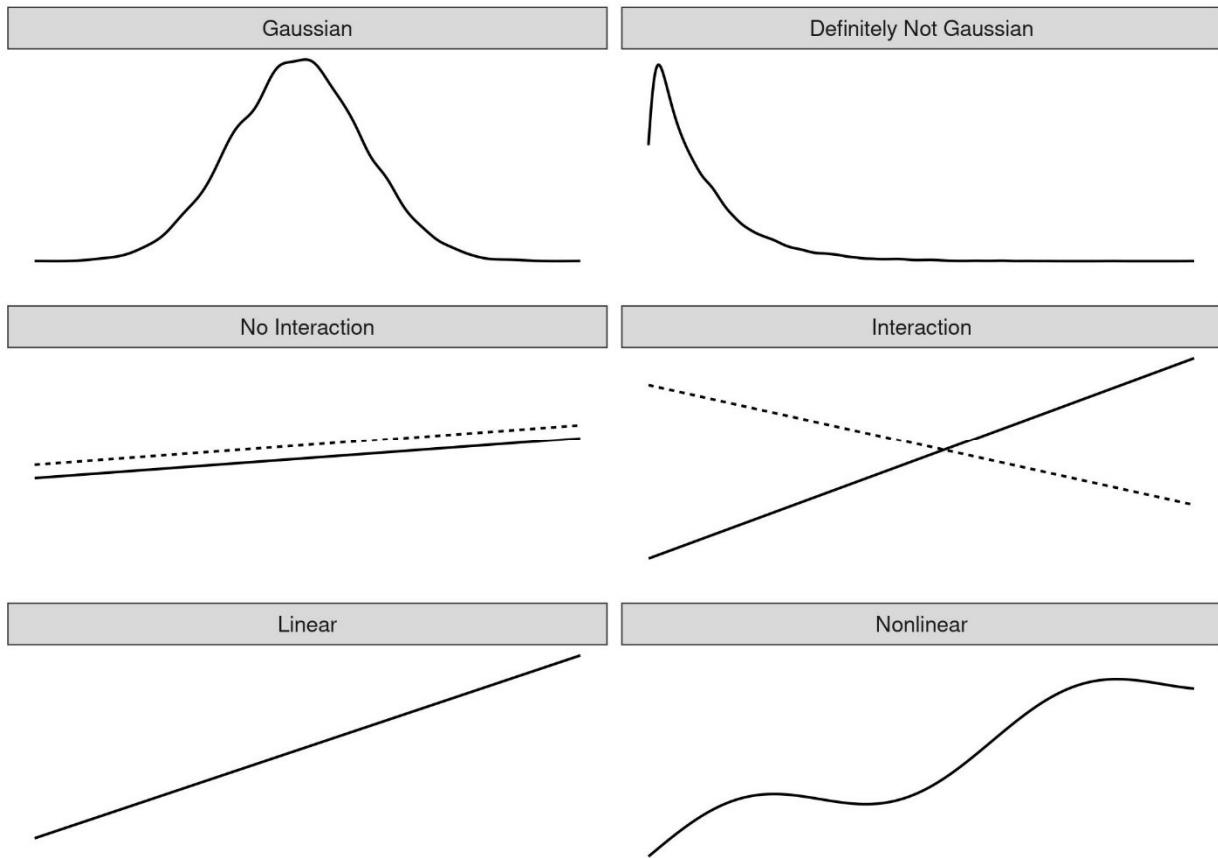
این فصل قطعاً راهنمای قطعی شما برای توسعه مدل‌های خطی نیست. بلکه به عنوان یک نمای کلی از برنامه‌های تعمیمی مانند مدل‌های خطی تعمیم یافته (GLMs) و مدل‌های افزودنی تعمیم یافته (GAMs) عمل می‌کند و کمی شهود به شما می‌دهد. پس از مطالعه، باید یک دید کلی از نحوه تعمیم مدل‌های خطی داشته باشید. اگر می‌خواهید ابتدا درباره مدل رگرسیون خطی بیشتر بدانید، پیشنهاد می‌کنم فصل مدل‌های رگرسیون خطی را مطالعه کنید، اگر قبلًاً این کار را نکرده‌اید.

بیایید فرمول یک مدل رگرسیون خطی را به خاطر بیاوریم:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

مدل رگرسیون خطی فرض می‌کند که نتیجه y یک نمونه را می‌توان با مجموع وزنی از p ویژگی‌های آن با یک خطای فردی مشخص ϵ بیان کرد، که از توزیع گاووسی پیروی می‌کند. با وارد کردن داده‌ها به این فرمول، قابلیت تفسیر مدل زیادی را به دست می‌آوریم. اثرات ویژگی افزایشی هستند، به این معنی که هیچ تعاملی وجود ندارد، و رابطه خطی است، به این معنی که افزایش یک ویژگی به اندازه یک واحد می‌تواند مستقیماً به افزایش/کاهش نتیجه پیش‌بینی شده منجر شود. مدل خطی به ما اجازه می‌دهد تا رابطه بین یک ویژگی و نتیجه مورد انتظار را در یک عدد واحد، یعنی وزن تخمینی، فشرده کنیم.

اما یک جمع وزنی ساده، برای بسیاری از مسائل پیش‌بینی دنیای واقعی بسیار محدود است. در این فصل با سه مسئله مدل رگرسیون خطی کلاسیک و نحوه حل آنها آشنا خواهیم شد. مسائل بسیاری وجود دارند که فرضیات در نظر گرفته شده را نقض می‌کنند، اما ما بر روی سه مورد نشان داده شده در شکل زیر تمرکز خواهیم کرد:



شکل ۵.۸: سه فرض مدل خطی (سمت چپ): توزیع گاوسی خروجی با توجه به ویژگی‌ها، افزایش (= بدون تعامل) و رابطه خطی. در واقعیت معمولاً این مفروضات نقض می‌شوند (سمت راست): نتایج ممکن است دارای توزیع‌های غیر گاوسی باشند، ویژگی‌ها ممکن است تعامل داشته باشند و رابطه ممکن است غیرخطی باشد.
برای همه این مشکلات، راه حلی وجود دارد:

مشکل: نتیجه هدف y با توجه به ویژگی‌ها از توزیع گاوسی پیروی نمی‌کند.

مثال: فرض کنید می‌خواهم پیش‌بینی کنم که در یک روز معین چند دقیقه دوچرخه‌سواری خواهم کرد. به عنوان ویژگی من نوع روز، آب و هوا و غیره را دارم. اگر از یک مدل خطی استفاده کنم، می‌تواند دقیقه‌های منفی را نیز پیش‌بینی کند، زیرا توزیع را گاوسی فرض می‌کند که در دقیقه ۰ متوقف نمی‌شود. همچنین اگر بخواهم احتمالات را با یک مدل خطی پیش‌بینی کنم، می‌توانم احتمالات منفی یا بزرگ‌تر از ۱ را به دست بیاورم.

راه حل: مدل‌های خطی تعمیم‌یافته (GLMs).

مشکل: ویژگی‌ها با هم تعامل دارند.

مثال: به طور متوسط، باران ملایم تأثیر منفی جزئی بر تمایل من به دوچرخه سواری دارد. اما در تابستان، در ساعات شلوغی، از باران استقبال می‌کنم، زیرا در این صورت تمام دوچرخه‌سواران هوای مطبوع در خانه می‌مانند

و من مسیرهای دوچرخه را برای خودم دارم! این یک تعامل بین زمان و آب و هوا است که با یک مدل صرفاً افزودنی قابل درک نیست.

راه حل: افزودن تعاملات به صورت دستی.

مشکل: رابطه واقعی بین ویژگی‌ها و y خطی نیست.

مثال: بین 0° تا 25° درجه سانتیگراد، تأثیر دما بر تمایل من به دوچرخه سواری می‌تواند خطی باشد، به این معنی که افزایش از 0° به 1° درجه باعث افزایش همان افزایش میل دوچرخه سواری با افزایش از 20° به 21° می‌شود. اما در دماهای بالاتر انگیزه من برای دوچرخه سواری کاهش می‌یابد و حتی کاهش می‌یابد - من دوست ندارم وقتی هوا خیلی گرم است دوچرخه سواری کنم.

راه حل‌ها: مدل‌های افزایشی تعمیم یافته (GAMs)، تبدیل ویژگی‌ها.

راه حل‌های این سه مشکل در این فصل ارائه شده است. بسیاری از تعمیمات دیگر مدل خطی حذف شده‌اند. اگر بخواهم همه چیز را در اینجا پوشش دهم، این فصل به سرعت تبدیل به کتابی درباره این موضوع می‌شود که قبل از بسیاری از کتاب‌های دیگر پوشش داده شده است. اما از آنجایی که شما در حال حاضر اینجا هستید، من یک بررسی اجمالی مشکل، به اضافه یک راه حل برای تعمیم مدل خطی آورده ام که می‌توانید در انتهای فصل آن را مشاهده کنید. نام راه حل به عنوان نقطه شروع برای جستجو است.

۵.۳.۱ خروجی غیر گاوی - GLMs

مدل رگرسیون خطی فرض می‌کند که نتیجه با توجه به ویژگی‌های ورودی از یک توزیع گاوی پیروی می‌کند. این فرض شامل حال بسیاری از موارد می‌شود: خروجی می‌تواند یک دسته (سرطان در مقابل سالم)، شمارش (تعداد فرزندان)، زمان وقوع یک رویداد (زمان تا خرابی یک دستگاه) یا یک نتیجه بسیار نامتقارن (skewed outcome) با تعداد کمی مقادیر بسیار زیاد (درآمد خانوار) باشد. مدل رگرسیون خطی را می‌توان برای مدل سازی همه این نوع خروجی‌ها گسترش داد. این تعمیم، مدل‌های خطی تعمیم یافته یا به اختصار GLM نامیده می‌شود. در طول این فصل، من از نام GLM هم برای چارچوب کلی و هم برای مدل‌های خاص این چارچوب استفاده خواهم کرد. مفهوم اصلی هر GLM این است: جمع وزنی ویژگی‌ها را حفظ کنید، اما توزیع‌های نتیجه غیر گاوی را مجاز کنید و میانگین مورد انتظار این توزیع و مجموع وزنی را از طریق یکتابع احتمالاً غیرخطی به هم مرتبط کنید. به عنوان مثال، مدل رگرسیون لجستیک توزیع برنولی را برای خروجی فرض می‌کند و میانگین مورد انتظار و مجموع وزنی را با استفاده از تابع لجستیک به هم مرتبط می‌کند.

GLM به صورت ریاضی، جمع وزنی ویژگی‌ها را با مقدار میانگین توزیع فرضی با استفاده از تابع اتصال (link function) g متصل می‌کند، که بسته به نوع خروجی می‌تواند به طور انعطاف‌پذیر انتخاب شود.

$$g(E_Y(y|x)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

GLM‌ها از سه جزء تشکیل شده‌اند: تابع اتصال g , مجموع وزنی $X^T \beta$ (گاهی اوقات پیش‌بینی کننده خطی نامیده می‌شود) و یک توزیع احتمال از خانواده نمایی که E_Y تعریف می‌شود.

خانواده نمایی مجموعه‌ای از توزیع‌هایی است که می‌توان با همان فرمول (پارامتری شده) نوشت که شامل یک توان، میانگین و واریانس توزیع و برخی پارامترهای دیگر است. من وارد جزئیات ریاضی نمی‌شوم زیرا این جهان بسیار بزرگی است که من نمی‌خواهم وارد آن شوم. ویکی‌پدیا فهرست دقیقی از توزیع‌ها از خانواده نمایی دارد. هر توزیعی از این لیست می‌تواند برای GLM شما انتخاب شود. بر اساس نوع خروجی که می‌خواهید پیش‌بینی کنید، توزیع مناسبی را انتخاب کنید. آیا خروجی، شمارش چیزی است (مثلاً تعداد کودکانی که در یک خانواده زندگی می‌کنند)? در این حالت، توزیع پواسون می‌تواند انتخاب خوبی باشد. آیا نتیجه همیشه مثبت است (مثلاً زمان بین دو رویداد)? در این حالت، توزیع نمایی می‌تواند انتخاب خوبی باشد.

اجازه دهید مدل خطی کلاسیک را به عنوان یک مورد خاص از یک GLM در نظر بگیریم. تابع اتصال برای توزیع گاووسی در مدل خطی کلاسیک به سادگی تابع تشخیص است. توزیع گاووسی با میانگین و واریانس پارامتری می‌شود. میانگین مقداری را که به طور متوسط انتظار داریم و واریانس نشان می‌دهد که مقادیر در حدود این میانگین چقدر تغییر می‌کنند. در مدل خطی، تابع اتصال، مجموع وزنی ویژگی‌ها را به میانگین توزیع گاووسی مربوط می‌کند.

در چارچوب GLM، این مفهوم، به هر توزیع (از خانواده نمایی) و توابع اتصال دلخواه تعمیم می‌یابد. اگر y شمارشی از چیزی باشد، مانند تعداد قهوه‌هایی که فرد در یک روز خاص می‌نوشد، می‌توانیم آن را با GLM با توزیع پواسون و لگاریتم طبیعی به عنوان تابع اتصال مدل‌سازی کنیم:

$$\ln(E_Y(y|x)) = X^T \beta$$

مدل رگرسیون لجستیک نیز یک GLM است که توزیع برنولی را فرض می‌کند و از تابع لاجیت (logit) به عنوان تابع اتصال استفاده می‌کند. میانگین توزیع دوچشمی ای مورد استفاده در رگرسیون لجستیک احتمال y است که مقدارش ۱ می‌باشد.

$$X^T \beta = \ln\left(\frac{E_Y(y|x)}{1 - E_Y(y|x)}\right) = \ln\left(\frac{P(y = 1|x)}{1 - P(y = 1|x)}\right)$$

و اگر این معادله را بنحوی حل کنیم که در یک طرف ($y = 1$) باشد، فرمول رگرسیون لجستیک به دست می‌آید:

$$P(y = 1) = \frac{1}{1 + \exp(-x^T \beta)}$$

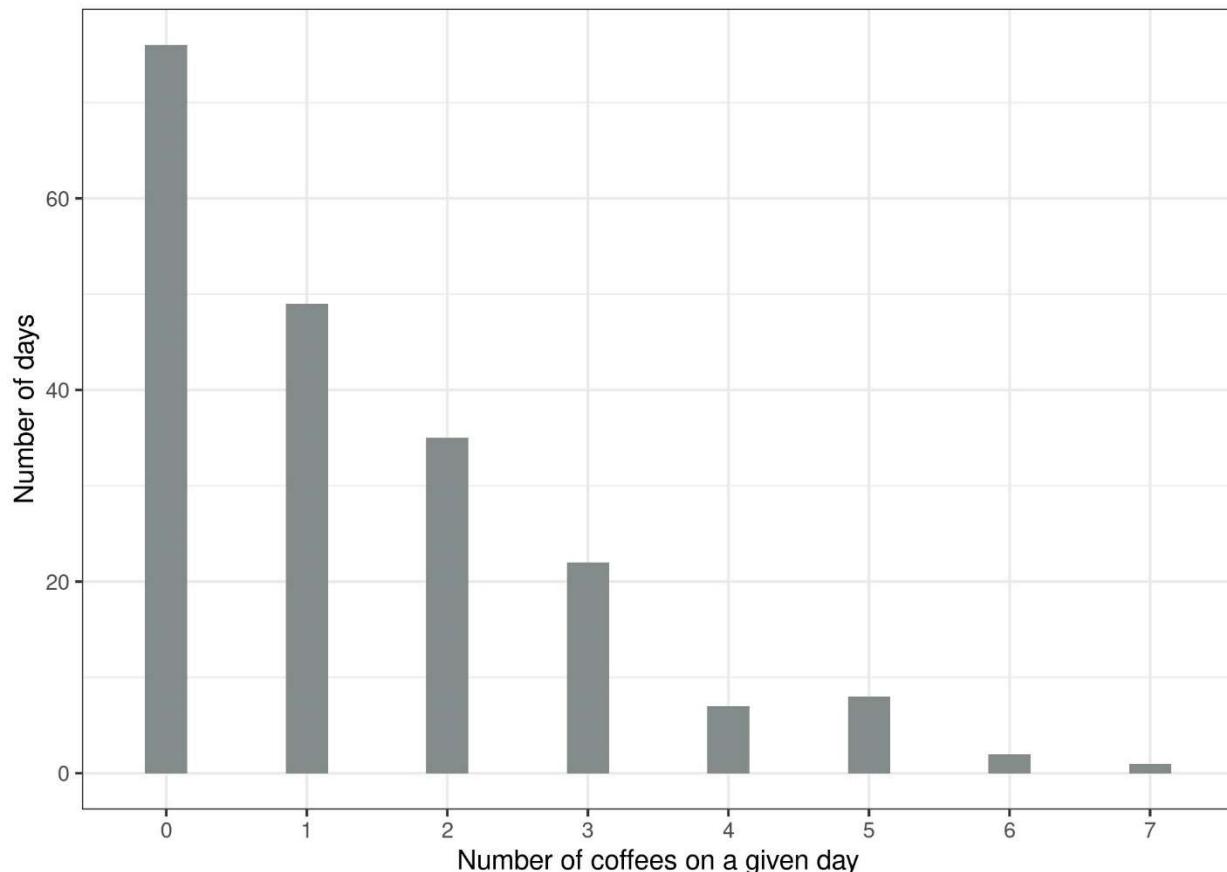
هر توزیع از خانواده نمایی دارای یک تابع اتصال متعارف (canonical link function) است که می‌تواند به صورت ریاضی از توزیع استخراج شود. چارچوب GLM امکان انتخاب تابع اتصال را مستقل از توزیع فراهم می‌کند. چگونه تابع اتصال مناسب را انتخاب کنیم؟ هیچ دستور العمل کاملی وجود ندارد. شما دانش خود را در مورد توزیع هدف

در نظر می‌گیرید، اما ملاحظات نظری و اینکه مدل چقدر با داده‌های واقعی شما مطابقت دارد را نیز در نظر داشته باشید. برای برخی از توزیع‌ها، تابع اتصال متعارف می‌تواند به مقادیری منجر شود که برای آن توزیع نامعتبر هستند. در مورد توزیع نمایی، تابع اتصال متعارف، معکوس منفی است که می‌تواند منجر به پیش‌بینی‌های منفی شود که خارج از دامنه توزیع نمایی هستند. از آنجایی که می‌توانید هر تابع اتصالی را انتخاب کنید، راه حل ساده این است که تابع دیگری را انتخاب کنید که به دامنه توزیع احترام بگذارد.

مثال‌ها

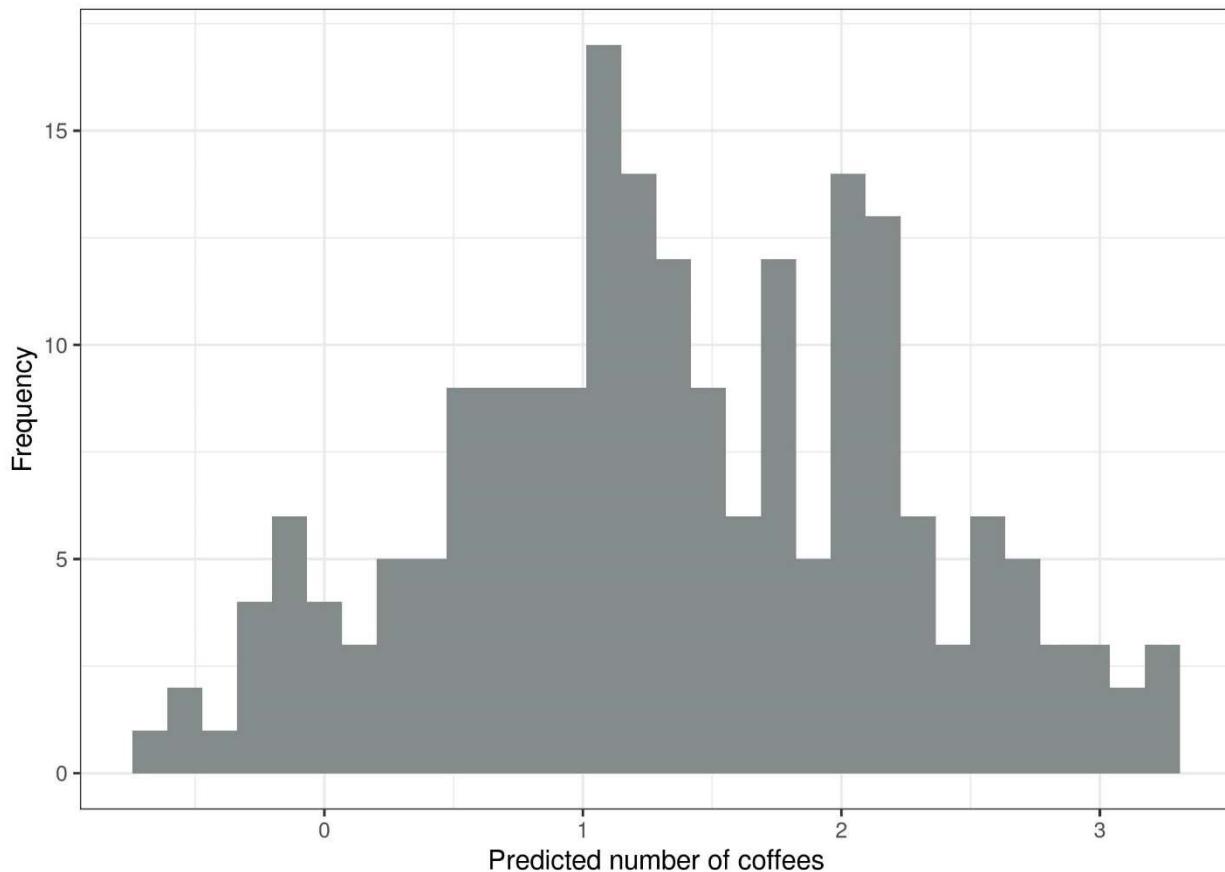
من مجموعه‌داده‌ای را در مورد رفتار نوشیدن قهقهه شبیه سازی کرده ام تا نیاز به GLM‌ها را برجسته کنم. فرض کنید اطلاعاتی در مورد رفتار نوشیدن قهقهه روزانه خود جمع‌آوری کرده‌اید. اگر قهقهه دوست ندارید، در مورد چای این کار را انجام دهد. همراه با تعداد فنجان‌ها، سطح استرس فعلی خود را در مقیاس ۱ تا ۱۰ ثبت می‌کنید، شب قبل چقدر خوب خوابیده اید در مقیاس ۱ تا ۱۰ و اینکه آیا باید در آن روز کار کنید یا خیر. هدف پیش‌بینی تعداد قهقهه‌ها با توجه به ویژگی‌های استرس، خواب و کار است. من داده‌ها را برای ۲۰۰ روز شبیه سازی کردم. استرس و خواب به طور یکنواخت بین ۱ تا ۱۰ تجسم شد و بله/نه کار با شانس ۵۰/۵۰ تجسم شد (عجب زندگی!). سپس برای هر روز، تعداد قهقهه‌ها از توزیع پواسون گرفته شد و مدل‌سازی شدت آن λ (که همچنین مقدار مورد انتظار توزیع پواسون است) به عنوان تابعی از ویژگی‌های خواب، استرس و کار انجام شد. می‌توانید حدس بزنید که این داستان به کجا ختم می‌شود: "اجازه دهید این داده‌ها را با یک مدل خطی مدل‌سازی کنیم. متاسفانه مدل خطی کار نمی‌کند. حالا اجازه دهید یک GLM با توزیع پواسون را امتحان کنیم. حالا کار می‌کند!". امیدوارم داستان را زیاد برای شما لو نداده باشم.

بیایید به توزیع متغیر هدف، تعداد قهقهه در یک روز معین نگاه کنیم:



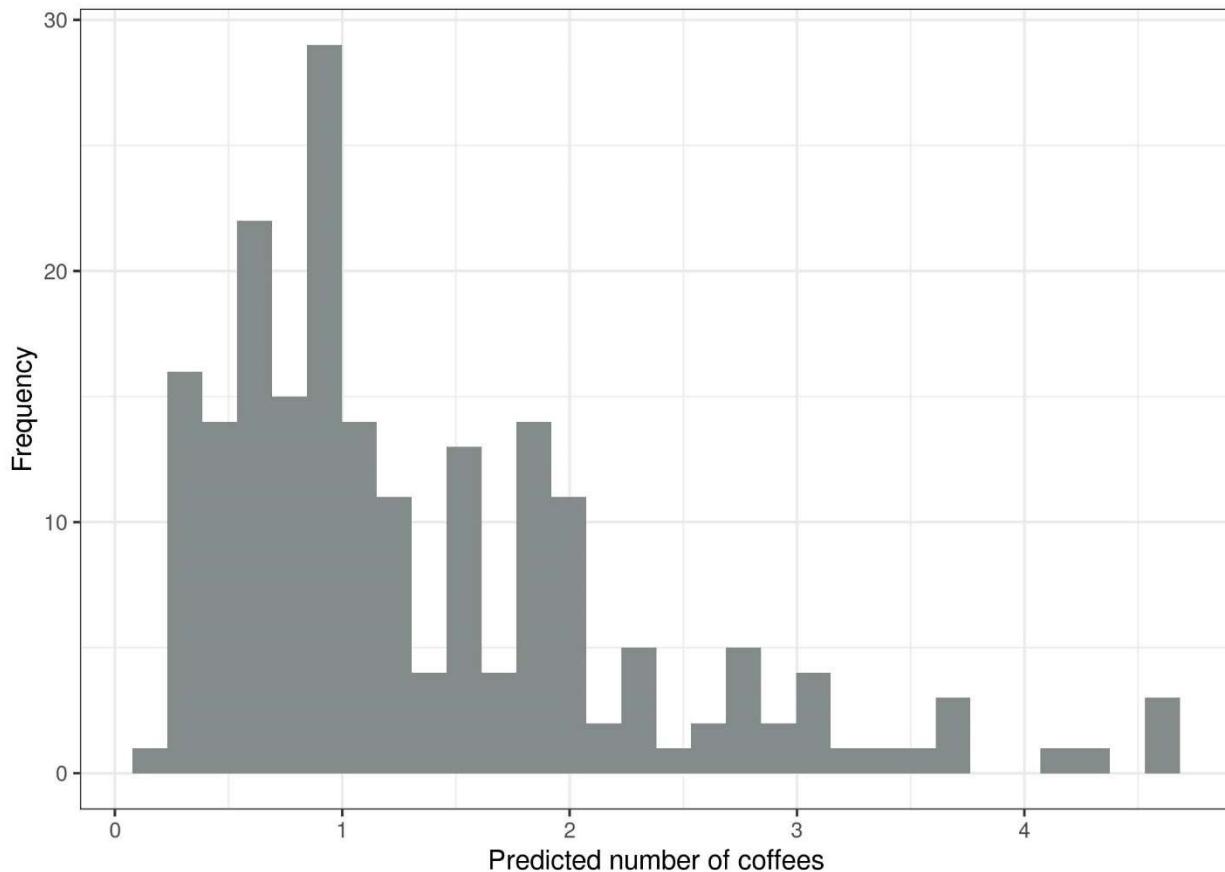
شکل ۵: توزیع شبیه سازی شده تعداد قهوههای روزانه برای ۲۰۰ روز.

در ۷۶ روز از ۲۰۰ روز، اصلاً قهوه نخوردید و در شدیدترین روز، ۷ قهوه خوردید. اجازه دهید ساده‌لوحانه از یک مدل خطی برای پیش‌بینی تعداد قهوه‌ها با استفاده از سطح خواب، سطح استرس و کار به/خیر به عنوان ویژگی‌ها استفاده کنیم. وقتی به اشتباه توزیع گاووسی را فرض می‌کنیم مرتكب چه اشتباهی شده ایم؟ یک فرض اشتباه می‌تواند منجر به نامعتبر شدن تخمین‌ها، به ویژه فواصل اطمینان وزن‌ها، شود. مشکل واضح‌تر این است که پیش‌بینی‌ها، با دامنه «مجاز» خروجی واقعی مطابقت ندارند، همان‌طور که شکل زیر نشان می‌دهد.



شکل ۵.۱۰: تعداد قهوه‌های پیش‌بینی شده با استفاده از ویژگی‌های استرس، خواب و کار. مدل خطی مقادیر منفی را پیش‌بینی می‌کند.

مدل خطی منطقی نیست، زیرا تعداد قهوه‌های منفی را پیش‌بینی می‌کند. این مشکل را می‌توان با مدل‌های خطی تعمیم یافته (GLM) حل کرد. ما می‌توانیمتابع اتصال و توزیع فرضی را تغییر دهیم. یکی از امکان‌ها حفظ توزیع گاوسی و استفاده ازتابع اتصال است که همیشه به پیش‌بینی‌های مثبتی مانندتابع exp -link (معکوستابع log-) است) به جایتابع تشخیص منجر می‌شود. حتی بهتر: ما توزیعی را انتخاب می‌کنیم که با فرآیند تولید داده و یکتابع اتصال مناسب مطابقت دارد. از آنجایی که نتیجه یک شمارش است، توزیع پواسون یک انتخاب مناسب به همراه لگاریتم به عنوانتابع اتصال است. در این مورد بخصوص، که داده‌ها با توزیع پواسون تولید شده‌اند، پواسون GLM انتخابی عالی است. پواسون GLM برازش داده شده منجر به توزیع زیر از مقادیر پیش‌بینی شده می‌شود:



شکل ۵.۱۱: تعداد قهوه‌های پیش‌بینی شده با توجه به استرس، خواب و کار. GLM با فرض پواسون و اتصال \log مدل مناسبی برای این مجموعه داده است. بدون مقادیر منفی قهوه، اکنون مدل بسیار بهتر به نظر می‌رسد.

تفسیر وزن‌های GLM

توزیع مفروض همراه باتابع اتصال تعیین می‌کند که وزن ویژگی‌های برآورده شده چگونه تفسیر می‌شوند. در مثال شمارش قهوه، من از یک GLM با توزیع پواسون و اتصال لگاریتمی استفاده کردم که دلالت بر رابطه زیر بین نتیجه مورد انتظار و ویژگی‌های استرس(str)، خواب(slp) و کار(wrk) دارد.

$$\ln(E(coffee|str - slp - wrk)) = \beta_0 + \beta_{str}x_{str} + \beta_{slp}x_{slp} + \beta_{wrk}x_{wrk}$$

برای تفسیر وزن‌ها،تابع اتصال را معکوس می‌کنیم تا بتوانیم تأثیر ویژگی‌ها را بر نتیجه مورد انتظار تفسیر کنیم و نه بر لگاریتم نتیجه مورد انتظار.

$$E(coffee|str - slp - wrk) = \exp(\beta_0 + \beta_{str}x_{str} + \beta_{slp}x_{slp} + \beta_{wrk}x_{wrk})$$

از آنجایی که همه وزن‌ها در تابع نمایی هستند، تفسیر اثر جمعی نیست، بلکه ضربی است، زیرا $\exp(a + b)$ برابر با $\exp(a) \exp(b)$ است. آخرین عنصر برای تفسیر، وزن‌های واقعی داده‌های ساختگی است. جدول زیر وزن‌های تخمینی و $\exp(\text{weight})$ را همراه با فاصله اطمینان ۹۵ درصد فهرست می‌کند:

جدول ۵.۳: وزن‌ها در مدل پواسون

	weight	$\exp(\text{weight}) [2.5\%, 97.5\%]$
(Intercept)	-0.16	0.85 [0.54, 1.32]
stress	0.12	1.12 [1.07, 1.18]
sleep	-0.15	0.86 [0.82, 0.90]
workYES	0.80	2.23 [1.72, 2.93]

افزایش سطح استرس به اندازه یک واحد، تعداد قهوه مورد انتظار را در ضرب ۱.۱۲ ضرب می‌کند. افزایش کیفیت خواب یک واحد، تعداد قهوه مورد انتظار را در ضرب ۰.۸۶ ضرب می‌کند. تعداد قهوه‌های پیش‌بینی شده در یک روز کاری به طور متوسط ۲.۲۳ برابر تعداد قهوه‌های یک روز تعطیل است. به طور خلاصه، هر چه استرس بیشتر، خواب کمتر و کار بیشتر باشد، قهوه بیشتری مصرف می‌شود.

در این بخش شما کمی در مورد مدل‌های خطی تعمیم یافته یاد گرفتید که زمانی مفید هستند که خروجی از توزیع گاووسی پیروی نمی‌کند. در مرحله بعد، به نحوه ادغام تعاملات بین دو ویژگی در مدل رگرسیون خطی می‌پردازیم.

۵.۳.۲ فعل و انفعالات

مدل رگرسیون خطی فرض می‌کند که تأثیر یک ویژگی بدون توجه به مقادیر سایر ویژگی‌ها یکسان است (= بدون تعامل). اما اغلب در داده‌ها تعاملاتی وجود دارد. برای پیش‌بینی تعداد دوچرخه‌های کرایه شده، ممکن است بین دما و اینکه روز کاری است یا نه، تعاملی وجود داشته باشد. شاید وقتی مردم مجبور به کار هستند، دما زیاد روی تعداد دوچرخه‌های اجاره‌ای تأثیر نمی‌گذارد، زیرا مردم هر اتفاقی بیفتند با دوچرخه اجاره‌ای به محل کار خود می‌روند. در روزهای تعطیل، بسیاری از مردم برای لذت سوار می‌شوند، اما فقط زمانی که هوا به اندازه کافی گرم باشد. وقتی صحبت از دوچرخه‌های کرایه‌ای می‌شود، ممکن است انتظار تعامل بین دما و روز کاری را داشته باشید.

چگونه می‌توانیم مدل خطی را شامل تعاملات کنیم؟ قبل از اینکه مدل خطی را برازش کنید، یک ستون به ماتریس ویژگی اضافه کنید که نشان دهنده تعامل بین ویژگی‌ها است و در ادامه مطابق معمول مدل را برازش دهید. راه حل به نوعی با سلیقه انتخاب شده است، زیرا به هیچ تغییری در مدل خطی نیاز ندارد، فقط به ستون‌های

اضافی در داده‌ها نیاز دارد. در مثال روز کاری و دما، یک ویژگی جدید اضافه می‌کنیم که برای روزهای بدون کار صفر دارد، در غیر این صورت با فرض اینکه روز کاری مقوله مرجع باشد، مقدار ویژگی دما را دارد. فرض کنید داده‌های ما به این شکل است:

work	temp
Y	25
N	12
N	30
Y	5

ماتریس داده استفاده شده توسط مدل خطی کمی متفاوت به نظر می‌رسد. جدول زیر نشان می‌دهد که اگر هیچ گونه تعاملی را مشخص نکنیم، داده‌های تهیه شده برای مدل چگونه به نظر می‌رسند. به طور معمول، این تبدیل به طور خودکار توسط هر نرم افزار آماری انجام می‌شود.

Intercept	workY	temp
1	1	25
1	0	12
1	0	30
1	1	5

ستون اول عبارت عرض از مبدا است. ستون دوم ویژگی طبقه‌بندی را با ۰ برای دسته مرجع و ۱ برای دسته دیگر رمزگذاری می‌کند. ستون سوم شامل دما است.

اگر بخواهیم مدل خطی تعامل بین دما و ویژگی روز کاری را در نظر بگیرد، باید یک ستون برای برهمنکش اضافه کنیم:

Intercept	workY	temp	workY.temp
1	1	25	25
1	0	12	0
1	0	30	0
1	1	5	5

ستون جدید "workY.temp" تعامل بین ویژگی‌های روز کاری (work) و دما (time) را نشان می‌دهد. برای مثال اگر ویژگی کار در رده مرجع ("N" برای روز غیرکاری) باشد، این ستون ویژگی جدید صفر است، در غیر این صورت مقادیر ویژگی دمای نمونه‌ها را در نظر می‌گیرد. با این نوع رمزگذاری، مدل خطی می‌تواند یک اثر خطی متفاوت دما را برای هر دو نوع روز یاد بگیرد. این اثر متقابل بین دو ویژگی است. بدون یک عبارت تعاملی، اثر ترکیبی یک ویژگی دسته‌ای و عددی را می‌توان با خطی توصیف کرد که برای دسته‌های مختلف به صورت عمودی جابجا شده است. اگر تعامل را لحاظ کنیم، اجازه می‌دهیم اثر ویژگی‌های عددی (شیب) در هر دسته مقدار متفاوتی داشته باشد.

برای تعامل دو ویژگی دسته‌ای به طور مشابه عمل می‌شود. ما ویژگی‌هایی اضافی ایجاد می‌کنیم که ترکیبی از دسته‌ها را نشان می‌دهد. در اینجا برخی از داده‌های مصنوعی حاوی روز کاری (work) و یک ویژگی دسته‌ای آب و هوا (wthr) آمده است:

work	wthr
Y	2
N	0
N	1
Y	2

در مرحله بعد، ما عبارات تعامل را اعمال می‌کنیم:

Intercept	workY	wthr1	wthr2	workY.wthr1	workY.wthr2
1	1	0	1	0	1
1	0	0	0	0	0
1	0	1	0	0	0
1	1	0	1	0	1

ستون اول برای تخمین عرض از مبدا است. ستون دوم ویژگی کار کدگذاری شده است. ستون‌های سه و چهار برای ویژگی آب و هوا هستند که به دو ستون نیاز دارند زیرا برای ثبت تاثیر برای سه دسته نیاز به دو وزن دارد که یکی از آنها دسته مرجع است. بقیه ستون‌ها تعاملات را نشان می‌دهند. برای هر دسته از هر دو ویژگی (به جز دسته‌های مرجع)، یک ستون ویژگی جدید ایجاد می‌کنیم که اگر هر دو ویژگی یک دسته خاص داشته باشند، ۱ است، در غیر این صورت ۰ است.

برای دو ویژگی عددی، ساخت ستون تعامل آسان‌تر است: ما به سادگی هر دو ویژگی عددی را ضرب می‌کنیم.

رویکردهایی برای شناسایی خودکار و افزودن اصطلاحات تعاملی وجود دارد. یکی از آنها را می‌توان در فصل RuleFit یافت. الگوریتم RuleFit ابتدا عبارات تعامل را استخراج می‌کند و سپس یک مدل رگرسیون خطی شامل برهمکنش‌ها را تخمین می‌زند.

مثال

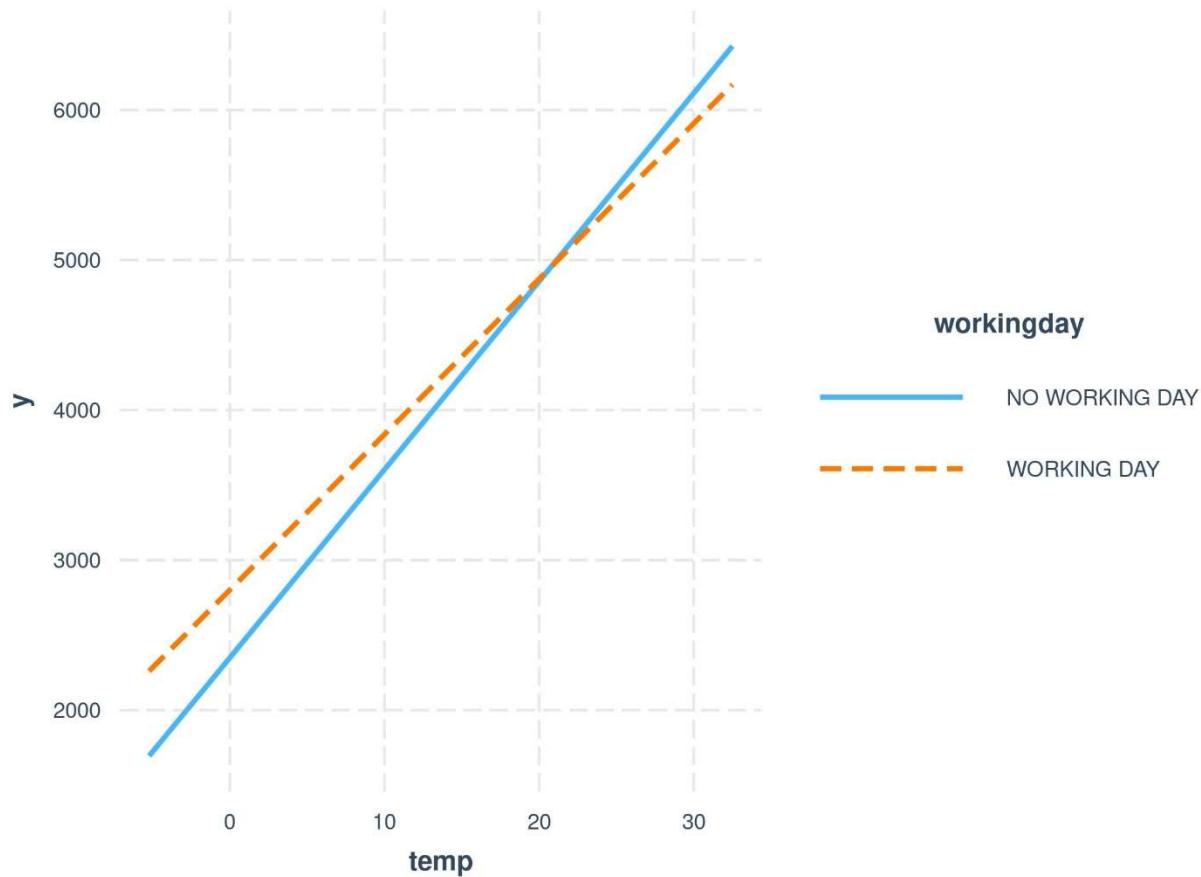
اجازه دهید به مساله پیش‌بینی اجاره دوچرخه که قبلاً در فصل مدل خطی مدل‌سازی کردۀ‌ایم بازگردیم. این بار، علاوه بر این موارد قبلی، تعامل بین ویژگی‌های دما و روز کاری را در نظر می‌گیریم. این منجر به وزن‌های تخمینی و فواصل اطمینان زیر می‌شود.

	Weight	Std. Error	2.5%	97.5%
(Intercept)	2185.8	250.2	1694.6	2677.1
seasonSPRING	893.8	121.8	654.7	1132.9
seasonSUMMER	137.1	161.0	-179.0	453.2
seasonFALL	426.5	110.3	209.9	643.2
holidayHOLIDAY	-674.4	202.5	-1071.9	-276.9
workingdayWORKING DAY	451.9	141.7	173.7	730.1
weathersitMISTY	-382.1	87.2	-553.3	-211.0
weathersitRAIN/...	-1898.2	222.7	-2335.4	-1461.0
temp	125.4	8.9	108.0	142.9
hum	-17.5	3.2	-23.7	-11.3
windspeed	-42.1	6.9	-55.5	-28.6
days_since_2011	4.9	0.2	4.6	5.3
workingdayWORKING DAY:temp	-21.8	8.1	-37.7	-5.9

اثر مقابل اضافی منفی است (-۲۱.۸) و به طور قابل توجهی با صفر متفاوت است، همان‌طور که مشاهده می‌شود فاصله اطمینان ۹۵٪ شامل صفر نمی‌شود. به هر حال، داده‌های مستقل با توزیع یکسان (Independent and identically distributed iid) نیستند، زیرا روزهای نزدیک به یکدیگر مستقل از یکدیگر نیستند. فواصل اطمینان ممکن است گمراه کننده باشد، در نتیجه زیاد آن را جدی نگیرید. عبارت تعامل، تفسیر وزن ویژگی‌های درگیر را تغییر می‌دهد. آیا دما در یک روز کاری است تأثیر منفی دارد؟ پاسخ منفی است، حتی اگر جدول آن را به یک

کاربر آموزش ندیده پیشنهاد کند. ما نمی توانیم وزن تعامل «WORKING DAY:temp» را به صورت مجزا تفسیر کنیم، زیرا این تفسیر به این صورت خواهد بود: «در حالی که همه مقادیر ویژگی های دیگر بدون تغییر باقی می مانند، افزایش اثر تعامل دما برای روز کاری، تعداد پیش بینی شده دوچرخه ها را کاهش می دهد. اما اثر متقابل فقط به اثر اصلی دما می افزاید. فرض کنید یک روز کاری است و می خواهیم بدانیم اگر امروز دمای هوا یک درجه گرمتر بود چه اتفاقی می افتد. سپس باید هر دو وزن "temp" و "workingday WORKING DAY:temp" را جمع کنیم تا تعیین کنیم تخمین چقدر افزایش می یابد.

در ک تعامل به صورت بصری آسان تر است. با معرفی یک عبارت تعاملی بین یک ویژگی دسته ای و عددی، به جای یک شبیب، دو شبیب برای دما به دست می آوریم. شبیب دما برای روزهایی که افراد مجبور به کار نیستند ("NO WORKING DAY") مستقیماً از جدول (۱۲۵.۴) قابل خواندن است. شبیب دما برای روزهایی که افراد باید در آن کار کنند ("روز کاری") مجموع هر دو وزن دما ($125.4 - 21.8 = 103.6$) است. عرض از مبدا خط NO WORKING DAY در دمای $= 0$ توسط عبارت عرض از مبدا مدل خطی (2185.8) تعیین می شود. عرض از مبدا خط "روز کاری" در دمای $= 0$ با عبارت عرض از مبدا $+ \text{اثر روز کاری}$ ($451.9 + 2185.8 = 2637.7$) تعیین می شود.



شکل ۵.۱۲: تأثیر (شامل برهمنکنش) دما و روز کاری بر تعداد پیش‌بینی شده دوچرخه‌ها با استفاده از یک مدل خطی. به طور موثر، ما دو شیب برای دما داریم، برای هر دسته از ویژگی روز کاری، یک شیب.

۵.۳.۳ GAM تأثیرات غیر خطی

دنیا خطی نیست. خطی بودن در مدل‌های خطی به این معنی است که صرفنظر از مقداری که یک نمونه در یک ویژگی خاص داشته باشد، افزایش مقدار به اندازه یک واحد همیشه همان اثر را بر خروجی پیش‌بینی شده دارد. آیا منطقی است که فرض کنیم افزایش یک درجه دما در ۱۰ درجه سانتیگراد همان تأثیری را بر تعداد دوچرخه‌های اجاره‌ای دارد که افزایش دما در حال حاضر ۴۰ درجه است؟ به طور شهودی، انتظار می‌رود که افزایش دما از ۱۰ به ۱۱ درجه سانتیگراد تأثیر مثبتی بر اجاره دوچرخه داشته باشد و از ۴۰ به ۴۱ تأثیر منفی داشته باشد، که همان‌طور که خواهید دید در بسیاری از نمونه‌ها نیز وجود دارد. ویژگی دما تأثیر خطی و مثبتی بر تعداد دوچرخه‌های اجاره‌ای دارد، اما در برخی مواقع صاف می‌شود و حتی در دماهای بالا تأثیر منفی می‌گذارد. مدل خطی به این موضوع، اهمیتی نمی‌دهد و با وظیفه‌شناسی بهترین صفحه خطی را (با به حداقل رساندن فاصله اقلیدسی) پیدا می‌کند.

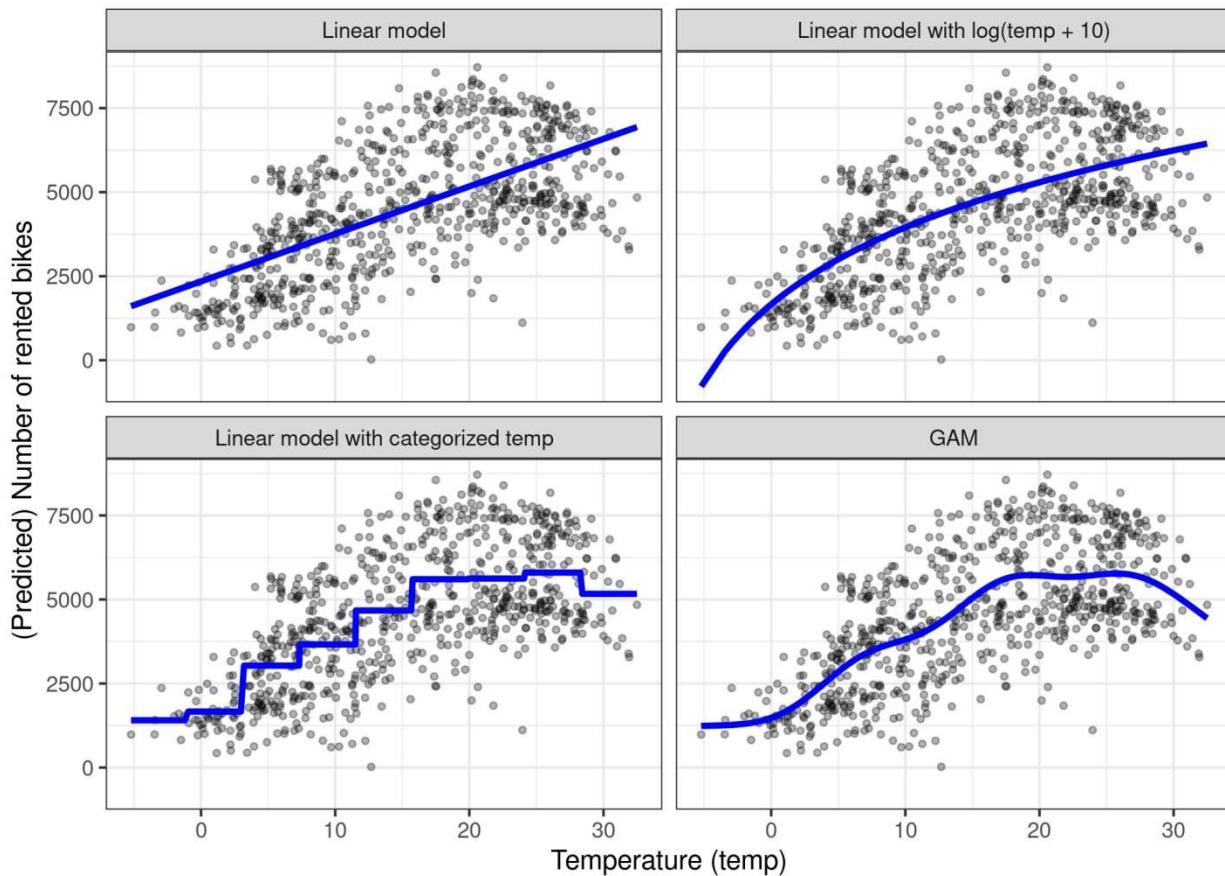
می‌توانید روابط غیرخطی را با استفاده از یکی از تکنیک‌های زیر مدل سازی کنید:

تبديل ساده ویژگی (مثلاً لگاریتم)

دسته بندی ویژگی

مدل‌های افزایشی تعمیم یافته (GAMs)

قبل از اینکه به جزئیات هر روش بپردازم، اجازه دهید با مثالی شروع کنیم که هر سه روش را نشان می‌دهد. من مجموعه‌داده اجاره دوچرخه را در نظر گرفتم و یک مدل خطی با فقط ویژگی دما برای پیش‌بینی تعداد دوچرخه‌های اجاره‌ای آموزش دادم. شکل زیر شیب برآورد شده را نشان می‌دهد: مدل خطی استاندارد، مدل خطی با دمای تبدیل شده (لگاریتم)، مدل خطی با دما به عنوان ویژگی طبقه‌بندی شده و با استفاده از خطوط رگرسیون (GAM).



شکل ۱۳.۵: پیش‌بینی تعداد دوچرخه‌های اجاره‌ای تنها با استفاده از ویژگی دما. یک مدل خطی (بالا سمت چپ) به خوبی با داده‌ها مطابقت ندارد. راه حل این است که ویژگی را با لگاریتم ویژگی (بالا سمت راست) جایگزین کنید، یا ویژگی را دسته بندی کنید (پایین سمت چپ)، که معمولاً یک تصمیم اشتباه است، یا استفاده از مدل‌های افزودنی تعمیم یافته که می‌تواند به طور خودکار یک منحنی نرم را برای دما تنظیم کند (پایین سمت راست).

تبدیل ویژگی

اغلب از لگاریتم ویژگی به عنوان تبدیل استفاده می‌شود. استفاده از لگاریتم نشان می‌دهد که هر 10 برابر افزایش دما تأثیر خطی یکسانی بر تعداد دوچرخه‌ها دارد، بنابراین تغییر از 1 درجه سانتیگراد همان تأثیر تغییر از 0.1 به 1 را دارد (که به نظر اشتباه می‌رسد). مثال‌های دیگر برای تبدیل ویژگی عبارت‌اند از: ریشه مربع، تابع مربع و تابع نمایی. استفاده از تبدیل ویژگی به این معنی است که ستون این ویژگی را در داده‌ها با تابعی از ویژگی مانند لگاریتم جایگزین می‌کنید و طبق معمول مدل خطی را برازش می‌کنید. برخی از برنامه‌های آماری همچنین به شما اجازه می‌دهند که تبدیل‌ها را در فراخوانی مدل خطی مشخص کنید. وقتی ویژگی را تغییر می‌دهید می‌توانید خلاق باشید. تفسیر ویژگی با توجه به تبدیل انتخاب شده تغییر می‌کند. اگر از تبدیل \log استفاده می‌کنید، تفسیر در یک مدل خطی به این صورت می‌شود: "اگر لگاریتم ویژگی یک افزایش یابد، پیش‌بینی با وزن

مربوطه افزایش می‌یابد." وقتی از GLM باتابع اتصال استفاده می‌کنید که تابع تشخیص نیست، تفسیر پیچیده‌تر می‌شود، زیرا باید هر دو تبدیل را در تفسیر بگنجانید (به جز زمانی که یکدیگر را خنثی می‌کنند، مانند \log و \exp ، که در این حالت تفسیر راحت‌تر می‌شود).

دسته‌بندی ویژگی‌ها

امکان دیگر برای مواجه با یک اثر غیرخطی، گسسته کردن ویژگی است. آن ویژگی را به یک ویژگی طبقه‌بندی تبدیل کنید. به عنوان مثال، می‌توانید ویژگی دما را به ۲۰ بازه با سطوح [۱۰، -۵)، [-۵، ۰)، و ... و غیره تقسیم کنید. هنگامی که شما از دمای طبقه‌بندی شده به جای دمای پیوسته استفاده می‌کنید، مدل خطی یک تابع پله ای را تخمین می‌زند زیرا هر سطح تخمین خاص خود را دارد. مشکل این رویکرد این است که به داده‌های بیشتری نیاز دارد، احتمال بیشتری وجود دارد که بیش از حد برازش رخ دهد و مشخص نیست که چگونه ویژگی را به طور معناداری گسسته کنیم (فاصله‌های مساوی یا چندک؟ چه تعداد بازه؟). من فقط در صورتی از گسسته سازی استفاده می‌کنم که یک دلیل بسیار قوی برای آن وجود داشته باشد. به عنوان مثال، برای این که بتوان مدل را با مطالعه دیگری مقایسه نمود.

مدل‌های افزایشی تعمیم یافته (GAM)

چرا به مدل خطی (تعمیم یافته) اجازه نمی‌دهیم روابط غیرخطی را یاد بگیرد؟ این انگیزه پشت GAM‌ها است. GAM‌ها این محدودیت را کاهش می‌دهند که رابطه باید یک جمع وزنی ساده باشد، و در عوض فرض می‌کنند که نتیجه می‌تواند با مجموع توابع دلخواه هر ویژگی مدل شود. از نظر ریاضی، رابطه در یک GAM به شکل زیر است:

$$g(E_Y(y|x)) = \beta_0 + f_1(x_1) + f_2(x_2) + \dots + f_p(x_p)$$

فرمول مشابه فرمول GLM است با این تفاوت که عبارت خطی $\beta_j x_j$ با یک تابع انعطاف پذیرتر ($f_j(x_j)$) جایگزین می‌شود. هسته یک GAM هنوز مجموع اثرات ویژگی است، اما شما این گزینه را دارید که اجازه دهید روابط غیرخطی بین برخی ویژگی‌ها و خروجی وجود داشته باشد. اثرات خطی نیز توسط چارچوب پوشش داده می‌شوند، برای اینکه ویژگی‌ها به صورت خطی مدیریت شوند، می‌توانید $(x_j)_j f_j$ را به شکل $\beta_j x_j$ محدود کنید.

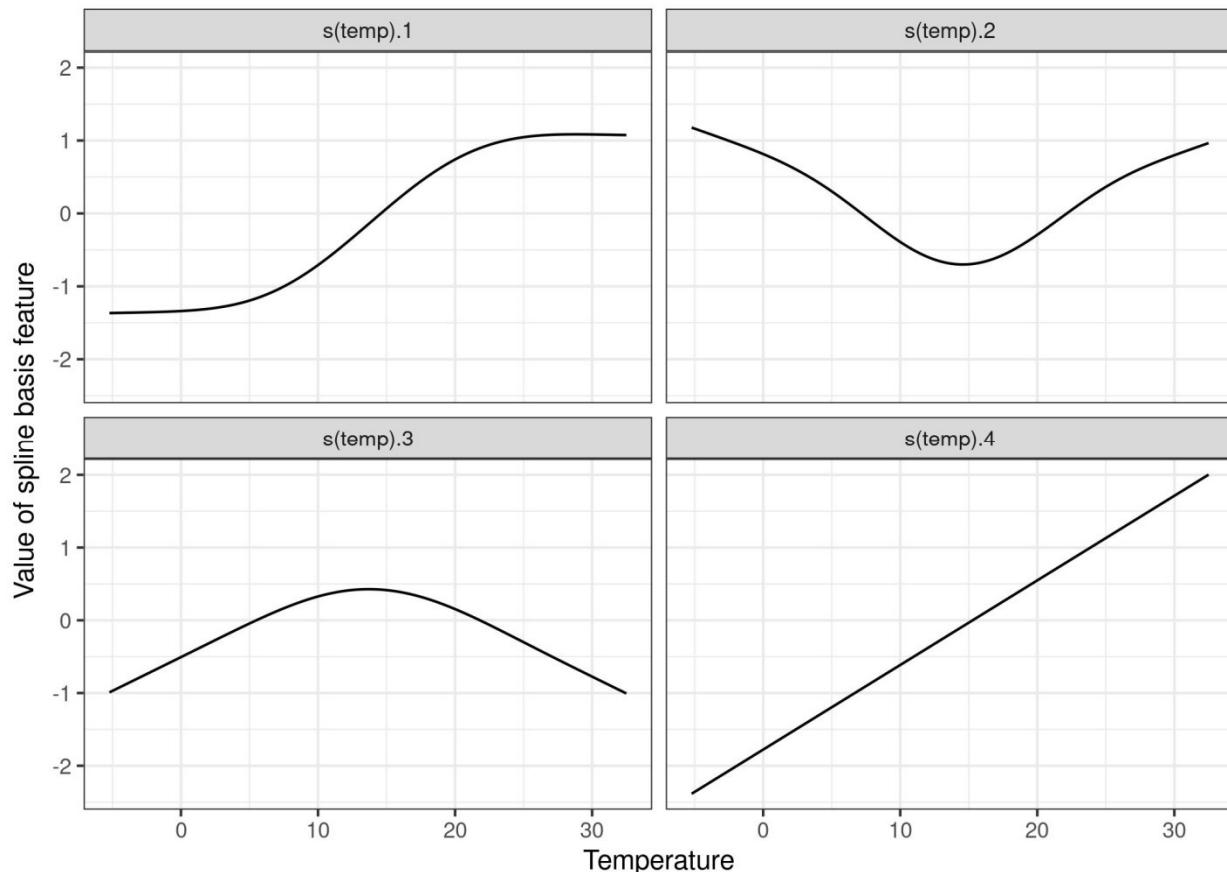
سوال بزرگ این است که چگونه توابع غیرخطی را یاد بگیریم. به این پاسخ «اسپلاین» یا «توابع اسپلاین» می‌گویند. اسپلاین‌ها توابعی هستند که از توابع پایه ساده‌تر ساخته می‌شوند. اسپلاین‌ها را می‌توان برای تقریب سایر توابع پیچیده‌تر استفاده کرد. کمی شبیه چیدن آجرهای لگو برای ساختن چیزی پیچیده‌تر. راههای گیج کننده‌ای برای تعریف این توابع پایه اسپلاین وجود دارد. اگر علاقه‌مند به کسب اطلاعات بیشتر در مورد تمام روش‌های تعریف توابع پایه هستید، برای شما در سفرтан آرزوی موفقیت می‌کنم. من قصد ندارم در اینجا وارد جزئیات شوم، من فقط قصد دارم یک شهود بسازم. چیزی که شخصاً بیشترین کمک را به من برای درک اسپلاین کرد، تجسم توابع

پایه فردی و بررسی چگونگی اصلاح ماتریس داده بود. به عنوان مثال، برای مدل سازی دما با اسپلاین، ما ویژگی دما را از داده‌ها حذف می‌کنیم و مثلاً ۴ ستون را جایگزین آن می‌کنیم که هر کدام یک تابع پایه اسپلاین را نشان می‌دهند. معمولاً توابع پایه‌ای اسپلاین بیشتری خواهید داشت، من فقط برای تجسم بهتر، تعداد را کاهش دادم. مقدار هر نمونه از این ویژگی‌های جدید پایه اسپلاین به مقادیر دمای نمونه‌ها بستگی دارد. همراه با تمام اثرات خطی، GAM سپس این وزن‌های اسپلاین را نیز تخمین می‌زند. GAM‌ها همچنین برای اوزان یک عبارت جریمه در نظر می‌گیرند تا آنها را نزدیک به صفر نگه دارد. این کار به طور موثر، انعطاف پذیری اسپلاین‌ها و برازش بیش از حد را کاهش می‌دهد. سپس یک پارامتر صافی که معمولاً برای کنترل انعطاف پذیری منحنی استفاده می‌شود، از طریق اعتبارسنجی متقطع (cross-validation) تنظیم می‌شود. با نادیده گرفتن عبارت جریمه، مدل سازی غیرخطی با اسپلاین، مهندسی ویژگی‌های فانتزی است.

در مثالی که ما تعداد دوچرخه‌ها را با GAM فقط با استفاده از دما پیش‌بینی می‌کنیم، ماتریس ویژگی مدل به این صورت است:

(Intercept)	s(temp).1	s(temp).2	s(temp).3	s(temp).4
1	-0.93	-0.14	0.21	-0.83
1	-0.83	-0.27	0.27	-0.72
1	-1.32	0.71	-0.39	-1.63
1	-1.32	0.70	-0.38	-1.61
1	-1.29	0.58	-0.26	-1.47
1	-1.32	0.68	-0.36	-1.59

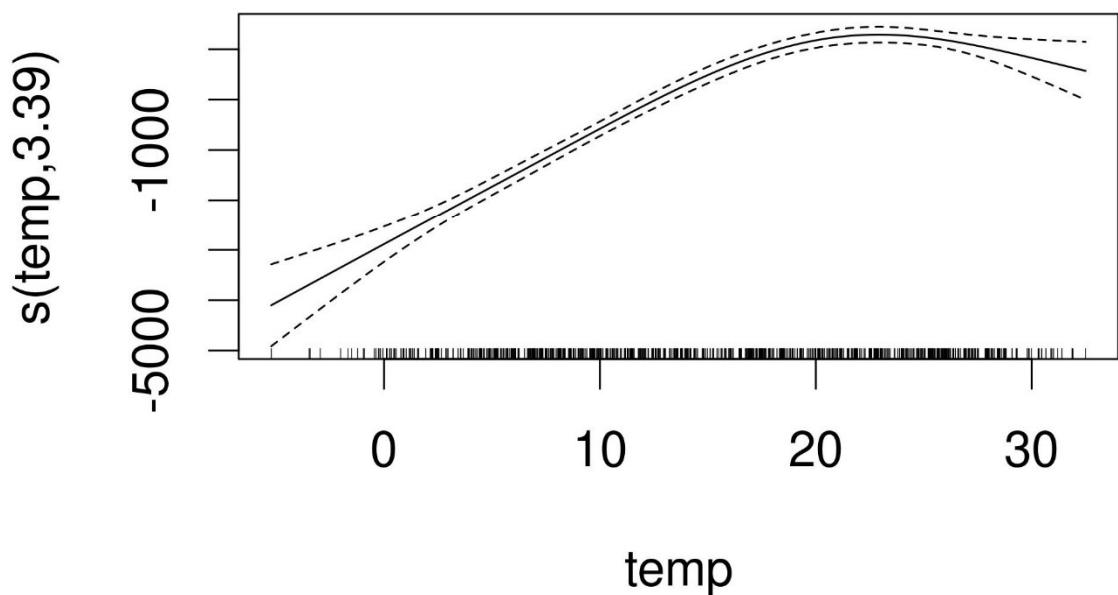
هر ردیف نشان دهنده یک نمونه از داده‌ها (یک روز) است. هر ستون پایه اسپلاین حاوی مقدار تابع پایه اسپلاین در مقادیر دمایی خاص است. شکل زیر نشان می‌دهد که این توابع پایه اسپلاین چگونه هستند:



شکل ۵.۱۴: برای مدل سازی هموار اثر دما، از ۴ تابع پایه اسپلاین استفاده می‌کنیم. در اینجا هر مقدار دما به ۴ مقدار پایه اسپلاین نگاشت می‌شود. اگر دمای یک نمونه ۳۰ درجه سانتیگراد باشد، مقدار اولین ویژگی پایه اسپلاین ۱- می‌شود، برای دومی ۰.۷، برای سومی -۰.۸ و برای چهارمین ۱.۷. GAM وزن‌هایی را به هر ویژگی پایه درجه حرارت اختصاص می‌دهد:

	weight
(Intercept)	4504.35
s(temp).1	989.34
s(temp).2	740.08
s(temp).3	2309.84
s(temp).4	558.27

و منحنی اسپلاین واقعی که از مجموع وزن دار توابع پایه اسپلاین با وزن‌های تخمینی حاصل می‌شود، به شکل زیر است:



شکل ۵.۱۵: اثر ویژگی GAM دما برای پیش‌بینی تعداد دوچرخه‌های کرایه شده (دماهی به عنوان تنها ویژگی استفاده شده است).

تفسیر اثرات هموار نیاز به بررسی بصری منحنی برآذش دارد. اسپلاین‌ها معمولاً حول میانگین پیش‌بینی متتمرکز می‌شوند، بنابراین یک نقطه در منحنی تفاوت با پیش‌بینی میانگین است. به عنوان مثال، در دماهی صفر درجه سانتیگراد، تعداد دوچرخه‌های پیش‌بینی شده ۳۰۰۰ واحد کمتر از میانگین پیش‌بینی است.

۵.۳.۴ مزایا

همه این تعمیمات مدل خطی به خودی خود، مقداری جهان را توصیف می‌کنند. با هر مشکلی که با مدل‌های خطی مواجه می‌شوید، احتمالاً تعمیمی خواهید یافت که آن را برطرف می‌کند.

بیشتر روش‌ها برای چندین دهه مورد استفاده قرار گرفته‌اند. به عنوان مثال، GAM‌ها تقریباً ۳۰ سال از عمر شان می‌گذرد. بسیاری از محققان و دست اندکاران صنعت در استفاده از مدل‌های خطی بسیار با تجربه هستند و این روش‌ها در بسیاری از جوامع به عنوان وضعیت موجود برای مدل سازی پذیرفته شده است.

علاوه بر پیش‌بینی، می‌توانید از مدل‌ها برای استنتاج، نتیجه‌گیری در مورد داده‌ها استفاده کنید – با توجه به اینکه مفروضات مدل نقض نمی‌شوند. شما فواصل اطمینان برای وزن‌ها، آزمون‌های معناداری، فواصل پیش‌بینی و موارد دیگر را دریافت می‌کنید.

نرم‌افزارهای آماری معمولاً دارای رابطه‌ای واقعاً خوبی برای برازش با GLM، GAM و مدل‌های خطی خاص‌تر هستند.

مشکل بسیاری از مدل‌های یادگیری ماشین ناشی از ۱) عدم تنکی (sparseness) است، به این معنی که تعداد زیادی از ویژگی‌ها استفاده می‌شوند، ۲) با ویژگی‌ها به صورت غیرخطی رفتار می‌شوند، به این معنی که برای توصیف اثر به بیش از یک وزن نیاز دارید، و ۳) مدل سازی تعاملات بین ویژگی‌ها. با فرض اینکه مدل‌های خطی بسیار قابل تفسیر هستند، اما اغلب با واقعیت مناسب نیستند، تعمیمات شرح داده شده در این فصل راه خوبی برای دستیابی به یک انتقال هموار به مدل‌های انعطاف‌پذیرتر ارائه می‌دهند، در حالی که برخی از قابلیت تفسیر را حفظ می‌کنند.

۵.۳.۵ معایب

به عنوان مزیت گفتم که مدل‌های خطی در جهان خودشان زندگی می‌کنند. تعداد زیادی از روش‌هایی که می‌توانید برای مدل خطی ساده تعمیم دهید، نه فقط برای مبتدیان، بلکه برای همگان بسیار زیاد است. در عمل، جهان‌های موازی متعددی وجود دارد، زیرا بسیاری از جوامع محققان و پزشکان نامهای خاص خود را برای روش‌هایی دارند که کم و بیش یک کار را انجام می‌دهند، که می‌تواند بسیار گیج‌کننده باشد.

اکثر اصلاحات مدل خطی باعث می‌شود که مدل کمتر قابل تفسیر باشد. هر تابع اتصال (در GLM) که تابع تشخیص نباشد، تفسیر را پیچیده می‌کند. تعاملات ویژگی‌ها نیز تفسیر را پیچیده می‌کند. تاثیرات ویژگی غیرخطی یا کمتر بصری هستند (مانند تبدیل \log) یا دیگر نمی‌توان آنها را در یک عدد خلاصه کرد (مثلاً توابع اسپیلاین). GAMها، GLMها و غیره بر فرضیات مربوط به فرآیند تولید داده تکیه دارند. اگر آنها نقض شوند، دیگر تفسیر اوزان معتبر نیست.

عملکرد مجموعه‌های مبتنی بر درخت مانند جنگل تصادفی یا تقویت درخت گرادیان (gradient tree boosting) در بسیاری موارد بهتر از پیچیده‌ترین مدل‌های خطی است. این بخشی از تجربه شخصی من و بخشی از مشاهدات از مدل‌های برنده در سیستم عامل‌هایی مانند kaggle.com است.

۵.۳.۶ نرم افزار

تمام مثال‌های این فصل با استفاده از زبان R پیاده سازی شده‌اند. برای GAMها از gam پکیج استفاده شد، اما نرم افزارهای متعدد دیگری نیز موجود می‌باشد. R دارای تعداد بسیار زیادی پکیج برای گسترش مدل‌های رگرسیون خطی است. بدون پیشی گرفتن از هر زبان تجزیه و تحلیل دیگری، R اولین گزینه برای هر تعمیمی از مدل رگرسیون

خطی است. شما پیاده سازی هایی از GAMها را در پایتون پیدا خواهید کرد (مانند pyGAM)، اما این پیاده سازی ها آنقدرها هم کامل نیستند.

۵.۳.۷ تعمیمات بیشتر

همان طور که وعده داده شده بود، در اینجا لیستی از مشکلاتی که ممکن است در مدل های خطی با آنها مواجه شوید، همراه با نام راه حلی که برای این مشکل وجود دارد، آورده شده است. می توانید نام روش را کپی و در موتور جستجوی مورد علاقه خود فراخوانی کنید.

داده های من فرض مستقل بودن و توزیع یکسان (iid) را نقض می کند.
به عنوان مثال، اندازه گیری های مکرر روی یک بیمار.

جستجوی مدل های ترکیبی (mixed models) یا معادلات تخمین تعمیم یافته (generalized estimating equations) است.

مدل من دارای خطاهای هم واریانس (heteroscedastic) است.
به عنوان مثال، هنگام پیش بینی ارزش یک خانه، خطاهای مدل معمولاً در خانه های گران قیمت بیشتر است، که همسانی مدل خطی را نقض می کند.

جستجوی رگرسیون مقاوم (robust regression) من نقاط پرت دارم که به شدت بر مدل من تأثیر می گذارد.
جستجوی رگرسیون مقاوم (robust regression) من می خواهم زمان رخ دادن یک رویداد را پیش بینی کنم.

داده های زمان تا رویداد معمولاً با اندازه گیری های سانسور شده ارائه می شوند، به این معنی که در برخی موارد زمان کافی برای مشاهده رویداد وجود نداشت. به عنوان مثال، یک شرکت می خواهد خرابی ماشین های یخ خود را پیش بینی کند، اما فقط برای دو سال اطلاعات دارد. برخی از ماشین ها پس از دو سال هنوز سالم هستند، اما ممکن است بعداً از کار بیفتند.

جستجو برای مدل های پارامتریک بقا (cox regression)، رگرسیون کاکس (parametric survival models)، تجزیه و تحلیل بقا (survival analysis) است.

اگر نتیجه دارای دو دسته است از مدل رگرسیون لجستیک استفاده کنید که احتمال را برای دسته ها مدل می کند.
اگر دسته های بیشتری دارید، رگرسیون چند جمله ای (multinomial regression) را جستجو کنید.
رگرسیون لجستیک و رگرسیون چند جمله ای هر دو GLM هستند.
من می خواهم دسته بندی های مرتب شده را پیش بینی کنم.

به عنوان مثال نمرات مدرسه.

مدل بخت‌های متناسب (proportional odds model) را جستجو کنید.

نتیجه من یک شمارش است (مانند تعداد فرزندان در یک خانواده).

جستجوی رگرسیون پواسون (Poisson regression)

مدل پواسون نیز GLM است. همچنین ممکن است این مشکل را داشته باشید که مقادیر شمارش شده + بسیار زیاد است.

جستجوی مدل رگرسیون پواسون با صفر انباسته (zero-inflated Poisson regression model)، مدل هاردل (hurdle model).

من مطمئن نیستم که چه ویژگی‌هایی باید در مدل گنجانده شود تا نتیجه گیری‌های علی درست انجام شود. به عنوان مثال، می‌خواهم بدانم اثر یک دارو بر فشار خون چیست. این دارو بر برخی از ارزش‌های خونی تأثیر

مستقیم دارد و این ارزش خونی بر نتیجه تأثیر می‌گذارد. آیا باید مقدار خون را در مدل رگرسیون لحاظ کنم؟

جستجو برای استنباط علیت (causal inference)، تحلیل میانجی (mediation analysis) من داده‌های گم شده دارم

جست و جو برای انتساب جانه‌ی چندگانه (multiple imputation).

من می‌خواهم دانش قبلی را در مدل‌های خود ادغام کنم.

جستجو برای استنباط بیزی (Bayesian inference) من اخیراً کمی احساس ضعف دارم.

جستجو برای "Amazon Alexa Gone Wild!!!!" نسخه کامل از ابتدا تا انتهای

۵.۴ درخت تصمیم

مدل‌های رگرسیون خطی و رگرسیون لجستیک در شرایطی که رابطه بین ویژگی‌ها و نتیجه غیرخطی است یا جایی که ویژگی‌ها با یکدیگر تعامل دارند، شکست می‌خورند. زمان درخشش برای درخت تصمیم است! مدل‌های

مبتنی بر درخت، داده‌ها را چندین بار بر اساس مقادیر قطعی مشخص در ویژگی‌ها تقسیم می‌کنند. از طریق تقسیم، زیرمجموعه‌های مختلفی از مجموعه داده ایجاد می‌شود که هر نمونه متعلق به یک زیر مجموعه است. زیر

مجموعه‌های نهایی را گره‌های پایانی یا برگ و زیر مجموعه‌های میانی را گره‌های داخلی یا گره‌های تقسیم می‌نامند. برای پیش‌بینی نتیجه در هر گره برگ، از میانگین نتیجه داده‌های آموزشی در این گره استفاده می‌شود.

درختان را می‌توان برای طبقه‌بندی و رگرسیون استفاده کرد.

الگوریتم‌های مختلفی وجود دارند که می‌تواند درخت را گسترش دهنند. آنها در ساختار احتمالی درخت (به عنوان مثال تعداد شکاف در هر گره)، معیارهای چگونگی یافتن شکاف‌ها، زمان توقف تقسیم و نحوه تخمین