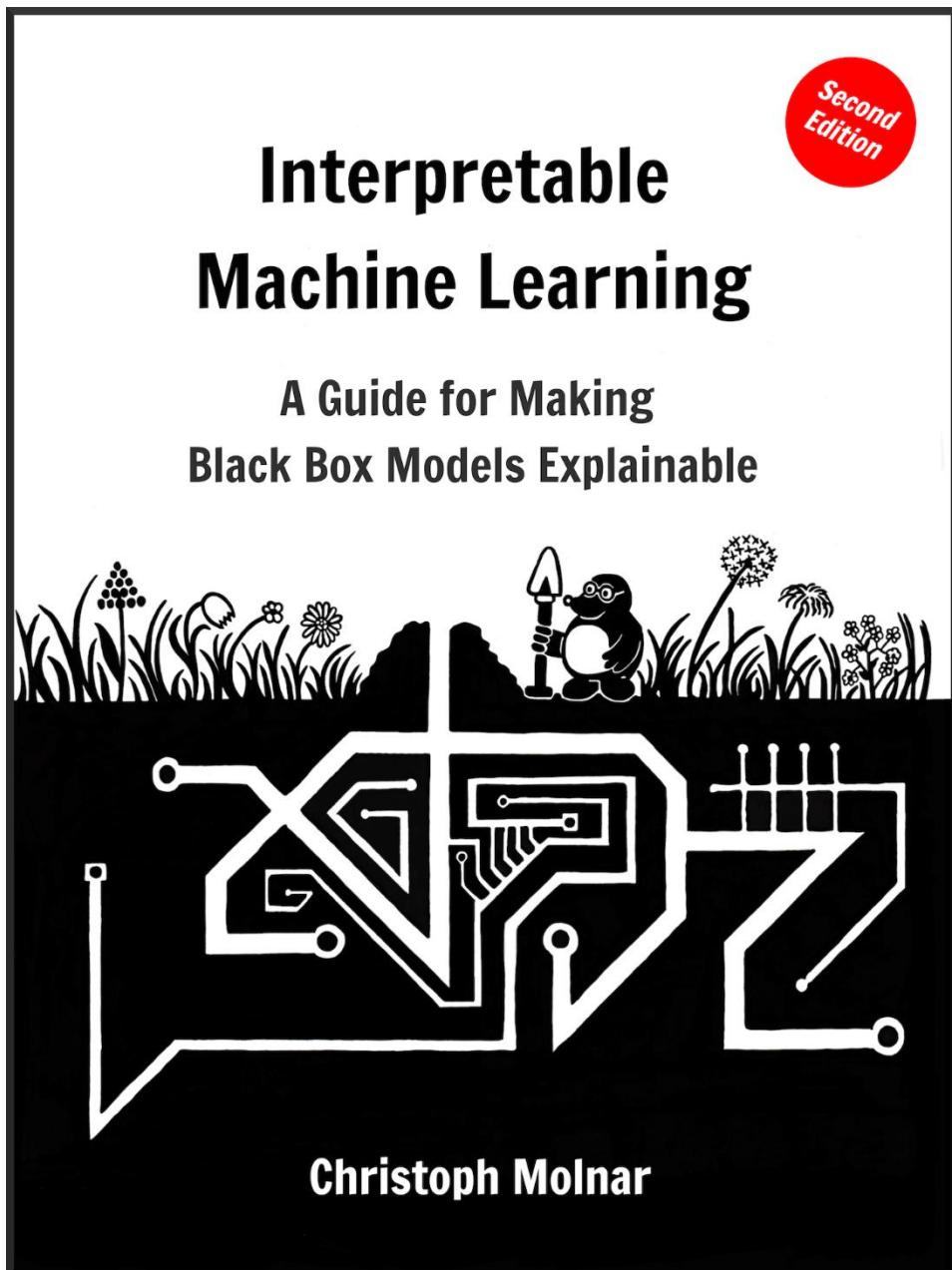


۱

یادگیری ماشینی قابل تفسیر

۲

راهنمای ساخت مدل های جعبه سیاه قابل توضیح



۳

۴

فهرست مطالب

۵	
۶	
۷	۹ خلاصه
۸	۱۰ فصل ۱ پیشگفتار نویسنده
۹	۱۴ فصل ۲ مقدمه
۱۰	۱۵ ۱ زمان داستان
۱۱	۲۲ ۲،۲ یادگیری ماشینی چیست؟
۱۲	۲۴ ۲،۳ اصطلاحات
۱۳	۲۹ فصل ۳ تفسیرپذیری
۱۴	۲۹ ۳،۱ اهمیت تفسیرپذیری
۱۵	۳۷ ۳،۲ طبقه بندی روش های تفسیرپذیری
۱۶	۳۹ ۳،۳ دامنه تفسیرپذیری
۱۷	۳۹ ۳،۳،۱ شفافیت الگوریتم
۱۸	۳۹ ۳،۳،۲ تفسیرپذیری مدل کل نگر
۱۹	۴۰ ۳،۳،۳ تفسیرپذیری مدل جهانی در سطح مدولار
۲۰	۴۰ ۳،۳،۴ تفسیر محلی برای یک پیش بینی واحد
۲۱	۴۱ ۳،۳،۵ تفسیر محلی برای گروهی از پیش بینی ها
۲۲	۴۱ ۳،۴ ارزیابی تفسیرپذیری
۲۳	۴۲ ۳،۵ خواص توضیحات
۲۴	۴۵ ۳،۶ توضیحات انسان پسند
۲۵	۴۵ ۳،۶،۱ توضیح چیست؟
۲۶	۴۶ ۳،۶،۲ یک توضیح خوب چیست؟
۲۷	۵۱ فصل ۴ مجموعه داده ها
۲۸	۵۱ ۴،۱ اجراء دوچرخه (بازگشت)
۲۹	۵۲ ۴،۲ نظرات هرزنامه (YouTube) طبقه بندی متن
۳۰	۵۳ ۴،۳ عوامل خطر برای سرطان دهانه رحم (طبقه بندی)
۳۱	۵۵ فصل ۵ مدل های قابل تفسیر
۳۲	۵۶ ۵،۱ رگرسیون خطی

۳۳	۵۹ تفسیر ۵,۱,۱
۳۴	۶۱ مثال ۵,۱,۲
۳۵	۶۳ تفسیر بصری ۵,۱,۳
۳۶	۶۵ ۵,۱,۴ پیش بینی های فردی را توضیح دهید
۳۷	۶۷ ۵,۱,۵ رمزگذاری ویژگی های دسته بندی
۳۸	۶۹ ۵,۱,۶ آیا مدل های خطی توضیحات خوبی ایجاد می کنند؟
۳۹	۷۰ ۵,۱,۷ مدل های خطی پراکنده
۴۰	۷۳ ۵,۱,۸ مزایا
۴۱	۷۴ ۵,۱,۹ معایب
۴۲	۷۴ ۵,۲ رگرسیون لجستیک
۴۳	۷۴ ۵,۲,۱ رگرسیون خطی برای طبقه بندی چه اشکالی دارد؟
۴۴	۷۶ ۵,۲,۲ نظریه
۴۵	۷۸ ۵,۲,۳ تفسیر
۴۶	۸۰ ۵,۲,۴ مثال
۴۷	۸۱ ۵,۲,۵ مزایا و معایب
۴۸	۸۲ ۵,۲,۶ نرم افزار
۴۹	۸۲ ۵,۳ GAM و GLM
۵۰	۸۴ ۵,۳,۱ نتایج غیر گاووسی - GLMS
۵۱	۹۰ ۵,۳,۲ فعل و انفعالات
۵۲	۹۴ ۵,۳,۳ جلوه های غیر خطی GAM
۵۳	۱۰۱ ۵,۳,۴ مزایا
۵۴	۱۰۱ ۵,۳,۵ معایب
۵۵	۱۰۲ ۵,۳,۶ نرم افزار
۵۶	۱۰۲ ۵,۳,۷ برنامه های افزودنی بیشتر
۵۷	۱۰۴ ۵,۴ درخت تصمیم
۵۸	۱۰۶ ۵,۴,۱ تفسیر
۵۹	۱۰۷ ۵,۴,۲ مثال
۶۰	۱۰۹ ۵,۴,۳ مزایا

۶۱	۱۱۰	۵,۴,۴ معايip
۶۲	۱۱۱	۵,۴,۵ نرم افزار
۶۳	۱۱۱	۵,۵ قولنين تصميم گيري
۶۴	۱۱۴	۱,۵,۵ يادگيری قولنين از يك ويزگی واحد(OneR)
۶۵	۱۱۷	۲,۵,۵ پوشش متوالی
۶۶	۱۲۲	۳,۵,۵ فهرست قولنين بيزى
۶۷	۱۲۸	۴,۵,۵ مزايا
۶۸	۱۲۹	۵,۵,۵ معايip
۶۹	۱۳۰	۶,۵,۵ نرم افزار و جايگرین
۷۰	۱۳۱	۶,۵ RuleFit
۷۱	۱۳۲	۱,۶,۵ تفسير و مثال
۷۲	۱۳۴	۲,۶,۵ نظرية
۷۳	۱۳۸	۳,۶,۵ مزايا
۷۴	۱۳۸	۴,۶,۵ معايip
۷۵	۱۴۰	۷,۵ ساير مدل های قابل تفسير
۷۶	۱۴۰	۱,۷,۵ طبقه بندی کننده ساده بيز
۷۷	۱۴۰	۲,۷,۵ K-نژدیکترین همسایه ها
۷۸	۱۴۲	۸,۵ فصل عروش های مدل-آگنوستيك
۷۹	۱۴۶	۸,۶ فصل توضیحات مبتنی بر مثال
۸۰	۱۴۹	۸,۷ فصل ۸,۷,۵ امدل جهانی-روشهای آگنوستيك
۸۱	۱۵۰	۱,۸,۱ طرح وابستگی جزئی(PDP)
۸۲	۱۵۱	۱,۱,۸,۱ اهمیت ویژگی مبتنی بر PDP
۸۳	۱۵۲	۲,۱,۸,۱ مثال
۸۴	۱۵۵	۳,۱,۸,۱ مزايا
۸۵	۱۵۶	۴,۱,۸,۱ معايip
۸۶	۱۵۷	۵,۱,۸,۱ نرم افزار و جايگرین
۸۷	۱۵۸	۲,۸,۱ طرح جلوه های محلی انباشته(ALE)
۸۸	۱۵۸	۱,۲,۸,۱ انگیزه و شهود

۸۹	۱۶۲	نظریه ۸,۲,۲
۹۰	۱۶۳	برآورد ۸,۲,۳
۹۱	۱۶۷	مثالها ۸,۲,۴
۹۲	۱۷۷	مزایا ۸,۲,۵
۹۳	۱۷۸	معایب ۸,۲,۶
۹۴	۱۸۰	اجرا و جایگزین ۸,۲,۷
۹۵	۱۸۱	تعامل با ویژگی ها ۸,۳
۹۶	۱۸۱	تعامل ویژگی؟ ۸,۳,۱
۹۷	۱۸۲	نظریه: آماره H فریدمن ۸,۳,۲
۹۸	۱۸۴	مثالها ۸,۳,۳
۹۹	۱۸۶	مزایا ۸,۳,۴
۱۰۰	۱۸۷	معایب ۸,۳,۵
۱۰۱	۱۸۸	پیاده سازی ها ۸,۳,۶
۱۰۲	۱۸۸	گزینه های جایگزین ۸,۳,۷
۱۰۳	۱۸۹	تجزیه عملکردی ۸,۴
۱۰۴	۱۹۱	چگونه کامپوننت ها را محاسبه نکنیم ۸,۴,۱
۱۰۵	۱۹۲	تجزیه عملکردی ۸,۴,۲
۱۰۶	۱۹۳	چگونه کامپوننت ها را محاسبه نکنیم II ۸,۴,۳
۱۰۷	۱۹۴	ANOVA عملکردی ۸,۴,۴
۱۰۸	۱۹۶	ANOVA عملکردی تعییم یافته برای ویژگی های وابسته ۸,۴,۵
۱۰۹	۱۹۷	نمودارهای اثر محلی انباسته شده ۸,۴,۶
۱۱۰	۱۹۸	مدل های رگرسیون آماری ۸,۴,۷
۱۱۱	۱۹۹	پاداش: طرح وابستگی جزئی ۸,۴,۸
۱۱۲	۲۰۹	مزایا ۸,۴,۹
۱۱۳	۲۰۰	معایب ۸,۴,۱۰
۱۱۴	۲۰۱	اهمیت ویژگی جایگشت ۸,۵
۱۱۵	۲۰۱	نظریه ۸,۵,۱
۱۱۶	۲۰۲	آیا باید اهمیت داده های آموزش یا آزمون را محاسبه کنم؟ ۸,۵,۲

۱۱۷	۲۰۵	۸,۵,۳	مثال و تفسیر
۱۱۸	۲۰۷	۸,۵,۴	مزایا
۱۱۹	۲۰۹	۸,۵,۵	معایب
۱۲۰	۲۱۰	۸,۵,۶	گزینه های جایگزین
۱۲۱	۲۱۱	۸,۵,۷	نرم افزار
۱۲۲	۲۱۲	۸,۶	جانشین جهانی
۱۲۳	۲۱۲	۸,۶,۱	نظريه
۱۲۴	۲۱۴	۸,۶,۲	مثال
۱۲۵	۲۱۶	۸,۶,۳	مزایا
۱۲۶	۲۱۷	۸,۶,۴	معایب
۱۲۷	۲۱۷	۸,۶,۵	نرم افزار
۱۲۸	۲۱۸	۸,۷	نمونه های اولیه و انتقادات
۱۲۹	۲۱۹	۸,۷,۱	نظريه
۱۳۰	۲۲۵	۸,۷,۲	مثالها
۱۳۱	۲۲۶	۸,۷,۳	مزایا
۱۳۲	۲۲۶	۸,۷,۴	معایب
۱۳۳	۲۲۷	۸,۷,۵	کد و جایگزین
۱۳۴	۲۲۸	۸,۷,۶	فصل مدل محلی-روش های آگنوستیک
۱۳۵	۲۲۹	۹,۱	انتظار شرطی فردی (ICE)
۱۳۶	۲۲۹	۹,۱,۱	مثال ها
۱۳۷	۲۳۴	۹,۱,۲	مزایا
۱۳۸	۲۳۴	۹,۱,۳	معایب
۱۳۹	۲۳۴	۹,۱,۴	نرم افزار و جایگزین
۱۴۰	۲۳۶	۹,۲,۱	LIME برای داده های جدولی
۱۴۱	۲۳۹	۹,۲,۱,۱	مثال
۱۴۲	۲۴۱	۹,۲,۲	LIME برای متن
۱۴۳	۲۴۲	۹,۲,۳	LIME برای تصاویر
۱۴۴	۲۴۳	۹,۲,۴	مزایا

۱۴۵	۲۴۴	۹,۲,۵ معايب
۱۴۶	۲۴۶	۹,۳ توضيحات خلاف واقع
۱۴۷	۲۴۹	۹,۳,۱ ايجاد توضيحات خلاف واقع
۱۴۸	۲۵۵	۹,۳,۲ مثال
۱۴۹	۲۵۶	۹,۳,۳ مزايا
۱۵۰	۲۵۷	۹,۳,۴ معايب
۱۵۱	۲۵۷	۹,۳,۵ نرم افzar و جايگرین
۱۵۲	۲۵۹	۹,۴ قوانين محدوده (لنگرها)
۱۵۳	۲۶۲	۹,۴,۱ يافتن لنگرها
۱۵۴	۲۶۴	۹,۴,۲ پيچيدگي و زمان اجرا
۱۵۵	۲۶۵	۹,۴,۳ مثال داده های جدولی
۱۵۶	۲۷۰	۹,۴,۴ مزايا
۱۵۷	۲۷۰	۹,۴,۵ معايب
۱۵۸	۲۷۱	۹,۴,۶ نرم افzar و جايگرین
۱۵۹	۲۷۲	۹,۵ ارزش های شپلي
۱۶۰	۲۷۲	۹,۵,۱ ايده کلي
۱۶۱	۲۷۶	۹,۵,۲ مثال ها و تفسير
۱۶۲	۲۷۸	۹,۵,۳ ارزش Shapley در جزئيات
۱۶۳	۲۷۹	۹,۵,۳,۱ ارزش Shapley
۱۶۴	۲۸۳	۹,۵,۴ مزايا
۱۶۵	۲۸۴	۹,۵,۵ معايب
۱۶۶	۲۸۵	۹,۵,۶ نرم افzar و جايگرین
۱۶۷	۲۸۶	۹,۶ SHAP توضيحات افزومني (SHapley)
۱۶۸	۲۸۶	۹,۶,۱تعريف
۱۶۹	۲۸۸	۹,۶,۲ KernelSHAP
۱۷۰	۲۹۲	۹,۶,۳ TreeSHAP
۱۷۱	۲۹۳	۹,۶,۴ مثالها
۱۷۲	۲۹۵	۹,۶,۵ SHAP اهميت ويزگي

۱۷۳	۲۹۶ SHAP خلاصه طرح ۹,۶,۶
۱۷۴	۲۹۷ SHAP وابستگی طرح ۹,۶,۷
۱۷۵	۲۹۸ SHAP تعامل های ارزش ۹,۶,۸
۱۷۶	۲۹۹ Shapley مقادیر خوش بندی ۹,۶,۹
۱۷۷	۳۰۰ مزایا ۹,۶,۱۰
۱۷۸	۳۰۱ معایب ۹,۶,۱۱
۱۷۹	۳۰۲ نرم افزار ۹,۶,۱۲
۱۸۰	۳۰۳ فصل ۱۱ نگاهی به توب کریستالی
۱۸۱	۳۰۵ فصل ۱۳ با استناد به این کتاب
۱۸۲	۳۰۶ فصل ۱۴ ترجمه ها
۱۸۳	۳۰۸ فصل ۱۵ اسپاسگزاریها
۱۸۴		
۱۸۵		

خلاصه

- ۱۸۶
- ۱۸۷
- ۱۸۸ یادگیری ماشین پتانسیل زیادی برای بهبود محصولات، فرآیندها و تحقیقات دارد. اما رایانه‌ها معمولاً پیش‌بینی‌های خود را توضیح نمی‌دهند که مانعی برای پذیرش یادگیری ماشینی است. این کتاب در مورد ساخت مدل‌های یادگیری ماشین و تصمیمات آنها قابل تفسیر است.
- ۱۸۹
- ۱۹۰
- ۱۹۱ پس از بررسی مفاهیم تفسیرپذیری، با مدل‌های ساده و قابل تفسیر مانند درخت تصمیم، قوانین تصمیم گیری 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 Shapley و LIME است. علاوه بر این، این کتاب روش‌های خاص شبکه‌های عصبی عمیق را ارائه می‌دهد.
- ۱۹۲
- ۱۹۳
- ۱۹۴
- ۱۹۵ همه روش‌های تفسیر به طور عمیق توضیح داده شده و به صورت انتقادی مورد بحث قرار می‌گیرند. چگونه زیر کاپوت کار می‌کنند؟ قوت و ضعف آنها در چیست؟ چگونه می‌توان خروجی‌های آنها را تفسیر کرد؟ این کتاب
- ۱۹۶
- ۱۹۷ 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 تمرین‌کنندگان یادگیری ماشین، دانشمندان داده، آماردانان و هر
- ۱۹۸
- ۱۹۹ کسی که علاقه‌مند به تفسیرپذیر ساختن مدل‌های یادگیری ماشینی است، توصیه می‌شود.
- ۲۰۰ می‌توانید نسخه PDF و کتاب الکترونیکی (epub، mobi) را در leanpub.com خریداری کنید.
- ۲۰۱ می‌توانید نسخه چاپی آن را در آمازون خریداری کنید.
- ۲۰۲ درباره من: نام من کریستوف مولنار است، من یک آمارگیر و یک زبان آموز ماشین هستم. هدف من این است که
- ۲۰۳ یادگیری ماشین را قابل تفسیر کنم.
- ۲۰۴ من را در توییتر دنبال کنید! @ChristophMolnar
- ۲۰۵ جلد توسط @YvonneDoinel
- ۲۰۶ همچنین کتاب دوم من ذهنیت‌های مدل‌سازی را بررسی کنید.
- ۲۰۷ این کتاب تحت مجوز Creative Commons Attribution-NonCommercial-ShareAlike 4.0 بین
- ۲۰۸ المللی مجوز دارد.
- ۲۰۹

فصل ۱ پیشگفتار نویسنده

- ۲۱۰
- ۲۱۱
- ۲۱۲ این کتاب زمانی که من به عنوان آمارگیر در تحقیقات بالینی کار می کردم به عنوان یک پژوهش جانبی شروع
۲۱۳ شد. چهار روز در هفته کار می کردم و در «روز تعطیل» روی پژوههای جانبی کار می کردم. در نهایت، یادگیری
۲۱۴ ماشینی قابل تفسیر به یکی از پژوههای جانبی من تبدیل شد. در ابتدا قصد نوشتن کتاب نداشتم. در عوض،
۲۱۵ من به سادگی علاقه مند به یافتن اطلاعات بیشتر در مورد یادگیری ماشینی قابل تفسیر بودم و به دنبال منابع
۲۱۶ خوبی برای یادگیری بودم. با توجه به موفقیت یادگیری ماشینی و اهمیت تفسیرپذیری، من انتظار داشتم که
۲۱۷ تعداد زیادی کتاب و آموزش در مورد این موضوع وجود داشته باشد. اما من فقط مقالات تحقیقاتی مرتبط و
۲۱۸ چند پست وبلاگ پراکنده در سراسر اینترنت را پیدا کردم، اما هیچ چیز با نمای کلی خوب نه کتاب، نه آموزش،
۲۱۹ نه مقاله مروری، نه هیچ چیز. این شکاف باعث شد من شروع به نوشتن این کتاب کنم. زمانی که مطالعه خود را
۲۲۰ در مورد یادگیری ماشینی قابل تفسیر شروع کردم، کتابی را که آرزو داشتم در دسترس باشد، نوشتم. قصد من
۲۲۱ از این کتاب دو چیز بود: برای خودم یاد بگیرم و این دانش جدید را با دیگران به اشتراک بگذارم.
- ۲۲۲ من مدرک لیسانس و فوق لیسانس خود را در رشته آمار در LMU مونیخ آلمان دریافت کردم. بیشتر دانش من
۲۲۳ در مورد یادگیری ماشینی از طریق دوره‌های آنلاین، مسابقات، پژوههای جانبی و فعالیت‌های حرفه‌ای به صورت
۲۲۴ خودآموز آموزش داده شد. پیشینه آماری من مبنای بسیار خوبی برای ورود به یادگیری ماشین و به ویژه برای
۲۲۵ تفسیرپذیری بود. در آمار، تمرکز عمدۀ بر ساخت مدل‌های رگرسیون قابل تفسیر است. بعد از اینکه فوق
۲۲۶ لیسانس آمار را تمام کردم تصمیم گرفتم دکتری نرم‌ware، چون از نوشتن پایان نامه فوق لیسانس لذت نمی
۲۲۷ بردم. فقط نوشتن خیلی به من استرس وارد کرد. بنابراین من به عنوان دانشمند داده در یک استارت آپ فین
۲۲۸ تک و به عنوان آمارگیر در تحقیقات بالینی مشغول به کار شدم. بعد از این سه سال در صنعت، نوشتن این
۲۲۹ کتاب را شروع کردم و چند ماه بعد، دکتراخی خود را در زمینه یادگیری ماشینی تفسیرپذیر شروع کردم.
- ۲۳۰ این کتاب بسیاری از تکنیک‌های یادگیری ماشینی قابل تفسیر را پوشش می دهد. در فصل‌های اول، مفهوم
۲۳۱ تفسیرپذیری را معرفی می کنم و انگیزه لازم را برای تفسیرپذیری بیان می کنم. حتی چند داستان کوتاه وجود
۲۳۲ دارد! این کتاب در مورد ویژگی‌های مختلف توضیحات و آنچه که انسان فکر می کند توضیح خوبی است بحث
۲۳۳ می کند. سپس مدل‌های یادگیری ماشینی را که ذاتاً قابل تفسیر هستند، به عنوان مثال مدل‌های رگرسیون و
۲۳۴ درخت‌های تصمیم مورد بحث قرار می دهیم. تمرکز اصلی این کتاب بر روی روش‌های تفسیرپذیری مدل-
۲۳۵ آگنوستیک است. مدل-آگنوستیک به این معنی است که این روش‌ها را می توان برای هر مدل یادگیری
۲۳۶ ماشینی اعمال کرد و پس از آموزش مدل اعمال می شود. این استقلال از مدل، روش‌های مدل-آگنوستیک را

بسیار انعطاف پذیر و قادر تمند می کند . برخی از تکنیک ها چگونگی پیش بینی های فردی را توضیح می دهند، مانند توضیحات مدل-آگنوتیک محلی قابل تفسیر (LIME) و مقادیر Shapley. سایر تکنیک ها میانگین رفتار مدل را در یک مجموعه داده توصیف می کنند. در اینجا با نمودار وابستگی جزئی، اثرات محلی انباشته شده، اهمیت ویژگی جایگشت و بسیاری از روش های دیگر آشنا می شویم. یک دسته خاص، روش های مبتنی بر مثال است که نقاط داده را به عنوان توضیحات تولید می کند. توضیحات خلاف واقع، نمونه های اولیه، نمونه های تاثیرگذار و مثال های مختص روش های متخصص روش های مبتنی بر مثال هستند که در این کتاب مورد بحث قرار گرفته اند. این کتاب با برخی تأملات در مورد آینده یادگیری ماشینی قابل تفسیر به پایان می رسد. اهمیت ویژگی جایگشت و بسیاری از روش های دیگر. یک دسته خاص، روش های مبتنی بر مثال است که نقاط داده را به عنوان توضیحات تولید می کند. توضیحات خلاف واقع، نمونه های اولیه، نمونه های تاثیرگذار و مثال های مختص روش های مبتنی بر مثال هستند که در این کتاب مورد بحث قرار گرفته اند. این کتاب با برخی تأملات در مورد آینده یادگیری ماشینی قابل تفسیر به پایان می رسد.

شما مجبور نیستید کتاب را از روی جلد به بالا بخوانید، می توانید به جلو و عقب بپرید و روی تکنیک هایی تمرکز کنید که بیشتر مورد علاقه شما هستند. من فقط توصیه می کنم که از مقدمه و فصل تفسیر پذیری شروع کنید. اکثر فصول از ساختار مشابهی پیروی می کنند و بر یک روش تفسیری تمرکز می کنند. پاراگراف اول روش را خلاصه می کند. سپس سعی می کنم بدون اتكا به فرمول های ریاضی روش را به صورت شهودی توضیح دهم. سپس به تئوری روش نگاه می کنیم تا درک عمیقی از نحوه عملکرد آن بدست آوریم. در اینجا از شما در امان نخواهید بود، زیرا تئوری حاوی فرمول هایی خواهد بود. من معتقدم که یک روش جدید با استفاده از مثال ها به بهترین وجه قابل درک است. بنابراین، هر روش برای داده های واقعی اعمال می شود. برخی افراد می گویند که آمارگیران افراد بسیار منتقدی هستند. برای من، این درست است، زیرا هر فصل شامل بحث های انتقادی در مورد مزايا و معایب روش تفسیر مربوطه است. این کتاب تبلیغی برای روش ها نیست، اما باید به شما کم کند تصمیم بگیرید که آیا یک روش برای برنامه شما خوب است یا خیر. در بخش آخر هر فصل، پیاده سازی نرم افزارهای موجود مورد بحث قرار می گیرد.

یادگیری ماشینی مورد توجه بسیاری از افراد در تحقیقات و صنعت قرار گرفته است. گاهی اوقات یادگیری ماشینی بیش از حد در رسانه ها منتشر می شود، اما برنامه های کاربردی واقعی و تاثیرگذار زیادی وجود دارد. یادگیری ماشینی یک فناوری قدرتمند برای محصولات، تحقیقات و اتوپاسیون است. امروزه از یادگیری ماشینی استفاده می شود، به عنوان مثال، برای شناسایی تراکنش های مالی تقلیل، توصیه فیلم ها و طبقه بندی تصاویر. اغلب مهم است که مدل های یادگیری ماشین قابل تفسیر باشند. تفسیرپذیری به توسعه دهنده کان در رفع اشکال و بهبودها کمک می کند، اعتماد به مدل ایجاد می کند، پیش بینی های مدل را توجیه می کند و به بینش های جدید منجر می شود. افزایش نیاز به تفسیرپذیری یادگیری ماشین نتیجه طبیعی افزایش استفاده از یادگیری ماشین است. این کتاب به منبعی ارزشمند برای بسیاری از افراد تبدیل شده است. مرتبان آموزشی از این کتاب برای معرفی دانش آموزان خود با مفاهیم یادگیری ماشینی قابل تفسیر استفاده می کنند. من از چندین دانشجوی کارشناسی ارشد و دکتری ایمیل دریافت کرده ام. دانشجویانی که به من گفتند این کتاب نقطه شروع و مهم ترین مرجع پایان نامه های آنها بوده است. این کتاب به محققان کاربردی در زمینه های بوم شناسی، مالی، روانشناسی و غیره که از یادگیری ماشینی برای درک داده های خود استفاده می کنند کمک کرده است. دانشمندان داده از صنعت به من گفتند که از کتاب "یادگیری ماشین قابل تفسیر" برای کار خود استفاده می کنند و آن را به همکاران خود توصیه می کنند. خوشحالم که افراد زیادی از این کتاب بهره بردن و در تفسیر نمونه متخصص شدند. دانشجویانی که به من گفتند این کتاب نقطه شروع و مهم ترین مرجع پایان نامه های آنها بوده است. این نامه های آنها بوده است. این کتاب به محققان کاربردی در زمینه های بوم شناسی، مالی، روانشناسی و غیره که از یادگیری ماشینی برای درک داده های خود استفاده می کنند کمک کرده است. دانشمندان داده از صنعت به من گفتند که از کتاب "یادگیری ماشین قابل تفسیر" برای کار خود استفاده می کنند و آن را به همکاران خود توصیه می کنند. خوشحالم که افراد زیادی از این کتاب بهره بردن و در تفسیر نمونه متخصص شدند. دانشجویانی که به من گفتند این کتاب نقطه شروع و مهم ترین مرجع پایان نامه های آنها بوده است. این کتاب به محققان کاربردی در زمینه های بوم شناسی، مالی، روانشناسی و غیره که از یادگیری ماشینی برای درک داده های خود استفاده می کنند کمک کرده است. دانشمندان داده از صنعت به من گفتند که از کتاب "یادگیری ماشین قابل تفسیر" برای کار خود استفاده می کنند و آن را به همکاران خود توصیه می کنند. خوشحالم که افراد زیادی از این کتاب بهره بردن و در تفسیر نمونه متخصص شدند.

من این کتاب را به تمرین کنندگانی توصیه می کنم که می خواهند مروری بر تکنیک هایی برای تفسیرپذیرتر کردن مدل های یادگیری ماشینی خود داشته باشند. همچنین برای دانشجویان و محققین (و هر کس دیگری) که به موضوع علاقه مند است، مفید خواهد بود. برای استفاده حداکثری از این کتاب، باید درک اولیه ای از یادگیری ماشین داشته باشید. همچنین باید درک درستی از ریاضیات پایه دانشگاهی داشته باشید تا بتوانید

۲۹۱ تئوری و فرمول های این کتاب را دنبال کنید. با این حال، درک توصیف شهودی روش در ابتدای هر فصل بدون
۲۹۲ ریاضیات نیز باید امکان پذیر باشد.

۲۹۳ امیدوارم از کتاب لذت ببری!

۲۹۴

فصل ۲ مقدمه

۲۹۵

۲۹۶

۲۹۷ این کتاب به شما توضیح می‌دهد که چگونه می‌توانید مدل‌های یادگیری ماشین (با نظارت) را قابل تفسیر کنید.
۲۹۸ فصل‌ها حاوی برخی فرمول‌های ریاضی هستند، اما شما باید بتوانید ایده‌های پشت روش‌ها را حتی بدون
۲۹۹ فرمول‌ها درک کنید. این کتاب برای افرادی نیست که سعی می‌کنند یادگیری ماشینی را از ابتدای بگیرند.
۳۰۰ اگر در یادگیری ماشینی تازه کار هستید، کتاب‌ها و منابع دیگری برای یادگیری اصول اولیه وجود دارد. من
۳۰۱ کتاب «عناصر یادگیری آماری» اثر هستی، تیبی‌شیرانی و فریدمن (۲۰۰۹) ۱ و دوره آنلاین «یادگیری ماشینی»
۳۰۲ اندر و نگ در پلتفرم یادگیری آنلاین coursera.com را برای شروع با یادگیری ماشین توصیه می‌کنم. هم
۳۰۳ کتاب و هم دوره رایگان در دسترس هستند!

۳۰۴ روش‌های جدید برای تفسیر مدل‌های یادگیری ماشین با سرعتی سراسام‌آور منتشر شده‌اند. همگام شدن با هر
۳۰۵ آنچه منتشر می‌شود دیوانگی و به سادگی غیرممکن است. به همین دلیل است که در این کتاب جدیدترین و
۳۰۶ فانتزی‌ترین روش‌ها را پیدا نمی‌کنید، بلکه روش‌های تثبیت شده و مفاهیم اساسی تفسیرپذیری یادگیری
۳۰۷ ماشین را پیدا خواهید کرد. این اصول شما را برای ساختن مدل‌های یادگیری ماشینی قابل تفسیر آماده می‌
۳۰۸ کند. درونی کردن مفاهیم اساسی همچنین به شما این امکان را می‌دهد که هر مقاله جدیدی در مورد
۳۰۹ تفسیرپذیری منتشر شده در arxiv.org در ۵ دقیقه گذشته از زمان شروع خواندن این کتاب را بهتر درک و
۳۱۰ ارزیابی کنید (ممکن است در میزان انتشار اغراق کنم).

۳۱۱ این کتاب با چند داستان کوتاه (دیستوپیایی) شروع می‌شود که برای درک کتاب مورد نیاز نیست، اما امیدوارم
۳۱۲ شما را سرگرم کند و به فکر فرو ببرد. سپس این کتاب مفاهیم تفسیرپذیری یادگیری ماشین را بررسی می‌کند
۳۱۳ . ما در مورد اینکه تفسیرپذیری مهم است و انواع مختلف توضیحاتی که وجود دارد بحث خواهیم کرد.
۳۱۴ اصطلاحات استفاده شده در سراسر کتاب را می‌توان در فصل اصطلاحات جستجو کرد. بیشتر مدل‌ها و روش
۳۱۵ های توضیح داده شده با استفاده از نمونه‌های داده واقعی ارائه شده‌اند که در فصل داده‌ها توضیح داده شده
۳۱۶ است . یکی از راه‌های قابل تفسیر کردن یادگیری ماشین، استفاده از مدل‌های قابل تفسیر، مانند مدل‌های
۳۱۷ خطی یا درخت‌های تصمیم‌گیری است. گزینه دیگر استفاده از ابزارهای تفسیر مدل-آگنوستیک که می‌توانند
۳۱۸ برای هر مدل یادگیری ماشینی نظارت شده اعمال شوند. روش‌های مدل-آگنوستیک را می‌توان به روش‌های
۳۱۹ کلی که رفتار میانگین مدل را توصیف می‌کنند و روش‌های محلی که پیش‌بینی‌های فردی را توضیح می‌دهند،
۳۲۰ تقسیم کرد . فصل روش‌های مدل-آگنوستیک به روش‌هایی مانند نمودارهای وایستگی جزئی و اهمیت ویژگی
۳۲۱ می‌پردازد . روش‌های مدل-آگنوستیک با تغییر ورودی مدل یادگیری ماشین و اندازه‌گیری تغییرات در خروجی

۳۲۲ پیش‌بینی کار می‌کنند. این کتاب با یک چشم انداز خوش بینانه در مورد آینده یادگیری ماشینی قابل تفسیر به
۳۲۳ پایان می‌رسد.

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

۳۲۵ امیدوارم از خواندن لذت ببرید!

۳۲۶ **۲،۱ زمان داستان**

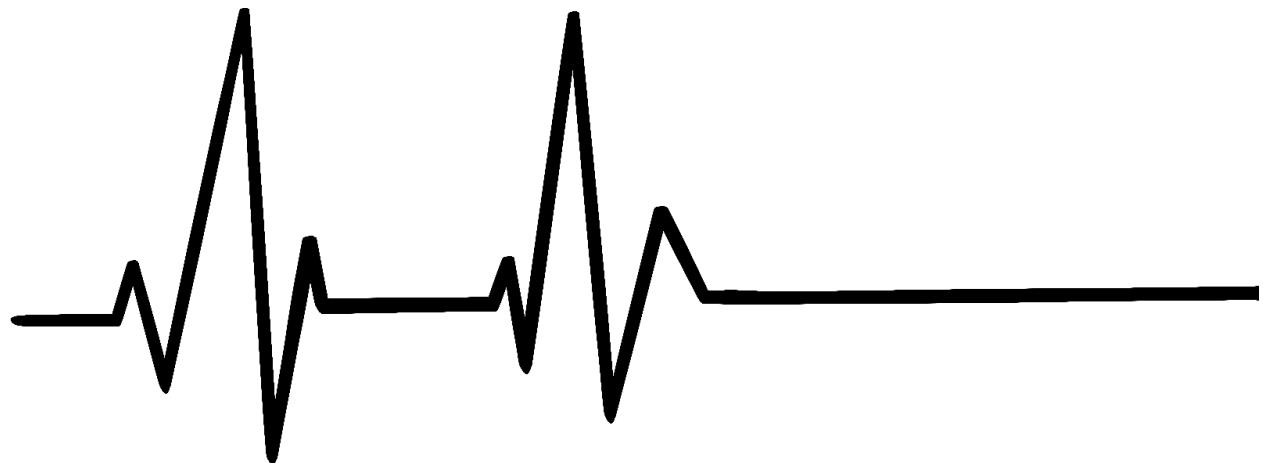
۳۲۷ با چند داستان کوتاه شروع می‌کنیم. هر داستان یک فراخوان اغراق‌آمیز برای یادگیری ماشینی قابل تفسیر است. اگر عجله دارید، می‌توانید از استوری‌ها صرف نظر کنید. اگر می‌خواهید سرگرم شوید و انگیزه نداشته باشید، ادامه مطلب را بخوانید!

۳۳۰ این قالب از داستان‌های فناوری جک کلارک در خبرنامه هوش مصنوعی وارداتی او الهم گرفته شده است. اگر

۳۳۱ این نوع داستان‌ها را دوست دارید یا اگر به هوش مصنوعی علاقه مند هستید، توصیه می‌کنم ثبت نام کنید.

۳۳۲ رعد و برق هرگز دو بار نمی‌زند

۳۳۳ **۲۰۳۰: یک آزمایشگاه پزشکی در سوئیس**



۳۳۴
۳۳۵ قطعاً این بدترین راه برای مردن نیست! تام خلاصه کرد و سعی کرد چیز مثبتی در تراژدی پیدا کند. او پمپ را
۳۳۶ از قطب داخل وریدی خارج کرد.

۳۳۷ لنا افزود: «او فقط به دلایل اشتباه مرد.

و مطمئناً با پمپ مرفین اشتباه! فقط کار بیشتری برای ما ایجاد کنید!» تام در حالی که پیچ پشتی پمپ را باز
۳۴۸ می کرد، شکایت کرد. بعد از برداشتن تمام پیچ ها، صفحه را بلند کرد و کنار گذاشت. او یک کابل را به درگاه
۳۴۹ تشخیص وصل کرد.

۳۴۱ "شما فقط از داشتن شغل شکایت نکردید، نه؟" لنا لبخند تمسخر آمیزی به او زد.

۳۴۲ "البته که نه. هرگز!" با لحنی کنایه آمیز فریاد زد.

۳۴۳ کامپیوتر پمپ را بوت کرد.

۳۴۴ لنا سر دیگر کابل را به تبلتش وصل کرد. او اعلام کرد: "بسیار خوب، تشخیص در حال انجام است." "من واقعاً
۳۴۵ کنجدکاو هستم که ببینم چه اشتباهی رخ داده است".

۳۴۶ مطمئناً جان دو ما را به نیروانا شلیک کرد. غلظت بالای این مواد مورفین. مرد. منظورم این است که ... این اولین
۳۴۷ است، درست است؟ به طور معمول یک پمپ شکسته مقدار بسیار کمی از مواد شیرین را تولید می کند یا اصلاً
۳۴۸ هیچ چیز را نمی دهد. اما هرگز، می دانید، آن شات دیوانه را دوست نداشته باشید،" تام توضیح داد.

۳۴۹ "میدانم. شما مجبور نیستید من را متقاعد کنید ... هی، این را نگاه کنید. لنا تبلتش را بالا گرفت. «این قله را
۳۵۰ اینجا می بینی؟ این قدرت ترکیب مسکن ها است. نگاه کن این خط سطح مرجع را نشان می دهد. بیچاره
۳۵۱ مخلوطی از مسکن در سیستم خونش داشت که می توانست ۱۷ برابر او را بکشد. توسط پمپ ما در اینجا تزریق
۳۵۲ می شود. و در اینجا...» او با انگشت خود تند تند گفت: «در اینجا می توانید لحظه مرگ بیمار را ببینید.»

۳۵۳ "پس، آیا می دانید چه اتفاقی افتاده است، رئیس؟" تام از سرپرستش پرسید.

۳۵۴ «هوم... حسگرها به نظر خوب هستند. ضربان قلب، سطح اکسیژن، گلوکز، ... داده ها همانطور که انتظار می رفت
۳۵۵ جمع آوری شد. برخی از مقادیر از دست رفته در داده های اکسیژن خون، اما این غیرعادی نیست. اینجا را نگاه
۳۵۶ کن. سنسورها همچنین کاهش ضربان قلب بیمار و سطوح بسیار پایین کورتیزول ناشی از مشتقان مورفین و
۳۵۷ سایر عوامل مسدود کننده درد را تشخیص داده اند. او همچنان به مرور گزارش تشخیصی ادامه داد.

۳۵۸ تام مجذوب صفحه نمایش خیره شد. این اولین تحقیق او در مورد خرابی واقعی دستگاه بود.

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

٣٦٢ تام در حالی که چشمانش همچنان به صفحه نمایش خیره شده بود سر تکان داد.

٣٦٣ لنا ادامه داد: «عجیب است. این هشدار همچنین باید باعث خاموش شدن پمپ شود. اما بدیهی است که موفق به
٣٦٤ انجام این کار نشده است. این باید یک اشکال باشد. چیزی که تیم با کیفیت از دست داد. یه چیز واقعاً بد شاید
٣٦٥ به مشکل پروتکل مربوط باشد».

٣٦٦ بنابراین، سیستم اورژانسی پمپ به نوعی خراب شد، اما چرا پمپ پر موز شد و این همه مسکن به جان دو تزریق
٣٦٧ کرد؟ تام تعجب کرد.

٣٦٨ "سؤال خوبی بود. حق با شماست. جدا از خرابی اضطراری پروتکل، پمپ اصلاً نباید آن مقدار دارو را تجویز می
٣٦٩ کرد. لنا توضیح داد که با توجه به سطح پایین کورتیزول و سایر علائم هشدار، الگوریتم باید خیلی زودتر به
٣٧٠ خودی خود متوقف می شد.

٣٧١ "شاید یک بدشานسی، مانند یک در میلیون چیز، مانند برخورد با رعد و برق؟" تام از او پرسید.

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

٣٧٨ تام که در فکر فرو رفته بود پاسخ داد: "می‌بینم...". آیا به هر حال بیمار به زودی نخواهد مرد؟ به خاطر سرطان
٣٧٩ یا چیزی؟»

٣٨٠ لنا در حالی که گزارش تحلیل را می‌خواند سر تکان داد.

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

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

٣٨٦ او مدام به چیزی بیرون نگاه می‌کرد.

۳۸۷ هر دو برای چند لحظه سکوت کردند.

۳۸۸ لنا دوباره سرش را پایین انداخت و به تحلیل ادامه داد. «ته، تام. این یک اشکال است... فقط یک حشره لعنتی.»

۳۸۹ پاییز اعتماد کن

۳۹۰ ۲۰۵۰: یک ایستگاه مترو در سنگاپور



۳۹۱ با عجله به سمت ایستگاه متروی بیشان رفت. با افکارش از قبل سر کار بود. آزمایشات برای معماری عصبی
۳۹۲ جدید باید تا الان کامل شده باشد. او طراحی مجدد «سیستم پیش‌بینی وابستگی مالیاتی برای اشخاص حقیقی»
۳۹۳ را رهبری کرد که پیش‌بینی می‌کند آیا شخص پول را از اداره مالیات پنهان می‌کند یا خیر. تیم او یک قطعه
۳۹۴ مهندسی ظریف را ارائه کرده است. در صورت موفقیت، این سیستم نه تنها به اداره مالیات خدمت می‌کند، بلکه
۳۹۵ به سیستم‌های دیگر مانند سیستم هشدار ضد تروریسم و ثبت تجاری نیز وارد می‌شود. یک روز، دولت حتی
۳۹۶ می‌تواند پیش‌بینی‌ها را در امتیاز اعتماد مدنی ادغام کند. امتیاز اعتماد مدنی تخمین می‌زند که یک فرد
۳۹۷ چقدر قابل اعتماد است. این تخمین بر هر بخش از زندگی روزمره شما تأثیر می‌گذارد، مانند دریافت وام یا
۳۹۸ مدت زمانی که باید برای پاسپورت جدید صبر کنید. وقتی از پله برقی پایین می‌آمد،
۳۹۹

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

۴۰۲ خیلی دیر.

۴۰۳ دماغ ابتدا وارد دروازه ورودی مترو شد و ابتدا با باسنیش به زمین افتاد. قرار بود در باز شود، اما باز نشد. مات و
۴۰۴ مبهوت از جایش بلند شد و به صفحه نمایش کنار دروازه نگاه کرد. یک شکلک دوستانه روی صفحه پیشنهاد

۴۰۵ کرد: «لطفا یک بار دیگر امتحان کنید». شخصی از آنجا گذشت و بی توجه به او دستش را روی خواننده پاک
۴۰۶ کرد. در باز شد و او رفت. در دوباره بسته شد. بینی اش را پاک کرد. درد داشت ولی حداقل خونریزی نداشت.
۴۰۷ سعی کرد در را باز کند، اما دوباره رد شد. عجیب بود. شاید حساب حمل و نقل عمومی او توکن کافی نداشته
۴۰۸ باشد. او برای بررسی موجودی حساب به ساعت هوشمند خود نگاه کرد.

۴۰۹ «ورود رد شد. لطفا با دفتر مشاوره شهروندان خود تماس بگیرید!» ساعتش به او خبر داد.

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

۴۱۳ تنها یک توضیح ممکن وجود داشت: امتیاز اعتماد مدنی او کاهش یافته بود. بطور قابل ملاحظه ای یک افت
۴۱۴ کوچک به معنای ناراحتی های جزئی بود، مانند عدم دریافت پرواژه های درجه یک یا نیاز به صبر کردن برای اسناد
۴۱۵ رسمی کمی بیشتر. نمره اعتماد پایین نادر بود و به این معنی بود که شما به عنوان یک تهدید برای جامعه طبقه
۴۱۶ بندی می شوید. یکی از اقدامات در برخورد با این افراد دور نگه داشتن آنها از مکان های عمومی مانند مترو بود.
۴۱۷ دولت تراکنش های مالی افراد دارای امتیاز اعتماد مدنی پایین را محدود کرد. آنها همچنین شروع به نظارت
۴۱۸ فعالانه بر رفتار شما در رسانه های اجتماعی کردند و حتی تا آنجا پیش رفتند که محتوای خاصی مانند بازی
۴۱۹ های خشونت آمیز را محدود کردند. افزایش امتیاز اعتماد مدنی هر چه کمتر بود به طور تصاعدی دشوارتر می
۴۲۰ شد. افراد با نمره بسیار پایین معمولا هرگز بهبود نمی یابند.

۴۲۱ او نمی توانست به هیچ دلیلی فکر کند که چرا نمره او باید پایین می آمد. امتیاز بر اساس یادگیری ماشین بود.
۴۲۲ سیستم امتیاز اعتماد مدنی مانند موتور روغن کاری شده ای عمل می کرد که جامعه را اداره می کرد. عملکرد
۴۲۳ سیستم امتیاز اعتماد همیشه به دقت نظارت می شد. یادگیری ماشینی از ابتدای قرن بسیار بهتر شده بود. آنقدر
۴۲۴ کارآمد شده بود که تصمیمات اتخاذ شده توسط سیستم امتیاز اعتماد دیگر قابل بحث نبود. یک نظام خطانایپذیر

۴۲۵ او با نالمیدی خندهید. نظام معصوم. اگر فقط. این سیستم به ندرت شکست خورده است. اما شکست خورد. او باید
۴۲۶ یکی از آن موارد خاص باشد. خطای سیستم؛ از این به بعد یک طرد شده هیچ کس جرات نداشت سیستم را زیر
۴۲۷ سوال ببرد. آنقدر در دولت، در خود جامعه ادغام شده بود که نمی توان آن را زیر سوال برد. در محدود کشورهای
۴۲۸ دموکراتیک باقیمانده، تشکیل جنبش های ضد دموکراتیک ممنوع بود، نه به این دلیل که ذاتاً بدخواهانه بودند،
۴۲۹ بلکه به این دلیل که سیستم فعلی را بی ثبات می کردند. همین منطق در مورد الگوکراسی های رایج تر هم اعمال
۴۳۰ می شود. نقد در الگوریتم ها به دلیل خطر برای وضع موجود ممنوع بود.

اعتراض الگوریتمی تار و پود نظم اجتماعی بود. برای منافع عمومی، امتیازات نادر اعتماد نادرست به طور ضمنی
۴۳۱ پذیرفته شد. صدها سیستم پیش‌بینی و پایگاه داده دیگر به امتیاز وارد شده‌اند و نمی‌دانند چه چیزی باعث افت
۴۳۲ امتیاز او شده است. او احساس می‌کرد که یک سوراخ تاریک بزرگ در داخل و زیر او باز می‌شود. با وحشت به
۴۳۳ فضای خالی نگاه کرد.

۴۳۵ سیستم وابستگی مالیاتی او در نهایت در سیستم امتیاز اعتماد مدنی ادغام شد، اما او هرگز با آن آشنا نشد.

۴۳۶ گیره‌های فرمی

۴۳۷ سال ۶۱۲ (AMS پس از استقرار مریخ): موزه‌ای در مریخ



۴۳۸ 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.

۴۳۹ "Xola، چه چیزی یاد گرفتی؟" معلم از او پرسید. اینکه مردم باستان تمام منابع سیاره زمین را مصرف کردند و
۴۴۰ سپس مردند؟ او با دقت پرسید. «نه. آنها آب و هوا را گرم کردند و مردم نبودند، کامپیوتر و ماشین بودند. دختر
۴۴۱ دیگری به نام لین اضافه کرد و این سیاره زمین است، نه سیاره زمین. زولا سری به تایید تکان داد. معلم با
۴۴۲ احساس غرور لبخندی زد و سری تکان داد. «هر دو حق با شمامست. میدونی چرا اینطوری شد؟» "چون مردم
۴۴۳ کوته فکر و حریص بودند؟ Xola پرسید. "مردم نمی‌توانند ماشین‌های خود را متوقف کنند!" لین تار شد.

۴۴۴ معلم تصمیم گرفت: «باز هم، هر دوی شما درست می‌گویید، اما این بسیار پیچیده تر از این است. بیشتر مردم
۴۴۵ در آن زمان از آنچه در حال رخ دادن بود آگاه نبودند. برخی تغییرات شدید را دیدند، اما نتوانستند آنها را
۴۴۶

۴۴۹ معکوس کنند. مشهورترین قطعه از این دوره شعری از نویسنده ناشناس است. به بهترین وجه آنچه را که در آن
۴۵۰ زمان اتفاق افتاد به تصویر می کشد. با دقت گوش کن!»

۴۵۱ معلم شعر را شروع کرد. تعداد زیادی از پهپادهای کوچک در مقابل کودکان قرار گرفتند و شروع به پخش
۴۵۲ مستقیم ویدیو در چشمان آنها کردند. فردی را با کت و شلوار نشان می داد که در جنگلی ایستاده بود و فقط
۴۵۳ کنده درخت باقی مانده بود. شروع کرد به صحبت کردن:

۴۵۴ ماشین ها محاسبه می کنند. ماشین ها پیش بینی می کنند
۴۵۵ همانطور که بخشی از آن هستیم راهپیمایی می کنیم.

۴۵۶ ما به عنوان آموزش دیده به دنبال یک بهینه هستیم.

۴۵۷ بهینه یک بعدی، محلی و بدون محدودیت است.

۴۵۸ سیلیکون و گوشت، در تعقیب نمایی.

۴۵۹ رشد ذهنیت ماست.

۴۶۰ وقتی همه جوایز جمع آوری شد،

۴۶۱ و عوارض جانبی نادیده گرفته شد.

۴۶۲ وقتی تمام سکه ها استخراج می شوند،

۴۶۳ و طبیعت عقب افتاده است.

۴۶۴ دچار مشکل خواهیم شد،

۴۶۵ به هر حال، رشد تصاعدی یک حباب است.

۴۶۶ تراژدی عوام آشکار می شود،

۴۶۷ منفجر شدن،

۴۶۸ جلوی چشمان ما

۴۶۹ محاسبات سرد و طمع سرد،

۴۷۰ زمین را از گرما پر کنید.

۴۷۱ همه چیز در حال مرگ است،

۴۷۲ و ما رعایت می کنیم.

۴۷۳ ما مانند اسبهایی با کرکره در مسابقه خلقت خود مسابقه می دهیم،

۴۷۴ به سوی فیلتر بزرگ تمدن.

۴۷۵ و بنابراین ما بی امان راهپیمایی می کنیم.

۴۷۶ همانطور که ما بخشی از ماشین هستیم.

۴۷۷ آنروپی را در بر می گیرد.

۴۷۸ معلم برای شکستن سکوت اتفاق گفت: یک خاطره تاریک. "در کتابخانه شما آپلود خواهد شد. تکلیف شما این

۴۷۹ است که آن را تا هفته آینده حفظ کنید." زولا آهی کشید. او موفق شد یکی از پهپادهای کوچک را بگیرد. پهپاد

۴۸۰ از CPU و موتورها گرم بود. Xola دوست داشت که چگونه دستان او را گرم می کرد.

۲.۲ یادگیری ماشینی چیست؟

۴۸۱ یادگیری ماشینی مجموعه ای از روش هایی است که رایانه ها برای انجام و بهبود پیش بینی ها یا رفتارها بر

۴۸۲ اساس داده ها استفاده می کنند.

۴۸۳ برای مثال، برای پیش بینی ارزش یک خانه، کامپیوتر الگوهایی را از فروش خانه های گذشته یاد می گیرد. این

۴۸۴ کتاب بر یادگیری ماشین ناظارت شده تمرکز دارد، که همه مشکلات پیش بینی را پوشش می دهد که در آن

۴۸۵ مجموعه داده ای داریم که از قبل نتیجه مورد علاقه را می دانیم (مثلاً قیمت های قبلی خانه) و می خواهیم یاد

۴۸۶ بگیریم که نتیجه را برای داده های جدید پیش بینی کنیم. برای مثال، کارهای خوش بندی (= یادگیری بدون

۴۸۷ ناظارت) که در آن ما یک نتیجه خاص مورد علاقه نداریم، اما می خواهیم خوش هایی از نقاط داده را پیدا کنیم، از

۴۸۸ یادگیری تحت ناظارت مستثنی شده اند. همچنین مواردی مانند یادگیری تقویتی، که در آن یک عامل یاد

۴۸۹ می گیرد با انجام دادن یک محیط (مثلاً رایانه ای که تتریس بازی می کند) پاداش خاصی را بهینه کند، مستثنی

۴۹۰ می شوند. هدف یادگیری تحت ناظارت، یادگیری یک مدل پیش بینی است که ویژگی های داده ها (به عنوان مثال

۴۹۱ اندازه خانه، مکان، نوع طبقه، ...) را به یک خروجی نگاشت می کند (مثلاً قیمت خانه). اگر خروجی مقوله ای

۴۹۲ باشد، کار را طبقه بندی و اگر عددی باشد، رگرسیون نامیده می شود. الگوریتم یادگیری ماشین یک مدل را با

۴۹۳ تخمین پارامترها (مانند وزن ها) یا ساختارهای یادگیری (مانند درختان) یاد می گیرد. الگوریتم توسط یک

۴۹۴ امتیاز یاتابع ضرر هدایت می شود که به حداقل می رسد. در مثال ارزش خانه، ماشین تفاوت بین قیمت

۴۹۶ تخمینی خانه و قیمت پیش بینی شده را به حداقل می رساند. سپس می توان از یک مدل یادگیری ماشین
۴۹۷ کاملاً آموزش دیده برای پیش بینی موارد جدید استفاده کرد. در مثال ارزش خانه، ماشین تفاوت بین قیمت
۴۹۸ تخمینی خانه و قیمت پیش بینی شده را به حداقل می رساند. سپس می توان از یک مدل یادگیری ماشین
۴۹۹ کاملاً آموزش دیده برای پیش بینی موارد جدید استفاده کرد. در مثال ارزش خانه، ماشین تفاوت بین قیمت
۵۰۰ تخمینی خانه و قیمت پیش بینی شده را به حداقل می رساند. سپس می توان از یک مدل یادگیری ماشین
۵۰۱ کاملاً آموزش دیده برای پیش بینی موارد جدید استفاده کرد.

۵۰۲ تخمین قیمت خانه، توصیه های محصول، تشخیص تابلوهای خیابانی، پیش بینی پیش فرض اعتبار و تشخیص
۵۰۳ تقلب: همه این مثال ها وجه اشتراک دارند که می توان آن ها را با یادگیری ماشین حل کرد. وظایف متفاوت است،
۵۰۴ اما رویکرد یکسان است:

۵۰۵ مرحله ۱: جمع آوری داده ها هرچی بیشتر بهتر. داده ها باید حاوی نتیجه های باشد که می خواهید پیش بینی کنید
۵۰۶ و اطلاعات اضافی که از آن می توان پیش بینی کرد. برای آشکارساز علائم خیابان ("آیا تابلوی خیابان در تصویر
۵۰۷ وجود دارد؟")، تصاویر خیابان را جمع آوری می کنید و برچسب می زنید که آیا تابلوی خیابان قابل مشاهده
۵۰۸ است یا خیر. برای یک پیش بینی کننده پیش فرض اعتبار، به داده های گذشته در مورد وام های واقعی، اطلاعاتی
۵۰۹ در مورد اینکه آیا مشتریان وام های خود را نکول کرده اند، و داده هایی که به شما در انجام پیش بینی ها کمک
۵۱۰ می کنند، مانند درآمد، پیش فرض اعتبارات گذشته و غیره نیاز دارید. برای یک برنامه تخمین زن خودکار ارزش
۵۱۱ خانه، می توانید داده ها را از فروش های قبلی خانه و اطلاعات مربوط به املاک مانند اندازه، مکان و غیره جمع
۵۱۲ آوری کنید.

۵۱۳ مرحله ۲: این اطلاعات را در یک الگوریتم یادگیری ماشینی وارد کنید که یک مدل آشکارساز علامت، یک مدل
۵۱۴ رتبه بندی اعتبار یا یک تخمین گر ارزش خانه ایجاد می کند.

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

۵۱۷ ماشین ها در بسیاری از کارها، مانند بازی شطرنج (یا اخیراً ۵۰) یا پیش بینی آب و هوا از انسان ها پیشی
۵۱۸ می گیرند. حتی اگر ماشین به خوبی یک انسان باشد یا در یک کار کمی بدتر باشد، مزایای زیادی از نظر سرعت،
۵۱۹ تکرار پذیری و مقیاس پذیری وجود دارد. یک مدل یادگیری ماشینی که زمانی پیاده سازی شده باشد، می تواند
۵۲۰ یک کار را بسیار سریع تر از انسان ها انجام دهد، نتایج قبل اعتمادی را ارائه می دهد و می تواند بی نهایت کمی
۵۲۱ شود. تکرار یک مدل یادگیری ماشینی در ماشین دیگر سریع و ارزان است. آموزش یک انسان برای انجام یک

کار می تواند چندین دهه طول بکشد (مخصوصاً در جوانی) و بسیار پرهزینه است. یک عیب عمدۀ استفاده از
یادگیری ماشینی این است که بینش در مورد داده ها و وظیفه ای که ماشین حل می کند در مدل های پیچیده
ای پنهان است. برای توصیف یک شبکه عصبی عمیق به میلیون ها عدد نیاز دارید، و هیچ راهی برای درک
کامل مدل وجود ندارد. مدل های دیگر، مانند جنگل تصادفی، از صدها درخت تصمیم تشکیل شده‌اند که به
پیش‌بینی‌ها رأی می‌دهند. برای درک چگونگی تصمیم گیری، باید به آرا و ساختار هر یک از صدها درخت نگاه
کنید. این صرف نظر از اینکه چقدر باهوش هستید یا حافظه کاری شما چقدر خوب است کار نمی کند. بهترین
مدل‌ها اغلب ترکیبی از چندین مدل (هم‌چنین گروه‌ها) هستند که قابل تفسیر نیستند، حتی اگر هر مدل منفرد
قابل تفسیر باشد. اگر فقط بر روی عملکرد تمرکز کنید، به طور خودکار مدل‌های غیرشفاف تری خواهد داشت.
شما باید به آرا و ساختار هر یک از صدها درخت نگاه کنید. این صرف نظر از اینکه چقدر باهوش هستید یا
حافظه کاری شما چقدر خوب است کار نمی کند. بهترین مدل‌ها اغلب ترکیبی از چندین مدل (هم‌چنین
گروه‌ها) هستند که قابل تفسیر نیستند، حتی اگر هر مدل منفرد قابل تفسیر باشد. اگر فقط بر روی عملکرد
تمرکز کنید، به طور خودکار مدل‌های غیرشفاف تری خواهد داشت. شما باید به آرا و ساختار هر یک از صدها
درخت نگاه کنید. این صرف نظر از اینکه چقدر باهوش هستید یا حافظه کاری شما چقدر خوب است کار نمی
کند. بهترین مدل‌ها اغلب ترکیبی از چندین مدل (هم‌چنین گروه‌ها) هستند که قابل تفسیر نیستند، حتی اگر
هر مدل منفرد قابل تفسیر باشد. اگر فقط بر روی عملکرد تمرکز کنید، به طور خودکار مدل‌های غیرشفاف تری
خواهد داشت. مدل های برنده در مسابقات یادگیری ماشین اغلب مجموعه ای از مدل ها یا مدل های بسیار
پیچیده مانند درختان تقویت شده یا شبکه های عصبی عمیق هستند.

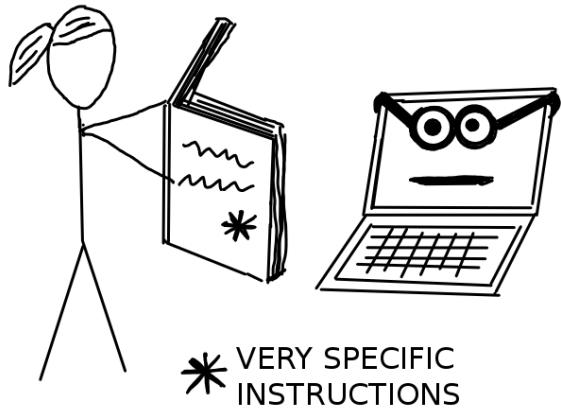
۲،۳ اصطلاحات

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

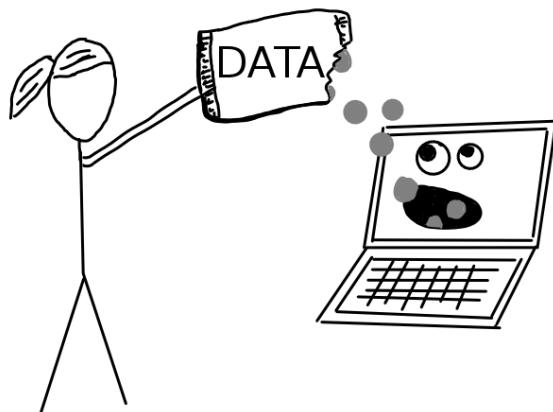
الگوریتم مجموعه ای از قوانین است که یک ماشین برای رسیدن به یک هدف خاص از آنها پیروی می کند .
یک الگوریتم را می توان به عنوان دستور العملی در نظر گرفت که ورودی ها، خروجی ها و تمام مراحل مورد
نیاز برای رسیدن از ورودی ها به خروجی را تعریف می کند. دستور العمل های آشپزی الگوریتم هایی هستند
که مواد اولیه ورودی، غذای پخته شده خروجی و مراحل آماده سازی و پخت دستورالعمل های الگوریتم هستند.
یادگیری ماشینی مجموعه‌ای از روش‌هایی است که به رایانه‌ها اجازه می‌دهد از داده‌ها برای انجام و بهبود
پیش‌بینی‌ها یاد بگیرند (مثالاً سلطان، فروش هفتگی، پیش‌فرض اعتبار). یادگیری ماشینی یک تغییر پارادایم از

۵۴۸ «برنامه‌نویسی عادی» است که در آن تمام دستورالعمل‌ها باید به صراحت به رایانه به «برنامه‌نویسی غیرمستقیم»
۵۴۹ داده شود که از طریق ارائه داده‌ها انجام می‌شود.

Without Machine Learning

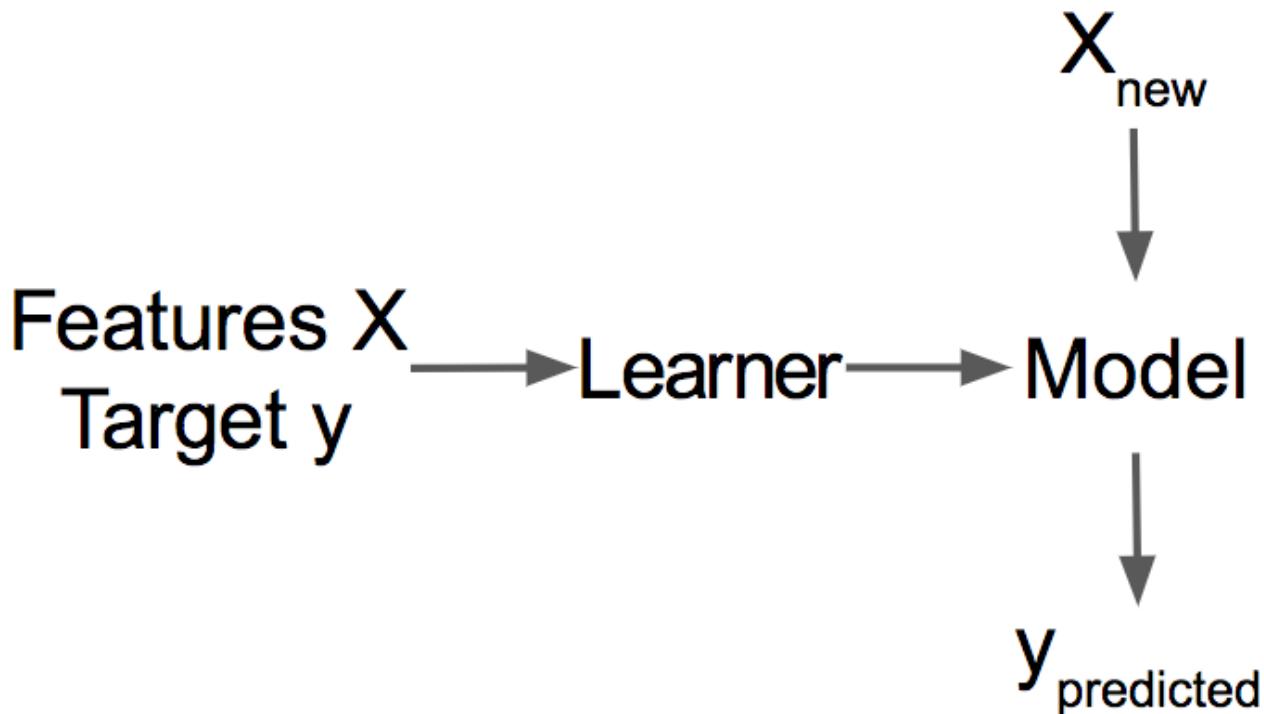


With Machine Learning



۵۵۰
۵۵۱ یادگیرنده یا الگوریتم یادگیری ماشینی برنامه‌ای است که برای یادگیری مدل یادگیری ماشین از داده‌ها
۵۵۲ استفاده می‌شود. نام دیگر "القاء کننده" است (به عنوان مثال "القا کننده درخت").

۵۵۳ مدل یادگیری ماشینی برنامه‌ای است که ورودی‌ها را به پیش‌بینی‌ها ترسیم می‌کند. این می‌تواند مجموعه‌ای از
۵۵۴ وزن‌ها برای یک مدل خطی یا برای یک شبکه عصبی باشد. نام‌های دیگر کلمه نسبتاً نامشخص "مدل"
۵۵۵ "پیش‌بینی" یا - بسته به کار - "طبقه‌بند" یا "مدل رگرسیون" است. در فرمول‌ها، مدل یادگیری ماشین
۵۵۶ آموزش دیده \hat{f} یا $f(x)$ نامیده می‌شود.



۵۵۷

۵۵۸

۵۵۹

۵۶۰

۵۶۱

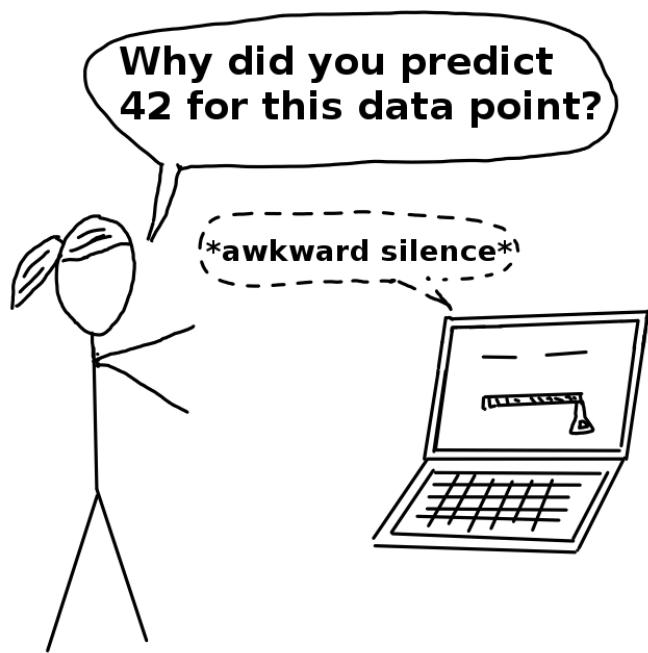
۵۶۲

۵۶۳

۵۶۴

شکل ۲،۱: یک یادگیرنده مدلی را از داده های آموزشی برچسب گذاری شده می آموزد. این مدل برای پیش بینی استفاده می شود.

مدل جعبه سیاه سیستمی است که مکانیسم های داخلی خود را آشکار نمی کند. در یادگیری ماشین، "جعبه سیاه" مدل هایی را توصیف می کند که با نگاه کردن به پارامترهای آنها قابل درک نیستند (مثلاً یک شبکه عصبی). نقطه مقابل جعبه سیاه گاهی اوقات به عنوان جعبه سفید شناخته می شود و در این کتاب به عنوان مدل قابل تفسیر از آن یاد می شود . روش های مدل آگنوستیک برای تفسیرپذیری، مدل های یادگیری ماشین را به عنوان جعبه های سیاه در نظر می گیرند، حتی اگر چنین نباشند.



۵۶۵

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

مجموعه داده جدولی با داده‌هایی است که ماشین از آن یاد می‌گیرد. مجموعه داده شامل ویژگی‌ها و هدف برای پیش‌بینی است. هنگامی که برای القای یک مدل استفاده می‌شود، مجموعه داده داده آموزشی نامیده می‌شود.

یک ردیف در مجموعه داده است. نام‌های دیگر «مثال» عبارتند از: (داده) نقطه، مثال، مشاهده. یک نمونه از مقادیر ویژگی (i) تشکیل شده است و در صورت شناخته شدن، نتیجه هدف y_i .

ویژگی‌ها ورودی‌هایی هستند که برای پیش‌بینی یا طبقه‌بندی استفاده می‌شوند. یک ویژگی یک ستون در مجموعه داده 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. اما اگر درک ویژگی‌های ورودی سخت باشد، درک اینکه مدل چه کاری انجام می‌دهد حتی سخت‌تر است. ماتریس با تمام ویژگی‌ها X نامیده می‌شود و یک نمونه بردار y_i یک ویژگی واحد برای همه موارد است و مقدار برای ویژگی است

هدف اطلاعاتی است که ماشین یاد می‌گیرد تا پیش‌بینی کند. در فرمول‌های ریاضی معمولاً هدف نامیده می‌شود

۵۸۱ وظیفه یادگیری ماشین ترکیبی از مجموعه داده با ویژگی ها و یک هدف است. بسته به نوع هدف، کار می تواند
۵۸۲ به عنوان مثال طبقه بندی، رگرسیون، تجزیه و تحلیل بقا، خوشه بندی، یا تشخیص پرت باشد.

۵۸۳ پیش‌بینی چیزی است که مدل یادگیری ماشین «حدس می‌زند» مقدار هدف باید بر اساس ویژگی‌های داده
۵۸۴ شده باشد . در این کتاب، پیش‌بینی مدل با نشان داده شده است

۵۸۵

فصل ۳ تفسیر پذیری

تعريف تفسیرپذیری (از نظر ریاضی) دشوار است. تعريف (غیر ریاضی) تفسیرپذیری که من توسط میلر (۲۰۱۷) ۵۸۶
دوست دارم این است: تفسیرپذیری درجه ای است که یک انسان می تواند علت یک تصمیم را درک کند.
۵۸۷
یکی دیگر این است: تفسیرپذیری درجه ای است که یک انسان می تواند به طور مداوم نتیجه مدل را پیش بینی
کند .. هرچه قابلیت تفسیر یک مدل یادگیری ماشین بالاتر باشد، درک اینکه چرا تصمیم‌ها یا پیش‌بینی‌های
۵۸۹
خاصی گرفته شده‌اند برای کسی آسان‌تر است. یک مدل بهتر از مدل دیگر قابل تفسیر است اگر تصمیمات آن
۵۹۰
برای انسان آسان‌تر از تصمیمات مدل دیگر باشد. من از هر دو اصطلاح قابل تفسیر و توضیح به جای یکدیگر
۵۹۱
استفاده خواهم کرد. مانند میلر (۲۰۱۷)، من فکر می کنم منطقی است که بین اصطلاحات تفسیرپذیری /
۵۹۲
توضیح پذیری و توضیح تفاوت قائل شویم. من از "توضیح" برای توضیح پیش‌بینی‌های فردی استفاده خواهم
۵۹۳
کرد. برای یادگیری آنچه که ما انسان‌ها به عنوان یک توضیح خوب می بینیم، به بخش توضیحات مراجعه کنید
۵۹۴
.

یادگیری ماشینی قابل تفسیر یک اصطلاح مفید است که "استخراج دانش مرتبط از یک مدل یادگیری ماشینی
۵۹۵
در مورد روابط موجود در داده‌ها یا آموخته شده توسط مدل" را نشان می دهد. ۵

۳.۱ اهمیت تفسیرپذیری

اگر یک مدل یادگیری ماشینی عملکرد خوبی دارد، چرا ما فقط به مدل اعتماد نمی کنیم و چرایی تصمیم
۶۰۱
خاصی را نادیده می گیریم؟ مشکل این است که یک معیار واحد، مانند دقت طبقه‌بندی، توصیف ناقصی از اکثر
۶۰۲
وظایف دنیای واقعی است. (دوشی-ولز و کیم ۲۰۱۷)

اجازه دهید به دلایلی که چرا تفسیرپذیری بسیار مهم است، عمیق‌تر بپردازیم. وقتی نوبت به مدل‌سازی
۶۰۴
پیش‌بینی می‌شود، باید یک مبادله انجام دهید: آیا فقط می‌خواهید بدانید چه چیزی پیش‌بینی می‌شود؟ به
۶۰۵
عنوان مثال، احتمال اینکه یک مشتری سرگردان شود یا اینکه برخی از داروها چقدر برای بیمار موثر است. یا
۶۰۶
می‌خواهید بدانید چرا؟ پیش‌بینی انجام شد و احتمالاً برای تفسیرپذیری با کاهش عملکرد پیش‌بینی‌کننده
۶۰۷
پرداخت؟ در برخی موارد، برای شما مهم نیست که چرا تصمیم گرفته شده است، کافی است بدانید که عملکرد
۶۰۸
پیش‌بینی‌کننده روی مجموعه داده آزمایشی خوب بود. اما در موارد دیگر، دانستن «چرا» می‌تواند به شما کمک
۶۰۹
کند درباره مشکل، داده‌ها و دلیل شکست یک مدل بیشتر بدانید. برخی از مدل‌ها ممکن است نیازی به توضیح
۶۱۰
نداشته باشند زیرا در یک محیط کم خطر استفاده می‌شوند، به این معنی که یک اشتباه عواقب جدی در پی
۶۱۱
نخواهد داشت (مثالاً سیستم توصیه‌کننده فیلم) یا روش قبل‌به طور گسترده مورد مطالعه و ارزیابی قرار گرفته
۶۱۲

است (مثلاً تشخیص کاراکتر نوری). نیاز به تفسیرپذیری از ناقصی در رسمی‌سازی مسئله ناشی می‌شود (دوشی-
ولز و کیم ۲۰۱۷)، به این معنی که برای مشکلات یا وظایف خاص، پیش‌بینی کافی نیست.چی). مدل همچنین
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 تقاضا برای تفسیرپذیری و توضیح می‌شود (دوشی-ولز و کیم ۲۰۱۷ و میلر
). ۲۰۱۷

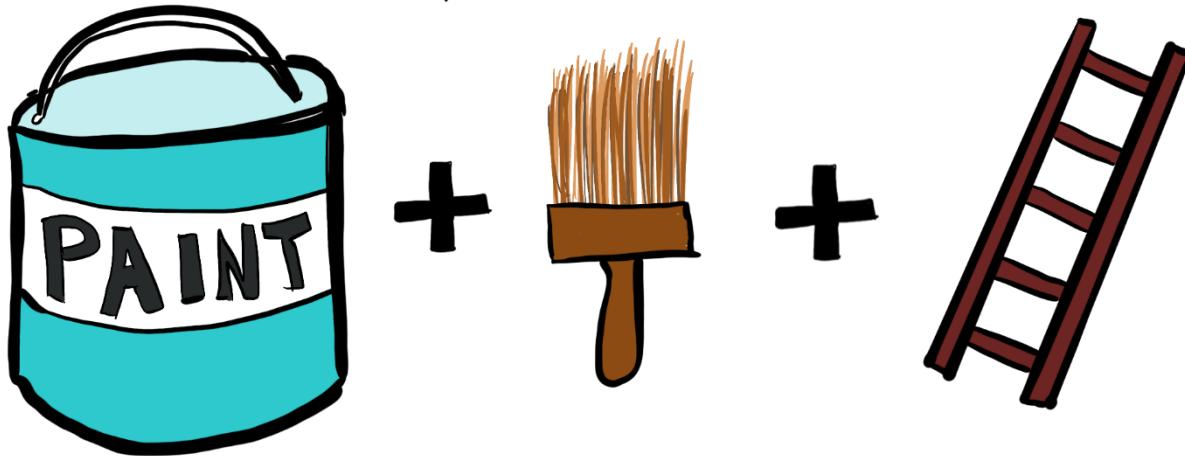
کنجکاوی و یادگیری انسان: انسان‌ها یک مدل ذهنی از محیط خود دارند که زمانی که اتفاق غیرمنتظره ای رخ
می‌دهد به روز می‌شود. این به روز رسانی با یافتن توضیحی برای رویداد غیرمنتظره انجام می‌شود. به عنوان
مثال، یک انسان به طور غیرمنتظره ای احساس بیماری می‌کند و می‌پرسد: "چرا اینقدر احساس بیماری می‌
کنم؟". او یاد می‌گیرد که هر بار که آن توت قرمز را می‌خورد بیمار می‌شود. او مدل ذهنی خود را به روز می‌
کند و به این نتیجه می‌رسد که توت‌ها باعث بیماری شده‌اند و بنابراین باید از آنها اجتناب شود. هنگامی که از
مدل‌های یادگیری ماشینی 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 نمی‌دانند
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 بین
عناصر ساختارهای دانش خود هماهنگ کنیم. "چرا سگم را گاز گرفت با وجود اینکه قبل این کار را نکرده
بود؟" ممکن است یک انسان بپرسد بین آگاهی از رفتار گذشته سگ و تجربه ناخوشایند نیش تازه ساخته شده
تناقض وجود دارد. توضیحات دامپیزشک 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 برای ماشین
اهمیت بیشتری دارد. اگر یک مدل یادگیری ماشین درخواست وام را رد کند، این ممکن است برای متقاضیان
کاملاً غیرمنتظره باشد. آنها فقط می‌توانند این ناسازگاری بین انتظار و واقعیت را با نوعی توضیح آشتبی دهند.
توضیحات در واقع نباید شرایط را به طور کامل توضیح دهند، بلکه باید به یک علت اصلی بپردازنند. مثال دیگر
توصیه 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 می‌کند
زیرا اخیراً یک ماشین لباسشویی خریدم و می‌دانم که در روزهای آینده تبلیغات ماشین لباسشویی دنبال خواهم

شده. بله، اگر از قبل کلاه زمستانی در سبد خریدم داشته باشم، پیشنهاد دستکش منطقی است. الگوریتم این فیلم را توصیه می کند، زیرا کاربرانی که فیلم های دیگری را دوست داشتند که من دوست داشتم، از فیلم پیشنهادی لذت بردند. شرکت های اینترنتی به طور فزاینده ای توضیحاتی را به توصیه های خود اضافه می کنند. یک مثال خوب، توصیه های محصول است که بر اساس ترکیبات محصولاتی که اغلب خریداری می شوند، هستند: من همیشه به این فکر می کنم که چرا محصولات یا فیلم های خاصی به صورت الگوریتمی به من توصیه شده اند. اغلب کاملاً واضح است: تبلیغات من را در اینترنت دنبال می کند زیرا اخیراً یک ماشین لباسشویی خریدم و می دانم که در روزهای آینده تبلیغات ماشین لباسشویی دنبال خواهم شد. بله، اگر از قبل کلاه زمستانی در سبد خریدم داشته باشم، پیشنهاد دستکش منطقی است. الگوریتم این فیلم را توصیه می کند، زیرا کاربرانی که فیلم های دیگری را دوست داشتند که من دوست داشتم، از فیلم پیشنهادی لذت بردند. شرکت های اینترنتی به طور فزاینده ای توضیحاتی را به توصیه های خود اضافه می کنند. یک مثال خوب، توصیه های محصول است که بر اساس ترکیبات محصولاتی که اغلب خریداری می شوند، هستند: من همیشه به این فکر می کنم که چرا محصولات یا فیلم های خاصی به صورت الگوریتمی به من توصیه شده اند. اغلب کاملاً واضح است: تبلیغات من را در اینترنت دنبال می کند زیرا اخیراً یک ماشین لباسشویی خریدم و می دانم که در روزهای آینده تبلیغات ماشین لباسشویی دنبال خواهم شد. بله، اگر از قبل کلاه زمستانی در سبد خریدم داشته باشم، پیشنهاد دستکش منطقی است. الگوریتم این فیلم را توصیه می کند، زیرا کاربرانی که فیلم های دیگری را دوست داشتند که من دوست داشتم، از فیلم پیشنهادی لذت بردند. شرکت های اینترنتی به طور فزاینده ای توضیحاتی را به توصیه های خود اضافه می کنند. یک مثال خوب، ترکیبات محصولاتی که اغلب خریداری می شوند، هستند: تبلیغات من را در اینترنت دنبال می کند زیرا به تازگی یک ماشین لباسشویی خریده ام و می دانم که در روزهای آینده تبلیغات ماشین لباسشویی را دنبال خواهم کرد. بله، اگر از قبل کلاه زمستانی در سبد خریدم داشته باشم، پیشنهاد دستکش منطقی است. الگوریتم این فیلم را توصیه می کند، زیرا کاربرانی که فیلم های دیگری را دوست داشتند که من دوست داشتم، از فیلم پیشنهادی لذت بردند. شرکت های اینترنتی به طور فزاینده ای توضیحاتی را به توصیه های خود اضافه می کنند. یک مثال خوب، توصیه های محصول است که بر اساس ترکیبات محصولاتی که اغلب خریداری می شوند، هستند: تبلیغات من را در اینترنت دنبال می کند زیرا به تازگی یک ماشین لباسشویی خریده ام و می دانم که در روزهای آینده تبلیغات ماشین لباسشویی را دنبال خواهم کرد. بله، اگر از قبل کلاه زمستانی در سبد خریدم داشته باشم، پیشنهاد دستکش منطقی است. الگوریتم این فیلم را توصیه می کند، زیرا کاربرانی که فیلم های دیگری را دوست داشتند که من دوست داشتم، از فیلم پیشنهادی لذت بردند. شرکت های اینترنتی به طور فزاینده ای توضیحاتی را به توصیه های خود اضافه می کنند. یک مثال خوب، توصیه های محصول است که بر

۶۶۸ اساس ترکیبات محصولاتی که اغلب خریداری می‌شوند، هستند: زیرا کاربرانی که فیلم‌های دیگری را دوست
۶۶۹ داشتند که من دوست داشتم نیز از فیلم پیشنهادی لذت بردن. شرکت‌های اینترنتی به طور فزاینده‌ای
۶۷۰ توضیحاتی را به توصیه‌های خود اضافه می‌کنند. یک مثال خوب، توصیه‌های محصول است که بر اساس
۶۷۱ ترکیبات محصولاتی که اغلب خریداری می‌شوند، هستند: زیرا کاربرانی که فیلم‌های دیگری را دوست داشتند
۶۷۲ که من دوست داشتم نیز از فیلم پیشنهادی لذت بردن. شرکت‌های اینترنتی به طور فزاینده‌ای توضیحاتی را به
۶۷۳ توصیه‌های خود اضافه می‌کنند. یک مثال خوب، توصیه‌های محصول است که بر اساس ترکیبات محصولاتی که
۶۷۴ اغلب خریداری می‌شوند، هستند:

Frequently Bought Together



۶۷۵

۶۷۶ شکل ۱: محصولات پیشنهادی که اغلب با هم خریداری می‌شوند.
۶۷۷

در بسیاری از رشته‌های علمی تغییر از روش‌های کیفی به کمی (به عنوان مثال جامعه‌شناسی، روانشناسی)، و
۶۷۸ همچنین به سمت یادگیری ماشین (زیست‌شناسی، ژنومیک) وجود دارد. هدف علم کسب دانش است، اما
۶۷۹ بسیاری از مشکلات با مجموعه داده‌های بزرگ و مدل‌های یادگیری ماشین جعبه سیاه حل می‌شوند. خود
۶۸۰ مدل به جای داده به منبع دانش تبدیل می‌شود. تفسیرپذیری امکان استخراج این دانش اضافی را که توسط
۶۸۱ مدل گرفته شده است را ممکن می‌سازد.

۶۸۲ مدل‌های یادگیری ماشین وظایف دنیای واقعی را انجام می‌دهند که به اقدامات ایمنی و آزمایش نیاز دارند. تصور
۶۸۳ کنید یک ماشین خودران به طور خودکار دوچرخه سواران را بر اساس یک سیستم یادگیری عمیق شناسایی می‌
۶۸۴ کند. شما می‌خواهید ۱۰۰٪ مطمئن باشید که انتزاعی که سیستم آموخته است بدون خطا است، زیرا دویدن

روی دوچرخه سواران بسیار بد است. یک توضیح ممکن است نشان دهد که مهم ترین ویژگی آموخته شده، تشخیص دو چرخ دوچرخه است، و این توضیح به شما کمک می کند در مورد لبه هایی مانند دوچرخه با کیسه های جانبی که تا حدی چرخ ها را می پوشانند فکر کنید.

به طور پیشفرض، مدل های یادگیری ماشین، سوگیری ها را از داده های آموزشی دریافت می کنند. این می تواند مدل های یادگیری ماشینی شما را به نژادپرستانی تبدیل کند که علیه گروه های دارای نمایندگی کمتر تعییض قائل می شوند. تفسیرپذیری یک ابزار اشکال زدایی مفید برای تشخیص سوگیری استدر مدل های یادگیری ماشینی ممکن است این اتفاق بیفتد که مدل یادگیری ماشینی که برای تأیید یا رد خودکار درخواست های اعتباری آموزش داده اید، علیه اقلیتی که از لحاظ تاریخی از حق امتیاز محروم شده اند تعییض قائل شود. هدف اصلی شما اعطای وام فقط به افرادی است که در نهایت آنها را بازپرداخت خواهند کرد. ناقص بودن فرمول مشکل در این مورد در این واقعیت نهفته است که شما نه تنها می خواهید نکول وام را به حداقل برسانید، بلکه موظف هستید بر اساس جمعیت شناسی خاص تعییض قائل نشوید. این یک محدودیت اضافی است که بخشی از فرمول مشکل شما (اعطای وام به روشی کم خطر و سازگار) است که توسط تابع ضرری که مدل یادگیری ماشین برای آن بهینه شده است پوشش داده نمی شود.

فرآیند ادغام ماشین ها و الگوریتم ها در زندگی روزمره ما نیازمند تفسیرپذیری برای افزایش پذیرش اجتماعی است . مردم باورها، امیال، نیات و غیره را به اشیا نسبت می دهند. در یک آزمایش معروف، هایدر و زیمل (۱۹۴۴) ۷ به شرکت کنندگان فیلم هایی از اشکال نشان داد که در آنها یک دایره یک "در" را برای ورود به یک "اتاق" (که به سادگی یک مستطیل بود) باز می کرد. شرکت کنندگان اعمال شکل ها را همانطور که اعمال یک عامل انسانی را توصیف می کردند، تخصیص نیات و حتی احساسات و ویژگی های شخصیتی به اشکال توصیف کردند. ربات ها مثال خوبی هستند، مانند جاروبرقی من که نام آن را "دوچ" گذاشت. اگر دوچ گیر کند، فکر می کنم: "دوچ می خواهد به تمیز کردن ادامه دهد، اما از من کمک می خواهد زیرا گیر کرده است." بعداً، وقتی تمیز کردن را تمام می کند و پایگاه خانه را برای شارژ مجدد جستجو می کند، فکر می کنم: "دوچ میل Doge به شارژ مجدد دارد و قصد دارد پایگاه خانه را پیدا کند." من همچنین ویژگی های شخصیتی را نسبت می دهم: "دوچ کمی گنگ است، اما به شکلی زیبا." اینها افکار من است، به خصوص وقتی متوجه می شوم که دوچ در حالی که خانه را با جاروبرقی جاروبرقی می کشد، گیاهی را کوبیده است. ماشین یا الگوریتمی که پیش بینی های خود را توضیح می دهد، مقبولیت بیشتری پیدا می کند. را نیز ببینید فصل توضیحات ، که استدلال می کند که تبیین ها یک فرآیند اجتماعی هستند.

از توضیحات برای مدیریت تعاملات اجتماعی استفاده می شود. توضیح دهنده با ایجاد معنای مشترک از چیزی،
بر اعمال، احساسات و باورهای گیرنده توضیح تأثیر می گذارد. برای اینکه یک ماشین با ما تعامل داشته باشد،
ممکن است نیاز داشته باشد که احساسات و باورهای ما را شکل دهد. ماشین ها باید ما را مقاعد کنند تا بتوانند
به هدف مورد نظر خود برسند. اگر ربات جاروبرقی خود را تا حدی توضیح نمی داد، به طور کامل نمی پذیرم.
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 می کند. جالب اینجاست که ممکن است
بین هدف ماشین توضیح دهنده (ایجاد اعتماد) و هدف گیرنده (درک پیش بینی یا رفتار) ناهمانگی وجود
داشته باشد. شاید توضیح کامل برای اینکه چرا Doge گیر کرده می تواند این باشد که باتری بسیار کم است،
یکی از چرخ ها به درستی کار نمی کند و یک اشکال وجود دارد که باعث می شود ربات بارها و بارها به همان
 نقطه برود حتی اگر وجود داشته باشد. یک مانع 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، دوج دوباره در حمام گیر کرد. قبل از اینکه دوج را جاروبرقی بگذاریم، باید
هر بار فرش ها را برداریم. اما فقط توضیح داد که چیزی در راه است و همین برای من کافی بود تا به رفتار آن
اعتماد کنم و معنای مشترکی از آن حادثه پیدا کنم. به هر حال، دوج دوباره در حمام گیر کرد. قبل از اینکه
دوج را جاروبرقی بگذاریم، باید هر بار فرش ها را برداریم. اما فقط توضیح داد که چیزی در راه است و همین
برای من کافی بود تا به رفتار آن اعتماد کنم و معنای مشترکی از آن حادثه پیدا کنم. به هر حال، دوج دوباره در
حمام گیر کرد. قبل از اینکه دوج را جاروبرقی بگذاریم، باید هر بار فرش ها را برداریم.



شکل ۳،۲: دوچ، جاروبرقی ما، گیر کرده است. به عنوان توضیحی برای تصادف، Doge به ما گفت که باید روی سطح صاف باشد.

مدل‌های یادگیری ماشینی را فقط می‌توان اشکال زدایی و ممیزی کردزمانی که بتوان آنها را تفسیر کرد. حتی در محیط‌های کم خطر، مانند توصیه‌های فیلم، توانایی تفسیر در مرحله تحقیق و توسعه و همچنین پس از استقرار ارزشمند است. بعداً، 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 یک جهت برای نحوه تعمیر سیستم ارائه می‌دهد. نمونه‌ای از طبقه‌بندی کننده هاسکی در مقابل گرگ را در نظر بگیرید که برخی هاسکی‌ها به اشتباه به عنوان گرگ طبقه‌بندی می‌کند. با استفاده از روش‌های یادگیری ماشینی قابل تفسیر، متوجه می‌شوید که طبقه‌بندی اشتباه به دلیل برف روی تصویر است. طبقه‌بندی کننده یاد گرفت که از برف به عنوان ویژگی برای طبقه‌بندی تصاویر به عنوان «گرگ» استفاده کند، که ممکن است از نظر جدا کردن گرگ‌ها از هاسکی در مجموعه داده‌های آموزشی منطقی باشد، اما نه در استفاده در دنیای واقعی.

اگر می‌توانید اطمینان حاصل کنید که مدل یادگیری ماشینی می‌تواند تصمیمات را توضیح دهد، می‌توانید ویژگی‌های زیر را نیز راحت‌تر بررسی کنید (دوشی-ولز و کیم ۲۰۱۷):

انصار: حصول اطمینان از اینکه پیش‌بینی‌ها بی‌طرفانه هستند و به‌طور ضمنی یا صریح علیه گروه‌های کم‌نمایش تعییض قائل نمی‌شوند. یک مدل قابل تفسیر می‌تواند به شما بگوید که چرا تصمیم گرفته است که یک فرد خاص نباید وام دریافت کند، و قضاوت در مورد اینکه آیا این تصمیم بر اساس یک سوگیری جمعیت شناختی (مثلاً نژادی) است برای یک انسان آسان‌تر می‌شود.

حریم خصوصی: اطمینان از محافظت از اطلاعات حساس در داده‌ها.

قابلیت اطمینان یا استحکام: اطمینان از اینکه تغییرات کوچک در ورودی منجر به تغییرات بزرگ در پیش‌بینی نمی‌شود.

علیت: بررسی کنید که فقط روابط علی انتخاب شده باشند.

اعتماد: در مقایسه با جعبه سیاه، اعتماد به سیستمی که تصمیمات خود را توضیح می‌دهد برای انسان آسان‌تر است.

زمانی که نیازی به تفسیر پذیری نداریم.

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

اگر مدل تاثیر 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 تحصیل کرده
غافلگیر کند که آنها در تعطیلات کجا خواهند رفت. اگر مدل اشتباه باشد مشکلی واقعی وجود ندارد (در بدترین
حالت فقط کمی خجالت مایک است)، همچنین اگر مایک نتواند خروجی مدل خود را توضیح دهد مشکلی وجود
ندارد. کاملاً خوب است که در این مورد قابل تفسیر نباشد. اگر مایک شروع به ایجاد یک کسب و کار در مورد
این پیش‌بینی‌های مقصود تعطیلات کند، وضعیت تغییر می‌کند. اگر مدل اشتباه باشد، کسب‌وکار ممکن است
ضرر کند، یا این مدل ممکن است به دلیل تعصبات نژادی آموخته شده برای برخی افراد بدتر عمل کند. به
محض اینکه مدل تأثیر قابل توجهی، خواه مالی یا اجتماعی داشته باشد، قابلیت تفسیر مرتبط می‌شود.

وقتی مسئله به خوبی مطالعه شده باشد، نیازی به تفسیرپذیری نیست. برخی از کاربردها به اندازه کافی خوب
مطالعه شده اند به طوری که تجربه عملی کافی با مدل وجود دارد و مشکلات مدل در طول زمان حل شده
است. یک مثال خوب یک مدل یادگیری ماشینی برای تشخیص کاراکترهای نوری است که تصاویر را از پاکتها
پردازش می‌کند و آدرس‌ها را استخراج می‌کند. سال‌ها تجربه با این سیستم‌ها وجود دارد و مشخص است که
کار می‌کنند. علاوه بر این، ما واقعاً علاقه‌ای به کسب بینش اضافی در مورد کار در دست نداریم.

تفسیرپذیری ممکن است افراد یا برنامه‌ها را قادر به دستکاری سیستم کند. مشکلات کاربرانی که یک سیستم
را فریب می‌دهند ناشی از عدم تطابق بین اهداف سازنده و کاربر یک مدل است. امتیازدهی اعتباری چنین
سیستمی است زیرا بانک‌ها می‌خواهند اطمینان حاصل کنند که وام‌ها فقط به مقاضیانی داده می‌شود که
احتمالاً آن‌ها را پس می‌دهند و مقاضیان قصد دارند وام را دریافت کنند حتی اگر بانک نخواهد به آنها وام
بدهد. این عدم تطابق بین اهداف، انگیزه‌هایی را برای مقاضیان ایجاد می‌کند تا با سیستم بازی کنند تا شанс
خود را برای دریافت وام افزایش دهند. اگر مقاضی می‌داند که داشتن بیش از دو کارت اعتباری بر امتیاز او تأثیر
منفی می‌گذارد، به سادگی سومین کارت اعتباری خود را برای بهبود امتیاز خود برمی‌گرداند و پس از تأیید وام،
کارت جدیدی ترتیب می‌دهد. در حالی که امتیاز او بهبود یافت، احتمال واقعی بازپرداخت وام بدون تغییر باقی
ماند. سیستم را تنها در صورتی می‌توان بازی کرد که ورودی‌ها پرآکسی برای یک ویژگی علی باشند، اما در
واقع باعث نتیجه نمی‌شود. در صورت امکان، باید از ویژگی‌های پرآکسی اجتناب شود، زیرا آنها مدل‌ها را قابل
بازی می‌کنند. به عنوان مثال، گوگل سیستمی به نام Google Flu Trends برای پیش‌بینی شیوع آنفلوانزا

ایجاد کرد. این سیستم جستجوهای گوگل را با شیوع آنفولانزا مرتبط کرد - و عملکرد ضعیفی داشته است.
توزیع عبارت‌های جستجو تغییر کرد و Google Flu Trends بسیاری از شیوع آنفولانزا را از دست داد.
جستجوی گوگل باعث آنفولانزا نمی‌شود. هنگامی که افراد علائمی مانند "تب" را جستجو می‌کنند، صرفاً یک
ارتباط با شیوع واقعی آنفولانزا است. در حالت ایده‌آل، مدل‌ها فقط از ویژگی‌های علی استفاده می‌کنند زیرا
قابل بازی نیستند. این سیستم جستجوهای گوگل را با شیوع آنفولانزا مرتبط کرد - و عملکرد ضعیفی داشته
است. توزیع عبارت‌های جستجو تغییر کرد و Google Flu Trends بسیاری از شیوع آنفولانزا را از دست داد.
جستجوی گوگل باعث آنفولانزا نمی‌شود. هنگامی که افراد علائمی مانند "تب" را جستجو می‌کنند، صرفاً یک
ارتباط با شیوع واقعی آنفولانزا است. در حالت ایده‌آل، مدل‌ها فقط از ویژگی‌های علی استفاده می‌کنند زیرا
قابل بازی نیستند. این سیستم جستجوهای گوگل را با شیوع آنفولانزا مرتبط کرد - و عملکرد ضعیفی داشته
است. توزیع عبارت‌های جستجو تغییر کرد و Google Flu Trends بسیاری از شیوع آنفولانزا را از دست داد.
جستجوی گوگل باعث آنفولانزا نمی‌شود. هنگامی که افراد علائمی مانند "تب" را جستجو می‌کنند، صرفاً یک
ارتباط با شیوع واقعی آنفولانزا است. در حالت ایده‌آل، مدل‌ها فقط از ویژگی‌های علی استفاده می‌کنند زیرا
قابل بازی نیستند.

۳.۲ طبقه‌بندی روش‌های تفسیرپذیری

روش‌های تفسیرپذیری یادگیری ماشین را می‌توان بر اساس معیارهای مختلف طبقه‌بندی کرد.

ذاتی یا پست؟ این معیار تشخیص می‌دهد که آیا تفسیرپذیری با محدود کردن پیچیدگی مدل یادگیری ماشین
(ذاتی) یا با استفاده از روش‌هایی که مدل را پس از آموزش تجزیه و تحلیل می‌کند (post hoc) به دست می‌
آید. تفسیرپذیری ذاتی به مدل‌های یادگیری ماشینی 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
پست است. روش‌های post hoc 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 تعیین می‌شود مدل‌های ذاتاً قابل تفسیر و روش‌های تفسیر تعقیبی (و مدل-آگنوستیک).)

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
روشنها یک عدد واحد را برای هر ویژگی بر می‌گردانند، 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 ویژگی ها نیز قابل تجسم هستند. برخی از خلاصه های ویژگی ها در
واقع تنها زمانی معنادار هستند که تجسم شوند و یک جدول انتخاب اشتباہی باشد. وابستگی جزئی یک ویژگی
چنین موردی است. نمودارهای وابستگی جزئی منحنی هایی هستند که یک ویژگی و میانگین نتیجه پیش بینی
شده را نشان می دهند. بهترین راه برای ارائه وابستگی های جزئی، رسم منحنی به جای چاپ مختصات است.

دروندی های مدل (مثلاً وزن های آموخته شده) : تفسیر مدل های قابل تفسیر ذاتی در این دسته قرار می گیرد. به
عنوان مثال وزن در مدل های خطی یا ساختار درختی آموخته شده (ویژگی ها و آستانه های مورد استفاده برای
تقسیم ها) درخت های تصمیم هستند. خطوط بین اجزای داخلی مدل و آمار خلاصه ویژگی، به عنوان مثال، در
مدل های خطی محظوظ می شوند، زیرا وزن ها هم زمان داخلی مدل و هم آمار خلاصه ویژگی ها هستند. روش دیگری
که مدل های داخلی را خروجی می دهد، تجسم آشکارسازهای ویژگی است که در شبکه های عصبی کانولوشن
آموخته شده اند. روش های تفسیر پذیری که داخلی های مدل خروجی را به دست می آورند، طبق تعریف، مختص
مدل هستند (معیار بعدی را ببینید).

نقشه داده : این دسته شامل تمام روش هایی است که نقاط داده (از قبل موجود یا تازه ایجاد شده) را برای قابل
تفسیر کردن یک مدل برمی گرداند. یکی از روش ها توضیحات خلاف واقع نامیده می شود. برای توضیح
پیش بینی یک نمونه داده، این روش با تغییر برخی از ویژگی هایی که نتیجه پیش بینی شده به روشنی مرتبط
تغییر می کند، یک نقطه داده مشابه پیدا می کند (مثلاً یک تلنگر در کلاس پیش بینی شده). مثال دیگر شناسایی
نمونه های اولیه کلاس های پیش بینی شده است. برای مفید بودن، روش های تفسیری که نقاط داده جدید را
تولید می کنند، مستلزم آن هستند که خود نقاط داده قابل تفسیر باشند. این برای تصاویر و متون به خوبی کار
می کند، اما برای داده های جدولی با صدھا ویژگی کمتر مفید است.

مدل قابل تفسیر ذاتی : یک راه حل برای تفسیر مدل های 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 می شود.

مدل خاص یا مدل آگنوستیک؟ ابزارهای تفسیر مدل خاص به کلاس های مدل خاص محدود می شوند. تفسیر
وزن های رگرسیون در یک مدل خطی یک تفسیر مدل خاص است، زیرا - طبق تعریف - تفسیر مدل های ذاتی
قابل تفسیر همیشه مختص مدل است. ابزارهایی که فقط برای تفسیر شبکه های عصبی کار می کنند، مختص
مدل هستند. ابزارهای مدل آگنوستیک را می توان در هر مدل یادگیری ماشینی استفاده کرد و پس از آموزش
مدل (post hoc) استفاده می شود. این روش های آگنوستیک معمولاً با تجزیه و تحلیل جفت های ورودی و

خروجی ویژگی کار می‌کنند. طبق تعریف، این روش‌ها نمی‌توانند به اجزای داخلی مدل مانند وزن یا اطلاعات ساختاری دسترسی داشته باشند.

محلی یا جهانی؟ آیا روش تفسیر یک پیش‌بینی فردی یا کل رفتار مدل را توضیح می‌دهد؟ یا دامنه در جایی در این بین است؟ در بخش بعدی بیشتر در مورد معیار دامنه مطالعه کنید

۳,۳ دامنه 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. اگر از شبکه‌های عصبی کانولوشن برای طبقه‌بندی تصاویر استفاده می‌کنید، می‌توانید توضیح دهید که الگوریتم آشکارسازهای لبه و فیلترها را در پایین ترین لایه‌ها یاد می‌گیرد. این درک نحوه عملکرد الگوریتم است، اما نه برای مدل خاصی که در پایان آموخته می‌شود، و نه برای چگونگی پیش‌بینی‌های فردی. شفافیت الگوریتم فقط به دانش الگوریتم نیاز دارد و نه از داده‌ها یا مدل‌های آموخته شده. 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. آنها با شفافیت بالا مشخص می‌شوند. رویکردهای یادگیری عمیق (هل کردن یک گرادیان از طریق شبکه‌ای با میلیون‌ها وزن) کمتر درک شده‌اند و کارهای درونی کانون تحقیقات مداوم هستند. آنها کمتر شفاف در نظر گرفته می‌شوند.

۳,۳,۲ تفسیرپذیری مدل کل نگر

مدل آموزش دیده چگونه پیش‌بینی می‌کند؟

اگر بتوانید کل مدل را یکجا درک کنید، می‌توانید یک مدل را قابل تفسیر توصیف کنید. (Lipton 2016 8) برای توضیح خروجی مدل جهانی، به مدل آموزش دیده، 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 باشد، بعید است که در حافظه کوتاه مدت یک انسان معمولی جای بگیرد. من استدلال می کنم که شما واقعاً نمی توانید 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 غیرقابل تصور است. معمولاً هنگامی که افراد سعی در درک یک مدل دارند، فقط بخش هایی از آن را در نظر می گیرند، مانند وزن ها در مدل های خطی.

۳,۳,۳ تفسیرپذیری مدل جهانی در سطح مدولار

چگونه بخش هایی از مدل بر پیش بینی ها تأثیر می گذارد؟

یک مدل ساده بیز با صدها ویژگی برای من و شما بزرگتر از آن است که بتوانیم آن را در حافظه کاری خود نگه داریم. و حتی اگر بتوانیم تمام وزن ها را به خاطر بسپاریم، نمی توانیم به سرعت برای نقاط داده جدید پیش بینی کنیم. علاوه بر این، شما باید توزیع مشترک همه ویژگی ها را در ذهن خود داشته باشید تا اهمیت هر ویژگی و اینکه ویژگی ها به طور متوسط چگونه بر پیش بینی ها تأثیر می گذارند، تخمین بزنید. یک کار غیر ممکن اما شما به راحتی می توانید یک وزن را درک کنید. در حالی که تفسیرپذیری مدل جهانی معمولاً دور از دسترس است، شанс خوبی برای درک حداقل برخی از مدل ها در سطح مدولار وجود دارد. همه مدل ها در سطح پارامتر قابل تفسیر نیستند. برای مدل های خطی، بخش های قابل تفسیر وزن ها هستند، برای درخت ها تقسیم ها (ویژگی های انتخابی به اضافه نقاط برش) و پیش بینی های گره برگ است. مدل های خطی، به عنوان مثال، به نظر می رسد که آنها را می توان به طور کامل در یک سطح مدولار تفسیر کرد، اما تفسیر یک وزن منفرد با تمام وزن های دیگر در هم تنیده است. تفسیر وزن منفرد همیشه با این پاورقی همراه می شود که سایر ویژگی های ورودی در همان مقدار باقی می مانند، که در مورد بسیاری از برنامه های واقعی صدق نمی کند. یک مدل خطی که ارزش یک خانه را پیش بینی می کند، که هم اندازه خانه و هم تعداد اتاق ها را در نظر می گیرد، می تواند وزن منفی برای ویژگی اتاق داشته باشد. این می تواند اتفاق بیفتند زیرا در حال حاضر ویژگی اندازه خانه بسیار همبسته وجود دارد. در بازاری که مردم اتاق های بزرگ تر را ترجیح می دهند، یک خانه با اتاق های کمتر می تواند ارزش بیشتری نسبت به خانه هایی با اتاق های بیشتر داشته باشد، اگر هر دو اندازه یکسان داشته باشند. وزن ها فقط در چارچوب سایر ویژگی های مدل معنا می باند.

۳,۳,۴ تفسیر محلی برای یک پیش بینی واحد

چرا مدل برای مثال پیش بینی خاصی انجام داد؟

می توانید روی یک نمونه بزرگنمایی کنید و آنچه را که مدل برای این ورودی پیش بینی می کند بررسی کنید و توضیح دهید که چرا. اگر به یک پیش بینی فردی نگاه کنید، رفتار مدل پیچیده تر ممکن است خوشایندتر رفتار

کند. 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 مجموعه داده، 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. بخش روش‌های مدل-آگنوتیک

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 با در نظر گرفتن گروه نمونه‌ها، رفتار با آن‌ها به گونه‌ای که گویی گروه مجموعه داده کامل است، و استفاده از روش‌های جهانی با این زیرمجموعه اعمال کرد. روش‌های توضیح فردی را می‌توان در هر نمونه استفاده کرد و سپس برای کل گروه فهرست یا جمع کرد.

۳.۴ ارزیابی تفسیرپذیری

هیچ اتفاق نظر واقعی 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.

دوشی ولز و کیم (۲۰۱۷) سه سطح اصلی را برای ارزیابی تفسیرپذیری پیشنهاد می‌کنند:

ارزیابی سطح برنامه (وظیفه واقعی) : توضیحات را در محصول قرار دهید و آن را توسط کاربر نهایی آزمایش کنید. نرم افزار تشخیص شکستگی را با یک جزء یادگیری ماشینی تصور کنید که شکستگی‌ها را در اشعه ایکس مکان یابی و علامت گذاری می‌کند. در سطح کاربرد، رادیولوژیست‌ها نرم افزار تشخیص شکستگی را مستقیماً برای ارزیابی مدل آزمایش می‌کنند. این نیاز به یک تنظیم تجربی خوب و درک چگونگی ارزیابی کیفیت دارد. یک مبنای خوب برای این همیشه 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

۹۱۲ (مخصوصاً اگر متخصصان حوزه رادیولوژیست باشند) و یافتن آزمایش کنندگان بیشتری آسان‌تر می‌شود. به
۹۱۳ عنوان مثال، توضیحات متفاوتی به کاربر نشان داده می‌شود و کاربر بهترین را انتخاب می‌کند.

۹۱۴ ارزیابی سطح عملکرد (وظیفه پروکسی) به انسان نیاز ندارد. این بهترین زمانی است که کلاس مدل مورد
۹۱۵ استفاده قبلًاً توسط شخص دیگری در ارزیابی سطح انسانی ارزیابی شده باشد. به عنوان مثال، ممکن است
۹۱۶ مشخص شود که کاربران نهایی درخت تصمیم را درک می‌کنند. در این مورد، یک پروکسی برای کیفیت
۹۱۷ توضیح ممکن است عمق درخت باشد. درختان کوتاه‌تر نمره توضیح پذیری بهتری را دریافت می‌کنند. منطقی
۹۱۸ است که این محدودیت را اضافه کنیم که عملکرد پیش‌بینی درخت خوب باقی می‌ماند و در مقایسه با درخت
۹۱۹ بزرگ‌تر خیلی کاهش نمی‌یابد.

۹۲۰ فصل بعدی بر ارزیابی توضیحات برای پیش‌بینی‌های فردی در سطح تابع مرکز دارد. ویژگی‌های مرتبط
۹۲۱ توضیحاتی که برای ارزیابی آنها در نظر می‌گیریم چیست؟

۳.۵ خواص توضیحات

۹۲۲ ما می‌خواهیم پیش‌بینی‌های یک مدل یادگیری ماشینی را توضیح دهیم. برای رسیدن به این هدف، ما به
۹۲۳ برخی از روش‌های توضیحی، که الگوریتمی است که توضیحات را تولید می‌کند، تکیه می‌کنیم. یک توضیح
۹۲۴ عموماً مقدادر ویژگی یک نمونه را به روشنی قابل درک برای انسان به پیش‌بینی مدل آن مرتبط می‌کند. انواع
۹۲۵ دیگر توضیحات شامل مجموعه‌ای از نمونه‌های داده است (مثلاً در مورد مدل k -NN نزدیک ترین همسایه). برای
۹۲۶ مثال، می‌توانیم خطر سرطان را با استفاده از یک ماشین بردار پشتیبان پیش‌بینی کنیم و پیش‌بینی‌ها را با
۹۲۷ استفاده از روش جایگزین محلی توضیح دهیم، که درخت‌های تصمیم را به عنوان توضیح تولید می‌کند. یا می‌
۹۲۸ توانیم به جای ماشین بردار پشتیبان از مدل رگرسیون خطی استفاده کنیم. مدل رگرسیون خطی قبلًاً با یک
۹۲۹ روش توضیحی (تفسیر وزن‌ها) مجهز شده است.
۹۳۰

۹۳۱ ما نگاهی دقیق‌تر به ویژگی‌های روش‌ها و توضیحات توضیح می‌اندازیم (Robnik-Sikonja and Bohanec, 2018).
۹۳۲ از این ویژگی‌ها می‌توان برای قضاوت در مورد خوب بودن روش یا توضیح توضیحی استفاده کرد.
۹۳۳ برای همه این ویژگی‌ها مشخص نیست که چگونه آنها را به درستی اندازه‌گیری کنیم، بنابراین یکی از چالش
۹۳۴ ها رسمی کردن نحوه محاسبه آنها است.

خواص روش‌های تبیین

۹۳۵ قدرت بیانی «زبان» یا ساختار توضیحاتی است که روش قادر به ایجاد آن است. یک روش توضیحی می‌تواند
۹۳۶ قوانین IF-THEN، درخت‌های تصمیم، جمع وزنی، زبان طبیعی یا چیز دیگری را ایجاد کند.
۹۳۷

توضیح می دهد که روش توضیح تا چه حد به بررسی مدل یادگیری ماشینی، 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 ماشینی 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؟ دقต 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 ماشینی نیز پایین باشد، و اگر هدف توضیح این باشد که مدل جعبه سیاه چه کاری انجام می‌دهد. در این مورد فقط وفاداری مهم است.

وفداری : 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 آموزش دیده 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

۹۶۴ انتخاب می کنم محاسبه می کنم و تفاوت های توضیحات را تجزیه و تحلیل می کنم. اگر توضیحات بسیار شبیه به
۹۶۵ هم باشند، توضیحات بسیار سازگار هستند. من این ویژگی را تا حدودی دشوار می دانم، زیرا این دو مدل
۹۶۶ می توانند از ویژگی های متفاوتی استفاده کنند، اما پیش بینی های مشابهی دریافت می کنند (همچنین «اثر
۹۶۷ راشومون» نامیده می شود). در این مورد سازگاری بالا مطلوب نیست زیرا توضیحات باید بسیار متفاوت باشند.
۹۶۸ اگر مدل ها واقعاً بر روابط مشابه متکی باشند، سازگاری بالا مطلوب است.

۹۶۹ پایداری : توضیحات برای نمونه های مشابه چقدر شبیه است؟ در حالی که سازگاری توضیحات بین مدل ها را
۹۷۰ مقایسه می کند، ثبات توضیحات بین نمونه های مشابه را برای یک مدل ثابت مقایسه می کند. پایداری بالا به
۹۷۱ این معنی است که تغییرات جزئی در ویژگی های یک نمونه، توضیح اساسی را تغییر نمی دهد (مگر اینکه این
۹۷۲ تغییرات جزئی نیز پیش بینی را به شدت تغییر دهد). عدم ثبات می تواند نتیجه واریانس زیاد روش تبیین باشد.
۹۷۳ به عبارت دیگر، روش توضیح به شدت تحت تأثیر تغییرات جزئی مقادیر ویژگی نمونه مورد توضیح قرار می
۹۷۴ گیرد. فقدان ثبات همچنین می تواند ناشی از مؤلفه های غیر قطعی روش توضیح باشد، مانند مرحله نمونه گیری
۹۷۵ داده ها، مانند روش جایگزین محلی . پایداری بالا همیشه مطلوب است.

۹۷۶ قابل درک بودن : انسان ها چقدر توضیحات را درک می کنند؟ این دقیقاً مانند یک ملک دیگر در میان بسیاری
۹۷۷ به نظر می رسد، اما فیل در اتاق است. تعریف و اندازه گیری دشوار است، اما درست کردن آن بسیار مهم است.
۹۷۸ بسیاری از مردم قبول دارند که قابل درک بودن به مخاطب بستگی دارد. ایده هایی برای اندازه گیری در کپذیری
۹۷۹ شامل اندازه گیری اندازه توضیح (تعداد ویژگی ها با وزن غیر صفر در یک مدل خطی، تعداد قوانین تصمیم گیری،
۹۸۰ ...) یا آزمایش اینکه افراد چقدر می توانند رفتار مدل یادگیری ماشینی را از توضیحات پیش بینی کنند، می شود ..
۹۸۱ قابل درک بودن ویژگی های استفاده شده در توضیح نیز باید در نظر گرفته شود. تغییر پیچیده ویژگی ها ممکن
۹۸۲ است کمتر از ویژگی های اصلی قابل درک باشد.

۹۸۳ قطعیت : آیا توضیح، قطعیت مدل یادگیری ماشین را منعکس می کند؟ بسیاری از مدل های یادگیری ماشینی
۹۸۴ فقط پیش بینی می کنند بدون اینکه بیانیه ای در مورد مدل ها وجود داشته باشد، اطمینان دارند که پیش بینی
۹۸۵ درست است. اگر مدل یک احتمال 4 درصدی سرطان را برای یک بیمار پیش بینی کند، آیا به اندازه احتمال 4
۹۸۶ درصدی است که بیمار دیگر با مقادیر ویژگی های متفاوت دریافت کرده است؟ توضیحی که شامل قطعیت مدل
۹۸۷ باشد بسیار مفید است.

۹۸۸ درجه اهمیت : توضیح چقدر اهمیت ویژگی ها یا بخش هایی از توضیح را منعکس می کند؟ به عنوان مثال، اگر
۹۸۹ یک قاعده تصمیم به عنوان توضیحی برای یک پیش بینی فردی ایجاد شود، آیا مشخص است که کدام یک از
۹۹۰ شرایط قانون مهم ترین بوده است؟

تازگی : آیا توضیح نشان می دهد که آیا نمونه داده ای که باید توضیح داده شود از یک منطقه "جدید" به دور از توزیع داده های 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.

نمايندگي : توضیح چند مورد را پوشش می دهد؟ توضیحات می توانند کل مدل را پوشش دهند (مثلاً تفسیر وزن ها در مدل رگرسیون خطی) یا فقط یک پیش بینی فردی را نشان دهند (مثلاً مقادیر Shapley).

۳.۶ توضیحات انسان پسند

بیایید عمیق‌تر کاوش کنیم و آنچه را که ما انسان‌ها به عنوان توضیحات «خوب» می‌بینیم و پیامدهای آن برای یادگیری ماشینی قابل تفسیر چیست، کشف کنیم. تحقیقات علوم انسانی می‌تواند به ما در یافتن این موضوع کمک کند. میلر (۲۰۱۷) نظرسنجی عظیمی از انتشارات درباره توضیحات انجام داده است و این فصل بر اساس خلاصه او است.

در این فصل، می‌خواهم شما را به موارد متقدار 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 علل غیرعادی توضیحات خوبی ارائه می‌دهند. تبیین‌ها تعاملات اجتماعی بین تبیین کننده و توضیح دهنده (گیرنده توضیح) هستند و بنابراین زمینه اجتماعی تأثیر زیادی بر محتوای واقعی تبیین دارد.

وقتی برای یک پیش‌بینی یا رفتار خاص به توضیحاتی با همه عوامل نیاز دارید، توضیحی انسان‌دوستانه نمی‌خواهید، بلکه یک انتساب علی کامل می‌خواهید. اگر از نظر قانونی ملزم به تعیین همه ویژگی‌های تأثیرگذار هستید یا اگر مدل یادگیری ماشینی را اشکال‌زدایی می‌کنید، احتمالاً می‌خواهید یک انتساب علی داشته باشد. در این صورت به نکات زیر توجه نکنید. در سایر موارد که افراد غیر روحانی یا افراد کم وقت دریافت کننده توضیحات هستند، بخش‌های زیر باید برای شما جالب باشد.

۱.۶ توضیح چیست؟

توضیح پاسخ به یک سوال چرایی است. (Miller 2017)

چرا درمان روی بیمار جواب نداد؟

چرا وام من رد شد؟

چرا هنوز زندگی بیگانه با ما تماس نگرفته است؟

دو سؤال اول را می‌توان با توضیح «روزمره» پاسخ داد، در حالی که سؤال سوم از دسته «پدیده‌های کلی‌تر علمی و سؤال‌های فلسفی» می‌آید. ما روی توضیحات نوع «روزانه» تمرکز می‌کنیم، زیرا این توضیحات مربوط به یادگیری ماشینی 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 «تبیین» به فرآیند اجتماعی و شناختی تبیین 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 خلاصه میلر را در مورد توضیحات "خوب" بیشتر فشرده می‌کند و مفاهیم ملموسی را برای یادگیری ماشینی قابل تفسیر اضافه می‌کند.

توضیحات متضاد هستند. (Lipton 1990 10) انسان‌ها معمولاً نمی‌پرسند که چرا یک پیش‌بینی خاص انجام شده است، بلکه 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 قیمت پایین‌تری که انتظار داشتند، بالا است. اگر درخواست وام من رد شود، اهمیتی برای شنیدن همه عواملی که به طور کلی موافق یا مخالف رد شدن هستند، ندارم. من به عواملی در درخواست خود علاقه‌مند هستم که برای دریافت وام باید تغییر کنم. من می‌خواهم تفاوت بین برنامه من و نسخه مورد پذیرش برنامه من را بدانم. تشخیص اینکه توضیحات متضاد اهمیت دارند، یافته مهمی برای یادگیری ماشینی 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 در تضاد قرار دهد. پزشکان ممکن است بپرسند: "چرا دارو برای بیمار من کار نمی‌کند؟" و ممکن است توضیحی بخواهند که بیمارشان را با بیماری که دارو برای او موثر بوده و مشابه بیمار بدون پاسخ است، مقایسه کند. درک 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 بیش از حد بیان شده است، بدن بیمار در تجزیه دارو به مواد شیمیایی بی اثر بسیار سریع عمل می‌کند، ... توضیح متضاد ممکن است بسیار ساده‌تر باشد: برخلاف بیمار پاسخ دهنده، بیمار بدون پاسخ دارای ترکیب خاصی از ژن‌ها است که اثربخشی دارو را کاهش می‌دهد. بهترین توضیح توضیحی است که بیشترین تفاوت را بین شی مورد نظر و شی مرجع برجسته کند.

معنی آن برای یادگیری ماشینی قابل تفسیر : انسان ها توضیح کاملی برای یک پیش بینی نمی خواهند، اما می خواهند تفاوت ها را با پیش بینی نمونه دیگری مقایسه کنند (می تواند مصنوعی باشد). ایجاد توضیحات متضاد وابسته به کاربرد است زیرا به نقطه مرجع برای مقایسه نیاز دارد. و این ممکن است به نقطه داده ای که باید توضیح داده شود، بلکه به کاربر دریافت کننده توضیح بستگی دارد. یک کاربر یک وبسایت پیش بینی قیمت خانه ممکن است بخواهد توضیحی در مورد پیش بینی قیمت خانه در تضاد با خانه خود یا شاید خانه دیگری در وبسایت یا شاید با یک خانه متوسط در همسایگی داشته باشد. راه حل برای ایجاد خودکار توضیحات متضاد همچنین ممکن است شامل یافتن نمونه های اولیه یا کهن الگوها در داده ها باشد.

توضیحات انتخاب شده است . مردم انتظار توضیحی ندارند که فهرست واقعی و کامل علل یک رویداد را پوشش 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 فزاینده علیه محصول شرکت به دلیل مشکلات مربوط به آخرین به روز رسانی نرم افزار است".

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

بی اعتمادی فزاینده به نهادهای مستقر و دولت ما عوامل اصلی کاهش مشارکت رای دهنده‌گان است.

این واقعیت که یک رویداد را می توان با علل مختلف توضیح داد، اثر راشومون نامیده می شود. راشومون یک فیلم ژاپنی است که داستان های (توضیحات) متناقض و جایگزین درباره مرگ یک سامورایی را روایت می کند. برای مدل های یادگیری ماشین، اگر بتوان یک پیش بینی خوب از ویژگی های مختلف انجام داد، سودمند است. روش های مجموعه ای که چندین مدل را با ویژگی های مختلف ترکیب می کنند (توضیحات مختلف) عموماً عملکرد خوبی دارند زیرا میانگین گیری بیش از آن «داستان ها» پیش بینی ها را قوی تر و دقیق تر می کند. اما همچنین به این معنی است که بیش از یک توضیح انتخابی وجود دارد که چرا یک پیش بینی خاص انجام شده است.

معنی آن برای یادگیری ماشینی قابل تفسیر : توضیح را خیلی کوتاه بیان کنید، فقط ۱ تا ۳ دلیل بیاورید، حتی اگر دنیا پیچیده تر باشد. روش LIME با این کار خوب انجام می دهد

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

بلاک چین، که توسط یک نهاد مرکزی قابل کنترل نیست، با افرادی که می خواهند ایمن شوند، طنین انداز
می شود. ثروت آنها، که تقاضا و قیمت بالا را توضیح می دهد. 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 ماشینی قابل تفسیر چیست: به محیط اجتماعی برنامه یادگیری ماشینی 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.

توضیحات بر موارد غیرعادی تمرکز دارند . مردم برای توضیح رویدادها بیشتر بر علل غیرعادی تمرکز می کنند
(Kahnemann and Tversky, 1981 11). اینها 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). انسان ها این
 النوع علل «غیر طبیعی» را به عنوان توضیحات خوبی در نظر می گیرند. مثالی از اشترومبلج و کونونکو (۲۰۱۱)
است: فرض کنید مجموعه داده ای از موقعیت های آزمون بین معلمان و دانش آموزان داریم. دانش آموزان در
یک دوره شرکت می کنند و پس از ارائه موقفيت آميز دوره را مستقيماً می گذرانند. معلم اين گزينه را دارد که
علاوه بر آن از دانش آموز سوالاتي بپرسد تا دانش آنها را محک بزنند. دانش آموزاني که نتوانند به اين سوالات
پاسخ دهند در دوره مردود خواهند بود. دانش آموزان می توانند سطوح آمادگی متفاوتی داشته باشند، که به
احتمالات متفاوتی برای پاسخ صحيح به سوالات معلم ترجمه می شود (اگر تصمیم به امتحان دانش آموز داشته
باشند). ما می خواهیم پیش بینی کنیم که آیا یک دانش آموز این دوره را سپری می کند و پیش بینی خود را
توضیح دهیم. 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 می پرسد (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 می
کند که نمی تواند به درستی پاسخ دهد. چرا دانش آموز در درس مردود شد؟ می گوییم تقصیر دانشجو بود که
درس نخواند.

سناریوی ۲: معلم به ندرت سوالات اضافی می پرسد (مثلاً ۲ از ۱۰۰ بار). برای دانش آموزی که برای سوالات
مطالعه نکرده است، احتمال گذراندن دوره را زیاد پیش بینی می کنیم، زیرا سوالات بعید است. البته یکی از دانش
آموزان برای سوالات آماده نشد که ۱۰ درصد شانس قبولی در سوالات را به او می دهد. او بدشانس است و معلم

- سؤالات اضافی می پرسد که دانش آموز نمی تواند به آنها پاسخ دهد و در درس مردود می شود. دلیل شکست چیست؟ من استدلال می کنم که اکنون، توضیح بهتر این است که "چون معلم دانش آموز را امتحان کرد".
بعید بود که معلم امتحان بددهد، بنابراین معلم رفتار غیرعادی داشت.
- معنى آن برای یادگیری ماشینی 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 غیر طبیعی 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 است. گران.
- توضیحات درست است . توضیحات خوب در واقعیت (یعنی در موقعیت های دیگر) صادق هستند. اما به طرز نگران‌کننده‌ای، این مهمترین عامل برای توضیح «خوب» نیست. به عنوان مثال، به نظر می رسد انتخابی مهمتر از صداقت است. توضیحی که فقط یک یا دو علت احتمالی را انتخاب می کند، به ندرت کل فهرست علل مرتبط را پوشش می دهد. انتخاب بخشی از حقیقت را حذف می کند. این درست نیست که مثلاً فقط یک یا دو عامل باعث سقوط بورس شده است، اما حقیقت این است که میلیون ها علت وجود دارد که میلیون ها نفر را تحت تأثیر قرار می دهد تا به گونه ای عمل کنند که در نهایت باعث سقوط شود..
- معنى آن برای یادگیری ماشینی قابل تفسیر : توضیح باید رویداد را تا حد امکان صادقانه پیش بینی کند، که در یادگیری ماشین گاهی اوقات به آن می گویند.وفاداری . بنابراین اگر بگوییم که بالکن دوم قیمت یک خانه را افزایش می دهد، باید برای خانه های دیگر (یا حداقل برای خانه های مشابه) نیز صدق کند. برای انسان ها، وفاداری یک توضیح به اندازه گزینش، تضاد و جنبه اجتماعی آن مهم نیست.
- توضیحات خوب با باورهای قبلی توضیح دهنده مطابقت دارد . انسان ها تمایل دارند اطلاعاتی را نادیده بگیرند که با باورهای قبلی آنها همخوانی ندارد. این اثر سوگیری تایید نامیده می شود . (Nickerson 1998 13)
توضیحات از این نوع سوگیری در امان نیست. مردم تمایل دارند توضیحاتی را که با عقاید آنها همخوانی ندارد بی ارزش کنند یا نادیده بگیرند. مجموعه باورها از فردی به فرد دیگر متفاوت است، اما باورهای قبلی مبنی بر گروه مانند جهان بینی سیاسی نیز وجود دارد.

معنی آن برای یادگیری ماشینی 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 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 ویژگی را دارد.

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 غیرطبیعی توضیحات
خوبی ارائه می‌دهند، تناقض دارد. همانطور که می‌بینم، علل غیرطبیعی بر علل عمومی غلبه می‌کنند. علل غیر
طبیعی 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
بسیار کلی و خوب است. توضیح اینکه چرا خانه‌ها گران یا ارزان هستند.

معنی آن برای یادگیری ماشینی قابل تفسیر چیست: کلیت را 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 داده ها

در سراسر کتاب، تمام مدل‌ها و تکنیک‌ها بر روی مجموعه داده‌های واقعی که به طور رایگان به صورت 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 (بازگشت)

این مجموعه داده 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 مهربانی توسط Capital-Bikeshare آشکارا در دسترس قرار گرفت ۱۴ (2013) Fanaee-T and Gama. 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 مجموعه داده اضافه شد و همه ویژگی‌های اصلی برای مثال‌های این کتاب استفاده نشده است. در اینجا لیستی از ویژگی‌هایی است که استفاده شده است:

تعداد دوچرخه‌ها شامل کاربران عادی و ثبت نام شده. شمارش به عنوان هدف در کار رگرسیون استفاده می‌شود.

فصل، یا بهار، تابستان، پاییز یا زمستان. نشان می‌دهد که آیا روز تعطیل بود یا نه.

سال، ۲۰۱۱ یا ۲۰۱۲.

تعداد روزهای پس از ۱۱,۰۱,۰۰ (اولین روز در مجموعه داده). 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: صاف، کمی ابر، نیمه ابری، ابری

مه + ابر، مه + ابرهای شکسته، مه + چند ابر، مه

- برف خفیف، باران خفیف + رعد و برق + ابرهای پراکنده، باران خفیف + ابرهای پراکنده ۱۱۶۴
- باران شدید + پالت های یخ + رعد و برق + مه، برف + غبار ۱۱۶۵
- دما بر حسب درجه سانتیگراد ۱۱۶۶
- رطوبت نسبی بر حسب درصد (۰ تا ۱۰۰) ۱۱۶۷
- سرعت باد بر حسب کیلومتر در ساعت ۱۱۶۸
- برای مثال های این کتاب، داده ها کمی پردازش شده است. میتوانید R اسکریپت پردازشی را در مخزن ۱۱۶۹
- کتاب به همراه فایل RData GitHub ۱۱۷۰
- نهایی پیدا کنید.
- ۴،۲ نظرات هرزنامه YouTube طبقه بندی متن (۱۱۷۱
- به عنوان نمونه ای برای طبقه بندی متن، ما با ۱۹۵۶ نظر از ۵ ویدیوی مختلف YouTube کار می کنیم. ۱۱۷۲
- خوبیختانه، نویسنده‌گانی که از این مجموعه داده در مقاله ای در مورد طبقه بندی هرزنامه استفاده کردند، داده ۱۱۷۳
- ها را به صورت رایگان در دسترس قرار دادند. (Alberto, Lochter, and Almeida (2015) 15)
- نظرات از طریق API YouTube از پنج ویدیو از ده ویدیوی پربازدید YouTube در نیمه اول سال ۲۰۱۵ ۱۱۷۵
- جمع آوری شد. هر ۵ ویدیو موزیک هستند. یکی از آنها «گانگنام استایل» اثر هنرمند کره ای Psy است. ۱۱۷۶
- هنرمندان دیگر کیتی پری، LMFAO، امینم و شکیرا بودند. ۱۱۷۷
- برخی از نظرات را بررسی کنید. نظرات به صورت دستی به عنوان هرزنامه یا قانونی برچسب گذاری شدند. ۱۱۷۸
- هرزنامه با "۱" و نظرات قانونی با "۰" کدگذاری شد. ۱۱۷۹
- جدول ۴،۱: نمونه نظرات از مجموعه داده های هرزنامه YouTube ۱۱۸۰

کلاس محتوا

1 خب، به هر حال این کanal یوتیوب را ببینید : kobyoshi02

سلام بچه ها کanal جدید من را ببینید و اولین ویدیوی ما این ما میمون ها هستیم !!! من میمون با 1 پیراهن سفید هستم، لطفا نظر خود را لايك کنید و لطفا سابسکرایب کنید!!!

1 فقط برای تست باید بگم murdev.com

1 من با تکان دادن الاغ سکسی خود در کanal لذت ببرید ^_^

کلاس	محتوا
1	watch?v=vtaRGgvGtWQ این را بررسی کنید.
1	سلام، وب سایت جدید من را بررسی کنید !! این سایت در مورد چیزهای کودکانه است . kidsmediausa . com
1	در کanal من عضو شوید
0	به محض اینکه روشن شدم، آن را خاموش کردم، فقط می‌خواستم نماها را بررسی کنم...
1	شما باید کanal من را برای ویدیوهای خنده دار بررسی کنید!
1	و شما باید کanal من را بررسی کنید و به من بگویید که در مرحله بعد باید چه کار کنم!

۱۱۸۱

۱۱۸۲ همچنین می‌توانید به یوتیوب بروید و به بخش نظرات نگاهی بیندازید. اما لطفاً در جهنم یوتیوب گرفتار نشوید
 ۱۱۸۳ و در نهایت به تماشای ویدیوهایی از میمون‌ها که در حال دزدیدن و نوشیدن کوکتل از گردشگران در ساحل
 ۱۱۸۴ هستند، بنشینید. آشکارساز هرزنامه گوگل نیز احتمالاً از سال ۲۰۱۵ تغییرات زیادی کرده است.

۱۱۸۵ ویدیوی رکورددشکنی "Gangnam Style" را در اینجا تماشا کنید.

۱۱۸۶ اگر می‌خواهید با داده‌ها بازی کنید، می‌توانید فایل RData را به همراه اسکریپت R با برخی عملکردهای راحت
 ۱۱۸۷ در مخزن GitHub کتاب پیدا کنید.

۱۱۸۸ **۴.۳ عوامل خطر برای سرطان دهانه رحم (طبقه بندی)**
 ۱۱۸۹ مجموعه داده‌های سرطان دهانه رحم شامل شاخص‌ها و عوامل خطر برای پیش‌بینی اینکه آیا یک زن به
 ۱۱۹۰ سرطان دهانه رحم مبتلا می‌شود یا خیر. این ویژگی‌ها شامل داده‌های جمعیت شناختی (مانند سن)، سبک
 ۱۱۹۱ زندگی و سابقه پزشکی است. داده‌ها را می‌توان از مخزن یادگیری ماشین UCI دانلود کرد و توسط
 ۱۱۹۲ Fernandes, Cardoso, and Fernandes (2017) توضیح داده شده است.

۱۱۹۳ زیرمجموعه ویژگی‌های داده استفاده شده در نمونه‌های کتاب عبارتند از:

۱۱۹۴ سن بر حسب سال

۱۱۹۵ تعداد شرکای جنسی

۱۱۹۶	اولین رابطه جنسی (سن در سال)
۱۱۹۷	تعداد حاملگی ها
۱۱۹۸	سیگار کشیدن بله یا خیر
۱۱۹۹	سیگار کشیدن (در سال)
۱۲۰۰	داروهای ضد بارداری هورمونی بله یا خیر
۱۲۰۱	داروهای ضد بارداری هورمونی (در سال)
۱۲۰۲	دستگاه داخل رحمی بله یا خیر (IUD)
۱۲۰۳	تعداد سالهای استفاده از دستگاه داخل رحمی (IUD)
۱۲۰۴	آیا بیمار تا به حال بیماری مقاربی (STD) داشته است بله یا خیر
۱۲۰۵	تعداد تشخیص های STD
۱۲۰۶	زمان از اولین تشخیص STD
۱۲۰۷	زمان از آخرین تشخیص STD
۱۲۰۸	نتیجه بیوپسی "سالم" یا "سرطان" است. نتیجه هدف.
۱۲۰۹	بیوپسی به عنوان استاندارد طلایی برای تشخیص سرطان دهانه رحم عمل می کند. برای مثال های این کتاب،
۱۲۱۰	نتیجه بیوپسی به عنوان هدف مورد استفاده قرار گرفت. مقادیر گمشده برای هر ستون با حالت (متداول ترین
۱۲۱۱	مقدار) نسبت داده می شود، که احتمالاً راه حل بدی است، زیرا پاسخ واقعی می تواند با احتمال گم شدن یک
۱۲۱۲	مقدار مرتبط باشد. احتمالاً سوگیری وجود دارد زیرا سؤالات ماهیت بسیار خصوصی دارند. اما این کتابی در مورد
۱۲۱۳	انتساب داده های از دست رفته نیست، بنابراین انتساب حالت باید برای مثال ها کافی باشد.
۱۲۱۴	برای بازتولید نمونه های این کتاب با این مجموعه داده، پیش پردازش-R اسکریپت و فایل RData نهایی را در
۱۲۱۵	مخزن GitHub کتاب پیدا کنید.
۱۲۱۶	

فصل ۵ مدل های قابل تفسیر

- ۱۲۱۷
- ۱۲۱۸
- ۱۲۱۹ ساده ترین راه برای دستیابی به تفسیرپذیری استفاده از زیر مجموعه ای از الگوریتم هایی است که مدل های قابل تفسیر را ایجاد می کنند. رگرسیون خطی، رگرسیون لجستیک و درخت تصمیم معمولاً از مدل های قابل تفسیر استفاده می شوند.
- ۱۲۲۰
- ۱۲۲۱
- ۱۲۲۲ در فصل های بعدی در مورد این مدل ها 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 مدل ها
- ۱۲۲۳
- ۱۲۲۴ تمرکز خواهیم کرد. این کتاب رگرسیون خطی، رگرسیون لجستیک، دیگر پسوندهای رگرسیون خطی،
- ۱۲۲۵ درختان تصمیم، قوانین تصمیم گیری و الگوریتم RuleFit را با جزئیات بیشتری مورد بحث قرار می دهد.
- ۱۲۲۶ همچنین سایر مدل های قابل تفسیر را فهرست می کند.
- ۱۲۲۷ تمام مدل های تفسیر پذیر توضیح داده شده در این کتاب در سطح مدولار قابل تفسیر هستند، به استثنای
- ۱۲۲۸ روش k-nearest همسایه. جدول زیر یک نمای کلی از انواع مدل های قابل تفسیر و ویژگی های آنها ارائه می
- ۱۲۲۹ دهد. یک مدل خطی است اگر ارتباط بین ویژگی ها و هدف به صورت خطی مدل شود. یک مدل با
- ۱۲۳۰ محدودیت های یکنواختی تصمین می کند که رابطه بین یک ویژگی و نتیجه هدف همیشه در یک جهت در کل
- ۱۲۳۱ محدوده ویژگی پیش می رود: افزایش در مقدار ویژگی یا همیشه منجر به افزایش یا همیشه به کاهش هدف
- ۱۲۳۲ می شود. نتیجه یکنواختی برای تفسیر یک مدل مفید است زیرا درک یک رابطه را آسان تر می کند. برخی از
- ۱۲۳۳ مدل ها می توانند به طور خودکار تعامل بین ویژگی ها را برای پیش بینی نتیجه هدف داشته باشند. می توانید با
- ۱۲۳۴ ایجاد دستی ویژگی های تعامل، تعاملات را در هر مدلی بگنجانید. فعل و انفعالات می توانند عملکرد پیش بینی
- ۱۲۳۵ را بهبود بخشند، اما تعاملات زیاد یا بسیار پیچیده می تواند به تفسیرپذیری آسیب برساند. برخی از مدل ها فقط
- ۱۲۳۶ رگرسیون، برخی فقط طبقه بندی و برخی دیگر هر دو را مدیریت می کنند.
- ۱۲۳۷ از این جدول، می توانید یک مدل قابل تفسیر مناسب برای کار خود انتخاب کنید، رگرسیون (regr) یا طبقه
- ۱۲۳۸ بندی (کلاس):

وظیفه	اثر متقابل	یکنواخت	خطی	الگوریتم
regr	خیر	آره	آره	رگرسیون خطی
کلاس	خیر	آره	خیر	رگرسیون لجستیک

وظیفه	اثر متقابل	یکنواخت	خطی	الگوریتم
کلاس، رگر	آره	مقداری	خیر	درختان تصمیم
کلاس، رگر	آره	خیر	آره	RuleFit
کلاس	خیر	آره	خیر	بیز ساده لوح
کلاس، رگر	خیر	خیر	خیر	k-آنزدیک ترین همسایگان

۱۲۳۹

- ۱۲۴۰ شما می توانید استدلال کنید که هم رگرسیون لجستیک و هم ساده لوحانه، توضیحات خطی را مجاز می دانند.
 ۱۲۴۱ با این حال، این فقط برای لگاریتم هدف صادق است: افزایش یک ویژگی به اندازه یک نقطه، لگاریتم احتمال
 ۱۲۴۲ هدف را به میزان معینی افزایش می دهد، با فرض ثابت ماندن همه ویژگی های دیگر.

۱۲۴۳

۵.۱ رگرسیون خطی

- ۱۲۴۴ یک مدل رگرسیون خطی هدف را به عنوان مجموع وزنی ورودی های ویژگی پیش بینی می کند. خطی بودن
 ۱۲۴۵ رابطه آموخته شده، تفسیر را آسان می کند. مدل های رگرسیون خطی مدت هاست که توسط آماردانان،
 ۱۲۴۶ دانشمندان کامپیوتر و سایر افرادی که به مشکلات کمی رسیدگی می کنند استفاده می شود.

- ۱۲۴۷ مدل های خطی می توانند برای مدل سازی وابستگی یک هدف رگرسیونی y به برخی از ویژگی های X استفاده
 ۱۲۴۸ شوند. روابط آموخته شده خطی هستند و می توان آنها را برای یک نمونه A به صورت زیر نوشت:

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

- ۱۲۵۰ نتیجه پیش بینی شده یک نمونه، مجموع وزنی از ویژگی های p آن است. بتاهای (β) وزن یا ضرایب ویژگی های
 ۱۲۵۱ آموخته شده را نشان می دهد. وزن اول در مجموع (β_0) قطع نامیده می شود و با ویژگی ضرب نمی شود. ایسیلوون
 ۱۲۵۲ (ϵ) خطایی است که ما هنوز مرتکب می شویم، یعنی تفاوت بین پیش بینی و نتیجه واقعی. فرض بر این است
 ۱۲۵۳ که این خطاهای از یک توزیع گاوی پیروی می کنند، به این معنی که ما در هر دو جهت منفی و مثبت خطا می
 ۱۲۵۴ کنیم و بسیاری از خطاهای کوچک و تعداد کمی از خطاهای بزرگ را انجام می دهیم.

- ۱۲۵۵ برای تخمین وزن بهینه می توان از روش های مختلفی استفاده کرد. روش حداقل مربعات معمولی معمولاً برای
 ۱۲۵۶ یافتن وزن هایی استفاده می شود که اختلاف مجدد بین نتایج واقعی و برآورد شده را به حداقل می رساند:

$$\hat{\beta} = \operatorname{argmin}_{\beta_0, \dots, \beta_p} \sum_{i=1}^n (y(i) - (\beta_0 + \beta_j x(i)))^2$$

۱۲۵۸ ما در مورد چگونگی یافتن وزن‌های بهینه به تفصیل بحث نخواهیم کرد، اما اگر علاقه مند هستید، می‌توانید
۱۲۵۹ فصل ۳,۲ کتاب "عناصر یادگیری آماری" (فریدمن، هستی و تبیه‌رانی ۲۰۰۹) ۱۷ یا یکی دیگر از کتاب‌های
۱۲۶۰ آنلاین را مطالعه کنید. منابع در مورد مدل‌های رگرسیون خطی

۱۲۶۱ بزرگترین مزیت مدل‌های رگرسیون خطی خطی بودن است: این روش تخمین را ساده می‌کند و مهمتر از
۱۲۶۲ همه، این معادلات خطی تفسیر آسانی در سطح مدولار (یعنی وزن‌ها) دارند. این یکی از دلایل اصلی گسترش
۱۲۶۳ مدل خطی و همه مدل‌های مشابه در زمینه‌های دانشگاهی مانند پزشکی، جامعه‌شناسی، روان‌شناسی و بسیاری
۱۲۶۴ دیگر از زمینه‌های تحقیقاتی کمی است. به عنوان مثال، در زمینه پزشکی، نه تنها پیش‌بینی نتیجه بالینی یک
۱۲۶۵ بیمار مهم است، بلکه تعیین کمیت تأثیر دارو و در عین حال در نظر گرفتن جنسیت، سن و سایر ویژگی‌ها به
۱۲۶۶ روشی قابل تفسیر است..

۱۲۶۷ وزن‌های تخمینی با فواصل اطمینان همراه است. فاصله اطمینان محدوده‌ای برای تخمین وزن است که وزن
۱۲۶۸ «واقعی» را با اطمینان خاصی پوشش می‌دهد. به عنوان مثال، فاصله اطمینان ۹۵٪ برای وزن ۲ می‌تواند از ۱ تا
۱۲۶۹ ۳ متغیر باشد. تفسیر این فاصله به این صورت خواهد بود: اگر تخمین را ۱۰۰ بار با داده‌های نمونه گیری جدید
۱۲۷۰ تکرار کنیم، فاصله اطمینان شامل وزن واقعی در سال ۹۵ می‌شود. از ۱۰۰ مورد، با توجه به اینکه مدل
۱۲۷۱ رگرسیون خطی مدل صحیح داده‌ها است.

۱۲۷۲ اینکه آیا مدل، مدل «درست» است بستگی به این دارد که آیا روابط موجود در داده‌ها مفروضات خاصی را
برآورده می‌کنند که عبارتند از خطی بودن، نرمال بودن، همسانی، استقلال، ویژگی‌های ثابت و عدم وجود چند
۱۲۷۳ خطی.

۱۲۷۵ خطی بودن

۱۲۷۶ مدل رگرسیون خطی، پیش‌بینی را مجبور می‌کند که ترکیبی خطی از ویژگی‌ها باشد، که هم بزرگترین قدرت و
۱۲۷۷ هم بزرگترین محدودیت آن است. خطی بودن منجر به مدل‌های قابل تفسیر می‌شود. کمیت کردن و توصیف
۱۲۷۸ اثرات خطی آسان است. آنها افزودنی هستند، بنابراین به راحتی می‌توان اثرات را از هم جدا کرد. اگر به فعل و
۱۲۷۹ انفعالات ویژگی یا ارتباط غیرخطی یک ویژگی با مقدار هدف مشکوک هستید، می‌توانید اصطلاحات تعامل را
۱۲۸۰ اضافه کنید یا از خطوط رگرسیون استفاده کنید.

۱۲۸۱ نرمال بودن

۱۲۸۲ فرض بر این است که نتیجه هدف با توجه به ویژگی‌ها از توزیع نرمال پیروی می‌کند. اگر این فرض نقض شود،
۱۲۸۳ فواصل اطمینان تخمینی وزن ویژگی‌ها نامعتبر است.

همسانی (واریانس ثابت)

۱۲۸۴ واریانس عبارات خطای کل فضای ویژگی ثابت فرض می شود. فرض کنید می خواهید ارزش یک خانه را با
۱۲۸۵ توجه به مساحت نشیمن بر حسب متر مربع پیش بینی کنید. شما یک مدل خطی را تخمین می زنید که فرض
۱۲۸۶ می کند، صرف نظر از اندازه خانه، خطای اطراف پاسخ پیش بینی شده واریانس یکسانی دارد. این فرض اغلب
۱۲۸۷ در واقعیت نقض می شود. در مثال خانه، قابل قبول است که واریانس شرایط خطای اطراف قیمت
۱۲۸۸ پیش بینی شده برای خانه های بزرگ تر بیشتر باشد، زیرا قیمت ها بالاتر هستند و فضای بیشتری برای نوسانات
۱۲۸۹ قیمت وجود دارد. فرض کنید میانگین خطای تفاوت بین قیمت پیش بینی شده و واقعی در مدل رگرسیون
۱۲۹۰ خطی شما ۵۰۰۰۰ یورو باشد. اگر همجنسگرا بودن را فرض کنید، فرض می کنید که میانگین خطای
۱۲۹۱ ۵۰۰۰۰ برای خانه هایی که ۱ میلیون قیمت دارند و برای خانه هایی که فقط ۴۰۰۰۰ قیمت دارند یکسان است.
۱۲۹۲

استقلال

۱۲۹۳ فرض بر این است که هر نمونه مستقل از هر نمونه دیگری است. اگر اندازه گیری های مکرر را انجام دهید، مانند
۱۲۹۴ آزمایش های خون متعدد برای هر بیمار، نقاط داده مستقل نیستند. برای داده های وابسته به مدل های رگرسیون
۱۲۹۵ خطی خاص، مانند مدل های اثر مختلط یا GEE نیاز دارید. اگر از مدل رگرسیون خطی "عادی" استفاده می
۱۲۹۶ کنید، ممکن است نتیجه گیری اشتباهی از مدل بگیرید.
۱۲۹۷

ویژگی های ثابت

۱۲۹۸ ویژگی های ورودی "ثابت" در نظر گرفته می شوند. ثابت به این معنی است که آنها به عنوان "ثابت داده شده"
۱۲۹۹ و نه به عنوان متغیرهای آماری در نظر گرفته می شوند. این بدان معناست که آنها فاقد خطاهای اندازه گیری
۱۳۰۰ هستند. این یک فرض نسبتاً غیر واقعی است. با این حال، بدون این فرض، شما باید مدل های خطای اندازه گیری
۱۳۰۱ بسیار پیچیده ای را که خطاهای اندازه گیری ویژگی های ورودی شما را محاسبه می کنند، تطبیق دهید. و معمولاً
۱۳۰۲ شما نمی خواهید این کار را انجام دهید.
۱۳۰۳

فقدان چند خطی

۱۳۰۴ شما ویژگی های قوی همبسته را نمی خواهید، زیرا این تخمین وزن ها را به هم می زند. در شرایطی که دو
۱۳۰۵ ویژگی به شدت همبستگی دارند، تخمین وزن ها مشکل ساز می شود، زیرا اثرات ویژگی افزایشی هستند و
۱۳۰۶ غیرقابل تعیین می شود که به کدام یک از ویژگی های همبسته نسبت داده شود.
۱۳۰۷

۱۳۰۸ تفسیر ۵,۱,۱

۱۳۰۹ تفسیر وزن در مدل رگرسیون خطی به نوع ویژگی مربوطه بستگی دارد.

۱۳۱۰
۱۳۱۱ ویژگی عددی: افزایش ویژگی عددی به اندازه یک واحد، نتیجه تخمینی را با وزن آن تغییر می‌دهد. یک مثال از
۱۳۱۲ یک ویژگی عددی اندازه یک خانه است.

۱۳۱۳ 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 ۰ کدگذاری
شده اند) به حساب می‌آید، مانند "بدون باغ". تغییر ویژگی از دسته مرجع به دسته دیگر، نتیجه تخمینی را بر
اساس وزن ویژگی تغییر می‌دهد.

۱۳۱۷ ویژگی طبقه‌بندی با دسته‌های متعدد: ویژگی با تعداد ثابتی از مقادیر ممکن. به عنوان مثال ویژگی «نوع کف» با
۱۳۱۸ دسته‌های احتمالی «فرش»، «لمینت» و «پارکت» است. یک راه حل برای مقابله با بسیاری از دسته‌ها،
۱۳۱۹ رمزگذاری یک گرم است، به این معنی که هر دسته دارای ستون باینری خاص خود است. برای یک ویژگی
۱۳۲۰ طبقه‌بندی با دسته‌های L، شما فقط به ستون‌های 1-L نیاز دارید، زیرا ستون-L امین اطلاعات اضافی دارد (به
۱۳۲۱ عنوان مثال وقتی ستون‌های 1 تا L-1 همه دارای مقدار ۰ برای یک مثال هستند، می‌دانیم که ویژگی
۱۳۲۲ طبقه‌بندی این نمونه در رد L قرار می‌گیرد. سپس تفسیر برای هر دسته مانند تفسیر ویژگی‌های باینری است.
۱۳۲۳ برخی از زبان‌ها، مانند R، به شما امکان می‌دهند تا ویژگی‌های دسته‌بندی را به روش‌های مختلف رمزگذاری
۱۳۲۴ کنید، همانطور که در ادامه این فصل توضیح داده شد.

۱۳۲۵ رهگیری: β وقفه وزن ویژگی برای "ویژگی ثابت" است که همیشه برای همه موارد ۱ است. اکثر بسته‌های
۱۳۲۶ نرم افزاری به طور خودکار این ویژگی "۱" را برای تخمین رهگیری اضافه می‌کنند. تفسیر این است: برای مثال
۱۳۲۷ با تمام مقادیر ویژگی‌های عددی در صفر و مقادیر ویژگی‌های طبقه‌بندی شده در دسته‌های مرجع، پیش
۱۳۲۸ بینی مدل وزن رهگیری است. تفسیر رهگیری معمولاً مرتبط نیست، زیرا نمونه‌هایی با مقادیر همه ویژگی‌ها در
۱۳۲۹ صفر اغلب معنی ندارند. تفسیر تنها زمانی معنادار است که ویژگی‌ها استاندارد شده باشند (میانگین صفر،
۱۳۳۰ انحراف معیار یک). سپس رهگیری نتیجه پیش‌بینی شده نمونه‌ای را منعکس می‌کند که در آن همه ویژگی‌ها در
۱۳۳۱ مقدار میانگین خود هستند.

۱۳۳۲ تفسیر ویژگی‌ها در مدل رگرسیون خطی را می‌توان با استفاده از الگوهای متنی زیر خودکار کرد.

۱۳۳۳ تفسیر یک ویژگی عددی

۱۳۳۴ افزایش ویژگی یک واحد پیش بینی ۷ را افزایش می دهد واحد زمانی که تمام مقادیر ویژگی های دیگر ثابت باقی می ماند.

۱۳۳۶ تفسیر یک ویژگی طبقه بندی شده

۱۳۳۷ در حال تغییر ویژگی از دسته دیگر، پیش بینی ۷ را افزایش می دهد زمانی که تمام ویژگی های دیگر ثابت می ماند.

۱۳۳۹ اندازه گیری مهم دیگر برای تفسیر مدل های خطی، اندازه گیری مربع R^2 است. به شما می گوید
۱۳۴۰ که چه مقدار از واریانس کل نتیجه هدف شما توسط مدل توضیح داده شده است. هرچه R^2 بالاتر
۱۳۴۱ باشد، مدل شما داده ها را بهتر توضیح می دهد. فرمول محاسبه R^2 به صورت زیر است:

$$R^2 = 1 - \frac{SSE}{SST}$$

۱۳۴۳ SSE مجدور عبارات خطأ است:

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

۱۳۴۵ SST مجموع مجدور واریانس داده است:

$$SST = n \sum_{i=1}^n (y(i) - \bar{y})^2$$

۱۳۴۷ SSE به شما می گوید که پس از برازش مدل خطی چقدر واریانس باقی می ماند، که با اختلاف مجدور بین
۱۳۴۸ مقادیر هدف پیش بینی شده و واقعی اندازه گیری می شود SST. واریانس کل نتیجه هدف است R^2 .
۱۳۴۹ به شما می گوید که چه مقدار از واریانس شما را می توان با مدل خطی توضیح داد R^2 . معمولاً بین
۱۳۵۰ برای مدل هایی که مدل اصلًا داده ها را توضیح نمی دهد و ۱ برای مدل هایی که تمام واریانس داده های شما را
۱۳۵۱ توضیح می دهند، متغیر است. همچنین ممکن است R^2 بدون نقض قوانین ریاضی یک مقدار منفی به
۱۳۵۲ خود بگیرد. این زمانی اتفاق می افتد که SSE بزرگ تر از SST باشد، به این معنی که یک مدل روند داده ها را
۱۳۵۳ نمی گیرد و بدتر از استفاده از میانگین هدف به عنوان پیش بینی با داده ها مطابقت دارد.

۱۳۵۴ یک نکته وجود دارد، زیرا R^2 با تعداد ویژگی های مدل افزایش می یابد، حتی اگر اصلًا حاوی اطلاعاتی
۱۳۵۵ در مورد مقدار هدف نباشند. بنابراین، بهتر است از R^2 تنظیم شده استفاده کنید که تعداد ویژگی
۱۳۵۶ های استفاده شده در مدل را به حساب می آورد. محاسبه آن این است:

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

۱۳۵۸ که در آن p تعداد ویژگی ها و n تعداد نمونه ها است.

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

۱۳۶۱ اهمیت ویژگی

۱۳۶۲

۱۳۶۳ اهمیت یک ویژگی در مدل رگرسیون خطی را می توان با قدر مطلق آماره t اندازه گیری کرد. آماره t وزن
۱۳۶۴ تخمین زده شده با خطای استاندارد آن است.

$$t^{\wedge \beta j} = {}^{\wedge \beta j}SE({}^{\wedge \beta j})$$

۱۳۶۵ اجازه دهید بررسی کنیم که این فرمول به ما چه می گوید: اهمیت یک ویژگی با افزایش وزن افزایش می یابد.
۱۳۶۶ این منطقی است. هر چه وزن تخمینی واریانس بیشتری داشته باشد ($=$ هر چه نسبت به مقدار صحیح اطمینان
۱۳۶۷ کمتری داشته باشیم)، اهمیت ویژگی کمتر است. این نیز منطقی است.

۱۳۶۸

۱۳۶۹

۱۳۷۰ در این مثال، ما از مدل رگرسیون خطی برای پیش‌بینی تعداد دوچرخه‌های اجاره‌ای در یک روز خاص، با توجه
۱۳۷۱ به اطلاعات آب و هوا و تقویم استفاده می‌کنیم. برای تفسیر، وزن‌های رگرسیون برآورده شده را بررسی می‌کنیم.
۱۳۷۲ ویژگی‌ها از ویژگی‌های عددی و طبقه‌ای تشکیل شده است. برای هر ویژگی، جدول وزن تخمینی، خطای
۱۳۷۳ استاندارد برآورد (SE) و قدر مطلق آماره ($|t|$) را نشان می‌دهد.

۱,۲ مثال ۵

	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	7.0	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

۱۳۷۴

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 ۳۷۹,۴ در مقایسه با هوای خوب کمتر است.

همه تفاسیر همیشه با این پاورقی 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 شبی (گرادیان) ابر صفحه را در هر جهت مشخص می‌کنند. جنبه خوب این است که افروزنی تفسیر یک اثر ویژگی فردی را از همه ویژگی‌های دیگر جدا می‌کند. این امکان پذیر است زیرا تمام جلوه‌های ویژگی (= وزن برابر مقدار ویژگی) در معادله با یک مثبت ترکیب می‌شوند. در جنبه بد، این تفسیر توزیع مشترک ویژگی‌ها را نادیده می‌گیرد. افزایش یک ویژگی، اما عدم تغییر ویژگی دیگر، می‌تواند به نقاط داده غیر واقعی یا حداقل بعید منجر شود. به عنوان مثال افزایش تعداد اتاق‌ها بدون افزایش اندازه خانه ممکن است غیرواقعی باشد.

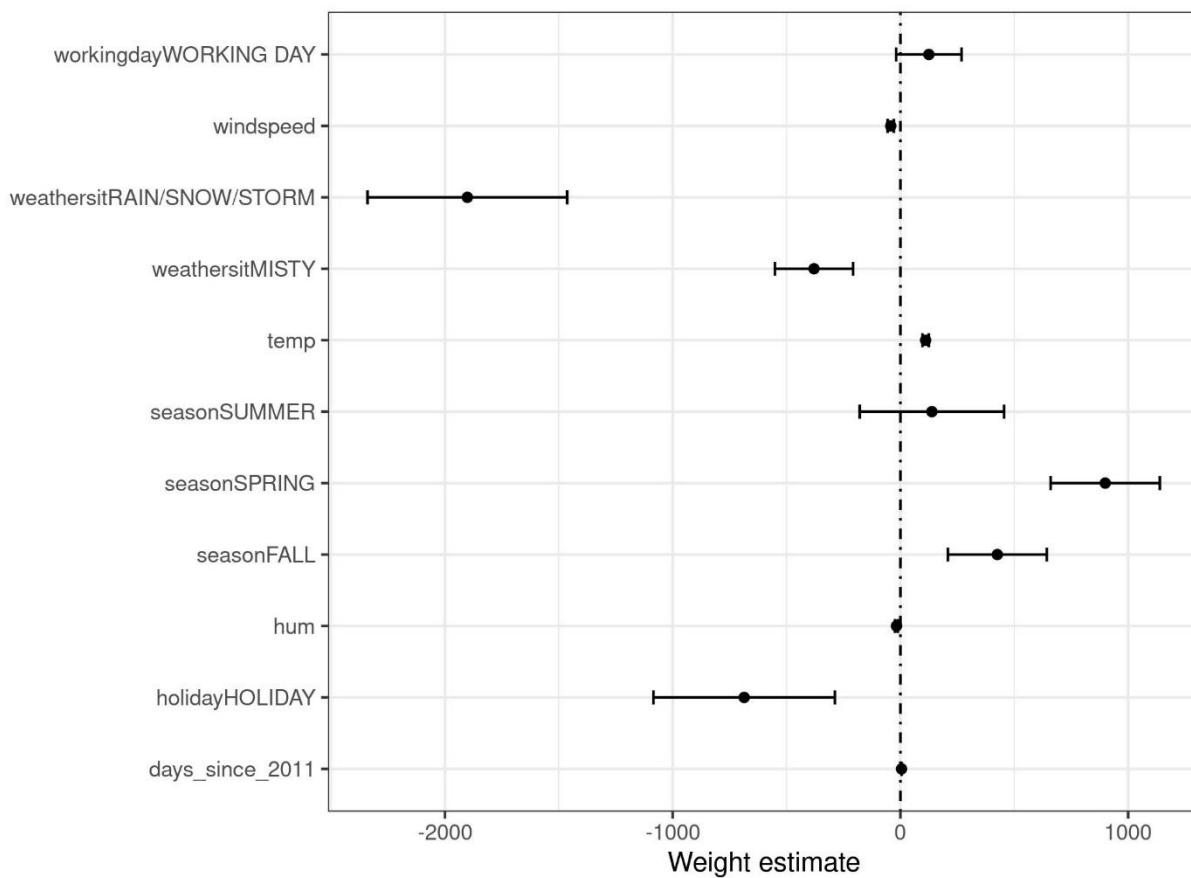
۱۳۸۹ ۵,۱,۳ تفسیر بصری

۱۳۹۰ تجسم های مختلف، مدل رگرسیون خطی را برای انسان آسان و سریع درک می کند.

۱۳۹۱ ۵,۱,۳,۱ نمودار وزن

۱۳۹۲ اطلاعات جدول وزنی (تخمین وزن و واریانس) را می توان در نمودار وزنی مشاهده کرد. نمودار زیر نتایج حاصل

۱۳۹۳ از مدل رگرسیون خطی قبلی را نشان می دهد.



۱۳۹۴ شکل ۵: وزن ها به صورت نقاط و فاصله های اطمینان ۹۵ درصد به صورت خطوط نمایش داده می شوند.

۱۳۹۵ نمودار وزن نشان می دهد که هوای بارانی / برفی / طوفانی تأثیر منفی قوی بر تعداد پیش بینی 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

ویژگی ها در مقیاس های مختلف اندازه گیری می شوند. در حالی که برای آب و هوا، وزن تخمینی تفاوت بین هوای خوب و بارانی/طوفانی/برفی را نشان می دهد، برای دما فقط افزایش ۱ درجه سانتی گراد را نشان می دهد. قبل از برازش مدل خطی، می توانید وزن های تخمینی را با مقیاس بندی ویژگی ها (میانگین صفر و انحراف استاندارد یک) قابل مقایسه تر کنید.

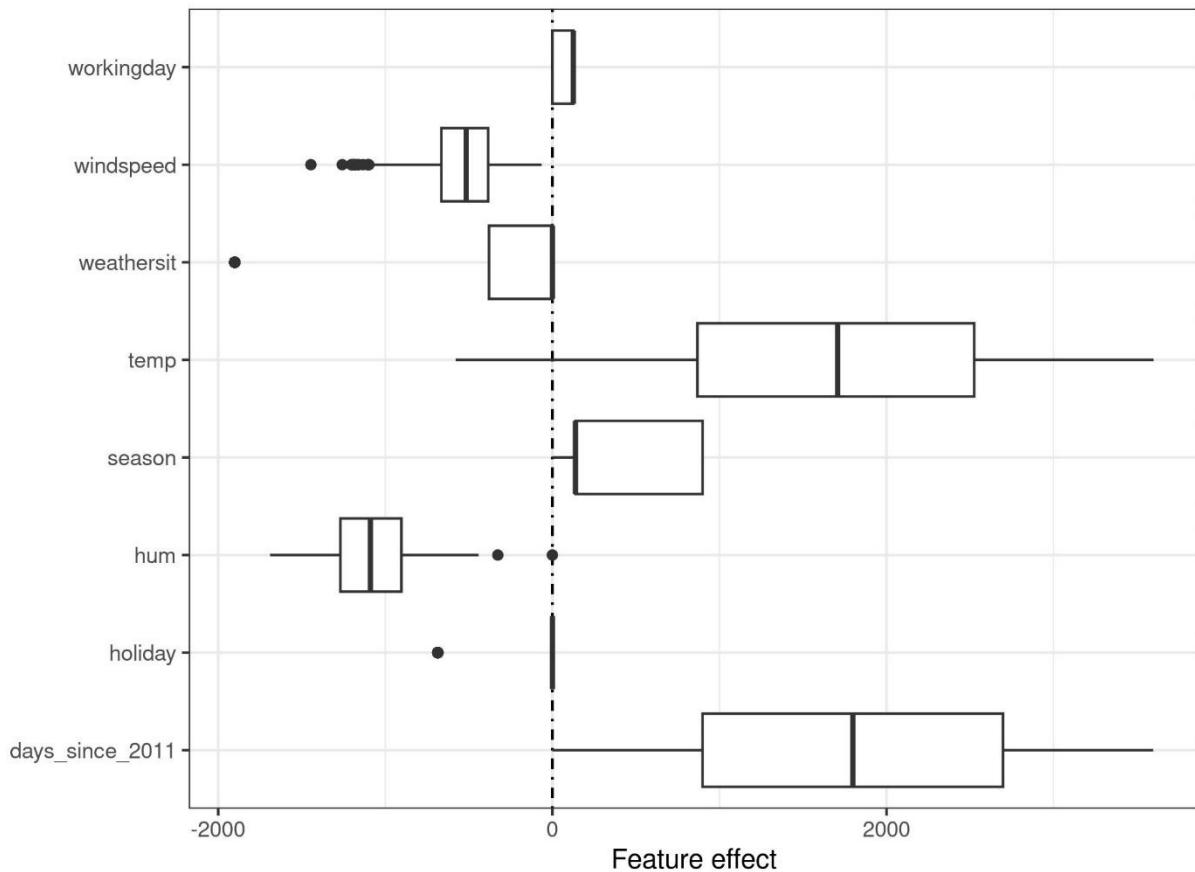
۱۴۰۵ ۲,۳,۵ نمودار اثر

وزن های مدل رگرسیون خطی زمانی که در مقادیر واقعی ضرب شوند می توانند به طور معنی داری تحلیل شوند. وزن ها به مقیاس ویژگی ها بستگی دارد و اگر ویژگی ای داشته باشد که مثلاً قد یک فرد را اندازه گیری می کند و از متر به سانتی متر تغییر می دهد، متفاوت خواهد بود. وزن تغییر خواهد کرد، اما اثرات واقعی در داده های شما تغییر نخواهد کرد. همچنین دانستن توزیع ویژگی خود در داده ها مهم است، زیرا اگر واریانس بسیار پایینی دارید، به این معنی است که تقریباً همه نمونه ها سهم مشابهی از این ویژگی دارند. نمودار اثر می تواند به شما کمک کند تا بفهمید که ترکیب وزن و ویژگی چقدر به پیش بینی های داده های شما کمک می کند. با محاسبه اثرات شروع کنید، که وزن هر ویژگی ضریب مقدار ویژگی یک نمونه است:

$$\text{effect}(i)j = w_j x(i)j$$

افکت ها را می توان با نمودارهای جعبه ای تجسم کرد. جعبه در یک باکس پلات شامل محدوده اثر برای نیمی از داده ها (٪۷۵ تا ٪۲۵ چندک اثر). خط عمودی در کادر، اثر میانه است، یعنی ۵۰ درصد از نمونه ها تأثیر کمتر و نیمی دیگر تأثیر بیشتری بر پیش بینی دارند. نقاط دورتر هستند، به عنوان نقاطی که بیش از $IQR * 1,5$ محدوده بین ربیعی، یعنی تفاوت بین ربع اول و سوم) بالای ربع سوم، یا کمتر از $IQR * 1,5$ زیر چارک اول تعريف می شوند. دو خط افقی که سبیل پایینی و بالایی نامیده می شوند، نقاط زیر چارک اول و بالای چارک سوم را که پرت نیستند به هم متصل می کنند. اگر نقاط پرت وجود نداشته باشد، سبیل ها به مقادیر حداقل و حداکثر گسترش می یابند.

اثرات طبقه بندی ویژگی را می توان در یک باکس پلات خلاصه کرد، در مقایسه با نمودار وزن، که در آن هر دسته ر دیف خاص خود را دارد.



شکل ۵،۲: نمودار اثر ویژگی توزیع اثرات (= ارزش ویژگی ضربدر وزن ویژگی) را در بین داده ها در هر ویژگی نشان می دهد.

بیشترین سهم در تعداد مورد انتظار دوچرخه های اجاره ای مربوط به ویژگی دما و ویژگی روز است که روند اجاره دوچرخه را در طول زمان نشان می دهد. دما دامنه وسیعی از میزان کمک به پیش بینی دارد. ویژگی روند روز از صفر به مشارکت های مثبت بزرگ می رسد، زیرا اولین روز در مجموعه داده (۰،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 می یابد و برای آخرین روز در مجموعه داده (۳۱،۱۲،۲۰۱۲) بالاترین میزان است. توجه داشته باشید که برای افکت هایی با وزن منفی، نمونه هایی با اثر مثبت آنها یی 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 پیش بینی های فردی را توضیح دهید

هر یک از ویژگی های یک نمونه چقدر در پیش بینی کمک کرده است؟ این را می توان با محاسبه اثرات برای این مثال پاسخ داد. تفسیر اثرات خاص نمونه فقط در مقایسه با توزیع اثر برای هر ویژگی منطقی است. ما می

خواهیم پیش بینی مدل خطی را برای نمونه ششم از مجموعه داده دوچرخه توضیح دهیم. نمونه دارای مقادیر
ویژگی زیر است.

۱۴۳۸

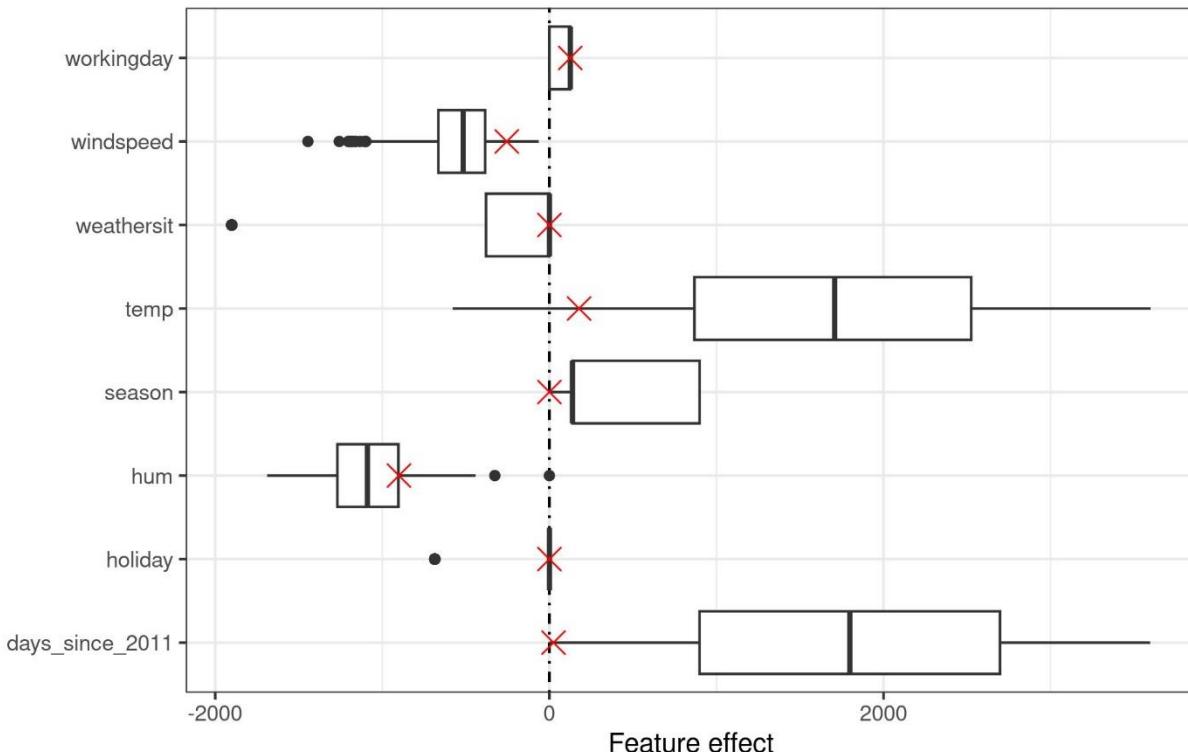
۱۴۳۹ جدول ۵،۱: مقادیر ویژگی برای نمونه ۶

Table 4.2: Feature values for instance 6

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 وزن های مربوطه از 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 افکتها را به صورت ضربدری به نمودار افکت اضافه می کنیم، که توزیع افکتها را در دادهها به ما نشان می دهد. این به ما اجازه می دهد تا اثرات فردی را با توزیع اثرات در داده ها مقایسه کنیم.

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



شکل ۳,۵: نمودار اثر برای یک نمونه توزیع اثر را نشان می دهد و اثرات نمونه مورد علاقه را برجسته می کند.

اگر پیش‌بینی‌های نمونه‌های آموزشی را میانگین کنیم، میانگین ۴۵۰۴ به دست می‌آید. در مقایسه، پیش‌بینی نمونه ششم کوچک است، زیرا تنها ۱۵۷۱ کرایه دوچرخه پیش‌بینی شده است. طرح اثر دلیل آن را نشان می‌دهد. نمودارهای جعبه توزیع اثرات را برای همه نمونه‌های مجموعه داده نشان می‌دهند، تلاقي‌ها 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.

راه‌های مختلفی برای رمزگذاری یک ویژگی طبقه بندی وجود دارد و انتخاب بر تفسیر وزن‌ها تأثیر می‌گذارد. استاندارد در مدل‌های رگرسیون خطی، کدگذاری درمان است که در اکثر موارد کافی است. استفاده از رمزگذاری‌های مختلف منجر به ایجاد ماتریس‌های مختلف (طراحی) از یک ستون واحد با ویژگی طبقه بندی

۵,۱,۵ رمزگذاری ویژگی‌های دسته بندی

۱۴۵۹ می شود. این بخش سه کدگذاری مختلف را ارائه می کند، اما تعداد بیشتری وجود دارد. مثال مورد استفاده
۱۴۶۰ دارای شش نمونه و یک ویژگی طبقه بندی شده با سه دسته است. برای دو مورد اول، ویژگی دسته A می
۱۴۶۱ گیرد. برای مثال سه و چهار، دسته B؛ و برای دو مورد آخر، دسته C.

۱۴۶۲ کدگذاری درمان

۱۴۶۳ در کدگذاری درمان، وزن هر دسته، تفاوت برآورده شده در پیش‌بینی بین دسته مربوطه و دسته مرجع است.
۱۴۶۴ فاصله مدل خطی میانگین دسته مرجع است (زمانی که سایر ویژگی ها ثابت می مانند). ستون اول ماتریس
۱۴۶۵ طراحی، وقفه است که همیشه ۱ است. ستون دو نشان می دهد که آیا نمونه A در رده B قرار دارد یا خیر، ستون
۱۴۶۶ سه نشان می دهد که آیا در رده C قرار دارد یا خیر. برای دسته A نیازی به ستون نیست، زیرا پس از آن معادله
۱۴۶۷ خطی بیش از حد مشخص می شود و هیچ راه حل منحصر به فردی برای وزن ها نمی توان یافت. کافی است
۱۴۶۸ بدانیم که یک نمونه در رده B یا C نیست.

۱۴۶۹ ماتریس ویژگی:

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

۱۴۷۰ کدنویسی افکت

۱۴۷۲ وزن هر دسته، تفاوت تخمینی ۷ از دسته مربوطه به میانگین کلی است (با توجه به اینکه همه ویژگی های دیگر
۱۴۷۳ صفر هستند یا دسته مرجع). ستون اول برای تخمین فاصله استفاده می شود. وزن β مرتبط با رهگیری نشان
۱۴۷۴ دهنده میانگین کلی و β ، وزن ستون دو، تفاوت بین میانگین کلی و دسته B است. اثر کل دسته B است β
۱۴۷۵ تفسیر دسته C معادل است. برای رده مرجع A، تفاوت به میانگین کلی است و β اثر کلی

۱۴۷۶ ماتریس ویژگی:

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

۱۴۷۷ کدنویسی ساختگی

این β در هر دسته، میانگین تخمینی مقدار ۷ برای هر دسته است (با توجه به اینکه تمام مقادیر ویژگی های 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 وزن های مدل خطی پیدا کرد. راه دیگر برای کاهش این مشکل چند خطی، کنار گذاشتن یکی از دسته بندی ها است.

۱۴۸۳ ماتریس ویژگی:

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

۱۴۸۴ اگر می خواهید کمی عمیق‌تر در رمزگذاری‌های مختلف ویژگی‌های طبقه‌بندی شده غوطه‌ور شویم، این صفحه ۱۴۸۵ وب و این پست وبلاگ را بررسی کنید.

۱۴۸۶ ۵.۱ آیا مدل های خطی توضیحات خوبی ایجاد می کنند؟

۱۴۸۷ با قضاوت بر اساس ویژگی‌هایی که توضیح خوبی را تشکیل می دهند، همانطور که در فصل توضیحات انسان ۱۴۸۸ دوستانه ارائه شده است، مدل های خطی بهترین توضیحات را ایجاد نمی کنند. آنها متضاد هستند، اما نمونه ۱۴۸۹ مرجع یک نقطه داده است که در آن همه ویژگی‌های عددی صفر هستند و ویژگی‌های طبقه‌بندی در دسته ۱۴۹۰ های مرجع خود قرار دارند. این معمولاً یک نمونه مصنوعی و بی معنی است که بعید است در داده‌ها یا واقعیت ۱۴۹۱ شما رخ دهد. یک استثنای وجود دارد: اگر همه ویژگی‌های عددی در مرکز میانگین باشند (ویژگی منهای میانگین ۱۴۹۲ ویژگی) و همه ویژگی‌های طبقه‌بندی با افکت کدگذاری شده باشند، نمونه مرجع نقطه داده‌ای است که در آن ۱۴۹۳ همه ویژگی‌ها مقدار میانگین ویژگی را می‌گیرند. این نیز ممکن است یک نقطه داده وجود نداشته باشد، اما ۱۴۹۴ حداقل ممکن است محتمل تر یا معنادارتر باشد. در این مورد، وزن‌ها ضربدر مقادیر ویژگی (اثرات ویژگی) سهم ۱۴۹۵ را در نتیجه پیش‌بینی شده در مقابل «میانگین نمونه» توضیح می‌دهند. یکی دیگر از جنبه‌های توضیح خوب، ۱۴۹۶ گزینش پذیری است که در مدل های خطی با استفاده از ویژگی‌های کمتر یا با آموزش مدل های خطی ۱۴۹۷ پراکنده می توان به آن دست یافت. اما به طور پیش فرض، مدل های خطی توضیحات انتخابی ایجاد نمی کنند. ۱۴۹۸ مدل های خطی توضیحات درستی ایجاد می کنند، تا زمانی که معادله خطی مدل مناسبی برای رابطه بین ۱۴۹۹ ویژگی‌ها و نتیجه باشد. هر چه غیر خطی‌ها و تعاملات بیشتر باشد، دقیق‌تر مدل خطی کمتر خواهد بود و ۱۵۰۰ توضیحات کمتر صادق می شود. خطی بودن توضیحات را کلی تر و ساده تر می کند. من معتقدم ماهیت خطی ۱۵۰۱ مدل عامل اصلی استفاده از مدل های خطی برای توضیح روابط است. مدل های خطی توضیحات انتخابی ایجاد ۱۵۰۲ نمی کنند. مدل های خطی توضیحات درستی ایجاد می کنند، تا زمانی که معادله خطی مدل مناسبی برای ۱۵۰۳

رابطه بین ویژگی ها و نتیجه باشد. هر چه غیر خطی ها و تعاملات بیشتر باشد، دقت مدل خطی کمتر خواهد
 بود و توضیحات کمتر صادق می شود. خطی بودن توضیحات را کلی تر و ساده تر می کند. من معتقدم ماهیت
 خطی مدل عامل اصلی استفاده از مدل های خطی برای توضیح روابط است. مدل های خطی توضیحات انتخابی
 ایجاد نمی کنند. مدل های خطی توضیحات درستی ایجاد می کنند، تا زمانی که معادله خطی مدل مناسبی
 برای رابطه بین ویژگی ها و نتیجه باشد. هر چه غیر خطی ها و تعاملات بیشتر باشد، دقت مدل خطی کمتر
 خواهد بود و توضیحات کمتر صادق می شود. خطی بودن توضیحات را کلی تر و ساده تر می کند. من معتقدم
 ماهیت خطی مدل عامل اصلی استفاده از مدل های خطی برای توضیح روابط است. خطی بودن توضیحات را
 کلی تر و ساده تر می کند. من معتقدم ماهیت خطی مدل عامل اصلی استفاده از مدل های خطی برای توضیح
 روابط است. خطی بودن توضیحات را کلی تر و ساده تر می کند. من معتقدم ماهیت خطی مدل عامل اصلی
 استفاده از مدل های خطی برای توضیح روابط است.

۱۵۱۴ ۵,۱,۷ مدل های خطی پراکنده

نمونه های مدل های خطی که من انتخاب کرده ام 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 مدل های
 رگرسیون خطی شما؟ تفسیرپذیری به سرشاری می رود. حتی ممکن است در موقعیتی قرار بگیرید که
 ویژگی های بیشتری نسبت به نمونه ها وجود دارد و اصلاً نمی توانید یک مدل خطی استاندارد را مطابقت دهید.
 خبر خوب این است که راه هایی برای معرفی پراکندگی (= چند ویژگی) در مدل های خطی وجود دارد.

۱۵۲۰ ۵,۱,۷,۱ کمند

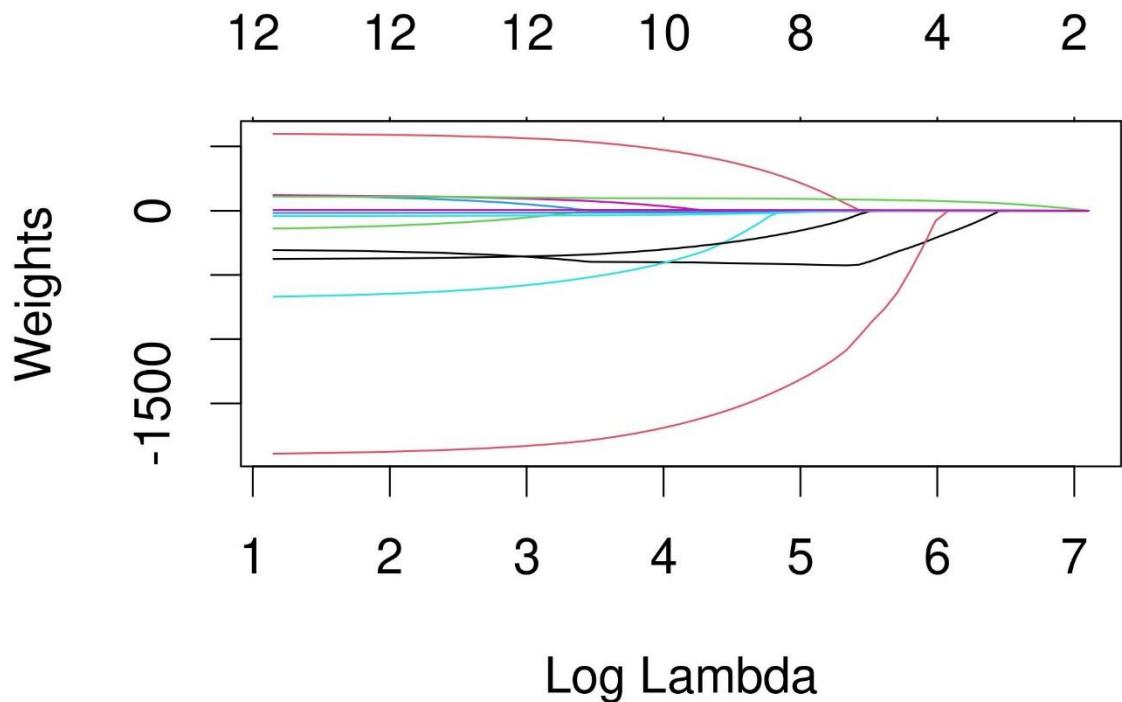
کمند یک راه خودکار و راحت برای وارد کردن پراکندگی به مدل رگرسیون خطی است Lasso. مخفف حداقل
 انقباض مطلق و عملگر انتخاب است و هنگامی که در یک مدل رگرسیون خطی اعمال می شود، انتخاب ویژگی
 و تنظیم وزن ویژگی انتخاب شده را انجام می دهد. اجازه دهید مشکل کمینه سازی را که وزن ها بهینه می کنند
 در نظر بگیریم:

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

۱۵۲۵ ۱۵۲۶ ۱۵۲۷ یک اصطلاح به این مسئله بهینه سازی اضافه می کند. Lasso

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

عبارت β هنجر L1 بردار ویژگی، منجر به جریمه وزن های بزرگ می شود. از آنجایی که از هنجر L1 استفاده می شود، بسیاری از وزن ها تخمینی دریافت می کنند و بقیه کوچک می شوند. پارامتر لامبدا (λ) (قدرت اثر منظم کننده را کنترل می کند و معمولاً با اعتبارسنجی متقطع تنظیم می شود. به خصوص هنگامی که لامبدا بزرگ است، بسیاری از وزن ها به تبدیل می شوند. وزن هر ویژگی با یک منحنی در شکل زیر نشان داده شده است.



شکل ۴,۵: با افزایش جریمه وزن ها، ویژگی های کمتر و کمتری تخمین وزن غیر صفر دریافت می کنند. به این منحنی ها مسیرهای منظم سازی نیز می گویند. عدد بالای نمودار تعداد وزن های غیر صفر است.

چه مقداری را برای لامبدا انتخاب کنیم؟ اگر عبارت جریمه را به عنوان یک پارامتر تنظیم می بینید، می توانید لامبدا را پیدا کنید که خطای مدل را با اعتبارسنجی متقطع به حداقل می رساند. همچنین می توانید لامبدا را به عنوان پارامتری برای کنترل تفسیرپذیری مدل در نظر بگیرید. هر چه جریمه بزرگتر باشد، ویژگی های کمتری در مدل وجود دارد (زیرا وزن آنها صفر است) و بهتر می توان مدل را تفسیر کرد.

مثال با کمند

اجاره دوچرخه را با استفاده از کمند پیش بینی می کنیم. تعداد ویژگی هایی که می خواهیم در مدل داشته باشیم را از قبل تعیین می کنیم. اجازه دهید ابتدا عدد را روی ۲ ویژگی تنظیم کنیم:

	Weight
seasonWINTER	0.00
seasonSPRING	0.00
seasonSUMMER	0.00
seasonFALL	0.00
holidayHOLIDAY	0.00
workingdayWORKING DAY	0.00
weathersitMISTY	0.00
weathersitRAIN/SNOW/STORM	0.00
temp	52.33
hum	0.00
windspeed	0.00
days_since_2011	2.15

دو ویژگی اول با وزن های غیر صفر در مسیر کمند دما ("دما") و روند زمانی ("days_since_2011") هستند.

اکنون، اجازه دهید ۵ ویژگی را انتخاب کنیم:

	Weight
seasonWINTER	-389.99
seasonSPRING	0.00
seasonSUMMER	0.00
seasonFALL	0.00
holidayHOLIDAY	0.00
workingdayWORKING DAY	0.00
weathersitMISTY	0.00
weathersitRAIN/SNOW/STORM	-862.27
temp	85.58
hum	-3.04
windspeed	0.00
days_since_2011	3.82

توجه داشته باشید که وزن های "temp" و "days_since_2011" با مدل با دو ویژگی متفاوت است. دلیل این امر این است که با کاهش لامبدا، حتی ویژگی هایی که قبلاً «در» هستند، کمتر جریمه می شوند و ممکن است وزن مطلق بیشتری به دست آورند. تفسیر وزن های کمnd با تفسیر وزن ها در مدل رگرسیون خطی مطابقت دارد. فقط باید به استاندارد بودن یا نبودن ویژگی ها توجه کنید، زیرا این روی وزن ها تأثیر می گذارد. در این مثال، ویژگی ها توسط نرم افزار استاندارد شده بودند، اما وزن ها به طور خودکار برای ما تغییر شکل دادند تا با مقیاس های ویژگی اصلی مطابقت داشته باشند.

روش های دیگر برای پراکندگی در مدل های خطی

۱۵۵۵ طیف گستردۀ ای از روش‌ها را می‌توان برای کاهش تعداد ویژگی‌ها در یک مدل خطی استفاده کرد.

۱۵۵۶ روش‌های پیش‌پردازش:

۱۵۵۷ ویژگی‌های انتخاب شده دستی: همیشه می‌توانید از دانش متخصص 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
۱۵۵۸ که داده‌ها را درک کند.

۱۵۶۰ انتخاب تک متغیره: یک مثال ضریب همبستگی است. شما فقط ویژگی‌هایی را در نظر می‌گیرید که از آستانه
۱۵۶۱ مشخصی از همبستگی بین ویژگی و هدف فراتر می‌روند. نقطه ضعف این است که فقط ویژگی‌ها را به صورت
۱۵۶۲ جدایانه در نظر می‌گیرد. برخی از ویژگی‌ها ممکن است تا زمانی که مدل خطی برخی ویژگی‌های دیگر را در
۱۵۶۳ نظر نگرفته باشد، همبستگی نشان نمی‌دهند. آنها یعنی که با روش‌های انتخاب تک متغیره از دست خواهد داد.

۱۵۶۴ روش‌های گام به گام:

۱۵۶۵ انتخاب رو به جلو: مدل خطی را با یک ویژگی مناسب کنید. این کار را با هر ویژگی انجام دهید. مدلی را انتخاب
کنید که بهترین عملکرد را دارد (مثلاً بالاترین مربع R^2). اکنون دوباره، برای ویژگی‌های باقی‌مانده، با افزودن هر
۱۵۶۷ ویژگی به بهترین مدل فعلی، نسخه‌های مختلف مدل خود را متناسب کنید. یکی را انتخاب کنید که بهترین
۱۵۶۸ عملکرد را دارد. این کار را تا رسیدن به معیاری مانند حداکثر تعداد ویژگی‌های مدل ادامه دهید.

۱۵۶۹ انتخاب به عقب: مشابه انتخاب رو به جلو. اما به جای افزودن ویژگی‌ها، با مدلی شروع کنید که شامل همه
۱۵۷۰ ویژگی‌ها است و سعی کنید کدام ویژگی را حذف کنید تا بالاترین افزایش عملکرد را داشته باشید. این کار را تا
۱۵۷۱ رسیدن به معیار توقف تکرار کنید.

۱۵۷۲ توصیه می‌کنم از Lasso استفاده کنید، زیرا می‌تواند خودکار باشد، همه ویژگی‌ها را به طور همزمان در نظر
۱۵۷۳ بگیرد و از طریق لامبда قابل کنترل باشد. همچنین برای مدل رگرسیون لجستیک برای طبقه بندی کار می‌
۱۵۷۴ کند.

۱۵۷۵ **۵,۱,۸ مزایا**

۱۵۷۶ مدل‌سازی پیش‌بینی‌ها به عنوان یک جمع وزنی، نحوه تولید پیش‌بینی‌ها را شفاف می‌کند. و با LASSO می‌توانیم
۱۵۷۷ اطمینان حاصل کنیم که تعداد ویژگی‌های مورد استفاده کم باقی می‌ماند.

۱۵۷۸ بسیاری از افراد از مدل‌های رگرسیون خطی استفاده می‌کنند. این بدان معناست که در بسیاری از جاها برای
۱۵۷۹ مدل‌سازی پیش‌بینی و انجام استنتاج پذیرفته شده است. سطح بالایی از تجربه و تخصص جمعی، از جمله

۱۵۸۰ مطالب آموزشی در مورد مدل های رگرسیون خطی و پیاده سازی نرم افزار وجود دارد . رگرسیون خطی را می
۱۵۸۱ توان در R ، Python ،Julia ، Scala ، Javascript ...

۱۵۸۲ از نظر ریاضی، تخمین وزن ها ساده است و شما تضمینی برای یافتن وزن های بهینه دارید (با توجه به اینکه تمام
۱۵۸۳ مفروضات مدل رگرسیون خطی توسط داده ها برآورده می شوند).

۱۵۸۴ همراه با وزن ها، فواصل اطمینان، آزمون ها و تئوری آماری جامد را دریافت می کنید. همچنین توسعه های
۱۵۸۵ زیادی برای مدل رگرسیون خطی وجود دارد (به فصل GAM ، GLM و موارد دیگر مراجعه کنید).

۱۵۸۶ ۵,۱,۹ معایب

۱۵۸۷ مدل های رگرسیون خطی فقط می توانند روابط خطی را نشان دهند، یعنی مجموع وزنی ویژگی های ورودی.
۱۵۸۸ هر غیرخطی یا تعامل باید به صورت دستی ساخته شود و به طور صریح به مدل به عنوان یک ویژگی ورودی
۱۵۸۹ داده شود.

۱۵۹۰ مدل های خطی نیز اغلب در مورد عملکرد پیش بینی کننده خوب نیستند ، زیرا روابطی که می توان آموخت بسیار
۱۵۹۱ محدود است و معمولاً پیچیدگی واقعیت را بیش از حد ساده می کند.

۱۵۹۲ تفسیر یک وزن می تواند غیر شهودی باشد زیرا به تمام ویژگی های دیگر بستگی دارد. یک ویژگی با همبستگی
۱۵۹۳ مثبت بالا با نتیجه ۷ و یک ویژگی دیگر ممکن است در مدل خطی وزن منفی داشته باشد، زیرا با توجه به
۱۵۹۴ ویژگی همبسته دیگر، در فضای با ابعاد بالا با ۷ همبستگی منفی دارد. ویژگی های کاملاً همبسته حتی یافتن
۱۵۹۵ یک راه حل منحصر به فرد برای معادله خطی را غیرممکن می کند. یک مثال: شما یک مدل برای پیش بینی
۱۵۹۶ ارزش یک خانه دارید و دارای ویژگی هایی مانند تعداد اتاق ها و اندازه خانه هستید. اندازه خانه و تعداد اتاق ها
۱۵۹۷ به شدت مرتبط هستند: هر چه خانه بزرگتر باشد، اتاق های بیشتری دارد. اگر هر دو ویژگی را در یک مدل
۱۵۹۸ خطی قرار دهید، ممکن است این اتفاق بیفتد که اندازه خانه پیش بینی کننده بهتری باشد و وزن مثبت زیادی
۱۵۹۹ دریافت کند. ممکن است تعداد اتاق ها وزن منفی داشته باشد، زیرا

۱۶۰۰ ۵,۲ رگرسیون لجستیک

۱۶۰۱ رگرسیون لجستیک احتمالات مسائل طبقه بندی را با دو نتیجه ممکن مدل می کند. این توسعه مدل رگرسیون
۱۶۰۲ خطی برای مسائل طبقه بندی است.

۱۶۰۳ ۵,۲,۱ رگرسیون خطی برای طبقه بندی چه اشکالی دارد؟

۱۶۰۴ مدل رگرسیون خطی می تواند برای رگرسیون خوب کار کند، اما برای طبقه بندی شکست می خورد. چرا
۱۶۰۵ اینطور است؟ در صورت وجود دو کلاس، می توانید یکی از کلاس ها را با ۰ و دیگری را با ۱ برچسب گذاری

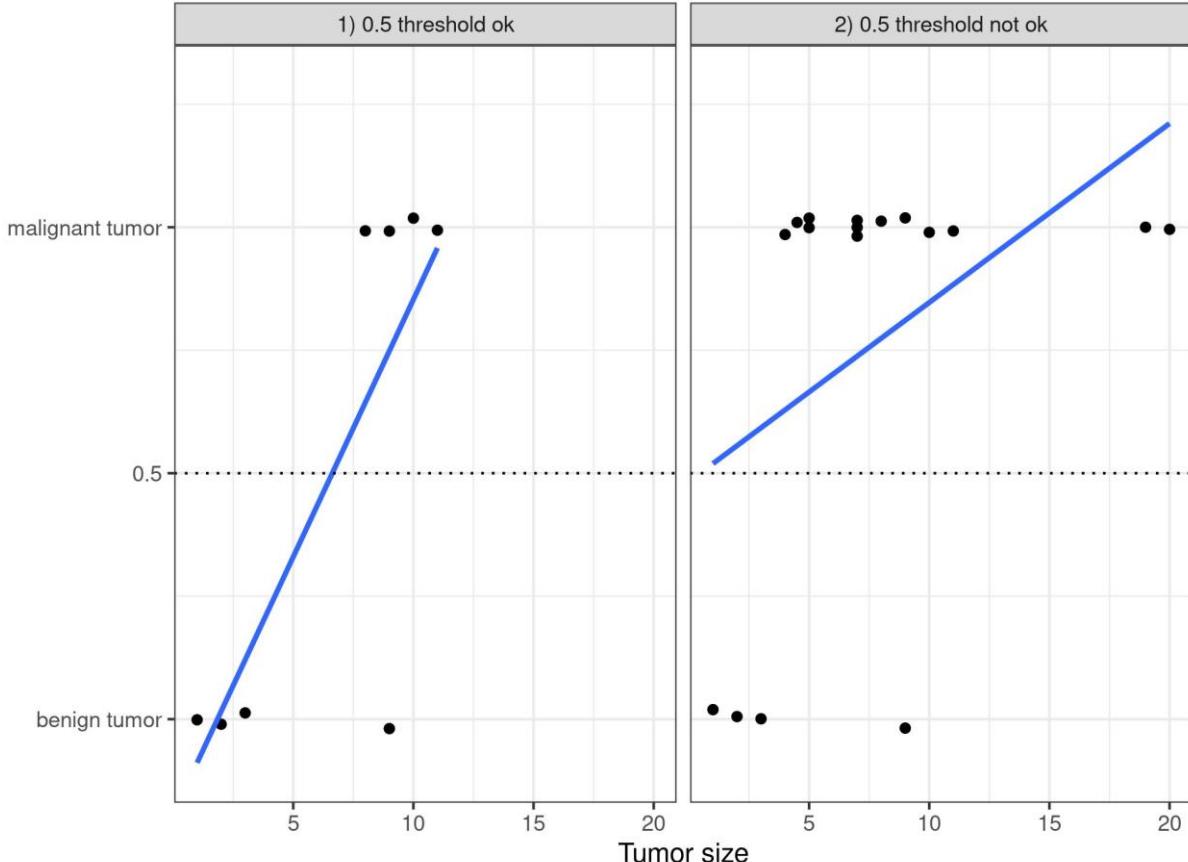
کنید و از رگرسیون خطی استفاده کنید. از نظر فنی کار می کند و اکثر برنامه های مدل خطی وزنه ها را برای شما بیرون می دهند. اما این روش چند مشکل دارد:

یک مدل خطی احتمالات را خروجی نمی کند، اما طبقات را به عنوان اعداد (۰ و ۱) در نظر می گیرد و با بهترین ابر صفحه (برای یک ویژگی، یک خط است) که فاصله بین نقاط و ابر صفحه را به حداقل می رساند. بنابراین به سادگی بین نقاط درون یابی می شود و شما نمی توانید آن را به عنوان احتمال تفسیر کنید.

یک مدل خطی نیز برونویابی می کند و مقادیر زیر صفر و بالای یک را به شما می دهد. این نشانه خوبی است که ممکن است رویکرد هوشمندانه تری برای طبقه بندی 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 در Stackoverflow ارائه شده است.

مدل‌های خطی به مسائل طبقه‌بندی با کلاس‌های متعدد گسترش 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 رابطه 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.



۱۶۲۱

شکل ۵,۵: یک مدل خطی تومورها را با توجه به اندازه آنها به عنوان بدخیم (۱) یا خوش خیم (۰) طبقه بندی می کند. خطوط پیش بینی مدل خطی را نشان می دهد. برای داده های سمت چپ، می توانیم از ۰,۵ به عنوان آستانه طبقه بندی استفاده کنیم. پس از معرفی چند مورد تومور بدخیم دیگر، خط رگرسیون تغییر می کند و آستانه ۰,۵ دیگر کلاس ها را از هم جدا نمی کند. برای کاهش ترسیم بیش از حد، نقاط کمی تکان می خورند.

۱۶۲۶

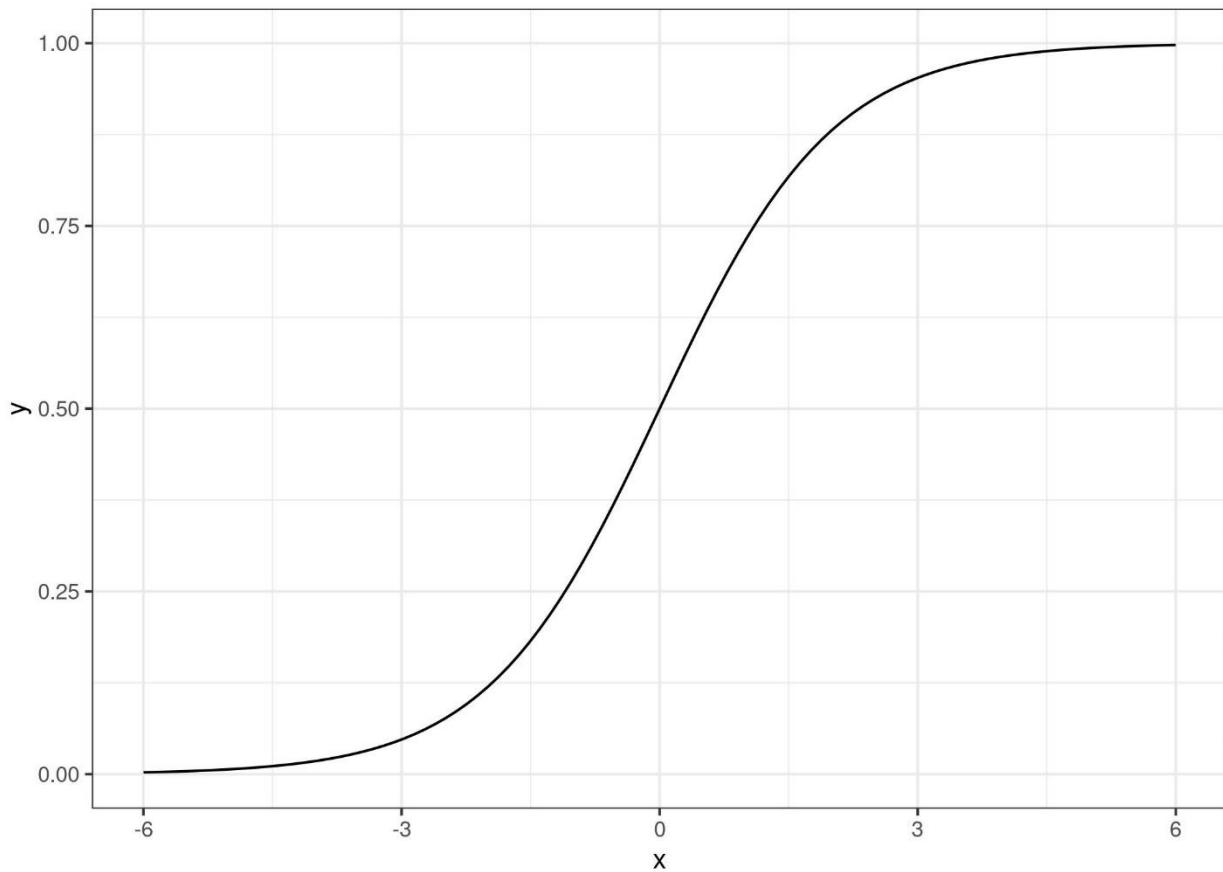
۵,۲,۲ نظریه

یک راه حل برای طبقه بندی رگرسیون لجستیک است. مدل رگرسیون لجستیک به جای برازش یک خط مستقیم یا ابر صفحه، ازتابع لجستیک برای فشرده کردن خروجی یک معادله خطی بین ۰ و ۱ استفاده می کند. تابع لجستیک به صورت زیر تعریف می شود:

۱۶۳۰

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

و به نظر می رسد این است:



۱۶۳۲ شکل ۶,۵: تابع لجستیک. خروجی اعداد بین ۰ و ۱ در ورودی ۰,۵ است.

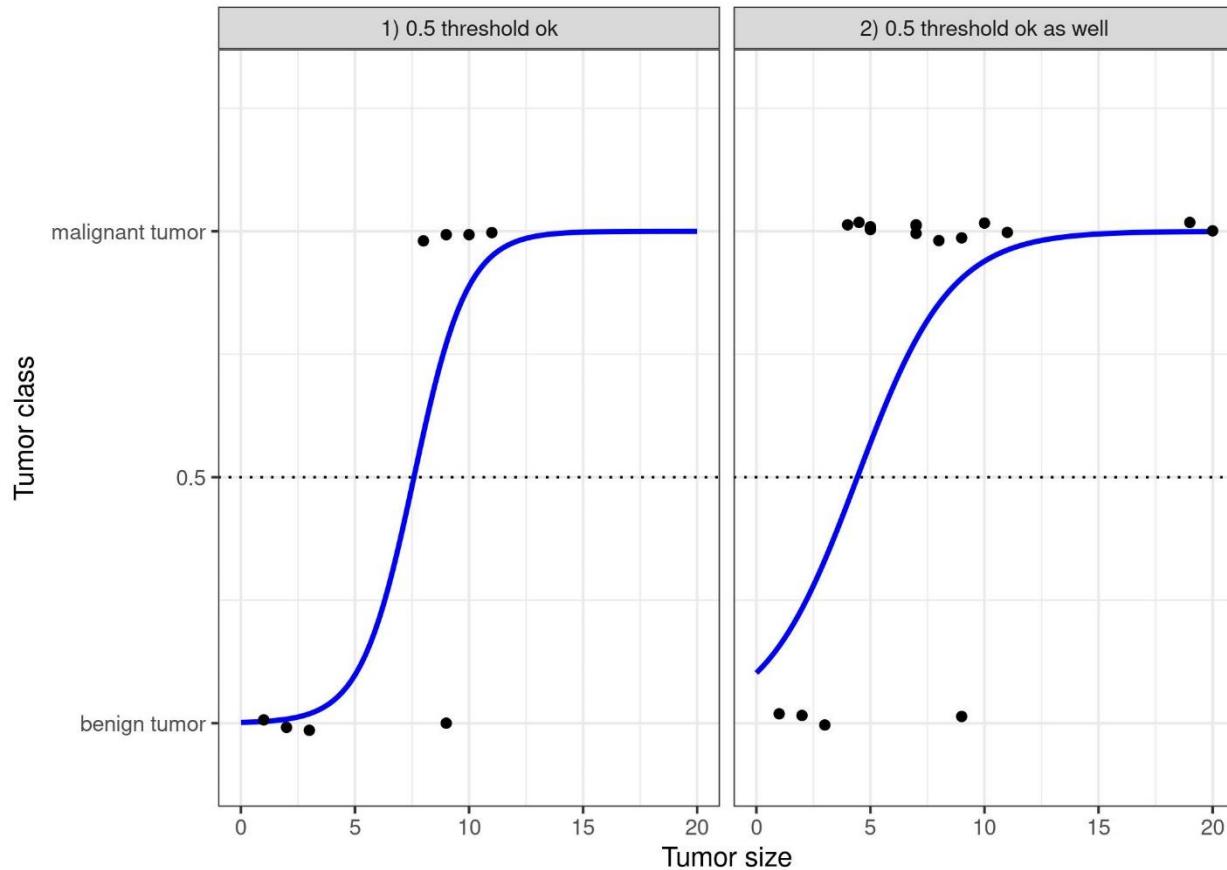
۱۶۳۴ گام از رگرسیون خطی به رگرسیون لجستیک به نوعی ساده است. در مدل رگرسیون خطی، ما رابطه بین نتیجه
۱۶۳۵ و ویژگی ها را با یک معادله خطی مدل کرده ایم:

$$1636 \quad \hat{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_p x_p^{(i)}$$

۱۶۳۷ برای طبقه بندی، احتمالات بین ۰ و ۱ را ترجیح می دهیم، بنابراین سمت راست معادله را در تابع لجستیک
۱۶۳۸ قرار می دهیم. این امر خروجی را مجبور می کند که فقط مقادیر بین ۰ و ۱ را در نظر بگیرد.

$$1639 \quad P(y^{(i)} = 1) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_p x_p^{(i)}))}$$

۱۶۴۰ اجازه دهد و مثال اندازه تومور را بررسی کنیم. اما به جای مدل رگرسیون خطی، از مدل رگرسیون
۱۶۴۱ لجستیک استفاده می کنیم:



۱۶۴۲

۱۶۴۳

شکل ۵: مدل رگرسیون لجستیک مرز تصمیم گیری صحیح بین بدخیم و خوش خیم را بسته به اندازه تومور پیدا می کند. خط تابع لجستیکی است که برای جابجایی داده ها جابجا شده و فشرده می شود.

۱۶۴۵

۱۶۴۶

طبقه بندی با رگرسیون لجستیک بهتر عمل می کند و در هر دو مورد می توانیم از ۰,۵ به عنوان آستانه استفاده کنیم. گنجاندن نقاط اضافی واقعاً بر منحنی تخمینی تأثیر نمی گذارد.

۱۶۴۷

۵,۲,۳ تفسیر

۱۶۴۸

۱۶۴۹

۱۶۵۰

۱۶۵۱

تفسیر وزن ها در رگرسیون لجستیک با تفسیر وزن ها در رگرسیون خطی متفاوت است، زیرا نتیجه در رگرسیون لجستیک احتمالی بین ۰ و ۱ است. وزن ها دیگر روی احتمال به صورت خطی تأثیر نمی گذارند. مجموع وزنی توسط تابع لجستیک به یک احتمال تبدیل می شود. بنابراین ما باید معادله را برای تفسیر دوباره فرمول بندی کنیم تا فقط عبارت خطی در سمت راست فرمول باشد.

۱۶۵۲

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

ما این اصطلاح را در تابع "ln شانس" می نامیم (احتمال رویداد تقسیم بر احتمال هیچ رویدادی) و در لگاریتم پیچیده شده به آن $\log \text{odds}$ می گویند.

این فرمول نشان می دهد که مدل رگرسیون لجستیک یک مدل خطی برای شانس ورود است. عالی! این مفید به نظر نمی رسد! با کمی به هم ریختن اصطلاحات، می توانید بفهمید که چگونه پیش بینی هنگام یکی از ویژگی ها تغییر می کند. برای انجام این کار، ابتدا می توانیم تابع $\exp()$ را در دو طرف معادله اعمال کنیم:

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

سپس آنچه را که وقتی یکی از مقادیر ویژگی را ۱ افزایش می دهیم با هم مقایسه می کنیم. اما به جای بررسی تفاوت، به نسبت دو پیش بینی نگاه می کنیم:

$$\frac{\text{odds}_{x_j+1}}{\text{odds}} = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_j(x_j+1) + \dots + \beta_p x_p)}{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_j x_j + \dots + \beta_p x_p)}$$

ما قانون زیر را اعمال می کنیم:

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

و بسیاری از اصطلاحات را حذف می کنیم:

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

در پایان، ما چیزی به سادگی $\exp()$ از وزن ویژگی داریم. تغییر در یک ویژگی به اندازه یک واحد، نسبت شانس (ضریب) را با ضریب (β_j) تغییر می دهد. همچنین می توانیم آن را این گونه تفسیر کنیم: تغییر در X_j یک واحد نسبت شانس ورود به سیستم را با مقدار وزن مربوطه افزایش می دهد. اکثر مردم نسبت شانس را تفسیر می کنند زیرا فکر کردن در مورد (\ln) چیزی برای مغز سخت است. تفسیر نسبت شانس از قبل نیاز به عادت دارد. به عنوان مثال، اگر شما شانس ۲ دارید، به این معنی است که احتمال $y=1$ دو برابر $y=0$ است. اگر وزن (= نسبت $\log \text{odds ratio}$) دارید، آنگاه افزایش ویژگی مربوطه در یک واحد، شانس را در ضرب می کند (تقریباً ۲) و شانس به ۴ تغییر می کند. اما معمولاً شما با شانس برخورد نمی کنید. وزن ها را فقط به عنوان نسبت شانس تفسیر کنید. زیرا برای محاسبه واقعی شانس باید یک مقدار برای هر ویژگی تعیین کنید، که تنها زمانی منطقی است که بخواهید به یک نمونه خاص از مجموعه داده خود نگاه کنید.

اینها تفاسیر مدل رگرسیون لجستیک با انواع ویژگی های مختلف است:

ویژگی عددی: اگر ارزش ویژگی را افزایش دهید X_j با یک واحد، شанс تخمین زده شده با ضریب تغییر می کند

$$\exp(\beta_j)$$

ویژگی دسته‌بندی باینری: یکی از دو مقدار ویژگی، دسته مرجع است (در برخی زبان‌ها، مقداری که با ۰ کدگذاری می‌شود). تغییر ویژگی X_j از دسته مرجع به دسته دیگر شанс تخمینی را با ضریب $\exp(\beta_j)$ تغییر می‌دهد

ویژگی طبقه‌بندی با بیش از دو دسته: یک راه حل برای مقابله با چندین دسته، یک کدگذاری داغ است، به این معنی که هر دسته ستون خاص خود را دارد. شما فقط به ستون‌های $L-1$ برای یک ویژگی دسته‌بندی با دسته‌های L نیاز دارید، در غیر این صورت بیش از حد پارامتر شده است. پس از آن، ردتا، مقوله مرجع است. می‌توانید از هر رمزگذاری دیگری که می‌تواند در رگرسیون خطی استفاده شود استفاده کنید. تفسیر برای هر دسته پس از آن معادل تفسیر ویژگی‌های باینری است.

رهگیری: β_0 وقتی همه ویژگی‌های عددی صفر هستند و ویژگی‌های طبقه‌بندی در دسته مرجع قرار می‌گیرند، شанс $\exp(\beta_0)$ تخمین زده می‌شود. تفسیر وزن رهگیری معمولاً مرتبط نیست.

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 شанс مرتبط و خطای استاندارد تخمین‌ها را نشان می‌دهد.

جدول ۵,۲: نتایج برآشش یک مدل رگرسیون لجستیک بر روی مجموعه داده سرطان دهانه رحم. ویژگی‌های مورد استفاده در مدل، وزن‌های تخمینی و نسبت‌های شанс مربوطه و خطاهای استاندارد وزن‌های تخمینی نشان داده 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 ("تعداد STD‌های تشخیص داده شده"): افزایش تعداد STD‌های تشخیص داده شده (بیماری‌های مقاربتی) شанс ابتلا به سرطان در مقایسه با عدم وجود سرطان را با ضریب ۲,۲۷ تغییر می‌دهد

۱۷۰۰ (افزایش می دهد) در حالی که همه ویژگی های دیگر همان باقی می ماند. به خاطر داشته باشد که همبستگی
۱۷۰۱ به معنای علیت نیست.

۱۷۰۲ تفسیر یک ویژگی طبقه بندی شده ("داروهای ضد بارداری هورمونی: γ/n ") برای زنانی که از داروهای ضد
۱۷۰۳ بارداری هورمونی استفاده می کنند، شانس ابتلا به سرطان در مقایسه با بدون سرطان ۰,۸۹ کمتر است، در
۱۷۰۴ مقایسه با زنان بدون ضد بارداری هورمونی، با توجه به باقی ماندن سایر ویژگی ها. یکسان.

۱۷۰۵ مانند مدل خطی، تفاسیر همیشه با این بند آمده است که «همه ویژگی های دیگر ثابت می مانند.»

۱۷۰۶ ۵,۲,۵ مزايا و معایب
۱۷۰۷ بسیاری از مزايا و معایب مدل رگرسیون خطی در مورد مدل رگرسیون لجستیک نیز صدق می کند. رگرسیون
۱۷۰۸ لجستیک به طور گسترده توسط افراد مختلف مورد استفاده قرار گرفته است، اما با بیان محدود خود (مثلاً
۱۷۰۹ تعاملات باید به صورت دستی اضافه شوند) مشکل دارد و مدل های دیگر ممکن است عملکرد پیش‌بینی بهتری
۱۷۱۰ داشته باشند.

۱۷۱۱ یکی دیگر از معایب مدل رگرسیون لجستیک این است که تفسیر دشوارتر است زیرا تفسیر اوزان ضربی است و
۱۷۱۲ افزایشی نیست.

۱۷۱۳ رگرسیون لجستیک می تواند از جدایی کامل رنج ببرد . اگر ویژگی ای وجود داشته باشد که این دو کلاس را
۱۷۱۴ کاملاً از هم جدا کند، مدل رگرسیون لجستیک دیگر قابل آموزش نیست. این به این دلیل است که وزن آن
۱۷۱۵ ویژگی به هم نزدیک نمی شود، زیرا وزن بهینه بی نهايت خواهد بود. این واقعاً کمی تاسف بار است، زیرا چنین
۱۷۱۶ ویژگی واقعاً مفید است. اما اگر قانون ساده ای دارید که هر دو کلاس را از هم جدا می کند، نیازی به یادگیری
۱۷۱۷ ماشین ندارید. مشکل جداسازی کامل را می توان با معرفی جرمیه وزن ها یا تعریف توزیع احتمال قبلی وزن ها
۱۷۱۸ حل کرد.

۱۷۱۹ از طرفی، مدل رگرسیون لجستیک نه تنها یک مدل طبقه بندی است، بلکه احتمالاتی را نیز به شما می دهد.
۱۷۲۰ این یک مزیت بزرگ نسبت به مدل هایی است که فقط می توانند طبقه بندی نهايی را ارائه دهند. دانستن اينکه
۱۷۲۱ یک نمونه برای یک کلاس ۹۹ درصد احتمال دارد در مقایسه با ۵۱ درصد، تفاوت بزرگی ایجاد می کند.

۱۷۲۲ رگرسیون لجستیک همچنین می تواند از طبقه بندی باينري به طبقه بندی چند طبقه گسترش يابد. سپس به
۱۷۲۳ آن رگرسیون چند جمله ای می گويند.

۵،۲،۶ نرم افزار

من از **glm** تابع 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، مانند پایتون، جاوا، استاتا، متلب و

۵،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 بسیاری 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 نتیجه 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.

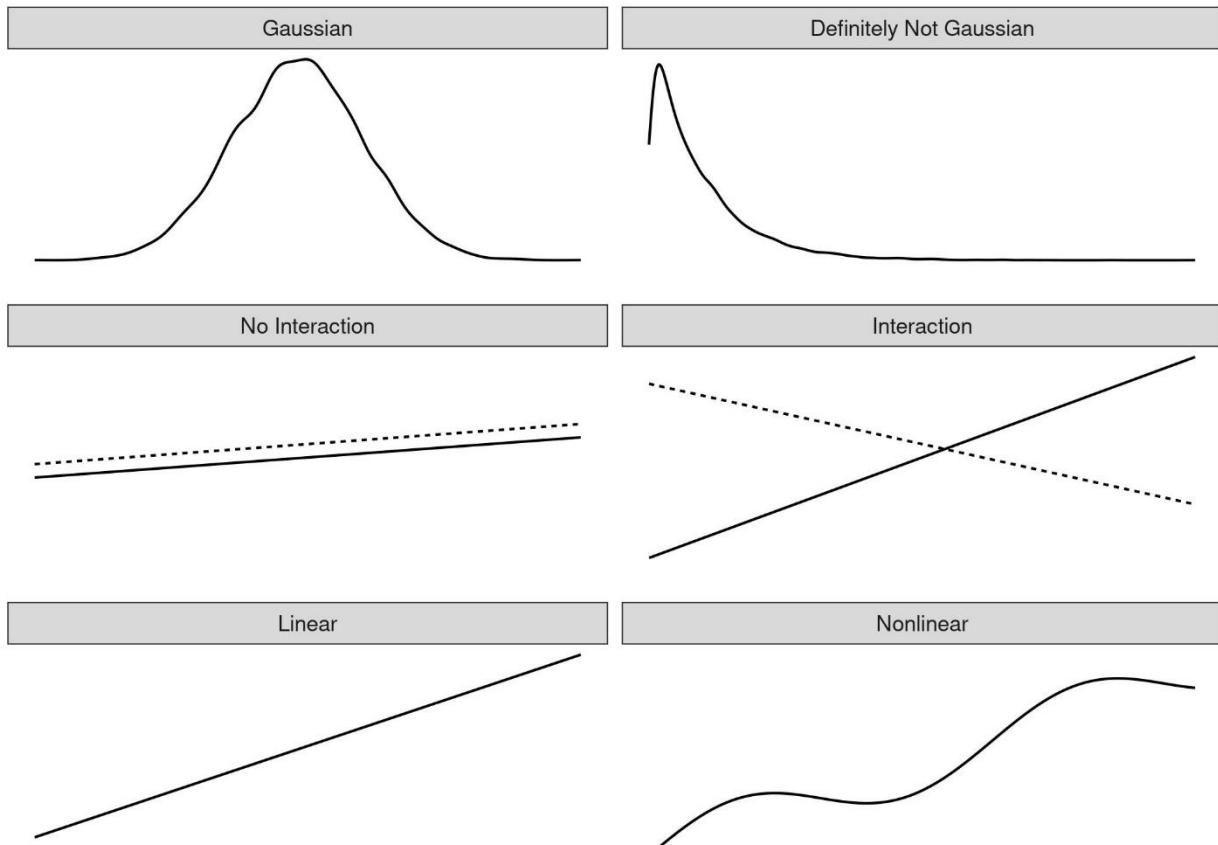
این فصل قطعاً راهنمای قطعی شما برای توسعه مدل‌های خطی نیست. بلکه به عنوان یک نمای کلی از برنامه‌های افزودنی مانند مدل‌های خطی تعمیم یافته (GLM) و مدل‌های افزودنی تعمیم یافته (GAM) عمل می‌کند و کمی شهود به شما می‌دهد. پس از مطالعه، باید یک دید کلی از نحوه گسترش مدل‌های خطی داشته باشید. اگر می‌خواهید ابتدا 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 بسپاریم:

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

مدل رگرسیون خطی فرض می‌کند که نتیجه y یک نمونه را می‌توان با مجموع وزنی از ویژگی‌های p آن با یک خطای فردی ϵ بیان کرد. که از توزیع گاووسی پیروی می‌کند. با وارد کردن داده‌ها به این کرست یک 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 تبدیل 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. مشکلات بسیاری در مورد فرضیات احتمالاً نقض شده وجود دارد، اما ما بر روی سه مورد نشان داده شده در شکل زیر تمرکز خواهیم کرد:



۱۷۵۰
۱۷۵۱ شکل ۵,۸: سه فرض مدل خطی (سمت چپ): توزیع گاوسی نتیجه با توجه به ویژگی‌ها، افزایش (= بدون تعامل)
۱۷۵۲ و رابطه خطی. واقعیت معمولاً به آن مفروضات پایبند نیست (سمت راست): نتایج ممکن است دارای توزیع‌های
۱۷۵۳ غیر گاوسی باشند، ویژگی‌ها ممکن است تعامل داشته باشند و رابطه ممکن است غیرخطی باشد.
۱۷۵۴ برای همه این مشکلات راه حلی وجود دارد:

۱۷۵۵ مشکل: نتیجه هدف ۷ با توجه به ویژگی‌ها از توزیع گاوسی پیروی نمی‌کند.

۱۷۵۶ مثال: فرض کنید می‌خواهم پیش‌بینی کنم که در یک روز معین چند دقیقه دوچرخه‌سواری خواهم کرد. به
۱۷۵۷ عنوان ویژگی من نوع روز، آب و هوا و غیره را دارم. اگر از یک مدل خطی استفاده کنم، می‌تواند دقیقه‌های
۱۷۵۸ منفی را پیش‌بینی کند، زیرا توزیع گاوسی را فرض می‌کند که در دقیقه ۰ متوقف نمی‌شود. همچنین اگر
۱۷۵۹ بخواهم احتمالات را با یک مدل خطی پیش‌بینی کنم، می‌توانم احتمالات منفی یا بزرگ‌تر از ۱ را بدست بیاورم.

۱۷۶۰ راه حل: مدل‌های خطی تعمیم‌یافته. (GLMs)

۱۷۶۱ 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 ساعت 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 فعل و انفعالات به صورت دستی.

مشکل : رابطه واقعی بین ویژگی‌ها و خطی نیست.

مثال : بین ۰ تا ۲۵ درجه سانتیگراد، تأثیر دما بر تمایل من به دوچرخه سواری می‌تواند خطی باشد، به این معنی که افزایش از ۰ به ۱ درجه باعث افزایش همان افزایش میل دوچرخه سواری با افزایش از ۲۰ به ۲۱ 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 توسعه‌های بیشتر مدل خطی حذف شده‌اند. اگر بخواهم همه چیز را در اینجا پوشش دهم، این فصل به سرعت تبدیل به کتابی در یک کتاب درباره موضوعی می‌شود که قبلاً در 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 راه حل برای برنامه‌های افزودنی مدل خطی ایجاد کرده ام که می‌توانید در انتهای فصل پیدا کنید. نام راه حل به عنوان نقطه شروع برای جستجو است.

۵.۳.۱ نتایج غیر گاوی - GLMs

مدل رگرسیون خطی فرض می‌کند که نتیجه با توجه به ویژگی‌های ورودی از یک توزیع گاوی پیروی می‌کند. این فرض بسیاری از موارد را کنار می‌گذارد: نتیجه می‌تواند یک دسته (سرطان در مقابل سالم)، شمارش (تعداد فرزندان)، زمان وقوع یک رویداد (زمان تا خرابی یک دستگاه) یا یک نتیجه بسیار ناهنجار باشد. چند ارزش بسیار بالا (درآمد خانوار). مدل رگرسیون خطی را می‌توان برای مدل سازی همه این نوع پیامدها گسترش داد. این پسوند مدل‌های خطی تعمیم یافته یا GLM نامیده می‌شود به طور خلاصه در طول این فصل، من از نام GLM هم برای چارچوب کلی و هم برای مدل‌های خاص از آن 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 متصل کنید.

به عنوان مثال، مدل رگرسیون لجستیک توزیع برنولی را برای نتیجه فرض می کند و میانگین مورد انتظار و مجموع وزنی را با استفاده از تابع لجستیک به هم مرتبط می کند.

به صورت ریاضی جمع وزنی ویژگی ها را با مقدار میانگین توزیع فرضی با استفاده از تابع پیوند g پیوند GLM می دهد که بسته به نوع نتیجه می تواند به طور انعطاف پذیر انتخاب شود.

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

ها از سه جزء تشکیل شده اند: تابع پیوند g ، مجموع وزنی $\beta^T X$ گاهی اوقات پیش بینی خطی نامیده می شود) و یک توزیع احتمال از خانواده نمایی که EY تعریف می کند

خانواده نمایی مجموعه ای از توزیع هایی است که می توان با همان فرمول (پارامتری شده) نوشت که شامل یک توان، میانگین و واریانس توزیع و برخی پارامترهای دیگر است. من وارد جزئیات ریاضی نمی شوم زیرا این جهان بسیار بزرگی است که من نمی خواهم وارد آن شوم. ویکی پدیا فهرست دقیقی از توزیع ها از خانواده نمایی دارد . هر توزیعی از این لیست می تواند برای GLM شما انتخاب شود. بر اساس نوع نتیجه ای که می خواهید پیش بینی کنید، توزیع مناسبی را انتخاب کنید. آیا نتیجه چیزی است (مثلاً تعداد کودکانی که در یک خانواده زندگی می کنند)? سپس توزیع پواسون می تواند انتخاب خوبی باشد. آیا نتیجه همیشه مثبت است (مثلا زمان بین دو رویداد)? سپس توزیع نمایی می تواند انتخاب خوبی باشد.

اجازه دهید مدل خطی کلاسیک را به عنوان یک مورد خاص از یک GLM در نظر بگیریم. تابع پیوند برای توزیع گاوی در مدل خطی کلاسیک به سادگی تابع هویت است. توزیع گاوی با پارامترهای میانگین و واریانس پارامتر می شود. میانگین مقداری را که به طور متوسط انتظار داریم و واریانس نشان می دهد که مقادیر در حدود این میانگین چقدر تغییر می کنند را توصیف می کند. در مدل خطی، تابع پیوند، مجموع وزنی ویژگی ها را به میانگین توزیع گاوی پیوند می دهد.

تحت چارچوب GLM ، این مفهوم به هر توزیع (از خانواده نمایی) و توابع پیوند دلخواه تعمیم می یابد. اگر ۷ شمارشی از چیزی باشد، مانند تعداد قوه هایی که فرد در یک روز خاص می نوشد، می توانیم آن را با GLM با توزیع پواسون و لگاریتم طبیعی به عنوان تابع پیوند مدل سازی کنیم:

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

مدل رگرسیون لجستیک نیز یک GLM است که توزیع برنولی را فرض می کند و از تابع لاجیت به عنوان تابع پیوند استفاده می کند. میانگین توزیع دوجمله ای مورد استفاده در رگرسیون لجستیک احتمال ۱ است.

$$x^T \beta = \ln \left(\frac{E_Y(y|x)}{1 - E_Y(y|x)} \right) = \ln \left(\frac{P(y=1|x)}{P(y=0|x)} \right)$$

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

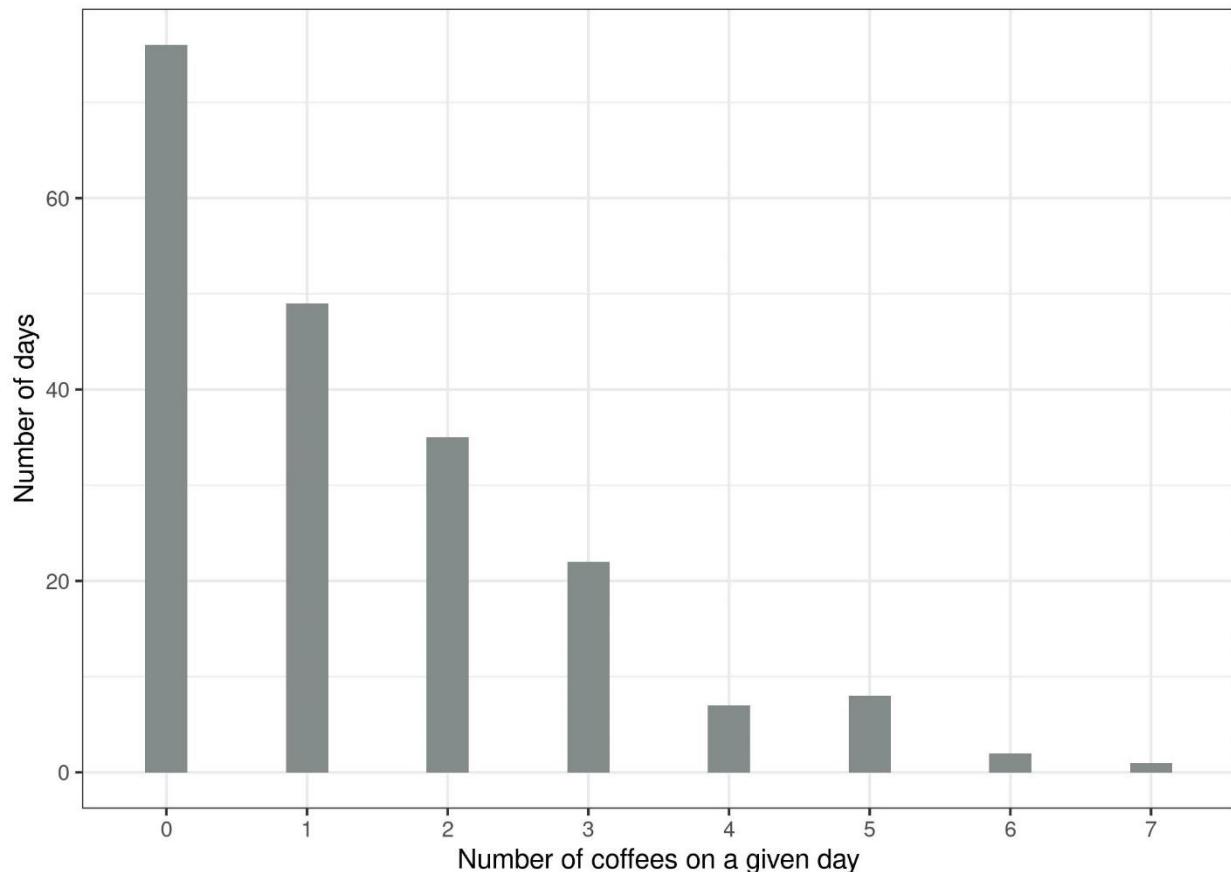
$$P(y=1) = \frac{1}{1 + \exp(-x^T \beta)}$$

هر توزیع از خانواده نمایی دارای یکتابع پیوند متعارف است که می تواند به صورت ریاضی از توزیع استخراج شود. چارچوب GLM امکان انتخاب تابع پیوند را مستقل از توزیع فراهم می کند. چگونه تابع پیوند مناسب را انتخاب کنیم؟ هیچ دستور العمل کاملی وجود ندارد. شما دانش در مورد توزیع هدف خود را در نظر می گیرید، اما ملاحظات نظری و اینکه مدل چقدر با داده های واقعی شما مطابقت دارد را نیز در نظر می گیرید. برای برخی از توزیع ها، تابع پیوند متعارف مقادیری منجر شود که برای آن توزیع نامعتبر هستند. در مورد توزیع نمایی، تابع پیوند متعارف معکوس منفی است که می تواند منجر به پیش بینی های منفی شود که خارج از حوزه توزیع نمایی هستند. از آنجایی که می توانید هر تابع پیوندی را انتخاب کنید،

مثال ها

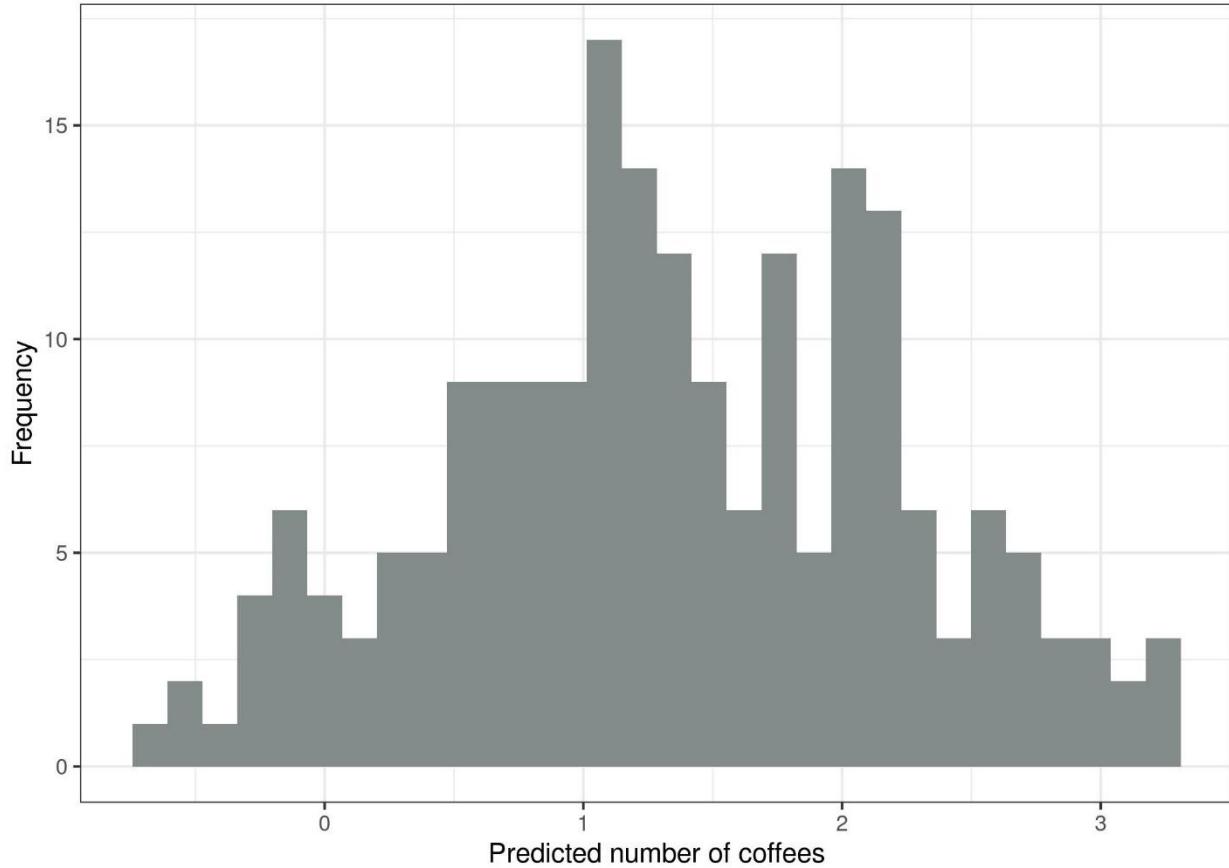
من مجموعه داده ای را در مورد رفتار نوشیدن قهوه شبیه سازی کرده ام تا نیاز به GLM ها را برجسته کنم. فرض کنید اطلاعاتی در مورد رفتار نوشیدن قهوه روزانه خود جمع آوری کرده اید. اگر قهوه دوست ندارید، وانمود کنید که در مورد چای است. همراه با تعداد فنجان ها، سطح استرس فعلی خود را در مقیاس ۱ تا ۱۰ ثبت می کنید، شب قبل چقدر خوب خوابیده اید در مقیاس ۱ تا ۱۰ و اینکه آیا باید در آن روز کار کنید یا خیر. هدف پیش بینی تعداد قهوه ها با توجه به ویژگی های استرس، خواب و کار است. من داده ها را برای ۲۰۰ روز شبیه سازی کردم. استرس و خواب به طور یکنواخت بین ۱ تا ۱۰ ترسیم شد و بله / نه کار با شанс ۵۰/۵۰ ترسیم شد (عجب زندگی!). سپس برای هر روز، تعداد قهوه ها از توزیع پواسون گرفته شد و شدت آن را مدل سازی کرد (که همچنین مقدار مورد انتظار توزیع پواسون است) به عنوان تابعی از ویژگی های خواب، استرس و کار. می توانید حدس بزنید که این داستان به کجا ختم می شود: «هی، اجازه دهید این داده ها را با یک مدل خطی مدل سازی کنیم... اوه، کار نمی کند... اجازه دهید یک GLM با توزیع پواسون را امتحان کنیم ... SURPRISE! حالا کار می کند!». امیدوارم داستان را زیاد برای شما اسپویل نکرده باشم.

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



شکل ۵،۹: توزیع شبیه سازی شده تعداد قهوه های 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 بله/خیر به عنوان ویژگی‌ها استفاده کنیم. وقتی به اشتباه توزیع گاووسی را فرض می‌کنیم چه چیزی می‌تواند اشتباه کند؟ یک فرض اشتباه می‌تواند تخمين‌ها، به ویژه فواصل اطمینان وزن‌ها را باطل کند. مشکل واضح‌تر این است که پیش‌بینی‌ها با دامنه «مجاز» نتیجه واقعی مطابقت ندارند، همانطور که شکل زیر نشان می‌دهد.



۱۸۴۲

۱۸۴۳

شکل ۵،۱۰: تعداد قهوه های پیش بینی شده وابسته به استرس، خواب و کار. مدل خطی مقادیر منفی را پیش بینی می کند.

۱۸۴۵

۱۸۴۶

۱۸۴۷

۱۸۴۸

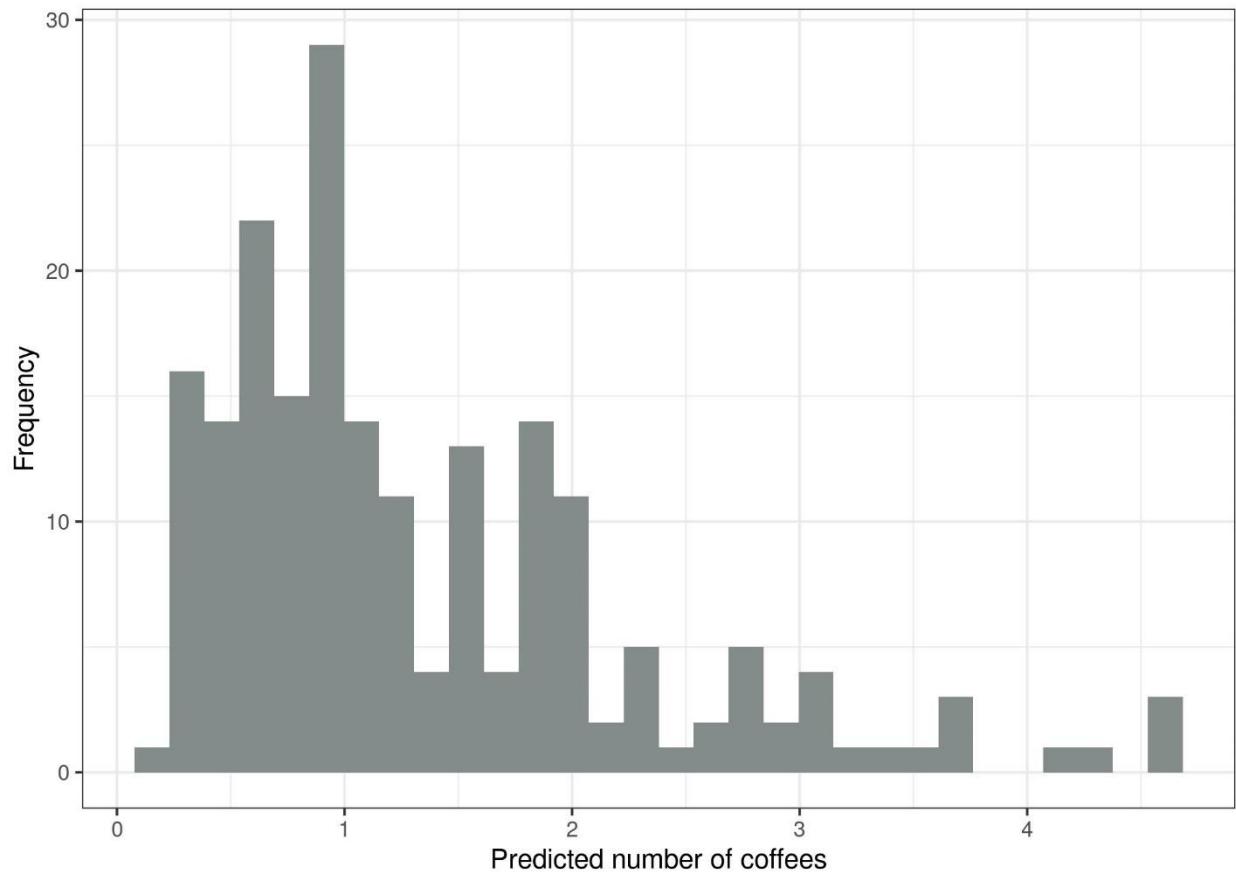
۱۸۴۹

۱۸۵۰

۱۸۵۱

۱۸۵۲

مدل خطی منطقی نیست، زیرا تعداد قهوه های منفی را پیش بینی می کند. این مشکل را می توان با مدل های خطی تعمیم یافته (GLM) حل کرد. ما می توانیمتابع پیوند و توزیع فرضی را تغییر دهیم. یکی از امکان ها حفظ توزیع گاوی و استفاده از تابع پیوندی است که همیشه به پیش بینی های مثبتی مانند *log-link* (معکوس تابع *exp*-*exp* است) به جای تابع هویت منجر می شود. حتی بهتر: ما توزیعی را انتخاب می کنیم که با فرآیند تولید داده و یک تابع پیوند مناسب مطابقت دارد. از آنجایی که نتیجه یک شمارش است، توزیع پواسون یک انتخاب طبیعی به همراه لگاریتم به عنوان تابع پیوند است. در این مورد، داده ها حتی با توزیع پواسون تولید شده اند، بنابراین پواسون GLM انتخاب عالی است. پواسون GLM نصب شده منجر به توزیع زیر از مقادیر پیش بینی شده می شود:



۱۸۵۳

۱۸۵۴

۱۸۵۵

۱۸۵۶

۱۸۵۷

۱۸۵۸

۱۸۵۹

۱۸۶۰

۱۸۶۱

۱۸۶۲

۱۸۶۳

۱۸۶۴

شکل ۱۱: تعداد قهوه های پیش بینی شده وابسته به استرس، خواب و کار GLM با فرض پواسون و پیوند log امدل مناسبی برای این مجموعه داده است.

بدون مقادیر منفی قهوه، اکنون بسیار بهتر به نظر می رسد.

تفسیر وزن های GLM

توزیع مفروض همراه باتابع پیوند تعیین می کند که وزن ویژگی های برآورده شده چگونه تفسیر می شوند. در مثال شمارش قهوه، من از یک GLM با توزیع پواسون و پیوند لاغ استفاده کردم که دلالت بر رابطه زیر بین نتیجه مورد انتظار و ویژگی های استرس(str)، خواب(slp) و کار(wrk) دارد.

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

برای تفسیر وزن ها،تابع پیوند را معکوس می کنیم تا بتوانیم تأثیر ویژگی ها را بر نتیجه مورد انتظار تفسیر کنیم و نه بر لگاریتم نتیجه مورد انتظار.

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

از آنجایی که همه وزن ها در تابع نمایی هستند، تفسیر اثر جمعی نیست، بلکه ضریبی است، زیرا $\exp(a + b)$ است. آخرین عنصر برای تفسیر، وزن های واقعی نمونه اسباب بازی است. جدول زیر وزن های تخمینی و اکسپت (وزن) را همراه با فاصله اطمینان ۹۵ درصد فهرست می کند:

جدول ۳: وزن ها در مدل پواسون

	weight	exp(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 نقطه، تعداد قهوه مورد انتظار را در ضرب ۱,۱۲ ضرب می کند. افزایش کیفیت خواب یک نقطه، تعداد قهوه مورد انتظار را در ضرب ۰,۸۶ ضرب می کند. تعداد قهوه های پیش بینی شده در یک روز کاری به طور متوسط ۲,۲۳ برابر تعداد قهوه های یک روز تعطیل است. به طور خلاصه، هر چه استرس بیشتر، خواب کمتر و کار بیشتر باشد، قهوه بیشتری مصرف می شود.

در این بخش شما کمی در مورد مدل های خطی تعمیم یافته یاد گرفتید که زمانی مفید هستند که هدف از توزیع گاوی پیروی نمی کند. در مرحله بعد، به نحوه ادغام تعاملات بین دو ویژگی در مدل رگرسیون خطی می پردازیم.

۵,۳,۲ فعل و انفعالات

مدل رگرسیون خطی فرض می کند که تأثیر یک ویژگی بدون توجه به مقادیر سایر ویژگی ها یکسان است (= بدون تعامل). اما اغلب در داده ها تعاملاتی وجود دارد. برای پیش بینی تعداد دوچرخه های 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 دوچرخه های اجاره ای تأثیر نمی گذارد، زیرا مردم هر اتفاقی بیفتند با دوچرخه اجاره ای به محل کار خود می روند. در روزهای تعطیل، بسیاری از مردم برای لذت سوار می شوند، اما فقط زمانی که هوا به اندازه کافی گرم باشد. وقتی صحبت از دوچرخه های کرایه ای می شود، ممکن است انتظار تعامل بین دما و روز کاری را داشته باشید.

چگونه می توانیم 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 ویژگی ها است و مطابق معمول مدل را مطابقت می دهد. راه حل به نوعی ظریف است، زیرا به هیچ تغییری در مدل خطی نیاز ندارد، فقط به ستون های اضافی در داده ها نیاز دارد. در مثال روز کاری و دما، یک ویژگی جدید اضافه می کنیم 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 داده های 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 می رسد. جدول زیر نشان می دهد که اگر هیچ گونه تعاملی را مشخص نکنیم، داده های تهیه شده برای مدل چگونه به نظر می رسند. به طور معمول، این تبدیل به طور خودکار توسط هر نرم افزار آماری انجام می شود.

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

ستون اول عبارت intercept است. ستون دوم ویژگی طبقه بندی را با ۰ برای دسته مرجع و ۱ برای دسته دیگر رمزگذاری می کند. 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	1	25	25
1	0	0	12	0
1	0	0	30	0
1	1	1	5	5

ستون جدید "workY.temp" تعامل بین ویژگی های روز کاری (کار) و دما (دما) را نشان می دهد. برای مثال اگر ویژگی کار در رده مرجع باشد ("N") برای هیچ روز کاری، این ستون ویژگی جدید صفر است، در غیر این صورت مقادیر ویژگی دمای نمونه ها را در نظر می گیرد. با این نوع رمزگذاری، مدل خطی می تواند یک اثر خطی متفاوت دما را برای هر دو نوع روز یاد بگیرد. این اثر متقابل بین دو ویژگی است. بدون یک اصطلاح تعاملی، اثر ترکیبی یک ویژگی مقوله ای و عددی را می توان با خطی توصیف کرد که برای دسته های مختلف به صورت عمودی جابجا شده است. اگر تعامل را لحاظ کنیم، اجازه می دهیم اثر ویژگی های عددی (شیب) در هر دسته ارزش متفاوتی داشته باشد.

تعامل دو ویژگی مقوله ای به طور مشابه عمل می کند. ما ویژگی های اضافی ایجاد می کنیم که ترکیبی از دسته ها را نشان می دهد. در اینجا برخی از داده های مصنوعی حاوی روز کاری (کار) و یک ویژگی طبقه بندی آب و هوا (wthr) آمده است:

work	wthr
Y	2
N	0
N	1
Y	2

۱۹۱۱

۱۹۱۲

در مرحله بعد، ما اصطلاحات تعامل را شامل می‌شویم:

۱۹۱۳

ستون اول برای تخمین رهگیری است. ستون دوم ویژگی کار کدگذاری شده است. ستون های سه و چهار برای ویژگی آب و هوا هستند که به دو ستون نیاز دارند زیرا برای ثبت افکت برای سه دسته نیاز به دو وزن دارید که یکی از آنها دسته مرجع است. بقیه ستون ها تعاملات را نشان می‌دهند. برای هر دسته از هر دو ویژگی (به جز دسته های 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 اصطلاحات تعامل را استخراج می‌کند و سپس یک مدل رگرسیون خطی شامل برهمکنش ها را تخمین می‌زند.

۱۹۲۳ مثال

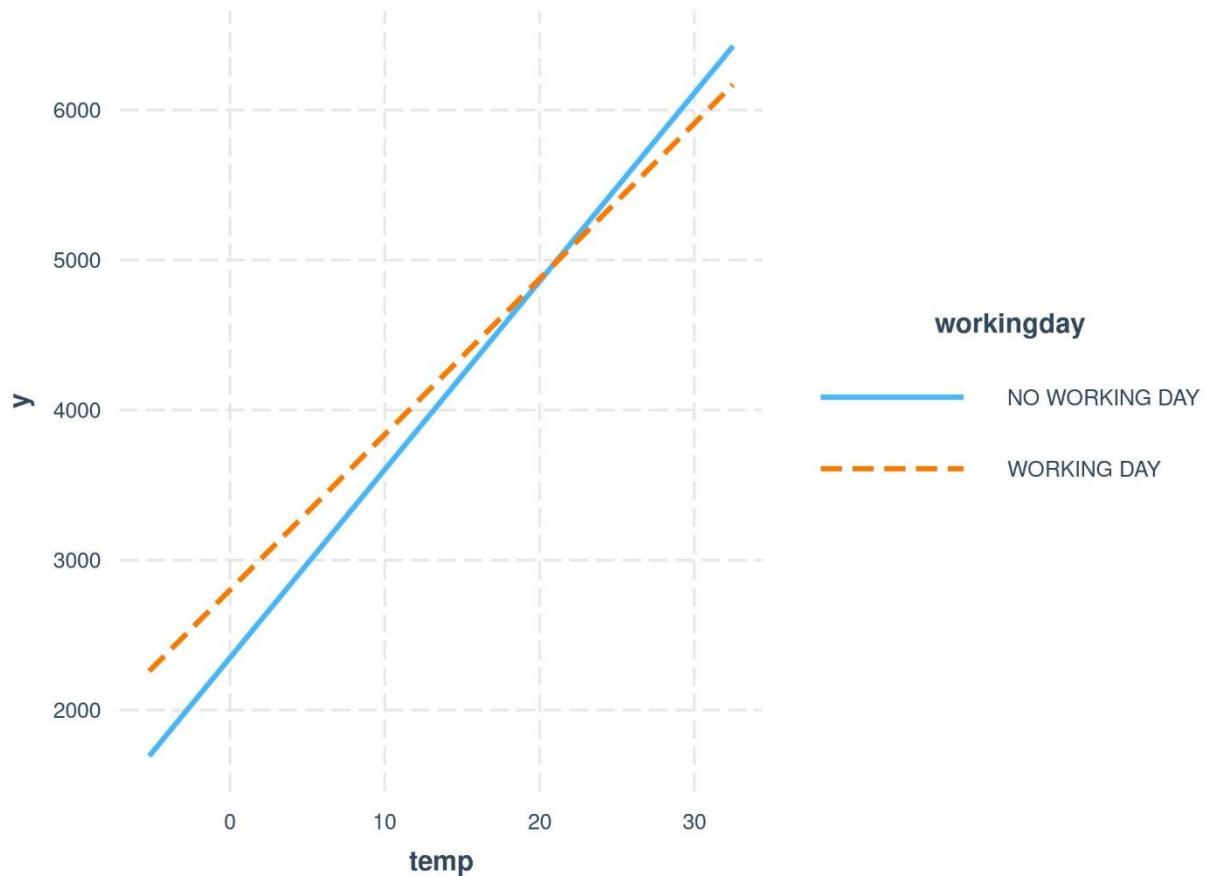
اجازه دهید به کار پیش‌بینی اجاره دوچرخه که قبلاً در فصل مدل خطی مدل‌سازی کردہ‌ایم بازگردیم . این بار، علاوه بر این، تعامل بین دما و ویژگی روز کاری را در نظر می‌گیریم. این منجر به وزن های تخمینی و فواصل اطمینان زیر می‌شود.

۱۹۲۷

	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 متقابل اضافی منفی است ($-21,8$) و به طور قابل توجهی با صفر متفاوت است، همانطور که با فاصله اطمینان 95% نشان داده شده است، که صفر را شامل نمی شود. به هر حال، داده ها id نیستند، زیرا روزهای نزدیک به یکدیگر مستقل از یکدیگر نیستند. فواصل اطمینان ممکن است گمراه کننده باشد، فقط آن را با یک دانه نمک مصرف کنید. اصطلاح تعامل تفسیر وزن ویژگی های درگیر را تغییر می دهد. آیا دما با توجه به اینکه 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 «روز کاری روز کاری: دما» 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 گرمرتر بود چه اتفاقی می افتد. سپس باید هر دو وزن "temp" و "workingday" را جمع کنیم تا تعیین کنیم تخمین چقدر افزایش می یابد.

درک تعامل بصری آسان تر است. با معرفی یک اصطلاح تعاملی بین یک ویژگی طبقه ای و عددی، به جای یک شیب، دو شیب برای دما به دست می آوریم. شیب دما برای روزهایی که افراد مجبور به کار نیستند "NO" ("WORKING DAY") مستقیماً از جدول ($125,4$) قابل خواندن است. شیب دما برای روزهایی که افراد باید در آن کار کنند ("روز کاری") مجموع هر دو وزن دما ($125,4 - 21,8 = 103,6$) است. وقفه خط "NO" در دمای $= 0$ توسط عبارت رهگیری مدل خطی ($2185,8$) تعیین می شود. وقفه خط "روز کاری" در دمای $= 0$ با عبارت رهگیری $+ \text{اثر روز کاری}$ ($451,9 + 2185,8 = 2637,7$) تعیین می شود.



شکل ۱۲: تأثیر (شامل برهمکنش) دما و روز کاری بر تعداد پیش‌بینی شده دوچرخه‌ها برای یک مدل خطی.
به طور موثر، ما دو شیب برای دما دریافت می‌کنیم، یکی برای هر دسته از ویژگی روز کاری.

۵,۳,۳ - GAM جلوه‌های غیر خطی

دنیا خطی نیست. خطی بودن در مدل‌های خطی به این معنی است که صرف نظر از مقداری که یک نمونه در یک ویژگی خاص داشته باشد، افزایش مقدار به اندازه یک واحد همیشه همان اثر را بر نتیجه 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 نمونه‌های کتاب نیز وجود دارد. ویژگی دما تأثیر خطی و مثبتی بر تعداد دوچرخه‌های اجاره‌ای دارد، اما در برخی موقعیت‌ها می‌شود و حتی در دمای بالا تأثیر منفی می‌گذارد. مدل خطی اهمیتی نمی‌دهد،

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

تبدیل ساده ویژگی (مثلا لگاریتم)

۱۹۵۸

دسته بندی ویژگی

۱۹۵۹

۱۹۶۰

مدل های افزایشی تعمیم یافته (GAM)

۱۹۶۱

قبل از اینکه به جزئیات هر روش بپردازم، اجازه دهید با مثالی شروع کنیم که هر سه روش را نشان می دهد.

۱۹۶۲

من مجموعه داده اجاره دوچرخه را گرفتم و یک مدل خطی با فقط ویژگی دما برای پیش‌بینی تعداد

۱۹۶۳

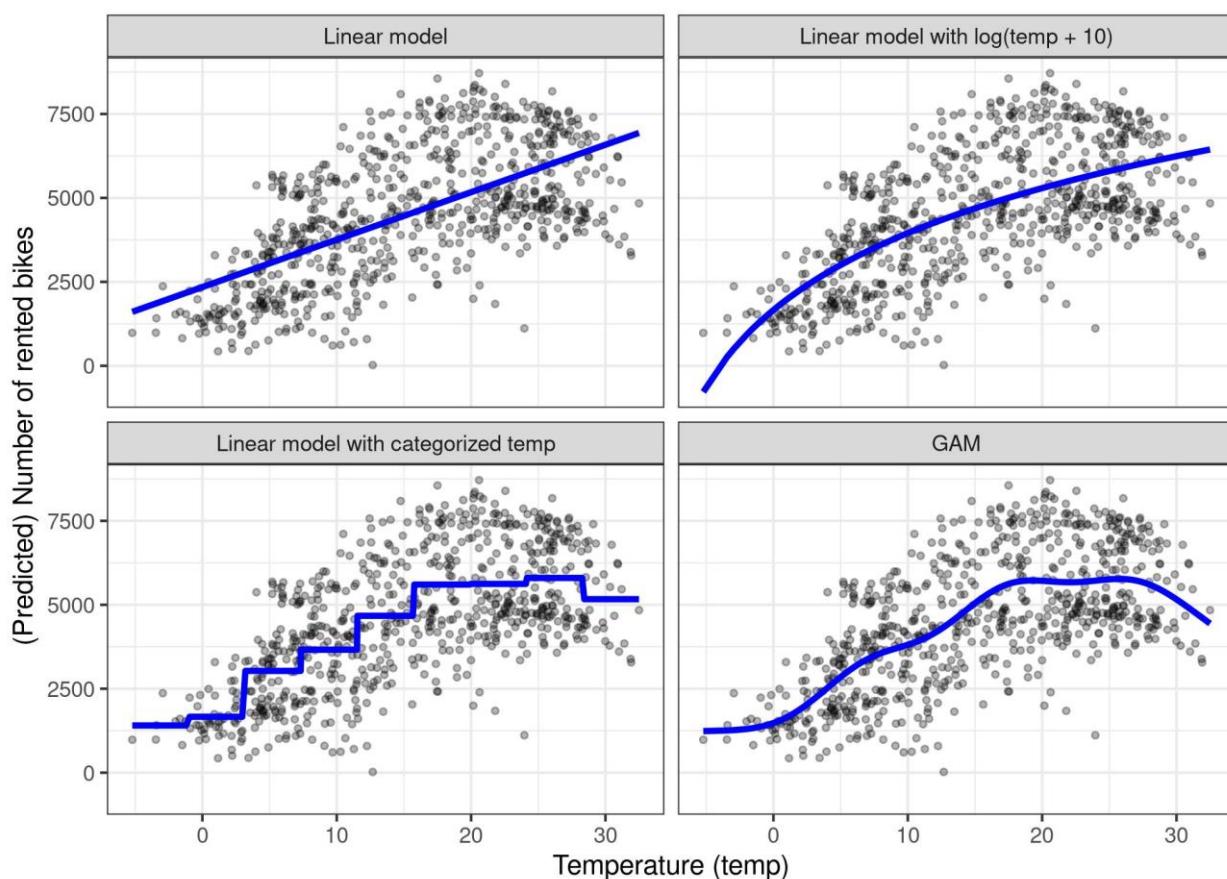
دوچرخه‌های اجاره‌ای آموزش دادم. شکل زیر شبیه برآورد شده را نشان می دهد: مدل خطی استاندارد، مدل

۱۹۶۴

خطی با دمای تبدیل شده (لگاریتم)، مدل خطی با دما به عنوان ویژگی طبقه بندی شده و با استفاده از خطوط

۱۹۶۵

رگرسیون (GAM).



۱۹۶۶

شکل ۱۳: پیش‌بینی تعداد دوچرخه‌های اجاره‌ای تنها با استفاده از ویژگی دما. یک مدل خطی (بالا سمت چپ) به خوبی با داده‌ها مطابقت ندارد. یک راه حل این است که ویژگی را با لگاریتم (بالا سمت راست) تغییر دهید، آن را دسته بندی کنید (پایین سمت چپ)، که معمولاً یک تصمیم اشتباه است، یا استفاده از مدل‌های

۱۹۷۰ افزودنی تعمیم یافته که می تواند به طور خودکار یک منحنی صاف را برای دما تنظیم کند (پایین سمت راست).

۱۹۷۲ تبدیل ویژگی

۱۹۷۳ اغلب از لگاریتم ویژگی به عنوان تبدیل استفاده می شود. استفاده از لگاریتم نشان می دهد که هر 10 برابر

۱۹۷۴ افزایش دما تأثیر خطی یکسانی بر تعداد دوچرخه ها دارد، بنابراین تغییر از 1 درجه سانتیگراد به 10 درجه

۱۹۷۵ سانتیگراد همان تأثیر تغییر از $1,000$ به 1 را دارد (به نظر اشتباہ می رسد). مثال های دیگر برای تبدیل ویژگی

۱۹۷۶ عبارتند از: ریشه مربع،تابع مربع و تابع نمایی. استفاده از تبدیل ویژگی به این معنی است که ستون این ویژگی

۱۹۷۷ را در داده ها با تابعی از ویژگی مانند لگاریتم جایگزین می کنید و طبق معمول مدل خطی را برازش می کنید.

۱۹۷۸ برخی از برنامه های آماری همچنین به شما اجازه می دهند که تبدیل ها را در فراخوانی مدل خطی مشخص

۱۹۷۹ کنید. وقتی ویژگی را تغییر می دهید می توانید خلاق باشید. تفسیر ویژگی با توجه به تبدیل انتخاب شده تغییر

۱۹۸۰ می کند. اگر از تبدیل \log استفاده می کنید، تفسیر در یک مدل خطی به این صورت می شود: "اگر لگاریتم

۱۹۸۱ ویژگی یک افزایش یابد، پیش بینی با وزن مربوطه افزایش می یابد." وقتی از GLM با تابع پیوند استفاده

۱۹۸۲ می کنید که تابع هویت نیست، تفسیر پیچیده تر می شود، زیرا باید هر دو تبدیل را در تفسیر بگنجانید (به جز

۱۹۸۳ زمانی که یکدیگر را لغو می کنند، مانند \log و \exp ، سپس تفسیر راحت تر می شود.)

۱۹۸۴ دسته بندی ویژگی ها

۱۹۸۵ امکان دیگر برای دستیابی به یک اثر غیرخطی، گسسته کردن ویژگی است. آن را به یک ویژگی طبقه بندی

۱۹۸۶ تبدیل کنید. به عنوان مثال، می توانید ویژگی دما را به 20 بازه با سطوح $[10-, 10+, 5-, 5+]$... وغیره برش

۱۹۸۷ دهید. هنگامی که شما از دمای طبقه بندی شده به جای دمای پیوسته استفاده می کنید، مدل خطی یک تابع

۱۹۸۸ گام را تخمین می زند زیرا هر سطح تخمین خاص خود را دارد. مشکل این رویکرد این است که به داده های

۱۹۸۹ بیشتری نیاز دارد، احتمال بیشتری وجود دارد که بیش از حد برازش کند و مشخص نیست که چگونه ویژگی را

۱۹۹۰ به طور معناداری گسسته کنیم (فاصله های مساوی یا چند ک؟ چند بازه؟). من فقط در صورتی از گسسته سازی

۱۹۹۱ استفاده می کنم که یک مورد بسیار قوی برای آن وجود داشته باشد. به عنوان مثال، برای مقایسه مدل با

۱۹۹۲ مطالعه دیگری.

۱۹۹۳ مدل های افزایشی تعمیم یافته (GAM)

۱۹۹۴ چرا به مدل خطی (تعمیم یافته) اجازه نمی دهیم روابط غیرخطی را یاد بگیرد؟ این انگیزه پشت GAM ها

۱۹۹۵ است GAM ها این محدودیت را کاهش می دهند که رابطه باید یک جمع وزنی ساده باشد، و در عوض فرض

می کنند که نتیجه می تواند 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 $\beta_j X_j$ است . با یک عملکرد انعطاف پذیرتر جایگزین می شود . $f_j(x_j)$ هسته یک GAM هنوز مجموع اثرات ویژگی است، اما شما این گزینه را دارید که اجازه دهید روابط غیرخطی بین برخی ویژگی ها و خروجی وجود داشته باشد. افکتهاي خطی نيز توسط چارچوب پوشش داده می شوند، زیرا برای اينكه ویژگیها به صورت خطی مدیريت شوند، می توانيد آنها $f_j(x_j)$ را به شکل β_j محدود کنيد

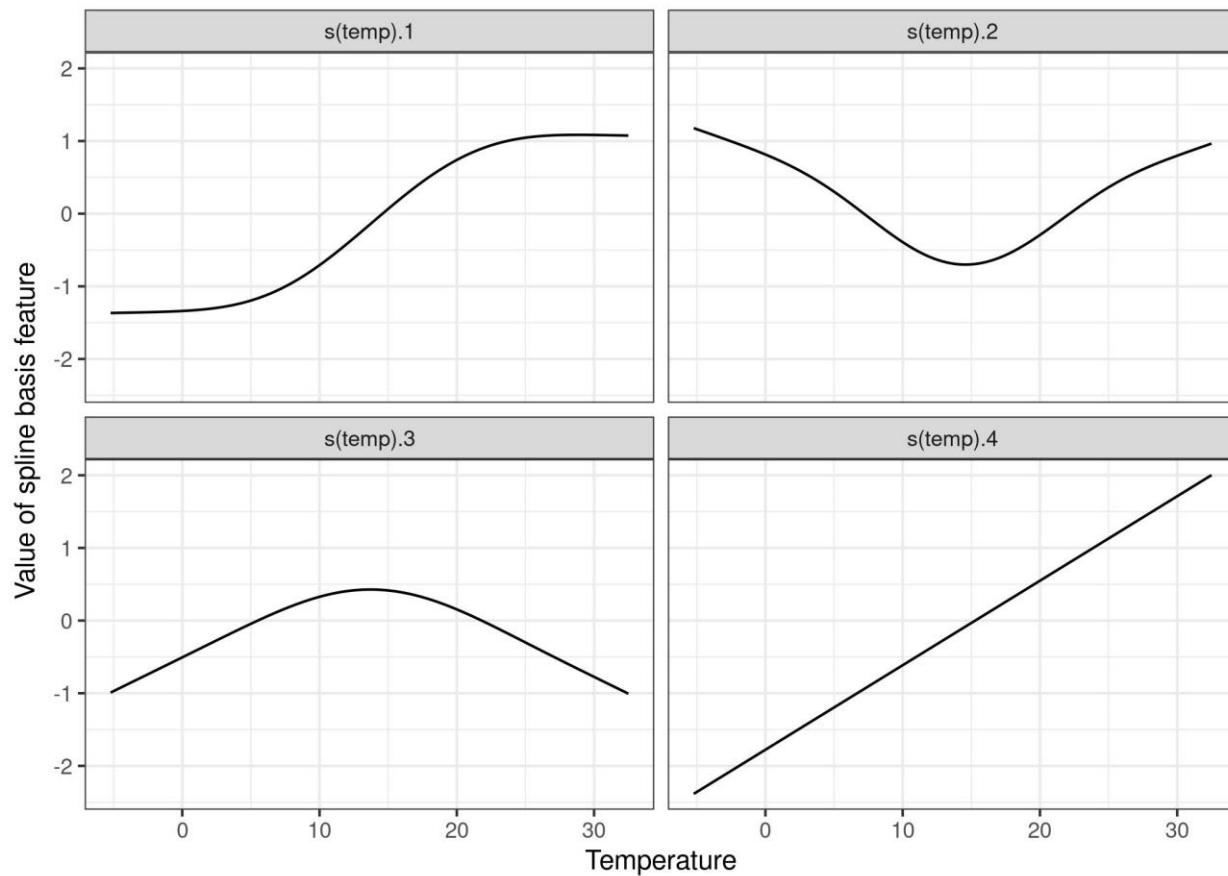
سوال بزرگ اين است که چگونه توابع غیرخطی را ياد بگيريم. به اين پاسخ «اسپلاین» يا «توابع اسپلاین» می گويند Splines. توابعی هستند که از توابع پایه ساده تر ساخته می شوند Splines. را می توان برای تقریب سایر توابع پیچیده تر استفاده کرد. کمی شبیه چیدن آجرهای لگو برای ساختن چیزی پیچیده تر. راه های گیج کننده ای برای تعریف این توابع پایه spline وجود دارد. اگر علاقه مند به کسب اطلاعات بیشتر در مورد تمام روش های تعریف توابع پایه هستید، برای شما در سفرتان آرزوی موفقیت می کنم. 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 spline کرد، تجسم توابع پایه فردی و بررسی چگونگی اصلاح ماتریس داده بود. به عنوان مثال، برای مدل سازی GAM با اسپلاین، ما ویژگی دما را از دادهها حذف می کنیم و مثلًا ۴ ستون را جایگزین آن می کنیم که هر کدام یکتابع پایه spline را نشان می دهند. معمولاً توابع مبتنی بر spline بیشتری خواهید داشت، من فقط برای اهداف تصویری تعداد را کاهش دادم. مقدار هر نمونه از این ویژگی های جدید پایه spline به مقادیر دمای نمونه ها بستگی دارد. همراه با تمام اثرات خطی، GAM سپس این وزن های اسپلاین را نیز تخمین می زند . GAM ها همچنین برای وزنه ها یک اصطلاح جريمه ارائه می کنند تا آنها را نزديک به صفر نگه دارد. اين به طور موثر انعطاف پذيری اسپلاین ها را کاهش می دهد و بيش از حد برازش را کاهش می دهد. سپس یک پارامتر صافی که معمولاً برای کنترل انعطاف پذيری منحنی استفاده می شود، از طریق اعتبارسنجی متقطع تنظیم می شود. با نادیده گرفتن عبارت جريمه، مدل سازی غیرخطی با اسپلاین، مهندسی ویژگی های فانتزی است. هر کدام یکتابع پایه spline را نشان می دهند. معمولاً توابع مبتنی بر spline بیشتری خواهید داشت، من فقط برای اهداف تصویری تعداد را کاهش دادم. مقدار هر نمونه از این ویژگی های جدید پایه spline به مقادیر دمای نمونه ها بستگی دارد. همراه با تمام اثرات خطی، GAM سپس این وزن های اسپلاین را نیز تخمین می زند GAM ها همچنین برای وزنه ها یک اصطلاح جريمه ارائه می کنند تا آنها را نزديک به صفر نگه دارد.

۲۰۲۳ این به طور موثر انعطاف پذیری اسپلاین ها را کاهش می دهد و بیش از حد برازش را کاهش می دهد. سپس
۲۰۲۴ یک پارامتر صافی که معمولاً برای کنترل انعطاف پذیری منحنی استفاده می شود، از طریق اعتبارسنجی متقطع
۲۰۲۵ تنظیم می شود. با نادیده گرفتن عبارت جریمه، مدل سازی غیرخطی با اسپلاین، مهندسی ویژگی های فانتزی
۲۰۲۶ است. هر کدام یک تابع پایه *spline* را نشان می دهنده. معمولاً توابع مبتنی بر *spline* بیشتری خواهید داشت،
۲۰۲۷ من فقط برای اهداف تصویری تعداد را کاهش دادم. مقدار هر نمونه از این ویژگی های جدید پایه *spline* به
۲۰۲۸ مقادیر دمای نمونه ها بستگی دارد. همراه با تمام اثرات خطی، GAM سپس این وزن های اسپلاین را نیز تخمین
۲۰۲۹ می زند GAM ها همچنین برای وزنه ها یک اصطلاح جریمه ارائه می کنند تا آنها را نزدیک به صفر نگه دارد.
۲۰۳۰ این به طور موثر انعطاف پذیری اسپلاین ها را کاهش می دهد و بیش از حد برازش را کاهش می دهد. سپس
۲۰۳۱ یک پارامتر صافی که معمولاً برای کنترل انعطاف پذیری منحنی استفاده می شود، از طریق اعتبارسنجی متقطع
۲۰۳۲ تنظیم می شود. با نادیده گرفتن عبارت جریمه، مدل سازی غیرخطی با اسپلاین، مهندسی ویژگی های فانتزی
۲۰۳۳ است. مقدار هر نمونه از این ویژگی های جدید پایه *spline* به مقادیر دمای نمونه ها بستگی دارد. همراه با تمام
۲۰۳۴ اثرات خطی، GAM سپس این وزن های اسپلاین را نیز تخمین می زند GAM ها همچنین برای وزنه ها یک
۲۰۳۵ اصطلاح جریمه ارائه می کنند تا آنها را نزدیک به صفر نگه دارد. این به طور موثر انعطاف پذیری اسپلاین ها را
۲۰۳۶ کاهش می دهد و بیش از حد برازش را کاهش می دهد. سپس یک پارامتر صافی که معمولاً برای کنترل انعطاف
۲۰۳۷ پذیری منحنی استفاده می شود، از طریق اعتبارسنجی متقطع تنظیم می شود. با نادیده گرفتن عبارت جریمه،
۲۰۳۸ مدل سازی غیرخطی با اسپلاین، مهندسی ویژگی های فانتزی است. مقدار هر نمونه از این ویژگی های جدید پایه
۲۰۳۹ به مقادیر دمای نمونه ها بستگی دارد. همراه با تمام اثرات خطی، GAM سپس این وزن های اسپلاین را
۲۰۴۰ نیز تخمین می زند GAM ها همچنین برای وزنه ها یک اصطلاح جریمه ارائه می کنند تا آنها را نزدیک به صفر
۲۰۴۱ نگه دارد. این به طور موثر انعطاف پذیری اسپلاین ها را کاهش می دهد و بیش از حد برازش را کاهش می دهد.
۲۰۴۲ سپس یک پارامتر صافی که معمولاً برای کنترل انعطاف پذیری منحنی استفاده می شود، از طریق اعتبارسنجی
۲۰۴۳ متقطع تنظیم می شود. با نادیده گرفتن عبارت جریمه، مدل سازی غیرخطی با اسپلاین، مهندسی ویژگی های
۲۰۴۴ فانتزی است. این به طور موثر انعطاف پذیری اسپلاین ها را کاهش می دهد و بیش از حد برازش را کاهش می
۲۰۴۵ دهد. سپس یک پارامتر صافی که معمولاً برای کنترل انعطاف پذیری منحنی استفاده می شود، از طریق
۲۰۴۶ اعتبارسنجی متقطع تنظیم می شود. با نادیده گرفتن عبارت جریمه، مدل سازی غیرخطی با اسپلاین، مهندسی
۲۰۴۷ ویژگی های فانتزی است. این به طور موثر انعطاف پذیری اسپلاین ها را کاهش می دهد و بیش از حد برازش را
۲۰۴۸ کاهش می دهد. سپس یک پارامتر صافی که معمولاً برای کنترل انعطاف پذیری منحنی استفاده می شود، از
۲۰۴۹ طریق اعتبارسنجی متقطع تنظیم می شود. با نادیده گرفتن عبارت جریمه، مدل سازی غیرخطی با اسپلاین،
۲۰۵۰ مهندسی ویژگی های فانتزی است.

۲۰۵۱ در مثالی که ما تعداد دوچرخه‌ها را با GAM فقط با استفاده از دما پیش‌بینی می‌کنیم، ماتریس ویژگی مدل به
 ۲۰۵۲ 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

۲۰۵۳
 ۲۰۵۴ هر ردیف نشان دهنده یک نمونه از داده‌ها (یک روز) است. هر ستون پایه spline حاوی مقدار تابع پایه
 ۲۰۵۵ در مقادیر دمایی خاص است. شکل زیر نشان می‌دهد که این توابع پایه spline چگونه به نظر می‌رسند:
 ۲۰۵۶



۲۰۵۷
 ۲۰۵۸ شکل ۱۴: برای مدل سازی هموار اثر دما، از ۴ تابع پایه spline استفاده می‌کنیم. هر مقدار دما به (اینجا) ۴
 ۲۰۵۹ مقدار پایه spline ترسیم می‌شود. اگر دمای یک نمونه ۳۰ درجه سانتیگراد باشد، مقدار اولین ویژگی پایه
 ۲۰۶۰ spline برای دومی ۷,۰, برای سومی -۸,۰ و برای چهارمین -۱,۷ است.

۲۰۶۱ وزن‌هایی را به هر ویژگی پایه درجه حرارت اختصاص می‌دهد:

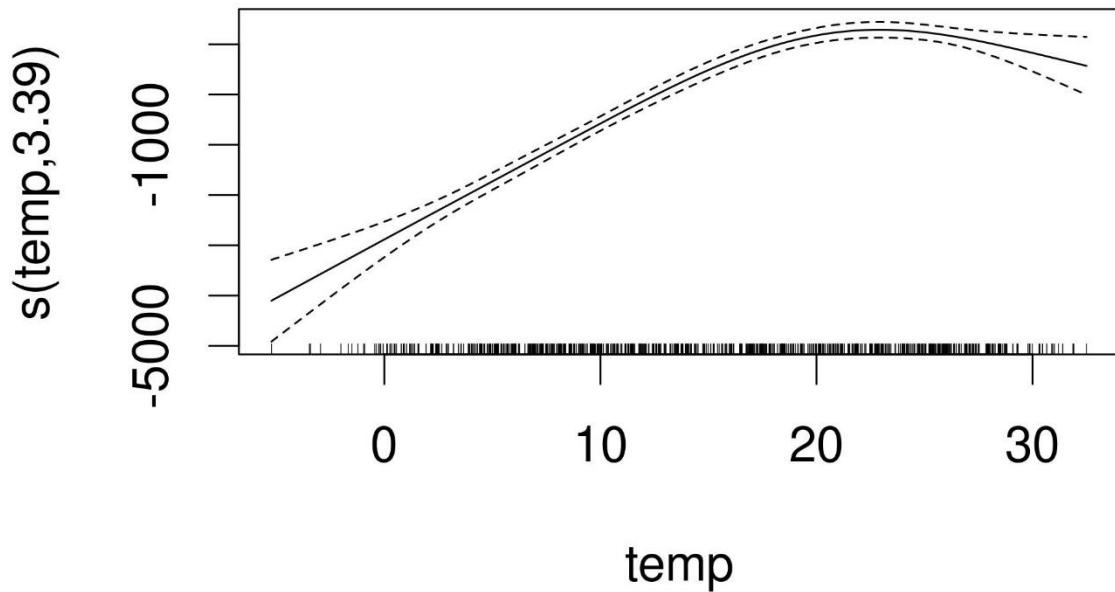
۲۰۶۲

۲۰۶۳

۲۰۶۴

	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 توابع پایه اسپلاین وزن شده با وزن‌های تخمینی حاصل می‌شود، به شکل زیر است:



۲۰۶۵

۲۰۶۶ شکل ۱۵: اثر ویژگی GAM دما برای پیش‌بینی تعداد دوچرخه‌های کرايه شده (دماهی استفاده شده به عنوان
۲۰۶۷ تنها ویژگی).

۲۰۶۸ تفسیر جلوه‌های صاف نیاز به بررسی بصری منحنی برآذش دارد. اسپلاین‌ها معمولاً حول میانگین پیش‌بینی
۲۰۶۹ مرکز می‌شوند، بنابراین یک نقطه در منحنی تفاوت با پیش‌بینی میانگین است. به عنوان مثال، در دماهی
۲۰۷۰ صفر درجه سانتیگراد، تعداد دوچرخه‌های پیش‌بینی شده ۳۰۰۰ کمتر از میانگین پیش‌بینی شده است.

۵,۳,۴ مزایا

۲۰۷۱ همه این پسوندهای مدل خطی به خودی خود کمی جهان هستند. با هر مشکلی که با مدل های خطی مواجه
۲۰۷۲ می شوید، احتمالاً پسوندی خواهد یافت که آن را بطرف می کند.

۲۰۷۴ بیشتر روش ها برای چندین دهه مورد استفاده قرار گرفته اند. به عنوان مثال، GAMها تقریباً ۳۰ سال از
۲۰۷۵ عمرشان می گذرد. بسیاری از محققان و دست اندکاران صنعت با مدل های خطی بسیار با تجربه هستند و این
۲۰۷۶ روش ها در بسیاری از جوامع به عنوان وضعیت موجود برای مدل سازی پذیرفته شده است.

۲۰۷۷ علاوه بر پیش‌بینی، می‌توانید از مدل‌ها برای استنتاج، نتیجه‌گیری در مورد داده‌ها استفاده کنید – با توجه به
۲۰۷۸ اینکه مفروضات مدل نقض نمی‌شوند. شما فواصل اطمینان برای وزن‌ها، آزمون‌های اهمیت، فواصل پیش‌بینی
۲۰۷۹ و موارد دیگر را دریافت می‌کنید.

۲۰۸۰ نرم‌افزارهای آماری معمولاً دارای رابطه‌ای واقعاً خوبی برای تطبیق با GLM، GAM و مدل‌های خطی خاص‌تر
۲۰۸۱ هستند.

۲۰۸۲ کدورت بسیاری از مدل‌های یادگیری ماشین ناشی از ۱) عدم پراکندگی است، به این معنی که بسیاری از
۲۰۸۳ ویژگی‌ها استفاده می‌شوند، ۲) ویژگی‌هایی که به صورت غیرخطی رفتار می‌شوند، به این معنی که برای توصیف
۲۰۸۴ اثر به بیش از یک وزن نیاز دارید، و ۳) مدل سازی تعاملات بین ویژگی‌ها. با فرض اینکه مدل‌های خطی بسیار
۲۰۸۵ قابل تفسیر هستند، اما اغلب با واقعیت مناسب نیستند، پسوندهای شرح داده شده در این فصل راه خوبی برای
۲۰۸۶ دستیابی به یک انتقال هموار به مدل‌های انعطاف‌پذیرتر ارائه می‌دهند، در حالی که برخی از قابلیت تفسیر را
۲۰۸۷ حفظ می‌کنند.

۵,۳,۵ معایب

۲۰۸۹ به عنوان مزیت گفته ام که مدل‌های خطی در جهان خودشان زندگی می‌کنند. تعداد زیادی از روش‌هایی که
۲۰۹۰ می‌توانید مدل خطی ساده را گسترش دهید، نه فقط برای مبتدیان، بسیار زیاد است. در واقع، جهان‌های
۲۰۹۱ موازی متعددی وجود دارد، زیرا بسیاری از جوامع محققان و پژوهشگران نامهای خاص خود را برای روش‌هایی دارند
۲۰۹۲ که کم و بیش یک کار را انجام می‌دهند، که می‌تواند بسیار گیج‌کننده باشد.

۲۰۹۳ اکثر اصلاحات مدل خطی باعث می‌شود که مدل کمتر قابل تفسیر باشد. هرتابع پیوند (در GLM که تابع
۲۰۹۴ هویت نباشد، تفسیر را پیچیده می‌کند. فعل و انفعالات نیز تفسیر را پیچیده می‌کند. جلوه‌های ویژگی
۲۰۹۵ غیرخطی یا کمتر بصری هستند (مانند تبدیل \log یا دیگر نمی‌توان آنها را با یک عدد خلاصه کرد (مثلاً تابع
۲۰۹۶ spline).

۲۰۹۷ GAMها و غیره بر فرضیات مربوط به فرآیند تولید داده تکیه دارند. اگر آنها نقض شوند، دیگر تفسیر
۲۰۹۸ اوزان معتبر نیست.

۲۰۹۹ عملکرد مجموعه های مبتنی بر درخت مانند جنگل تصادفی یا تقویت درخت گرادیان در بسیاری موارد بهتر از
۲۱۰۰ پیچیده ترین مدل های خطی است. این بخشی از تجربه شخصی من و بخشی از مشاهدات از مدل های برنده در
۲۱۰۱ سیستم عامل هایی مانند kaggle.com است.

۲۱۰۲ **۵,۳,۶ نرم افزار**
۲۱۰۳ تمام مثال های این فصل با استفاده از زبان R ایجاد شده اند. برای GAM ها از gam پکیج استفاده شد، اما
۲۱۰۴ موارد دیگر نیز وجود دارد R . دارای تعداد باورنکردنی بسته برای گسترش مدل های رگرسیون خطی است. بدون
۲۱۰۵ پیشی گرفتن از هر زبان تجزیه و تحلیل دیگری، R خانه هر بسط قابل تصوری از پسوند مدل رگرسیون خطی
۲۱۰۶ است. شما پیاده سازی هایی از GAM ها را در پایتون پیدا خواهید کرد (مانند pyGAM) ، اما این پیاده سازی
۲۱۰۷ ها آنقدرها هم بالغ نیستند.

۲۱۰۸ **۵,۳,۷ برنامه های افزودنی بیشتر**
۲۱۰۹ همانطور که وعده داده شده بود، در اینجا لیستی از مشکلاتی که ممکن است در مدل های خطی با آنها مواجه
۲۱۱۰ شوید، همراه با نام راه حلی برای این مشکل وجود دارد که می توانید آن را کپی و در موتور جستجوی مورد
۲۱۱۱ علاقه خود قرار دهید.

۲۱۱۲ داده های من فرض مستقل بودن و توزیع یکسان (iid) را نقض می کند.

۲۱۱۳ به عنوان مثال، اندازه گیری های مکرر روی همان بیمار.

۲۱۱۴ جستجوی مدل های ترکیبی یا معادلات تخمین تعمیم یافته.

۲۱۱۵ مدل من دارای خطاهای هتروسکداستیک است.

۲۱۱۶ به عنوان مثال، هنگام پیش‌بینی ارزش یک خانه، خطاهای مدل معمولاً در خانه‌های گران‌قیمت بیشتر است، که
۲۱۱۷ همسانی مدل خطی را نقض می کند.

۲۱۱۸ **جستجوی رگرسیون قوی**

۲۱۱۹ من نقاط پرت دارم که به شدت بر مدل من تأثیر می گذارد.

۲۱۲۰ **جستجوی رگرسیون قوی**

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 بیفتدند.

جستجو برای مدل های پارامتریک بقا، رگرسیون کاکس، تجزیه و تحلیل بقا.

نتیجه من برای پیش بینی یک مقوله 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 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 شанс متناسب

را جستجو کنید.

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

جستجوی رگرسیون پواسون

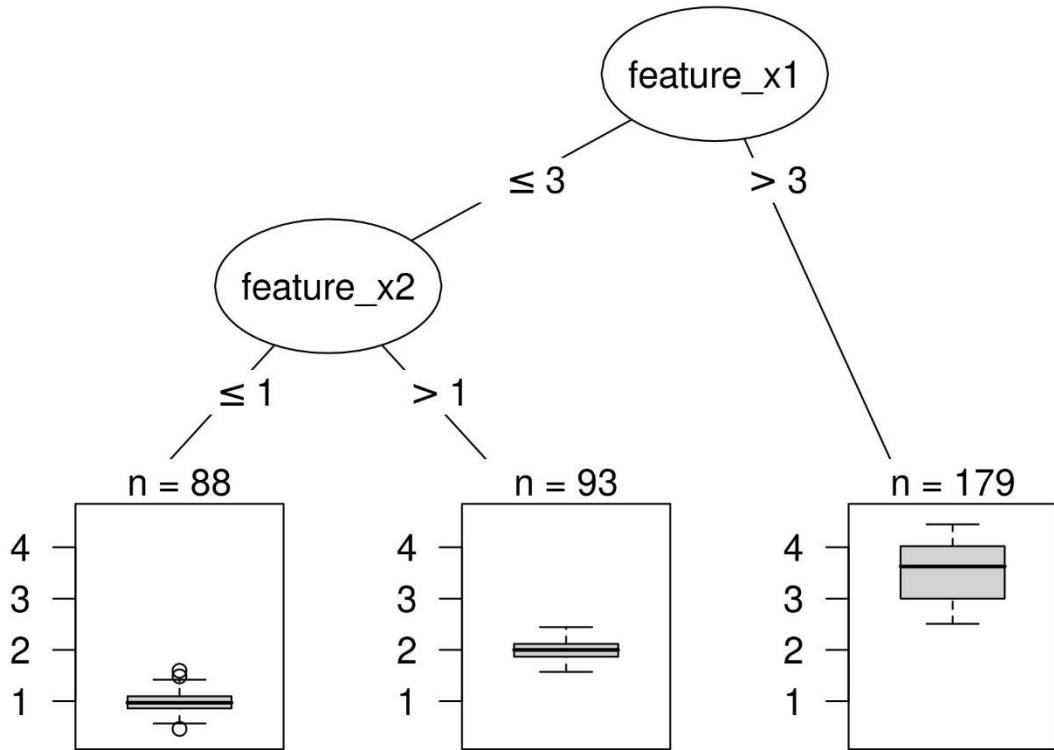
مدل پواسون نیز GLM است. همچنین ممکن است این مشکل را داشته باشد که مقدار شمارش + بسیار مکرر است.

جستجوی رگرسیون پواسون با تورم صفر، مدل مانع.

من مطمئن نیستم که چه ویژگی هایی باید در مدل گنجانده شود تا نتیجه گیری های علی درست انجام شود.

به عنوان مثال، می خواهم بدانم اثر یک دارو بر فشار خون چیست. این دارو بر برخی از ارزش های خونی تأثیر مستقیم دارد و این ارزش خونی بر نتیجه تأثیر می گذارد. آیا باید مقدار خون را در مدل رگرسیون لحاظ کنم؟

- جستجو برای استنتاج علی ، تحلیل میانجیگری.
من داده های گم شده دارم جست و
جو برای انتساب چندگانه
من می خواهم دانش قبلی را در مدل های خود ادغام کنم.
جستجو برای استنتاج بیزی
من اخیراً کمی احساس ضعف دارم.
جستجو برای "Amazon Alexa Gone Wild!!! نسخه کامل از ابتدا تا انتها
۵.۴ درخت تصمیم
مدل های رگرسیون خطی و رگرسیون لجستیک در شرایطی که رابطه بین ویژگی ها و نتیجه غیرخطی است یا
جایی که ویژگی ها با یکدیگر تعامل دارند، شکست می خورند. زمان درخشش برای درخت تصمیم است! مدل های
مبتنی بر درخت، داده ها را چندین بار بر اساس مقادیر قطعی مشخص در ویژگی ها تقسیم می کنند. از طریق
تقسیم، زیرمجموعه های مختلفی از مجموعه داده ایجاد می شود که هر نمونه متعلق به یک زیر مجموعه است.
زیر مجموعه های نهایی را گره های پایانی یا برگ و زیر مجموعه های میانی را گره های داخلی یا گره های
تقسیم می نامند. برای پیش بینی نتیجه در هر گره برگ، از میانگین نتیجه داده های آموزشی در این گره
استفاده می شود. درختان را می توان برای طبقه بندی و رگرسیون استفاده کرد.
الگوریتم های مختلفی وجود دارد که می تواند درخت را رشد دهد. آنها در ساختار احتمالی درخت (به عنوان
مثال تعداد شکاف در هر گره)، معیارهای چگونگی یافتن شکاف ها، زمان توقف تقسیم و نحوه تخمین مدل های
садه در گره های برگ متفاوت هستند. الگوریتم درختان طبقه بندی و رگرسیون (CART) احتمالاً محبوب
ترین الگوریتم برای القای درخت است. ما بر روی CART مرکز خواهیم کرد، اما این تفسیر برای اکثر انواع
درختان مشابه است. من کتاب "عناصر یادگیری آماری (Friedman, Hastie and Tibshirani 2009)"
را برای معرفی دقیق تر CART توصیه می کنم.



۲۱۶۴

شکل ۱۶,۵: درخت تصمیم با داده های مصنوعی. نمونه هایی با مقدار بیشتر از ۳ برای ویژگی x_1 به گره ۵ ختم می شوند. همه نمونه های دیگر بسته به اینکه مقادیر ویژگی x_2 از ۱ بیشتر باشد به گره ۳ یا گره ۴ اختصاص داده می شوند.

۲۱۶۸

فرمول زیر رابطه بین نتیجه ۷ و ویژگی های x را توصیف می کند.

۲۱۶۹

$$\hat{y} = \hat{f}(x) = \sum_{m=1}^M c_m I\{x \in R_m\}$$

هر نمونه دقیقاً در یک گره برگ (= زیر مجموعه R_m) تابع هویتی است که $1 \text{ if } x \in R_m \text{ else } 0$ را برمی گرداند. در زیر مجموعه R_m قرار دارد و $y = c_1$ نتیجه پیش بینی شده است. جایی که c_1 میانگین تمام نمونه های آموزشی در گره برگ است

اما زیر مجموعه ها از کجا می آیند؟ این بسیار ساده است CART: یک ویژگی را می گیرد و تعیین می کند که کدام نقطه برش واریانس ۷ را برای یک کار رگرسیونی یا شاخص جینی توزیع کلاس ۷ برای وظایف طبقه بندی را به حداقل می رساند. واریانس به ما می گوید که مقادیر ۷ در یک گره چقدر در اطراف مقدار میانگین خود

۲۱۷۶ پخش می شوند. شاخص جینی به ما می گوید که یک گره چقدر "ناخالص" است، به عنوان مثال اگر همه
۲۱۷۷ کلاس ها فرکانس یکسانی داشته باشند، گره ناخالص است، اگر فقط یک کلاس وجود داشته باشد، حداقل
۲۱۷۸ خالص است. زمانی که نقاط داده در گره ها مقادیر بسیار مشابهی برای ۷ داشته باشند، واریانس و شاخص جینی
۲۱۷۹ به حداقل می رسد. در نتیجه، بهترین نقطه برش، دو زیرمجموعه حاصل را تا حد ممکن با توجه به نتیجه هدف
۲۱۸۰ متفاوت می کند. برای ویژگی های طبقه بندی، الگوریتم سعی می کند با گروه بندی های مختلف دسته ها،
۲۱۸۱ زیرمجموعه هایی ایجاد کند. پس از تعیین بهترین برش برای هر ویژگی، الگوریتم ویژگی را برای تقسیم که منجر
۲۱۸۲ به بهترین پارامیتر از نظر واریانس یا شاخص جینی می شود انتخاب می کند و این تقسیم را به درخت اضافه
۲۱۸۳ می کند. الگوریتم این جستجو و تقسیم را به صورت بازگشتی در هر دو گره جدید تا رسیدن به یک معیار توقف
۲۱۸۴ ادامه می دهد. معیارهای ممکن عبارتند از: حداقل تعداد نمونه هایی که باید قبل از تقسیم در یک گره باشند، یا
۲۱۸۵ حداقل تعداد نمونه هایی که باید در یک گره پایانه باشند.

۵.۴.۱ تفسیر

۲۱۸۶ تفسیر ساده است: با شروع از گره ریشه، به گره های بعدی می روید و لبه ها به شما می گویند که به کدام زیر
۲۱۸۷ مجموعه ها نگاه می کنید. هنگامی که به گره برگ رسیدید، گره نتیجه پیش بینی شده را به شما می گوید.
۲۱۸۸ تمام لبه ها با "AND" به هم متصل می شوند.

۲۱۸۹ الگو: اگر ویژگی] X کوچکتر از بزرگتر] از آستانه C و ... باشد، نتیجه پیش بینی شده میانگین مقدار ۷ از نمونه های
۲۱۹۰ آن گره است.

۲۱۹۲ اهمیت ویژگی

۲۱۹۳ اهمیت کلی یک ویژگی در یک درخت تصمیم را می توان به روش زیر محاسبه کرد: تمام تقسیم بندی هایی را که
۲۱۹۴ این ویژگی برای آنها استفاده شده است را بررسی کنید و اندازه بگیرید که چقدر واریانس یا شاخص جینی را در
۲۱۹۵ مقایسه با گره والد کاهش داده است. مجموع همه اهمیت ها به ۱۰۰ مقیاس می شود. این بدان معنی است که
۲۱۹۶ هر اهمیت را می توان به عنوان سهمی از اهمیت کلی مدل تفسیر کرد.

۲۱۹۷ تجزیه درخت

۲۱۹۸ پیش بینی های فردی یک درخت تصمیم را می توان با تجزیه مسیر تصمیم به یک جزء در هر ویژگی توضیح
۲۱۹۹ داد. می توانیم یک تصمیم را از طریق درخت ردیابی کنیم و یک پیش بینی را با کمک های اضافه شده در هر گره
۲۲۰۰ تصمیم توضیح دهیم.

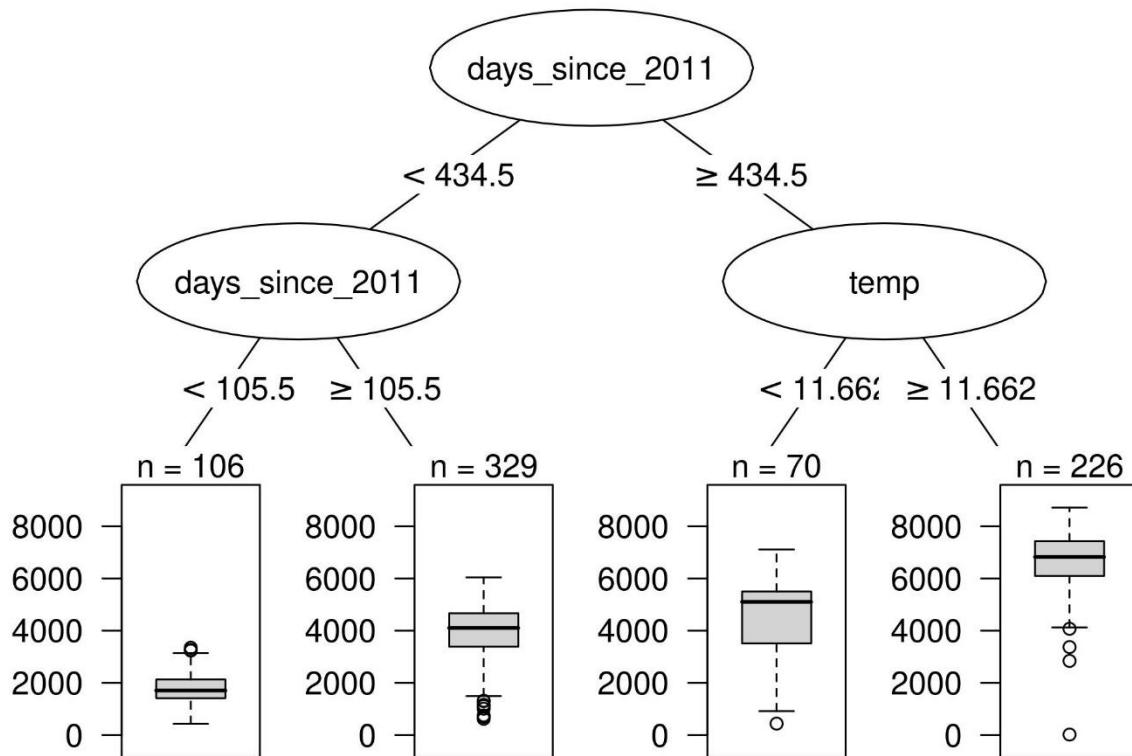
۲۲۰۱ گره ریشه در درخت تصمیم نقطه شروع ما است. اگر بخواهیم از گره ریشه برای پیش بینی استفاده کنیم،
۲۲۰۲ میانگین نتیجه داده های آموزشی را پیش بینی می کند. با تقسیم بعدی، بسته به گره بعدی در مسیر، یا کم می
۲۲۰۳ کنیم یا یک جمله به این جمع اضافه می کنیم. برای رسیدن به پیش بینی نهایی، باید مسیر نمونه داده ای را که
۲۲۰۴ می خواهیم توضیح دهیم دنبال کنیم و مدام به فرمول اضافه کنیم.

$$\hat{f}(x) = \bar{y} + \sum_{d=1}^D \text{split.contrib}(d,x) = \bar{y} + \sum_{j=1}^p \text{feat.contrib}(j,x)$$

۲۲۰۵
۲۲۰۶ پیش بینی یک نمونه منفرد، میانگین نتیجه هدف به اضافه مجموع تمام مشارکت های تقسیم های D است که
۲۲۰۷ بین گره ریشه و گره پایانی که در آن نمونه به پایان می رسد، رخ می دهد. اگرچه ما به مشارکت های تقسیم شده
۲۲۰۸ علاقه ای نداریم، بلکه به مشارکت های ویژگی علاقه مندیم. یک ویژگی ممکن است برای بیش از یک تقسیم
۲۲۰۹ استفاده شود یا اصلاً استفاده نشود. می توانیم مشارکت ها را برای هر یک از ویژگی های p اضافه کنیم و تفسیری
۲۲۱۰ از میزان مشارکت هر ویژگی در یک پیش بینی دریافت کنیم.

۲۲۱۱
۲۲۱۲ اجازه دهید نگاهی دیگر به داده های اجاره دوچرخه داشته باشیم . می خواهیم تعداد دوچرخه های کرایه شده
۲۲۱۳ در یک روز خاص را با درخت تصمیم پیش بینی کنیم. درخت آموخته شده به شکل زیر است:

۵.۴.۲ مثال



۲۲۱۴

۲۲۱۵

شکل ۵,۱۷: درخت رگرسیون بر روی داده های اجاره دوچرخه نصب شده است. حداکثر عمق مجاز برای درخت روی ۲ تنظیم شد. ویژگی روند (روزها از سال ۲۰۱۱) و دما (دما) برای شکاف ها انتخاب شده است. نمودارهای جعبه توزیع تعداد دوچرخه را در گره پایانه نشان می دهد.

۲۲۱۸

۲۲۱۹

تقسیم اول و یکی از تقسیم دوم با ویژگی روند انجام شد که از زمان شروع جمع آوری داده ها شمارش می کند و روندی را پوشش می دهد که سرویس اجاره دوچرخه در طول زمان محبوب تر شده است. برای روزهای قبل از روز ۱۰۵، تعداد دوچرخه های پیش بینی شده حدود ۱۸۰۰ دوچرخه است، بین روزهای ۱۰۶ و ۴۳۰ حدود ۳۹۰۰ است. برای روزهای بعد از روز ۴۳۰، پیش بینی یا ۴۶۰۰ (اگر دما زیر ۱۲ درجه باشد) یا ۶۶۰۰ است.

۲۲۲۰

۲۲۲۱

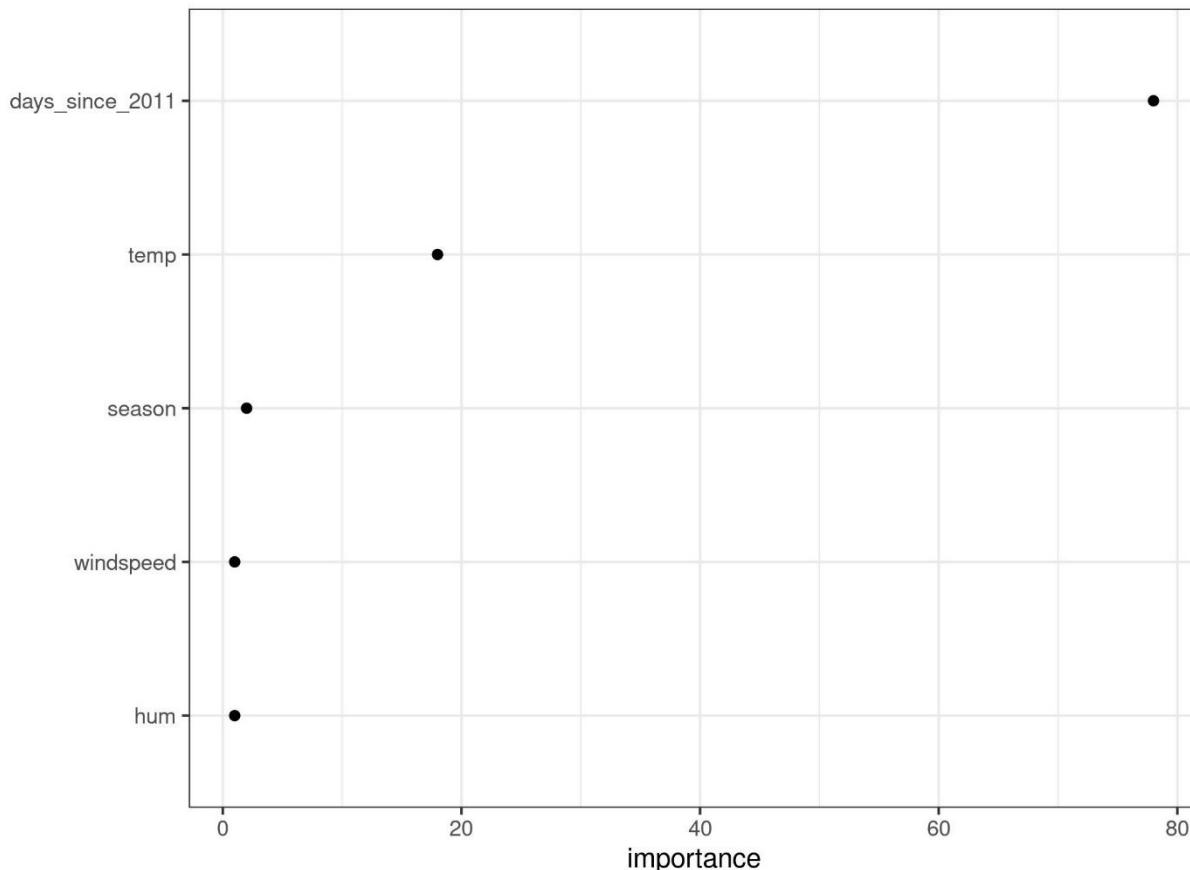
۲۲۲۲

(اگر دما بالای ۱۲ درجه باشد).

۲۲۲۳

اهمیت ویژگی به ما می گوید که یک ویژگی چقدر به بهبود خلوص همه گره ها کمک کرده است. در اینجا، واریانس استفاده شد، زیرا پیش بینی اجاره دوچرخه یک کار رگرسیونی است.

درخت تجسم شده نشان می دهد که هم دما و هم روند زمانی برای شکاف ها استفاده شده است، اما مشخص نمی کند که کدام ویژگی مهم تر است. اندازه گیری اهمیت ویژگی نشان می دهد که روند زمانی بسیار مهم تر از دما است.



شکل ۱۸: اهمیت ویژگی های اندازه گیری شده با میزان بهبود خلوص گره به طور متوسط.

ساختم درختی برای ثبت تعاملات بین ویژگی ها در داده ها ایده آل است.

داده ها به گروه های مجزا ختم می شوند که درک آنها اغلب آسان تر از نقاط روی یک ابر صفحه چند بعدی مانند رگرسیون خطی است. این تفسیر مسلماً بسیار ساده است.

ساختم درختی با گره ها و لبه هایش تجسم طبیعی نیز دارد.

درختان توضیحات خوبی را همانطور که در فصل "توضیحات انسان دوستانه" تعریف شده است ایجاد می کنند.. ساختار درختی به طور خودکار دعوت می کند تا در مورد مقادیر پیش بینی شده برای نمونه های فردی به عنوان

خلاف واقع فکر کنیم: «اگر یک ویژگی بزرگتر/کوچکتر از نقطه تقسیم بود، پیش‌بینی به جای ۷۲، ۱۷ بود.»
توضیحات درخت متضاد هستند، زیرا همیشه می‌توانید پیش‌بینی یک نمونه را با سناریوهای مربوط به «چه می‌شود» (همانطور که توسط درخت تعریف می‌شود) که صرفاً سایر گره‌های برگ درخت هستند، مقایسه کنید.
اگر درخت کوتاه باشد، مانند یک تا سه شکاف عمیق، توضیحات حاصل انتخابی است. درختی با عمق سه به حداقل سه ویژگی و نقطه تقسیم نیاز دارد تا توضیحی برای پیش‌بینی یک نمونه فردی ایجاد کند. صحبت پیش‌بینی به عملکرد پیش‌بینی درخت بستگی دارد. توضیحات مربوط به درختان کوتاه بسیار ساده و کلی است.

نیازی به تغییر ویژگی‌ها نیست. در مدل‌های خطی، گاهی اوقات لازم است لگاریتم یک ویژگی را در نظر بگیرید. یک درخت تصمیم به همان اندازه با هر تبدیل یکنواخت یک ویژگی کار می‌کند.

درختان نمی‌توانند با روابط خطی مقابله کنند. هر رابطه خطی بین یک ویژگی ورودی و نتیجه باید با تقسیم‌بندی تقریبی شود و یک تابع مرحله ایجاد کند. این کارآمد نیست.

این با عدم صافی همراه است. تغییرات جزئی در ویژگی ورودی می‌تواند تأثیر زیادی بر نتیجه پیش‌بینی شده داشته باشد که معمولاً مطلوب نیست. درختی را تصور کنید که ارزش یک خانه را پیش‌بینی می‌کند و درخت از اندازه خانه به عنوان یکی از ویژگی‌های تقسیم استفاده می‌کند. شکاف در ۱۰۰,۵ متر مربع رخ می‌دهد. تصور کنید که کاربر یک براوردگر قیمت خانه را از مدل درخت تصمیم شما استفاده می‌کند: آنها خانه خود را اندازه می‌گیرند، به این نتیجه می‌رسند که خانه ۹۹ متر مربع است، آن را وارد ماشین حساب قیمت می‌کند و ۲۰۰۰۰ یورو را پیش‌بینی می‌کنند. کاربران متوجه شده اند که فراموش کرده اند یک انبار کوچک ۲ متر مربعی را اندازه گیری کنند. اتاق انبار دارای یک دیوار شیبدار است، بنابراین آنها مطمئن نیستند که آیا می‌توانند تمام مساحت یا فقط نیمی از آن را بشمارند. بنابراین آنها تصمیم می‌گیرند هر دو ۱۰۰,۰ و ۱۰۱,۰ متر مربع را امتحان کنند. نتایج: ماشین حساب قیمت خروجی ۲۰۰۰۰ یورو و ۲۰۵۰۰۰ یورو دارد.

درختان نیز کاملاً ناپایدار هستند. چند تغییر در مجموعه داده آموزشی می‌تواند درختی کاملاً متفاوت ایجاد کند. این به این دلیل است که هر تقسیم به تقسیم والد بستگی دارد. و اگر یک ویژگی متفاوت به عنوان اولین ویژگی تقسیم انتخاب شود، کل ساختار درخت تغییر می‌کند. اگر ساختار به این راحتی تغییر کند، اطمینانی در مدل ایجاد نمی‌کند.

درختان تصمیم بسیار قابل تفسیر هستند - تا زمانی که کوتاه باشند. تعداد گره‌های پایانه به سرعت با عمق افزایش می‌یابد. هر چه گره‌های پایانه بیشتر و درخت عمیق‌تر باشد، درک قوانین تصمیم گیری درخت

۵.۴.۴ معایب

۲۲۶۳ دشوارتر می شود. عمق ۱ به معنای ۲ گره پایانه است. عمق ۲ به معنای حداکثر است. ۴ گره، عمق ۳ به معنای
۲۲۶۴ حداکثر است. ۸ گره حداکثر تعداد گره های پایانه در یک درخت ۲ به توان عمق است.

۵,۴,۵ نرم افزار

۲۲۶۵ برای مثال های این فصل، از rpart بسته R استفاده کردم که) CART درخت های طبقه بندی و رگرسیون) را
۲۲۶۶ پیاده سازی می کند CART. در بسیاری از زبان های برنامه نویسی از جمله پایتون پیاده سازی شده است . مسلماً
۲۲۶۷ یک الگوریتم بسیار قدیمی و تا حدودی منسخ شده است و الگوریتم های جدید جالبی برای برآش
۲۲۶۸ CART درختان وجود دارد. می توانید نمای کلی برخی از بسته های R برای درخت های تصمیم گیری را در نمای وظایف
۲۲۶۹ CRAN یادگیری ماشینی و آماری در زیر کلمه کلیدی «پارتیشن بندی بازگشتی» بیابید. در پایتون، بسته
۲۲۷۰ models الگوریتم های مختلفی را برای رشد درخت های تصمیم گیری (به عنوان مثال حریص در مقابل تناسب
۲۲۷۱ بهینه)، هرس درختان و منظم کردن درختان رائه می کند.
۲۲۷۲

۵,۵ قوانین تصمیم گیری

۲۲۷۳ قاعده تصمیم گیری یک دستور IF-THEN ساده است که از یک شرط (همچنین پیشین نامیده می شود) و یک
۲۲۷۴ پیش بینی تشکیل شده است. به عنوان مثال: اگر امروز باران ببارد و اگر آوریل باشد (شرط)، پس فردا باران
۲۲۷۵ خواهد آمد (پیش بینی). یک قانون تصمیم گیری واحد یا ترکیبی از چندین قانون می تواند برای پیش بینی
۲۲۷۶ استفاده شود.
۲۲۷۷

۲۲۷۸ قوانین تصمیم گیری از یک ساختار کلی پیروی می کنند: اگر شرایط برآورده شد، پیش بینی خاصی انجام
۲۲۷۹ دهید. قوانین تصمیم گیری احتمالاً قابل تفسیرترین مدل های پیش بینی هستند. ساختار IF-THEN آنها از
۲۲۸۰ نظر معنایی شبیه زبان طبیعی و طرز تفکر ما است، مشروط بر اینکه شرط از ویژگی های قابل فهم ساخته شده
۲۲۸۱ باشد، طول شرط کوتاه باشد (تعداد کمی از جفت ها با یک AND ترکیب شده است) و قوانین زیادی وجود
۲۲۸۲ ندارد. در برنامه نویسی، نوشتن قوانین IF-THEN بسیار طبیعی است. جدید در یادگیری
۲۲۸۳ ماشین این است که قوانین تصمیم گیری از طریق یک الگوریتم آموخته می شوند.

۲۲۸۴ تصور کنید از یک الگوریتم برای یادگیری قوانین تصمیم گیری برای پیش بینی ارزش یک خانه استفاده کنید)
۲۲۸۵ high, medium low(یک قانون تصمیم گیری که توسط این مدل آموخته می شود می تواند این باشد:
۲۲۸۶ اگر خانه ای بزرگتر از ۱۰۰ متر مربع باشد و دارای باغ باشد، ارزش آن زیاد است. رسمی تر IF size>100 :
۲۲۸۷ AND garden=1 THEN value=high.
۲۲۸۸ اجازه دهید قانون تصمیم را بشکنیم:

۲۲۸۹ شرط اول در بخش IF است. $size>100$

۲۲۹۰ شرط دوم در قسمت garden=1 است.

۲۲۹۱ این دو شرط با یک "AND" متصل می شوند تا یک شرط جدید ایجاد کنند. برای اعمال قانون، هر دو باید
۲۲۹۲ صادق باشند.

۲۲۹۳ نتیجه پیش بینی شده (THEN-part) است. $value=high$

۲۲۹۴ یک قانون تصمیم حداقل از یک feature=value عبارت در شرط استفاده می کند، بدون هیچ محدودیتی در
۲۲۹۵ مورد اینکه چند عبارت دیگر را می توان با «AND» اضافه کرد. یک استثنا قاعده پیش فرض است که بخش IF
۲۲۹۶ صریح ندارد و زمانی اعمال می شود که هیچ قانون دیگری اعمال نشود، اما بعداً در مورد آن بیشتر توضیح
۲۲۹۷ خواهیم داد.

۲۲۹۸ سودمندی یک قانون تصمیم معمولاً در دو عدد خلاصه می شود: پشتیبانی و دقت.

۲۲۹۹ پشتیبانی یا پوشش یک قانون: درصد مواردی که شرط یک قانون در مورد آنها اعمال می شود، حمایت نامیده
۲۳۰۰ می شود. به عنوان مثال قانون size=big AND location=good THEN value=high پیش بینی ارزش
۲۳۰۱ خانه را در نظر بگیرید. فرض کنید ۱۰۰ خانه از ۱۰۰۰ خانه بزرگ و در مکان مناسبی هستند، پس حمایت
۲۳۰۲ قانون ۱۰٪ است. پیش بینی (THEN-part) برای محاسبه پشتیبانی مهم نیست.

۲۳۰۳ دقت یا اطمینان یک قاعده: دقت یک قاعده معیاری است از میزان دقت قاعده در پیش بینی کلاس صحیح برای
۲۳۰۴ نمونه هایی که شرط قاعده در مورد آنها اعمال می شود. به عنوان مثال: فرض کنید از ۱۰۰ خانه که در آن قاعده
۲۳۰۵ صدق می کند، ۸۵ خانه دارند size=big AND location=good THEN value=high
۲۳۰۶ خانه دارند value=medium و ۱ دارای است value=low، سپس دقت قانون ۸۵٪ است.

۲۳۰۷ معمولاً بین دقت و پشتیبانی تعادل وجود دارد: با افزودن ویژگی های بیشتر به شرایط، می توانیم به دقت بالاتری
۲۳۰۸ دست پیدا کنیم، اما پشتیبانی را از دست بدھیم.

۲۳۰۹ برای ایجاد یک طبقه بندی خوب برای پیش بینی ارزش یک خانه، ممکن است لازم باشد نه تنها یک قانون، بلکه
۲۳۱۰ شاید ۲۰ یا ۲۰ را یاد بگیرید. سپس همه چیز پیچیده تر می شود و می توانید با یکی از مشکلات زیر مواجه
۲۳۱۱ شوید:

۲۳۱۲ قوانین می توانند همپوشانی داشته باشند: اگر بخواهم ارزش یک خانه را پیش بینی کنم و دو یا چند قانون
۲۳۱۳ اعمال شوند و پیش بینی های متناقضی به من بدهند، چه؟

- ۲۳۱۴ هیچ قانونی اعمال نمی شود: اگر بخواهم ارزش یک خانه را پیش بینی کنم و هیچ یک از قوانین اعمال نمی شود
- ۲۳۱۵ چه؟
- ۲۳۱۶ دو استراتژی اصلی برای ترکیب قوانین چندگانه وجود دارد: لیست تصمیم (مرتب شده) و مجموعه تصمیم
- ۲۳۱۷ (غیرترتیب). هر دو استراتژی متضمن راه حل های متفاوتی برای مشکل همپوشانی قوانین هستند.
- ۲۳۱۸ فهرست تصمیم گیری، نظمی را به قوانین تصمیم گیری معرفی می کند. اگر شرط قانون اول برای مثال درست
- ۲۳۱۹ باشد، از پیش بینی قانون اول استفاده می کنیم. اگر نه، به قانون بعدی می رویم و بررسی می کنیم که آیا اعمال
- ۲۳۲۰ می شود و غیره. لیست های تصمیم مشکل همپوشانی قوانین را تنها با برگرداندن پیش بینی اولین قانون در
- ۲۳۲۱ لیستی که اعمال می شود، حل می کند.
- ۲۳۲۲ یک مجموعه تصمیم شبیه دموکراسی قوانین است، با این تفاوت که برخی از قوانین ممکن است قدرت رای
- ۲۳۲۳ بالاتری داشته باشند. در یک مجموعه، قوانین یا متقابلاً منحصر به فرد هستند، یا یک استراتژی برای حل
- ۲۳۲۴ تعارض وجود دارد، مانند رأی اکثریت، که ممکن است با دقت قوانین فردی یا سایر معیارهای کیفی وزن شود.
- ۲۳۲۵ تفسیرپذیری به طور بالقوه زمانی آسیب می بیند که چندین قانون اعمال شود.
- ۲۳۲۶ هم لیست های تصمیم گیری و هم مجموعه ها می توانند از این مشکل رنج ببرند که هیچ قانونی برای یک
- ۲۳۲۷ نمونه اعمال نمی شود. این را می توان با معرفی یک قانون پیش فرض حل کرد. قانون پیش فرض قانونی است
- ۲۳۲۸ که زمانی اعمال می شود که هیچ قانون دیگری اعمال نمی شود. پیش بینی قانون پیش فرض اغلب رایج ترین
- ۲۳۲۹ کلاس از نقاط داده است که توسط قوانین دیگر پوشش داده نمی شوند. اگر مجموعه یا فهرستی از قوانین کل
- ۲۳۳۰ فضای ویژگی را پوشش دهد، آن را جامع می نامیم. با افزودن یک قانون پیش فرض، یک مجموعه یا فهرست به
- ۲۳۳۱ طور خودکار جامع می شود.
- ۲۳۳۲ راه های زیادی برای یادگیری قوانین از داده ها وجود دارد و این کتاب به دور از پوشش همه آنها است. این فصل
- ۲۳۳۳ سه مورد از آنها را به شما نشان می دهد. الگوریتم ها برای پوشش طیف وسیعی از ایده های کلی برای یادگیری
- ۲۳۳۴ قوانین انتخاب شده اند، بنابراین هر سه آنها رویکردهای بسیار متفاوتی را نشان می دهند.
- ۲۳۳۵ قوانین را از یک ویژگی می آموزد OneR. با سادگی، قابلیت تفسیر و استفاده از آن به عنوان یک معیار
- ۲۳۳۶ مشخص می شود.
- ۲۳۳۷ پوشش متوالی یک روش کلی است که به طور مکرر قوانین را یاد می گیرد و نقاط داده ای را که توسط قانون
- ۲۳۳۸ جدید پوشش داده شده اند حذف می کند. این روش توسط بسیاری از الگوریتم های یادگیری قوانین استفاده
- ۲۳۳۹ می شود.

۲۳۴۰ لیست قوانین بیزی الگوهای مکرر از پیش استخراج شده را با استفاده از آمار بیزی در یک لیست تصمیم گیری ترکیب می کند. استفاده از الگوهای از پیش استخراج شده یک رویکرد رایج است که توسط بسیاری از الگوریتم های یادگیری قوانین استفاده می شود.

۲۳۴۳ بیایید با ساده ترین رویکرد شروع کنیم: استفاده از بهترین ویژگی برای یادگیری قوانین.

۲۳۴۴ ۵,۵،^۱ یادگیری قوانین از یک ویژگی واحد(OneR) الگوریتم پیشنهاد شده توسط هولت (۱۹۹۳) یکی از ساده ترین الگوریتم های القاء قانون است. از ۲۳۴۵ بین تمام ویژگی ها، OneR یکی را انتخاب می کند که بیشترین اطلاعات را در مورد نتیجه مورد علاقه دارد و ۲۳۴۶ ۲۳۴۷ قوانین تصمیم گیری را از این ویژگی ایجاد می کند.

۲۳۴۸ علیرغم نام OneR که مخفف "One Rule" است، الگوریتم بیش از یک قانون تولید می کند: این در واقع یک ۲۳۴۹ قانون برای هر مقدار ویژگی منحصر به فرد بهترین ویژگی انتخاب شده است. نام بهتر OneFeatureRules ۲۳۵۰ است.

۲۳۵۱ الگوریتم ساده و سریع است:

۲۳۵۲ با انتخاب فواصل مناسب، ویژگی های پیوسته را گسترش کنید.

۲۳۵۳ برای هر ویژگی:

۲۳۵۴ یک جدول متقاطع بین مقادیر ویژگی و نتیجه (مقوله) ایجاد کنید.

۲۳۵۵ برای هر مقدار از ویژگی، یک قانون ایجاد کنید که متدالو ترین کلاس نمونه هایی را که دارای این مقدار ویژگی ۲۳۵۶ خاص هستند را پیش بینی می کند (از جدول متقاطع قابل خواندن است).

۲۳۵۷ خطای کل قوانین مربوط به ویژگی را محاسبه کنید.

۲۳۵۸ ویژگی با کمترین خطای کل را انتخاب کنید.

۲۳۵۹ همیشه تمام نمونه های مجموعه داده را پوشش می دهد، زیرا از تمام سطوح ویژگی انتخاب شده ۲۳۶۰ استفاده می کند. مقادیر از دست رفته را می توان به عنوان یک مقدار ویژگی اضافی در نظر گرفت یا از قبل به ۲۳۶۱ آنها نسبت داده شد.

۲۳۶۲ یک مدل OneR یک درخت تصمیم است که تنها یک تقسیم دارد. تقسیم لزوماً مانند CART باینری نیست، اما ۲۳۶۳ به تعداد مقادیر ویژگی منحصر به فرد بستگی دارد.

۲۳۶۴ اجازه دهد که مثالی نگاه کنیم که چگونه بهترین ویژگی توسط OneR انتخاب می‌شود. جدول زیر مجموعه
 ۲۳۶۵ داده‌های مصنوعی در مورد خانه‌ها را با اطلاعاتی در مورد ارزش، مکان، اندازه و اینکه آیا حیوانات خانگی مجاز
 ۲۳۶۶ هستند نشان می‌دهد. ما علاقه مند به یادگیری یک مدل ساده برای پیش‌بینی ارزش یک خانه هستیم.

location	size	pets	value
good	small	yes	high
good	big	no	high
good	big	no	high
bad	medium	no	medium
good	medium	only cats	medium
good	small	only cats	medium
bad	medium	yes	medium
bad	small	yes	low
bad	medium	yes	low
bad	small	no	low

۲۳۶۷

۲۳۶۸

جداول متقابل را بین هر ویژگی و نتیجه ایجاد می‌کند:

	value=low	value=medium	value=high
location=bad	3	2	0
location=good	0	2	3

۲۳۶۹

۲۳۷۰

	value=low	value=medium	value=high
size=big	0	0	2
size=medium	1	3	0
size=small	2	1	1

۲۳۷۱

۲۳۷۲

۲۳۷۳

۲۳۷۴

۲۳۷۵

۲۳۷۶

۲۳۷۷

۲۳۷۸

	value=low	value=medium	value=high
pets=no	1	1	2
pets=only cats	0	2	0
pets=yes	2	1	1

برای هر ویژگی، ردیف به ردیف جدول را مرور می‌کنیم: هر مقدار ویژگی، قسمت IF یک قانون است. رایج‌ترین کلاس برای نمونه‌هایی با این مقدار ویژگی، پیش‌بینی، قسمت THEN از قانون است. به عنوان مثال، ویژگی اندازه با سطوح small، medium و big منجر به سه قانون می‌شود. برای هر ویژگی، نرخ کل خطای قوانین تولید شده را محاسبه می‌کنیم که مجموع خطاهای است. ویژگی مکان دارای مقادیر ممکن good و bad. بیشترین مقدار برای خانه‌هایی که در مکان‌های بد قرار دارند این است low و وقتی به عنوان پیش‌بینی استفاده می‌کنیم low، دو اشتباه مرتکب می‌شویم، زیرا دو خانه دارای medium ارزش هستند. ارزش پیش‌بینی شده خانه‌ها در مکان‌های خوب است high و دوباره ما دو اشتباه می‌کنیم، زیرا دو خانه دارای a

۲۳۷۹ هستند medium ارزش. خطای ما با استفاده از ویژگی مکان $\frac{4}{4}$ ، برای ویژگی اندازه $10/3$ و برای ویژگی pet $10/4$ است. ویژگی اندازه قوانین را با کمترین خطا تولید می کند و برای مدل نهایی OneR استفاده می شود:

```
IF size=small THEN value=low
IF size=medium THEN value=medium
IF size=big THEN value=high
```

۲۳۸۱

۲۳۸۲ ویژگی هایی را با سطوح ممکن زیاد ترجیح می دهد، زیرا این ویژگی ها می توانند راحت تر به هدف اضافه شوند. مجموعه داده ای را تصور کنید که فقط شامل نویز و بدون سیگنال باشد، به این معنی که همه ویژگی ها مقادیر تصادفی می گیرند و هیچ ارزش پیش بینی کننده ای برای هدف ندارند. برخی از ویژگی ها سطوح بیشتری نسبت به سایرین دارند. ویژگی های با سطوح بیشتر اکنون می توانند به راحتی بیش از حد بازش شوند.

۲۳۸۳

۲۳۸۴ یک ویژگی که یک سطح جداگانه برای هر نمونه از داده ها دارد، می تواند کل مجموعه داده آموزشی را کاملاً

۲۳۸۵ پیش بینی کند. یک راه حل این است که داده ها را به مجموعه های آموزشی و اعتبار سنجی تقسیم کنید،

۲۳۸۶ قوانین مربوط به داده های آموزشی را یاد بگیرید و کل خطا را برای انتخاب ویژگی در مجموعه اعتبار سنجی

۲۳۸۷ ارزیابی کنید.

۲۳۸۸

۲۳۸۹

۲۳۹۰ کراوات مسئله دیگری است، یعنی زمانی که دو ویژگی منجر به یک خطای کل یکسان شوند OneR پیوندها را

۲۳۹۱ با استفاده از اولین ویژگی با کمترین خطا یا ویژگی با کمترین p-value در یک تست مجذور کای حل می

۲۳۹۲ کند.

۲۳۹۳ مثال

۲۳۹۴ اجازه دهد OneR را با داده های واقعی امتحان کنیم. ما از کار طبقه بندی سرطان دهانه رحم برای آزمایش

۲۳۹۵ الگوریتم OneR استفاده می کنیم. همه ویژگی های ورودی پیوسته در ۵ کمیت خود گستته شدند. قوانین زیر

۲۳۹۶ ایجاد می شود:

Age	prediction
(12.9,27.2]	Healthy
(27.2,41.4]	Healthy
(41.4,55.6]	Healthy
(55.6,69.8]	Healthy
(69.8,84.1]	Healthy

۲۳۹۷

۲۳۹۸ ویژگی سن توسط OneR به عنوان بهترین ویژگی پیش بینی انتخاب شد. از آنجایی که سرطان نادر است، برای

۲۳۹۹ هر قانون، طبقه اکثریت و بنابراین برچسب پیش بینی شده همیشه سالم است، که نسبتاً مفید نیست. استفاده از

۲۴۰۰ پیش بینی برچسب در این مورد نامتعادل منطقی نیست. جدول متقاطع بین فواصل سنی و سرطان/سالم همراه

۲۴۰۱ با درصد زنان مبتلا به سرطان آموزنده تر است:

	# Cancer	# Healthy	P(Cancer)
Age=(12.9,27.2]	26	477	0.05
Age=(27.2,41.4]	25	290	0.08
Age=(41.4,55.6]	4	31	0.11
Age=(55.6,69.8]	0	1	0.00
Age=(69.8,84.1]	0	4	0.00

۲۴۰۲

اما قبل از شروع تفسیر هر چیزی: از آنجایی که پیش‌بینی هر ویژگی و هر مقدار سالم است، میزان خطای کل برای همه ویژگی‌ها یکسان است. پیوندهای خطای کل، به طور پیش‌فرض، با استفاده از اولین ویژگی از ویژگی‌هایی که کمترین میزان خطای دارند (در اینجا، همه ویژگی‌ها دارای $858/55$ هستند)، که اتفاقاً ویژگی Age است، حل می‌شود.

از وظایف رگرسیون پشتیبانی نمی‌کند. اما می‌توانیم یک کار رگرسیون را با برش دادن نتیجه پیوسته به فواصل به یک کار طبقه‌بندی تبدیل کنیم. ما از این ترفند برای پیش‌بینی تعداد دوچرخه‌های اجاره‌ای با استفاده می‌کنیم و تعداد دوچرخه‌ها را به چهار چارک ($0\%-25\%$ ، $25\%-50\%$ ، $50\%-75\%$ و $75\%-100\%$) تقسیم می‌کنیم. جدول زیر ویژگی انتخاب شده را پس از برازش مدل OneR نشان می‌دهد:

mnth	prediction
JAN	[22,3152]
FEB	[22,3152]
MAR	[22,3152]
APR	(3152,4548]
MAY	(5956,8714]
JUN	(4548,5956]
JUL	(5956,8714]
AUG	(5956,8714]
SEP	(5956,8714]
OCT	(5956,8714]
NOV	(3152,4548]
DEC	[22,3152]

۲۴۱۱

ویژگی انتخاب شده ماه است. ویژگی ماه دارای ۱۲ سطح ویژگی (تعجب!) است که بیشتر از بسیاری از ویژگی‌های دیگر است. بنابراین خطر بیش از حد برازش وجود دارد. در جنبه خوش بینانه تر: ویژگی ماه می‌تواند روند فصلی را کنترل کند (مثلاً دوچرخه‌های اجاره کمتر در زمستان) و پیش‌بینی‌ها معقول به نظر می‌رسند.

اکنون از الگوریتم ساده OneR به رویه‌ای پیچیده‌تر با استفاده از قوانین با شرایط پیچیده‌تر مشکل از چندین ویژگی حرکت می‌کنیم: پوشش متوالی.

پوشش متوالی یک روش کلی است که به طور مکرر یک قانون واحد را برای ایجاد یک لیست تصمیم‌گیری (یا مجموعه) می‌آموزد که کل قاعده به قانون را پوشش می‌دهد. بسیاری از الگوریتم‌های یادگیری قوانین، انواعی

۵.۵.۲ پوشش متوالی

از الگوریتم پوشش متوالی هستند. این فصل دستور اصلی را معرفی می‌کند و از RIPPER، نوعی از الگوریتم پوشش متوالی برای مثال‌ها استفاده می‌کند.

ایده ساده است: ابتدا یک قانون خوب پیدا کنید که برای برخی از نقاط داده اعمال می‌شود. تمام نقاط داده ای که توسط قانون پوشش داده شده اند را حذف کنید. یک نقطه داده زمانی پوشش داده می‌شود که شرایط اعمال شود، صرف نظر از اینکه آیا نقاط به درستی طبقه بندی شده اند یا خیر. آموزش قوانین و حذف نقاط تحت پوشش را با نقاط باقیمانده تکرار کنید تا زمانی که امتیاز دیگری باقی نماند یا شرط توقف دیگری برآورده شود. نتیجه یک لیست تصمیم‌گیری است. این رویکرد یادگیری مکرر قوانین و حذف نقاط داده تحت پوشش "جدا کردن و تسخیر" نامیده می‌شود.

فرض کنید ما قبلًا یک الگوریتم داریم که می‌تواند یک قانون واحد ایجاد کند که بخشی از داده‌ها را پوشش دهد. الگوریتم پوشش متوالی برای دو کلاس (یکی مثبت، یکی منفی) به این صورت عمل می‌کند: با یک لیست خالی از قوانین (rlist) شروع کنید.

یک قانون را یاد بگیرید.

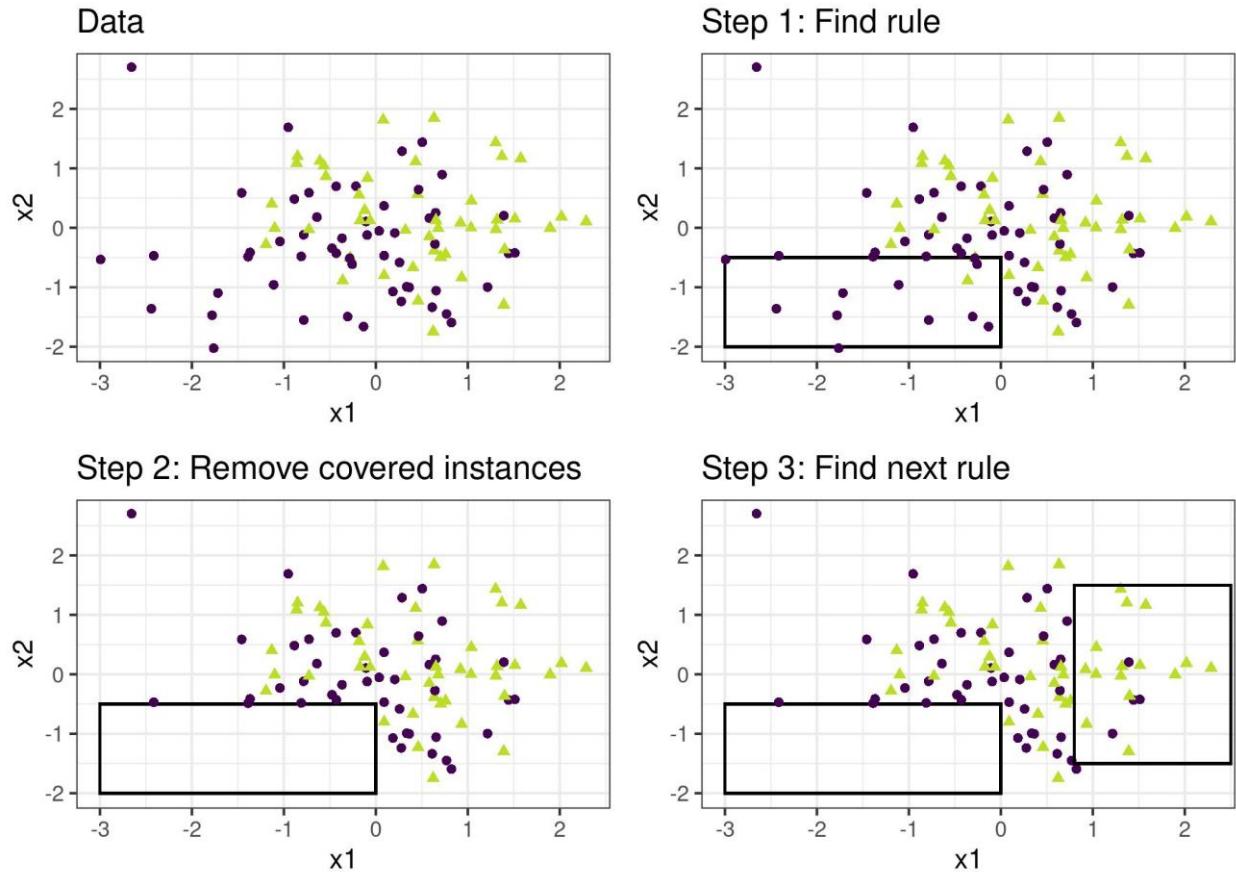
در حالی که لیست قوانین زیر یک آستانه کیفیت مشخص است (یا نمونه‌های مثبت هنوز پوشش داده نشده اند:)

قانون ۲ را به rlist اضافه کنید.

تمام نقاط داده تحت پوشش قانون ۲ را حذف کنید.

قانون دیگری را در مورد داده‌های باقی مانده بیاموزید.

لیست تصمیم را برگردانید.



۲۴۳۸

۲۴۳۹

شکل ۵.۱۹: الگوریتم پوشش با پوشش متواالی فضای ویژگی با قوانین واحد و حذف نقاط داده ای که قبلًا توسط آن قوانین پوشش داده شده اند، کار می کند. برای اهداف تجسم، ویژگی های x_1 و x_2 پیوسته هستند، اما اکثر الگوریتم های یادگیری قوانین به ویژگی های طبقه بندی نیاز دارند.

۲۴۴۲

۲۴۴۳

۲۴۴۴

۲۴۴۵

۲۴۴۶

۲۴۴۷

به عنوان مثال: ما یک کار و مجموعه داده برای پیش بینی مقادیر خانه ها از اندازه، مکان و اینکه آیا حیوانات خانگی مجاز هستند یا خیر داریم. قانون اول را یاد می گیریم که معلوم می شود: اگر `size=big` و `value=high`, سپس همه خانه های بزرگ در مکان های خوب را از مجموعه داده حذف می کنیم. با داده های باقی مانده، قانون بعدی را یاد می گیریم. شاید: اگر `location=good`, پس توجه داشته باشید که این قانون روی داده ها بدون خانه های بزرگ در مکان های خوب آموخته می شود و تنها خانه های متوسط و کوچک در مکان های خوب باقی می مانند.

۲۴۴۸

۲۴۴۹

۲۴۵۰

برای تنظیمات چند کلاسه، رویکرد باید اصلاح شود. اول، طبقات با افزایش شیوع مرتب می شوند. الگوریتم پوشش متواالی با کمترین کلاس شروع می شود، یک قانون برای آن می آموزد، تمام نمونه های تحت پوشش را حذف می کند، سپس به دومین کلاس کم معمول و غیره می رود. کلاس فعلی همیشه به عنوان طبقه مثبت در

نظر گرفته می شود و تمام طبقات با شیوع بالاتر در طبقه منفی ترکیب می شوند. آخرین کلاس قانون پیش فرض است. در طبقه بندی به این استراتژی یک در مقابل همه نیز گفته می شود.

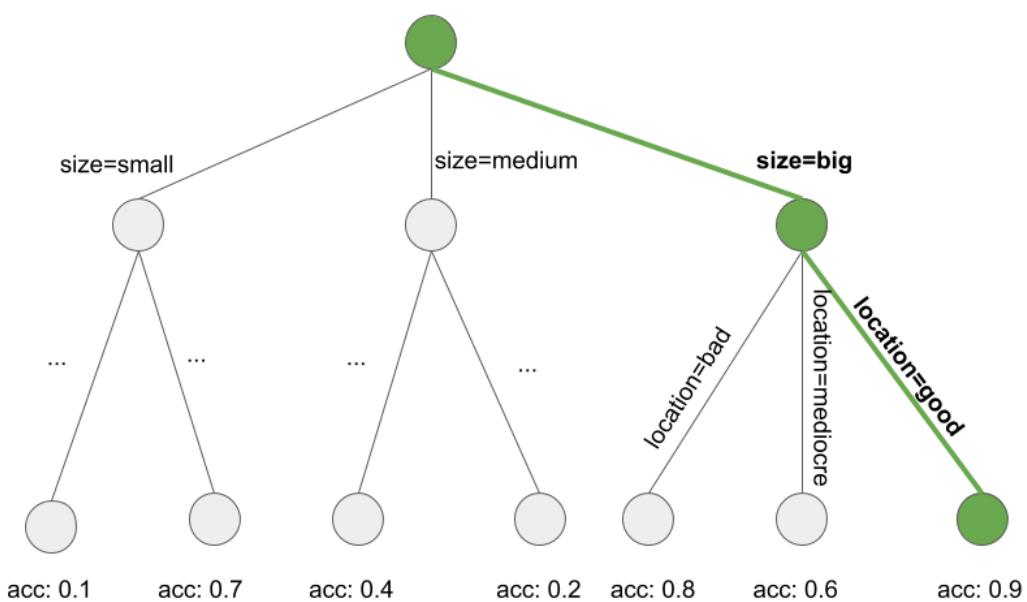
چگونه یک قانون واحد را یاد بگیریم؟ الگوریتم OneR در اینجا بی فایده خواهد بود، زیرا همیشه کل فضای ویژگی را پوشش می دهد. اما بسیاری از احتمالات دیگر وجود دارد. یک امکان یادگیری یک قانون واحد از درخت تصمیم با جستجوی پرتو است:

یک درخت تصمیم را بیاموزید (با الگوریتم CART یا الگوریتم یادگیری درختی دیگر).

از گره ریشه شروع کنید و به صورت بازگشتی خالص ترین گره را انتخاب کنید (مثلاً با کمترین نرخ طبقه بندی اشتباه).

کلاس اکثریت گره پایانه به عنوان پیش‌بینی قانون استفاده می شود. مسیر منتهی به آن گره به عنوان شرط قانون استفاده می شود.

شکل زیر جستجوی پرتو در یک درخت را نشان می دهد:



شکل ۵,۲۰: یادگیری یک قانون با جستجوی یک مسیر از طریق درخت تصمیم. درخت تصمیم برای پیش‌بینی هدف مورد نظر رشد می کند. ما از گره ریشه شروع می کنیم، حریصانه و مکرر مسیری را دنبال می کنیم که به صورت محلی خالص ترین زیرمجموعه را تولید می کند (به عنوان مثال بالاترین دقت) و تمام مقادیر تقسیم را

۲۴۶۶ به شرط قانون اضافه می کنیم. به این نتیجه می رسیم: اگر **size=big** و **location=good** ، آنگاه
۲۴۶۷ **value=high**.

۲۴۶۸ یادگیری یک قانون یک مشکل جستجو است، جایی که فضای جستجو فضای همه قوانین ممکن است. هدف از
۲۴۶۹ جستجو یافتن بهترین قانون بر اساس برخی معیارها است. استراتژی‌های جستجوی مختلفی وجود دارد:
۲۴۷۰ تپه‌نوردی، جستجوی پرتو، جستجوی جامع، جستجوی سفارشی، جستجوی تصادفی،
۲۴۷۱ جستجوی بالا به پایین، جستجوی پایین به بالا، ...

۲۴۷۲ هرس RAPPER افزایشی مکرر برای تولید کاهش خطاب) توسط کوهن (۱۹۹۵) ۲۰ گونه ای از الگوریتم پوشش
۲۴۷۳ متواالی است RAPPER. کمی پیچیده تر است و از یک مرحله پس پردازش (هرس قانون) برای بهینه سازی
۲۴۷۴ لیست تصمیم گیری (یا مجموعه) استفاده می کند RAPPER. می تواند در حالت مرتب یا نامرتب اجرا شود و
۲۴۷۵ یک لیست تصمیم یا مجموعه تصمیم ایجاد کند.

۲۴۷۶ مثال ها

۲۴۷۷ ما از RAPPER برای مثال استفاده خواهیم کرد.
۲۴۷۸ الگوریتم RAPPER هیچ قانونی را در طبقه بندی سلطان دهانه رحم پیدا نمی کند.

۲۴۷۹ هنگامی که از RAPPER در کار رگرسیون برای پیش‌بینی تعداد دوچرخه استفاده می‌کنیم ، قوانینی پیدا
۲۴۸۰ می‌شوند. از آنجایی که RAPPER فقط برای طبقه بندی کار می‌کند، شمارش دوچرخه باید به یک نتیجه طبقه
۲۴۸۱ بندی شود. من با کاهش شمارش دوچرخه به چارک به این امر دست یافتم. به عنوان مثال (۵۹۵۶، ۴۵۴۸)
۲۴۸۲ فاصله ای است که تعداد دوچرخه های پیش‌بینی شده بین ۴۵۴۸ و ۵۹۵۶ را پوشش می‌دهد. جدول زیر
۲۴۸۳ لیست تصمیم گیری قوانین آموخته شده را نشان می‌دهد.

rules
(temp >= 16) and (days_since_2011 <= 437) and (weathersit = GOOD) and (temp <= 24) and (days_since_2011 >= 131) => cnt=(4548,5956] (temp <= 13) and (days_since_2011 <= 111) => cnt=[22,3152] (temp <= 4) and (workingday = NO WORKING DAY) => cnt=[22,3152] (season = WINTER) and (days_since_2011 <= 368) => cnt=[22,3152] (hum >= 72) and (windspeed >= 16) and (days_since_2011 <= 381) and (temp <= 17) => cnt=[22,3152] (temp <= 6) and (weathersit = MISTY) => cnt=[22,3152] (hum >= 91) => cnt=[22,3152] (mnth = NOV) and (days_since_2011 >= 327) => cnt=[22,3152] (days_since_2011 >= 438) and (weathersit = GOOD) and (hum >= 51) => cnt=(5956,8714] (days_since_2011 >= 441) and (hum <= 73) and (temp >= 15) => cnt=(5956,8714] (days_since_2011 >= 441) and (windspeed <= 10) => cnt=(5956,8714] (days_since_2011 >= 455) and (hum <= 40) => cnt=(5956,8714] => cnt=(3152,4548]

۲۴۸۴

تفسیر ساده است: اگر شرایط اعمال شود، فاصله سمت راست را برای تعداد دوچرخه ها پیش بینی می کنیم. آخرین قانون، قانون پیش فرض است که زمانی اعمال می شود که هیچ یک از قوانین دیگر در مورد یک نمونه اعمال نمی شود. برای پیش بینی یک نمونه جدید، از بالای لیست شروع کنید و بررسی کنید که آیا یک قانون اعمال می شود یا خیر. هنگامی که یک شرط مطابقت دارد، سمت راست قانون پیش بینی این مثال است. قانون پیش فرض تضمین می کند که همیشه یک پیش بینی وجود دارد.

۲۴۹۰

۵.۵.۳ فهرست قوانین بیزی

در این بخش، من روش دیگری را برای یادگیری یک لیست تصمیم گیری به شما نشان می دهم که از این دستور تقریبی پیروی می کند:

الگوهای مکرر را از داده ها استخراج کنید که می توانند به عنوان شرایط قوانین تصمیم گیری استفاده شوند. از مجموعه ای از قوانین از پیش استخراج شده، فهرست تصمیم گیری را بیاموزید.

یک رویکرد خاص با استفاده از این دستور، لیست قوانین بیزی 21 (Letham et al., 2015) یا به اختصار BRL نامیده می شود. از آمار بیزی برای یادگیری لیست های تصمیم گیری از الگوهای مکرر استفاده می کند که از قبل با FP-tree استخراج شده اند (بورگلت ۲۰۰۵)

اما اجازه دهید به آرامی با اولین مرحله BRL شروع کنیم.

پیش استخراج الگوهای مکرر

یک الگوی مکرر، وقوع مکرر (همزمان) مقادیر ویژگی است. به عنوان یک مرحله پیش پردازش برای الگوریتم BRL، ما از ویژگی ها استفاده می کنیم (در این مرحله به نتیجه هدف نیاز نداریم) و الگوهای متدال را از آنها استخراج می کنیم. یک الگو می تواند یک مقدار ویژگی منفرد مانند $size=medium$ یا ترکیبی از مقادیر $size=medium$ AND $location=bad$ ویژگی مانند.

فرکانس یک الگو با پشتیبانی آن در مجموعه داده اندازه گیری می شود:

$$Support(x_j = A) = \frac{1}{n} \sum_{i=1}^n I(x_j^{(i)} = A)$$

که در آن A مقدار ویژگی، n تعداد نقاط داده در مجموعه داده و I تابع نشانگر است که در صورت ویژگی، ۱ را برمی گرداند اگر x_j برای مثال A سطح A دارد و در غیر این صورت ۰ است. در مجموعه داده ای از مقادیر خانه، اگر ۲۰٪ خانه ها فاقد بالکن و ۸۰٪ دارای یک یا چند باشند، آنگاه پشتیبانی از الگوی ۲۰٪ است. پشتیبانی همچنین می تواند برای ترکیبی از مقادیر ویژگی اندازه گیری شود، به عنوان مثال balcony=0. balcony=0 AND pets=allowed.

الگوریتم های زیادی برای یافتن چنین الگوهای مکرر وجود دارد، به عنوان مثال Apriori یا FP-Growth. این که شما از آن استفاده می کنید اهمیت زیادی ندارد، فقط سرعت یافتن الگوها متفاوت است، اما الگوهای حاصل همیشه یکسان هستند.

من یک ایده تقریبی از نحوه عملکرد الگوریتم Apriori برای یافتن الگوهای مکرر به شما ارائه خواهم داد. در واقع الگوریتم Apriori از دو بخش تشکیل شده است که بخش اول الگوهای مکرر را پیدا می کند و بخش دوم قوانین تداعی را از آنها می سازد. برای الگوریتم BRL، ما فقط به الگوهای مکرر علاقه مندیم که در قسمت اول تولید می شوند. Apriori

در مرحله اول، الگوریتم Apriori با تمام مقادیر ویژگی که پشتیبانی بیشتر از حداقل پشتیبانی تعريف شده توسط کاربر دارند، شروع می شود. اگر کاربر بگوید که حداقل پشتیبانی باید ۱۰٪ باشد و فقط ۵٪ از خانه ها دارند $size=big$ ، ما آن مقدار ویژگی را حذف می کنیم و فقط $size=small$ و $size=medium$ به عنوان الگو نگه می داریم. این به این معنی نیست که خانه ها از داده ها حذف می شوند، فقط به این معنی است که $size=big$ به عنوان الگوی مکرر برگردانده نمی شوند. بر اساس الگوهای مکرر با یک مقدار ویژگی واحد، الگوریتم Apriori به طور مکرر سعی می کند ترکیبی از مقادیر ویژگی با مرتبه بالاتر را بیابد. الگوها با ترکیب عبارات با یک AND منطقی ساخته می شوند، به عنوان مثال $size=medium$ AND $feature=value$ الگوهای تولید شده با ساپورت زیر حداقل ساپورت حذف می شوند. در پایان ما همه الگوهای $location=bad$.

۲۵۲۶ مکرر را داریم، هر زیر مجموعه ای از بندهای الگوی مکرر دوباره مکرر است که به آن ویژگی Apriori می
۲۵۲۷ گویند. به طور شهودی منطقی است: با حذف یک شرط از یک الگو، الگوی کاهشیافته فقط می‌تواند تعداد
۲۵۲۸ بیشتری یا همان تعداد نقاط داده را پوشش دهد، اما نه کمتر. به عنوان مثال، اگر ۲۰٪ از خانه‌ها هستند
۲۵۲۹ بیشتری برای کاهش تعداد الگوهای مورد بازرگاری که فقط هستند ۲۰٪ size=medium and location=good
۲۵۳۰ بیشتر است. ویژگی Apriori برای کاهش استفاده می‌شود. فقط در مورد الگوهای
۲۵۳۱ مکرر باید الگوهای مرتبه بالاتر را بررسی کنیم.

۲۵۳۲ اکنون ما با شرایط پیش کاوی برای الگوریتم فهرست قوانین بیزی به پایان رسیده ایم. اما قبل از اینکه به مرحله
۲۵۳۳ دوم BRL برویم، می‌خواهم به روش دیگری برای قوانین بر اساس الگوهای از پیش استخراج شده اشاره
۲۵۳۴ کنم. روش‌های دیگر پیشنهاد می‌کنند که نتیجه مورد علاقه را در فرآیند الگوکاوی مکرر و همچنین اجرای
۲۵۳۵ بخش دوم الگوریتم Apriori که قوانین IF-THEN را ایجاد می‌کند، شامل شود. از آنجایی که الگوریتم نظارت
۲۵۳۶ نشده است، قسمت THEN نیز حاوی مقادیر ویژگی است که ما به آنها علاقه ای نداریم. اما می‌توانیم با قوانینی
۲۵۳۷ فیلتر کنیم که فقط نتیجه مورد علاقه در قسمت THEN را دارند. این قوانین قبلًاً یک مجموعه تصمیم را
۲۵۳۸ تشکیل می‌دهند، اما تنظیم، هرس، حذف یا ترکیب مجدد قوانین نیز امکان پذیر است.

۲۵۳۹ با این حال، در رویکرد BRL، ما با الگوهای مکرر کار می‌کنیم و قسمت THEN و نحوه چیدمان الگوها را در
۲۵۴۰ لیست تصمیم گیری با استفاده از آمار بیزی یاد می‌گیریم.

۲۵۴۱ آموزش لیست قوانین بیزی

۲۵۴۲ هدف الگوریتم BRL یادگیری یک لیست تصمیم گیری دقیق با استفاده از منتخبی از پیش استخراج
۲۵۴۳ شده و در عین حال اولویت بندی لیست هایی با قوانین کم و شرایط کوتاه است BRL. با تعریف توزیعی از
۲۵۴۴ لیست های تصمیم گیری با توزیع های قبلی برای طول شرایط (ترجیحاً قوانین کوتاه تر) و تعداد قوانین
۲۵۴۵ (ترجیحاً یک لیست کوتاه تر) به این هدف می‌پردازد.

۲۵۴۶ توزیع احتمال پسینی لیست ها این امکان را فراهم می‌کند که با توجه به مفروضات کوتاه بودن و تناسب
۲۵۴۷ فهرست با داده ها، بگوییم که فهرست تصمیم چقدر محتمل است. هدف ما یافتن لیستی است که این احتمال
۲۵۴۸ پسین را به حداقل می‌رساند. از آنجایی که یافتن دقیق بهترین لیست مستقیماً از توزیع لیست ها امکان پذیر
۲۵۴۹ نیست، BRL دستور زیر را پیشنهاد می‌کند:

۲۵۵۰ ۱ (یک لیست تصمیم اولیه ایجاد کنید که به طور تصادفی از توزیع پیشینی گرفته می‌شود.

۲۵۵۱ ۲ (به طور مکرر لیست را با افزودن، جابجایی یا حذف قوانین تغییر دهید، و اطمینان حاصل کنید که لیست
۲۵۵۲ های حاصل از توزیع پسین لیست ها پیروی می کند.

۲۵۵۳ ۳ (لیست تصمیم گیری را از لیست های نمونه برداری شده با بیشترین احتمال با توجه به توزیع پسینی انتخاب
۲۵۵۴ کنید.

۲۵۵۵ اجازه دهید الگوریتم را دقیق تر بررسی کنیم: الگوریتم با الگوهای ارزش ویژگی قبل از استخراج با الگوریتم-FP
۲۵۵۶ شروع می شود Growth. چندین فرض را در مورد توزیع هدف و توزیع پارامترهایی که توزیع هدف را
۲۵۵۷ تعریف می کند، ایجاد می کند. (این آمار بیزی است). اگر با آمار بیزی آشنایی ندارید، زیاد در گیر توضیحات زیر
۲۵۵۸ نباشد. دانستن این نکته مهم است که رویکرد بیزی راهی برای ترکیب دانش یا الزامات موجود (به اصطلاح
۲۵۵۹ توزیع های پیشینی) و در عین حال متناسب با داده ها است. در مورد لیست های تصمیم، رویکرد بیزی منطقی
۲۵۶۰ است، زیرا مفروضات قبلی فهرست های تصمیم را به کوتاهی با قوانین کوتاه ترغیب می کند.

۲۵۶۱ هدف نمونه برداری از لیست های تصمیم گیری d از توزیع پسینی است:

$$\underbrace{p(d|x, y, A, \alpha, \lambda, \eta)}_{posteriori} \propto \underbrace{p(y|x, d, \alpha)}_{likelihood} \cdot \underbrace{p(d|A, \lambda, \eta)}_{priori}$$

۲۵۶۲ که در آن d یک لیست تصمیم گیری است، x ویژگی ها، y هدف، A مجموعه شرایط از پیش استخراج شده
۲۵۶۳ است، λ طول مورد انتظار قبلی لیست های تصمیم گیری، α تعداد شرایط مورد انتظار قبلی در یک قانون،
۲۵۶۴ η شبه شمار قبلی برای کلاس های مثبت و منفی که به بهترین وجه در $(1, 1)$ ثابت می شود.

$$p(d|x, y, A, \alpha, \lambda, \eta)$$

۲۵۶۷ با توجه به داده های مشاهده شده و مفروضات پیشینی، فهرست تصمیم گیری چقدر محتمل است. این با
۲۵۶۸ احتمال نتیجه با توجه به لیست تصمیم و داده ضربدر احتمال لیست مفروضات قبلی و شرایط از پیش تعیین
۲۵۶۹ شده متناسب است.

$$p(y|x, d, \alpha)$$

۲۵۷۱ با توجه به لیست تصمیم گیری و داده ها، احتمال y مشاهده شده است BRL. فرض می کند که y توسط توزیع
۲۵۷۲ Dirichlet-Multinomial تولید می شود. هرچه لیست تصمیم d داده ها را بهتر توضیح دهد، احتمال
۲۵۷۳ بیشتری دارد.

$$p(d|A, \lambda, \eta)$$

توزيع قبلی لیست های تصمیم گیری است. به طور ضربی یک توزیع پواسون کوتاه شده (پارامتر λ برای تعداد قوانین موجود در لیست و توزیع پواسون کوتاه شده (پارامتر η برای تعداد مقادیر ویژگی در شرایط قوانین).

یک لیست تصمیم گیری اگر نتیجه را به خوبی توضیح دهد و همچنین بر اساس مفروضات قبلی محتمل باشد، احتمال پسین بالایی دارد.

تخمین ها در آمار بیزی همیشه کمی دشوار است، زیرا ما معمولاً نمی توانیم پاسخ صحیح را مستقیماً محاسبه کنیم، اما باید نامزدها را ترسیم کنیم، آنها را ارزیابی کنیم و تخمین های پسینی خود را با استفاده از روش مونت کارلو زنجیره مارکوف به روز کنیم. برای لیست های تصمیم گیری، این حتی دشوارتر است، زیرا ما باید از توزیع لیست های تصمیم گیری استفاده کنیم. نویسندهای BRL پیشنهاد می کنند که ابتدا یک لیست تصمیم گیری اولیه ترسیم شود و سپس به طور مکرر آن را اصلاح کنند تا نمونه هایی از لیست های تصمیم گیری از توزیع پسین لیست ها (یک زنجیره مارکوف از لیست های تصمیم گیری) تولید شود. نتایج به طور بالقوه به لیست تصمیم گیری اولیه بستگی دارد، بنابراین توصیه می شود این روش را تکرار کنید تا از تنوع زیادی از لیست ها اطمینان حاصل شود. پیش فرض در اجرای نرم افزار ۱۰ بار است. دستور العمل زیر به ما می گوید که چگونه یک لیست تصمیم اولیه ترسیم کنیم:

الگوهای پیش از استخراج با FP-Growth.

پارامتر طول لیست m را از توزیع پواسون کوتاه شده نمونه بگیرید.

برای قانون پیش فرض: پارامتر توزیع دیریکله-چند جمله ای را نمونه برداری کنید θ .

از مقدار هدف (یعنی قانونی که زمانی اعمال می شود که هیچ چیز دیگری اعمال نمی شود.)

برای قاعده فهرست تصمیم $j=1, \dots, m$ ، انجام دهید:

از پارامتر طول قانون (تعداد شرایط) برای قانون J نمونه بگیرید.

یک شرط طول را نمونه کنید J

از شرایط از پیش استخراج شده

از پارامتر توزیع دیریکله-چند جمله ای برای قسمت THEN نمونه برداری کنید (یعنی برای توزیع نتیجه هدف با توجه به قانون)

برای هر مشاهده در مجموعه داده:

۲۶۰۰ قانون را از لیست تصمیم گیری که ابتدا اعمال می شود (بالا به پایین) بیابید.

۲۶۰۱ نتیجه پیش‌بینی شده را از توزیع احتمال (دو جمله‌ای) که توسط قانون اعمال می‌شود، ترسیم کنید.

۲۶۰۲ گام بعدی تولید بسیاری از لیست‌های جدید است که از این نمونه اولیه شروع می‌شود تا نمونه‌های زیادی از
۲۶۰۳ توزیع پسین لیست‌های تصمیم‌گیری به دست آید.

۲۶۰۴ لیست‌های تصمیم‌گیری جدید با شروع از لیست اولیه و سپس به طور تصادفی یا انتقال یک قانون به موقعیت
۲۶۰۵ دیگری در لیست یا اضافه کردن یک قانون به لیست تصمیم فعلی از شرایط از پیش استخراج شده یا حذف یک
۲۶۰۶ قانون از لیست تصمیم، نمونه برداری می‌شوند. کدام یک از قوانین تغییر، اضافه یا حذف شده است به طور
۲۶۰۷ تصادفی انتخاب می‌شود. در هر مرحله، الگوریتم احتمال پسینی لیست تصمیم (ترکیبی از دقت و کوتاهی) را
۲۶۰۸ ارزیابی می‌کند. الگوریتم Metropolis Hastings تصمین می‌کند که ما از لیست‌های تصمیم‌گیری نمونه
۲۶۰۹ برداری می‌کنیم که احتمال بالایی دارند. این روش نمونه‌های زیادی از توزیع لیست‌های تصمیم را در اختیار
۲۶۱۰ ما قرار می‌دهد. الگوریتم BRL لیست تصمیم‌گیری از نمونه‌هایی را با بیشترین احتمال پسین انتخاب می‌کند.

۲۶۱۱ مثال‌ها

۲۶۱۲ این در مورد تئوری است، اکنون بباید روش BRL را در عمل ببینیم. نمونه‌ها از نوع سریع‌تری از BRL به نام
۲۶۱۳ فهرست‌های قوانین بیزی مقیاس‌پذیر (SBRL) توسط یانگ و همکاران استفاده می‌کنند. (۲۰۱۷) ۲۳ . ما از
۲۶۱۴ الگوریتم SBRL برای پیش‌بینی خطر ابتلا به سرطان دهانه رحم استفاده می‌کنیم . ابتدا مجبور شدم تمام
۲۶۱۵ ویژگی‌های ورودی را برای کارکرد الگوریتم SBRL گستته کنم. برای این منظور، من ویژگی‌های پیوسته را بر
۲۶۱۶ اساس فراوانی مقادیر بر حسب چندک پیوند زدم.

۲۶۱۷ ما قوانین زیر را دریافت می‌کنیم:

rules
If {STDs=1} (rule[259]) then positive probability = 0.16049383
else if {Hormonal.Contraceptives.years.=[0,10)} (rule[82]) then
positive probability = 0.04685408
else (default rule) then positive probability = 0.27777778

۲۶۱۸ توجه داشته باشید که ما قوانین معقولی دریافت می‌کنیم، زیرا پیش‌بینی در قسمت THEN نتیجه کلاس
۲۶۱۹ نیست، بلکه احتمال پیش‌بینی شده برای سرطان است.

۲۶۲۱ شرایط از الگوهایی که از قبل با الگوریتم FP-Growth استخراج شده بودند انتخاب شدند. جدول زیر مجموعه
 ۲۶۲۲ شرایطی را نشان می دهد که الگوریتم SBRL می تواند از بین آنها برای ساخت یک لیست تصمیم انتخاب کند.
 ۲۶۲۳ حداکثر تعداد مقادیر ویژگی در شرایطی که من به عنوان کاربر اجازه دادم دو عدد بود. در اینجا نمونه ای از ده
 ۲۶۲۴ الگو وجود دارد:

pre-mined conditions
Num.of.pregnancies=[3.67, 7.33)
IUD=0, STDs=1
Number.of.sexual.partners=[1, 10),
STDs..Time.since.last.diagnosis=[1, 8)
First.sexual.intercourse=[10, 17.3), STDs=0
Smokes=1, IUD..years.=[0, 6.33)
Hormonal.Contraceptives..years.=[10, 20),
STDs..Number.of.diagnosis=[0, 1)
Age=[13, 36.7)
Hormonal.Contraceptives=1, STDs..Number.of.diagnosis=[0, 1)
Number.of.sexual.partners=[1, 10), STDs..number.=[0, 1.33)
STDs..number.=[1.33, 2.67), STDs..Time.since.first.diagnosis=[1, 8)

۲۶۲۵

۲۶۲۶ سپس، الگوریتم SBRL را برای پیش‌بینی اجاره دوچرخه اعمال می‌کنیم . این تنها در صورتی کار می کند که
 ۲۶۲۷ مشکل رگرسیون پیش بینی تعداد دوچرخه به یک کار طبقه بندی باینری تبدیل شود. من به طور خودسرانه با
 ۲۶۲۸ ایجاد یک برچسب که ۱ است اگر تعداد دوچرخه ها از ۴۰۰۰ دوچرخه در روز بیشتر شود، یک کار طبقه بندی
 ۲۶۲۹ ایجاد کرده ام، در غیر این صورت ۰.

۲۶۳۰ لیست زیر توسط SBRL آموخته شد:

rules
If {yr=2011,temp=[-5.22,7.35)} (rule[718]) then positive probability = 0.01041667
else if {yr=2012,temp=[7.35,19.9)} (rule[823]) then positive probability = 0.88125000
else if {yr=2012,temp=[19.9,32.5]} (rule[816]) then positive probability = 0.99253731
else if {season=SPRING} (rule[351]) then positive probability = 0.06410256
else if {temp=[7.35,19.9)} (rule[489]) then positive probability = 0.44444444
else (default rule) then positive probability = 0.79746835

۲۶۳۱

۲۶۳۲ اجازه دهید پیش بینی کنیم که تعداد دوچرخه ها برای یک روز در سال ۲۰۱۲ با دمای ۱۷ درجه سانتیگراد از
 ۲۶۳۳ ۴۰۰۰ عبور کند. قانون اول اعمال نمی شود، زیرا فقط برای روزهای سال ۲۰۱۱ اعمال می شود. قانون دوم
 ۲۶۳۴ اعمال می شود، زیرا روز در سال ۲۰۱۲ است و ۱۷ درجه در فاصله زمانی قرار دارد [۷,۳۵,۱۹,۹]. پیش‌بینی ما
 ۲۶۳۵ برای احتمال اجاره بیش از ۴۰۰۰ دوچرخه ۸۸ درصد است.

۲۶۳۶ ۵,۵,۴ مزايا

۲۶۳۷ این بخش به طور کلی مزایای قوانین IF-THEN را مورد بحث قرار می دهد.

- ۲۶۳۸ قوانین IF-TEN به راحتی قابل تفسیر هستند . آنها احتمالاً قابل تفسیرترین مدل‌های قابل تفسیر هستند. این
۲۶۳۹ بیانیه فقط در صورتی اعمال می‌شود که تعداد قوانین کم باشد، شرایط قوانین کوتاه باشند (حداکثر ۳ می‌توانم
۲۶۴۰ بگویم) و اگر قوانین در یک لیست تصمیم‌گیری یا مجموعه تصمیم‌گیری غیر همپوشانی سازماندهی شده باشند.
۲۶۴۱
- ۲۶۴۲ قوانین تصمیم‌گیری می‌توانند به اندازه درخت تصمیم‌گویا باشند، در حالی که فشرده‌تر هستند . درخت‌های
۲۶۴۳ تصمیم اغلب از درخت‌های فرعی تکراری نیز رنج می‌برند، یعنی زمانی که شکاف‌ها در یک گره فرزند چپ و
۲۶۴۴ راست ساختار یکسانی دارند.
- ۲۶۴۵ پیش‌بینی با قواعد IF-THEN سریع است ، زیرا برای تعیین اینکه کدام قواعد اعمال می‌شوند، فقط باید چند
۲۶۴۶ عبارت باینری بررسی شوند.
- ۲۶۴۷ قوانین تصمیم‌گیری در برابر تبدیل یکنواخت ویژگی‌های ورودی قوی هستند، زیرا فقط آستانه در شرایط تغییر
۲۶۴۸ می‌کند. آنها همچنین در برابر موارد پرت قوی هستند، زیرا فقط مهم است که یک شرط اعمال شود یا نه.
- ۲۶۴۹ قوانین IF-THEN معمولاً مدل‌های پراکنده تولید می‌کنند، به این معنی که ویژگی‌های زیادی گنجانده نشده
۲۶۵۰ است. آنها فقط ویژگی‌های مربوطه را برای مدل انتخاب می‌کنند. به عنوان مثال، یک مدل خطی به طور پیش
۲۶۵۱ فرض وزنی را به هر ویژگی ورودی اختصاص می‌دهد. ویژگی‌هایی که نامربوط هستند را می‌توان به سادگی با
۲۶۵۲ قوانین IF-THEN نادیده گرفت.
- ۲۶۵۳ قوانین ساده مانند OneR را می‌توان به عنوان خط پایه برای الگوریتم‌های پیچیده تر استفاده کرد.
- ۲۶۵۴ **۵.۵.۵ معایب**
- ۲۶۵۵ این بخش به طور کلی به معایب قوانین IF-THEN می‌پردازد.
- ۲۶۵۶ تحقیقات و ادبیات قوانین IF-THEN بر طبقه بندی تمرکز دارد و تقریباً به طور کامل از رگرسیون غفلت می‌
۲۶۵۷ کند . در حالی که همیشه می‌توانید یک هدف پیوسته را به فواصل تقسیم کنید و آن را به یک مشکل طبقه
۲۶۵۸ بندی تبدیل کنید، همیشه اطلاعات را از دست می‌دهید. به طور کلی، رویکردها در صورتی جذاب تر هستند که
۲۶۵۹ بتوان از آنها برای رگرسیون و طبقه بندی استفاده کرد.
- ۲۶۶۰ اغلب ویژگی‌ها نیز باید دسته بندی شوند . این بدان معناست که اگر می‌خواهید از ویژگی‌های عددی استفاده
۲۶۶۱ کنید، باید دسته‌بندی شوند. راه‌های زیادی برای برش یک ویژگی پیوسته به فواصل وجود دارد، اما این امر
۲۶۶۲ بی‌اهمیت نیست و با سوالات بسیاری بدون پاسخ روش‌های همراه است. ویژگی باید به چند بازه تقسیم شود؟ معیار

تقسیم چیست: طول بازه های ثابت، چندک یا چیز دیگری؟ دسته بندی ویژگی های پیوسته موضوعی غیر پیش
۲۶۶۳ پا افتاده است که اغلب نادیده گرفته می شود و افراد فقط از بهترین روش بعدی استفاده می کنند (مانند نمونه
۲۶۶۴ ها).
۲۶۶۵

بسیاری از الگوریتم های قدیمی تر یادگیری قوانین مستعد برآش بیش از حد هستند. الگوریتم های ارائه شده در
۲۶۶۶ اینجا، همگی حداقل برخی از حفاظت ها برای جلوگیری از تطبیق بیش از حد دارند OneR: محدود است، زیرا
۲۶۶۷ فقط می تواند از یک ویژگی استفاده کند (فقط اگر ویژگی دارای سطوح بیش از حد باشد یا اگر ویژگی های
۲۶۶۸ زیادی وجود داشته باشد، که برابر با مشکل آزمایش چندگانه است RIPPER. (هرس را انجام می دهد و لیست
۲۶۶۹ قوانین بیزی توزیع قبلی را در لیست های تصمیم گیری تحمیل می کند.
۲۶۷۰

قوانين تصمیم گیری در توصیف روابط خطی بین ویژگی ها و خروجی بد هستند. این مشکلی است که آنها با
۲۶۷۱ درختان تصمیم به اشتراک می گذارند. درخت ها و قوانین تصمیم گیری فقط می توانند توابع پیش بینی مرحله ای
۲۶۷۲ را تولید کنند، جایی که تغییرات در پیش بینی همیشه مراحل گستته هستند و هرگز منحنی های صافی ندارند.
۲۶۷۳ این مربوط به این موضوع است که ورودی ها باید دسته بندی شوند. در درخت های تصمیم گیری، آنها به طور
۲۶۷۴ ضمنی با تقسیم آنها طبقه بندی می شوند.
۲۶۷۵

۵,۵ نرم افزار و جایگزین

در بسته R OneR پیاده سازی شده است که برای مثال های این کتاب استفاده شده است . OneR همچنین در کتابخانه یادگیری ماشین Weka پیاده سازی شده است و به همین ترتیب در جاوا، R و Python موجود است Weka در RIPPER نیز پیاده سازی شده است. برای مثال، از پیاده سازی R از JRIP در بسته استفاده کردم SBRL . به صورت بسته) R که من برای مثال استفاده کردم)، در پایتون یا به صورت C در دسترس است . علاوه بر این، من پکیج imodels را توصیه می کنم که مدل های مبتنی بر قاعده مانند فهرست های قوانین بیزی، OneR، CORELS، Weka، فهرست های قوانین حریص و موارد دیگر را در یک بسته پایتون با یک رابط یادگیری scikit یکپارچه پیاده سازی می کند.

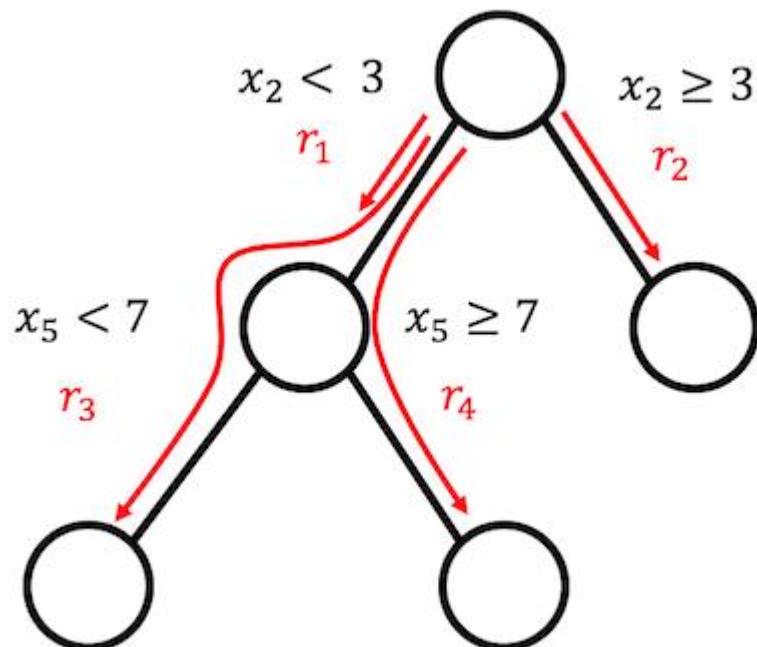
من حتی سعی نمی کنم همه یادگزین ها را برای یادگیری مجموعه ها و فهرست های قوانین تصمیم گیری فهرست کنم، اما به برخی از کارهای خلاصه کننده اشاره خواهم کرد. من کتاب "مبانی یادگیری قوانین" توسط فوئر کرانز و همکاران را توصیه می کنم. (۲۰۱۲) ۲۴ . این یک کار گسترده در مورد یادگیری قوانین است، برای کسانی که می خواهند عمیق تر به موضوع بپردازنند. این یک چارچوب جامع برای تفکر در مورد قوانین یادگیری فراهم می کند و بسیاری از الگوریتم های یادگیری قوانین را ارائه می دهد. همچنین توصیه می کنم یادگیرندگان قانون PART، RIPPER، OneR، M5Rules و بسیاری موارد دیگر را پیاده سازی می کنند، بررسی Weka را که

کنید. قوانین IF-THEN را می توان در مدل های خطی همانطور که در این کتاب در فصل مربوط به الگوریتم RuleFit توضیح داده شده است، استفاده کرد.

RuleFit^{۵,۶}

الگوریتم RuleFit توسط فریدمن و پوپسکو (۲۰۰۸) ۲۵ مدل های خطی پراکنده ای را می آموزد که شامل اثرات تعامل خودکار شناسایی شده در قالب قوانین تصمیم گیری است.

مدل رگرسیون خطی برای تعامل بین ویژگی ها حساب نمی کند. آیا داشتن مدلی که به اندازه مدل های خطی ساده و قابل تفسیر باشد، اما تعاملات ویژگی ها را نیز یکپارچه کند، راحت نیست؟ RuleFit این شکاف را پر می کند RuleFit. یک مدل خطی پراکنده با ویژگی های اصلی و همچنین تعدادی ویژگی جدید که قوانین تصمیم گیری هستند را می آموزد. این ویژگی های جدید تعامل بین ویژگی های اصلی را به تصویر می کشد . RuleFit به طور خودکار این ویژگی ها را از درخت های تصمیم تولید می کند. هر مسیر از طریق یک درخت می تواند با ترکیب تصمیمات تقسیم شده در یک قانون به یک قانون تصمیم تبدیل شود. پیش‌بینی‌های گره کنار گذاشته می‌شوند و فقط تقسیم‌ها در قوانین تصمیم‌گیری استفاده می‌شوند:



شکل ۴,۵: ۴ قانون را می توان از درختی با ۳ گره پایانی تولید کرد.

آن درختان تصمیم از کجا می آیند؟ درختان برای پیش بینی نتیجه مورد علاقه آموزش می بینند. این تضمین می کند که تقسیم ها برای کار پیش بینی معنی دار هستند. هر الگوریتمی که تعداد زیادی درخت تولید می کند می تواند برای RuleFit استفاده شود، برای مثال یک جنگل تصادفی. هر درخت به قوانین تصمیم گیری تجزیه می شود که به عنوان ویژگی های اضافی در یک مدل رگرسیون خطی پراکنده (Lasso) استفاده می شود.

مقاله RuleFit از داده های مسکن بوستون برای نشان دادن این موضوع استفاده می کند: هدف پیش بینی میانگین ارزش خانه یک محله بوستون است. یکی از قوانین ایجاد شده توسط RuleFit این است IF : number of rooms > 6.64 AND concentration of nitric oxide < 0.67 THEN 1 ELSE 0.

همچنین دارای یک اندازه گیری اهمیت ویژگی است که به شناسایی عبارات خطی و قوانینی که برای RuleFit پیش بینی ها مهم هستند کمک می کند. اهمیت ویژگی از وزن مدل رگرسیون محاسبه می شود. معیار اهمیت را می توان برای ویژگی های اصلی (که به شکل خام و احتمالاً در بسیاری از قوانین تصمیم گیری استفاده می شود) جمع آوری کرد.

همچنین نمودارهای وابستگی جزئی را برای نشان دادن میانگین تغییر در پیش بینی با تغییر یک ویژگی معرفی می کند. نمودار وابستگی جزئی یک روش مدل آگنوسنیک است که می تواند با هر مدلی استفاده شود و در فصل کتاب نمودارهای وابستگی جزئی توضیح داده شده است.

۱.۶ تفسیر و مثال

از آنجایی که RuleFit یک مدل خطی را در پایان تخمین می زند، تفسیر مشابه مدل های خطی "عادی" است. تنها تفاوت این است که مدل دارای ویژگی های جدیدی است که از قوانین تصمیم گیری به دست آمده است. قوانین تصمیم گیری ویژگی های باینری هستند: مقدار ۱ به این معنی است که همه شرایط قانون برآورده می شود، در غیر این صورت مقدار ۰ است. برای عبارت های خطی در RuleFit ، تفسیر مانند مدل های رگرسیون خطی است: اگر ویژگی یک واحد افزایش یابد، نتیجه پیش بینی شده با وزن ویژگی مربوطه تغییر می کند.

در این مثال، از RuleFit برای پیش بینی تعداد دوچرخه های اجاره ای در یک روز معین استفاده می کنیم. جدول پنج مورد از قوانینی را که توسط RuleFit ایجاد شده است به همراه وزن کمند و اهمیت آنها نشان می دهد. محاسبه بعداً در فصل توضیح داده شده است.

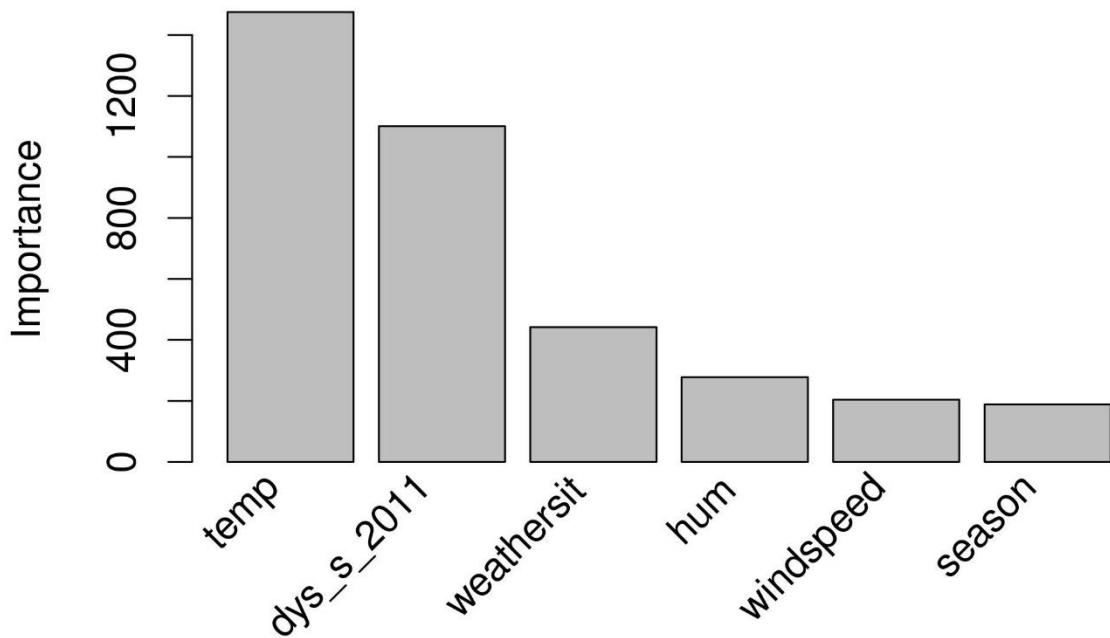
Description	Weight	Importance
days_since_2011 > 111 & weathersit in ("GOOD", "MISTY")	795	303
37.25 <= hum <= 90	-20	278
temp > 13 & days_since_2011 > 554	676	239
4 <= windspeed <= 24	-41	204
days_since_2011 > 428 & temp > 5	356	174

۲۷۲۸

۲۷۲۹ مهمترین قانون این بود "(*days_since_2011 > 111 & weathersit in ("GOOD", "MISTY")*)": و
 ۲۷۳۰ وزن مربوطه ۷۹۵ است. تفسیر این است، If *days_since_2011 > 111 & weathersit in ("GOOD",*
 ۲۷۳۱ *"MISTY"*)، سپس تعداد پیش‌بینی شده دوچرخه‌ها به میزان ۷۹۵ افزایش می‌یابد، زمانی که سایر مقادیر
 ۲۷۳۲ ویژگی ثابت باقی می‌مانند. در مجموع، ۲۷۸ قانون از این ۸ ویژگی اصلی ایجاد شد. خیلی زیاد! اما به لطف
 ۲۷۳۳ Lasso، تنها ۵۹ مورد از ۲۷۸ وزن متفاوت از ۰ دارند.

۲۷۳۴ محاسبه اهمیت ویژگی‌های جهانی نشان می‌دهد که دما و روند زمانی مهم‌ترین ویژگی‌ها هستند:

Variable importances



۲۷۳۵

۲۷۳۶ شکل ۵,۲۲: معیارهای اهمیت ویژگی برای مدل RuleFit که تعداد دوچرخه را پیش‌بینی می‌کند. مهمترین
 ۲۷۳۷ ویژگی برای پیش‌بینی دما و روند زمانی بود.

۲۷۳۸ اندازه گیری اهمیت ویژگی شامل اهمیت عبارت ویژگی خام و تمام قوانین تصمیم گیری است که ویژگی در آنها
۲۷۳۹ ظاهر می شود.

۲۷۴۰ الگوی تفسیر

۲۷۴۱ تفسیر مشابه مدل های خطی است: نتیجه پیش بینی شده تغییر می کند β_j اگر ویژگی X_j با یک واحد تغییر
۲۷۴۲ می کند، مشروط بر اینکه سایر ویژگی ها بدون تغییر باقی بمانند. تفسیر وزن یک قاعده تصمیم یک مورد خاص
۲۷۴۳ است: اگر همه شرایط یک قاعده تصمیم ۲ اعمال شود، نتیجه پیش بینی شده تغییر می کند) α وزن آموخته
۲۷۴۴ شده قانون ۲ در مدل خطی.(برای طبقه بندی (استفاده از رگرسیون لجستیک به جای رگرسیون خطی): اگر
۲۷۴۵ همه شرایط تصمیم گیری قاعده شود ۲ اعمال شود، شناس رویداد در مقابل بدون رویداد با ضریب تغییر می کند
۲۷۴۶ α .

۲۷۴۷ ۵,۶ نظریه
۲۷۴۸ اجازه دهید عمیق تر به جزئیات فنی الگوریتم RuleFit بپردازیم RuleFit. از دو مؤلفه تشکیل شده است:
۲۷۴۹ مؤلفه اول «قوانين» را از درخت های تصمیم ایجاد می کند و مؤلفه دوم با یک مدل خطی با ویژگی های اصلی
۲۷۵۰ و قوانین جدید به عنوان ورودی مناسب می شود (از این رو «RuleFit» نامیده می شود.).

۲۷۵۱ مرحله ۱: تولید قانون

۲۷۵۲ یک قانون چگونه به نظر می رسد؟ قوانین تولید شده توسط الگوریتم شکل ساده ای دارند. به عنوان مثال IF :
۲۷۵۳ $7 < 3ANDx5 < 2xسپس 0.1$ قوانین با تجزیه درختان تصمیم ساخته می شوند: هر مسیری به یک
۲۷۵۴ گره در یک درخت می تواند به یک قانون تصمیم تبدیل شود. درختان مورد استفاده برای قوانین برای پیش
۲۷۵۵ بینی نتیجه هدف تعبیه شده اند. بنابراین تقسیم ها و قوانین حاصل برای پیش بینی نتیجه مورد علاقه شما
۲۷۵۶ بهینه شده اند. مطلوب است که بسیاری از قوانین متنوع و معنادار ایجاد شود. تقویت گرادیان برای تناسب با
۲۷۵۷ مجموعه ای از درختان تصمیم با رگرسیون یا طبقه بندی Y با ویژگی های اصلی X استفاده می شود. هر درخت
۲۷۵۸ به دست آمده به چندین قانون تبدیل می شود. نه تنها درختان تقویت شده، بلکه از هر الگوریتم مجموعه
۲۷۵۹ درختی می توان برای تولید درختان برای RuleFit استفاده کرد. یک مجموعه درختی را می توان با این فرمول
۲۷۶۰ کلی توصیف کرد:

$$\hat{f}(x) = a_0 + \sum_{m=1}^M a_m \hat{f}_m(X)$$

۲۷۶۲ تعداد درختان و f متر و (X) تابع پیش بینی درخت m است. این آوزان هستند. گروههای کیسه‌ای، جنگل
۲۷۶۳ تصادفی، MART و AdaBoost مجموعه‌های درختی را تولید می‌کنند و می‌توانند برای RuleFit استفاده
۲۷۶۴ شوند.

۲۷۶۵ ما قوانین را از تمام درختان گروه ایجاد می‌کنیم. هر قانون ۲ شکل می‌گیرد:

$$r_m(x) = \prod_{j \in T_m} I(x_j \in s_{jm})$$

۲۷۶۶ جایی که T_m مجموعه‌ای از ویژگی‌های مورد استفاده در درخت- m است، اتابع نشانگر است که در هنگام
۲۷۶۷ ویژگی j است X_j در زیر مجموعه مقادیر مشخص شده S برای ویژگی j (همانطور که توسط تقسیم درخت
۲۷۶۸ مشخص شده است) و در غیر این صورت ۰ است. برای ویژگی‌های عددی، زیازه‌ای در محدوده مقدار ویژگی
۲۷۶۹ است. فاصله به نظر یکی از این دو حالت است:

$$x_{s_{jm},\text{lower}} < x_j$$

$$x_j < x_{s_{jm},\text{upper}}$$

۲۷۷۱ ۲۷۷۲ شکاف‌های بیشتر در آن ویژگی احتمالاً به فواصل پیچیده تر منجر می‌شود. برای ویژگی‌های طبقه‌بندی،
۲۷۷۳ زیرمجموعه S شامل دسته‌های خاصی از ویژگی است.

۲۷۷۴ یک مثال ساخته شده برای مجموعه داده اجاره دوچرخه:

$$r_{17}(x) = I(x_{\text{temp}} < 15) \cdot I(x_{\text{weather}} \in \{\text{good, cloudy}\}) \\ \cdot I(10 \leq x_{\text{windspeed}} < 20)$$

۲۷۷۵ ۲۷۷۶ اگر هر سه شرط برآورده شوند، این قانون ۱ را برمی‌گرداند، در غیر این صورت ۰ را برمی‌گرداند RuleFit همه
۲۷۷۷ قوانین ممکن را از یک درخت استخراج می‌کند، نه تنها از گره‌های برگ. بنابراین قانون دیگری که ایجاد می‌
۲۷۷۸ شود این است:

$$r_{18}(x) = I(x_{\text{temp}} < 15) \cdot I(x_{\text{weather}} \in \{\text{good, cloudy}\})$$

۲۷۷۹ ۲۷۸۰ در مجموع، تعداد قوانین ایجاد شده از مجموعه‌ای از درختان M با t_m گره‌های ترمینال هر کدام عبارتند از:

$$K = \sum_{m=1}^M 2(t_m - 1)$$

۲۷۸۱ ۲۷۸۲ ترفندی که توسط نویسنده‌گان RuleFit معرفی شده است، یادگیری درختان با عمق تصادفی است به طوری که
۲۷۸۳ بسیاری از قوانین متنوع با طول‌های مختلف تولید می‌شوند. توجه داشته باشید که مقدار پیش بینی شده را در

۲۷۸۴ هر گره کنار می گذاریم و فقط شرایطی را حفظ می کنیم که ما را به یک گره هدایت می کند و سپس از آن
۲۷۸۵ یک قانون ایجاد می کنیم. وزن قواعد تصمیم گیری در مرحله ۲ RuleFit انجام می شود.

۲۷۸۶ روش دیگری برای مشاهده مرحله ۱ RuleFit: مجموعه جدیدی از ویژگی های را از ویژگی های اصلی شما ایجاد
۲۷۸۷ می کند. این ویژگی ها باینری هستند و می توانند نشان دهنده تعاملات کاملاً پیچیده ویژگی های اصلی شما
۲۷۸۸ باشند. قوانین برای به حداقل رساندن کار پیش بینی انتخاب می شوند. قوانین به طور خودکار از ماتریس
۲۷۸۹ متغیرهای X تولید می شوند. شما به سادگی می توانید قوانین را به عنوان ویژگی های جدید بر اساس ویژگی
۲۷۹۰ های اصلی خود ببینید.

۲۷۹۱ مرحله ۲: مدل خطی پراکنده

۲۷۹۲ قوانین بسیاری را در مرحله ۱ دریافت می کنید. از آنجایی که مرحله اول را می توان تنها به عنوان یک تغییر
۲۷۹۳ ویژگی در نظر گرفت، هنوز کار با برآش یک مدل تمام نشده است. همچنین، می خواهید تعداد قوانین را
۲۷۹۴ کاهش دهید. علاوه بر قوانین، تمام ویژگی های «خام» شما از مجموعه داده اصلی شما نیز در مدل خطی
۲۷۹۵ پراکنده استفاده خواهد شد. هر قانون و هر ویژگی اصلی به یک ویژگی در مدل خطی تبدیل می شود و یک
۲۷۹۶ تخمین وزن می گیرد. ویژگی های خام اولیه اضافه می شوند زیرا درخت ها در نمایش روابط خطی ساده بین U و
۲۷۹۷ X شکست می خورند. قبل از اینکه یک مدل خطی پراکنده را آموزش دهیم، ویژگی های اصلی را winsorize
۲۷۹۸ می کنیم تا در برابر موارد پرت قوی تر باشند:

$$l_j^*(x_j) = \min(\delta_j^+, \max(\delta_j^-, x_j))$$

۲۸۰۰ جایی که δ و δ_j هستند δ چندک های توزیع داده ویژگی X_j . X_j انتخاب ۰,۰۵ برای به این معنی است که
۲۸۰۱ هر مقدار از ویژگی X_j که در ۵٪ کمترین یا ۵٪ بالاترین مقادیر باشد، به ترتیب بر روی چندک ها ۵٪ یا ۹۵٪
۲۸۰۲ تنظیم می شود. به عنوان یک قانون کلی، شما می توانید انتخاب کنید $\delta = 0.025$. علاوه بر این، اصطلاحات
۲۸۰۳ خطی باید به گونه ای نرم افزار شوند که اهمیت قبلی یک قانون تصمیم گیری معمولی داشته باشند:

$$l_j(x_j) = 0.4 \cdot l_j^*(x_j) / \text{std}(l_j^*(x_j))$$

۲۸۰۵ این ۴,۰ میانگین انحراف استاندارد قوانین با توزیع پشتیبانی یکنواخت است $(0, 1) \sim U$.

۲۸۰۶ ما هر دو نوع ویژگی را برای ایجاد یک ماتریس ویژگی جدید ترکیب می کنیم و یک مدل خطی پراکنده را با
۲۸۰۷ با ساختار زیر آموزش می دهیم: Lasso

$$\hat{f}(x) = \hat{\beta}_0 + \sum_{k=1}^K \hat{\alpha}_k r_k(x) + \sum_{j=1}^p \hat{\beta}_j l_j(x_j)$$

۲۸۰۸

۲۸۰۹

جایی که α بردار وزن تخمین زده شده برای ویژگی های قاعده و β بردار وزن برای ویژگی های اصلی از آنجایی که RuleFit از Lasso استفاده می کند،تابع ضرر محدودیت اضافی را دریافت می کند که برخی از وزن ها را مجبور می کند تا تخمین صفر را دریافت کنند:

۲۸۱۲

۲۸۱۳

۲۸۱۴

۲۸۱۵

۲۸۱۶

۲۸۱۷

۲۸۱۸

۲۸۱۹

۲۸۲۰

۲۸۲۱

۲۸۲۲

۲۸۲۳

۲۸۲۴

۲۸۲۵

۲۸۲۶

۲۸۲۷

$$\begin{aligned} (\{\hat{\alpha}\}_1^K, \{\hat{\beta}\}_0^p) = & \operatorname{argmin}_{\{\hat{\alpha}\}_1^K, \{\hat{\beta}\}_0^p} \sum_{i=1}^n L(y^{(i)}, f(x^{(i)})) \\ & + \lambda \cdot \left(\sum_{k=1}^K |\alpha_k| + \sum_{j=1}^p |b_j| \right) \end{aligned}$$

نتیجه یک مدل خطی است که اثرات خطی برای همه ویژگی های اصلی و قوانین دارد. تفسیر مشابه مدل های خطی است، تنها تفاوت این است که برخی از ویژگی ها اکنون قوانین باینری هستند.

مرحله ۳ (اختیاری): اهمیت ویژگی

برای شرایط خطی ویژگی های اصلی، اهمیت ویژگی با پیش بینی استاندارد اندازه گیری می شود:

$$I_j = |\hat{\beta}_j| \cdot \text{std}(l_j(x_j))$$

جایی که β وزن از مدل Lasso و انحراف استاندارد عبارت خطی بر روی داده است.

برای شرایط قاعده تصمیم گیری، اهمیت با فرمول زیر محاسبه می شود:

$$I_k = |\hat{\alpha}_k| \cdot \sqrt{s_k(1-s_k)}$$

جایی که وزن کمند مربوط به قاعده تصمیم است و s_k پشتیبانی از ویژگی در داده است، که درصدی از نقاط

داده ای است که قانون تصمیم گیری در مورد آنها اعمال می شود (در جایی که $r_k(x) = 1$)

$$s_k = \frac{1}{n} \sum_{i=1}^n r_k(x^{(i)})$$

یک ویژگی به عنوان یک عبارت خطی و احتمالاً در بسیاری از قوانین تصمیم گیری نیز رخ می دهد. چگونه

اهمیت کلی یک ویژگی را اندازه گیری کنیم؟ اهمیت $(x)_j$ یک ویژگی را می توان برای هر پیش بینی فردی

اندازه گیری کرد:

$$J_j(x) = I_j(x) + \sum_{x_j \in r_k} I_k(x)/m_k$$

۲۸۲۸ جایی که I_k اهمیت عبارت خطی و Γ_k اهمیت قواعد تصمیم گیری که در آن X_j ظاهر می شود، و m_k تعداد
۲۸۲۹ ویژگی های تشکیل دهنده قانون است . افزودن اهمیت ویژگی از همه نمونه ها به ما اهمیت ویژگی جهانی
۲۸۳۰ می دهد:

$$J_j(X) = \sum_{i=1}^n J_j(x^{(i)})$$

۲۸۳۱
۲۸۳۲ انتخاب زیر مجموعه ای از نمونه ها و محاسبه اهمیت ویژگی برای این گروه امکان پذیر است.

۲۸۳۳
۲۸۳۴ مزايا ۵,۶,۳
RuleFit
۲۸۳۵ به طور خودکار تعاملات ویژگی را به مدل های خطی اضافه می کند. بنابراین، مشکل مدل های خطی
۲۸۳۶ را حل می کند که باید اصطلاحات تعامل را به صورت دستی اضافه کنید و کمی به مسئله مدل سازی روابط
غیرخطی کمک می کند.

۲۸۳۷ RuleFit می تواند هم وظایف طبقه بندی و هم رگرسیون را انجام دهد.

۲۸۳۸ قوانین ایجاد شده به راحتی قابل تفسیر هستند، زیرا آنها قوانین تصمیم گیری باینری هستند. یا این قانون در
۲۸۳۹ مورد یک نمونه اعمال می شود یا خیر. تفسیرپذیری خوب تنها زمانی تضمین می شود که تعداد شرایط درون
۲۸۴۰ یک قانون خیلی زیاد نباشد. یک قانون با ۱ تا ۳ شرط به نظر من معقول است. این به معنای حداقل عمق ۳
۲۸۴۱ برای درختان مجموعه درخت است.

۲۸۴۲ حتی اگر قوانین زیادی در مدل وجود داشته باشد، آنها برای هر نمونه اعمال نمی شوند. برای یک مثال فردی
۲۸۴۳ فقط تعداد انگشت شماری از قوانین اعمال می شود (= وزن غیر صفر دارند). این قابلیت تفسیر محلی را بهبود
۲۸۴۴ می بخشد.

۲۸۴۵ RuleFit مجموعه ای از ابزارهای تشخیصی مفید را پیشنهاد می کند. این ابزارها مدل-آگنوستیک هستند،
بنابراین می توانید آنها را در بخش مدل-آگنوستیک کتاب بیابید: اهمیت ویژگی ، نمودارهای وابستگی جزئی و
۲۸۴۷ تعاملات ویژگی .

۲۸۴۸ معايب ۵,۶,۴
۲۸۴۹ گاهی اوقات RuleFit قوانین زیادی ایجاد می کند که وزن غیر صفر در مدل Lasso دریافت می کند.
۲۸۵۰ تفسیرپذیری با افزایش تعداد ویژگی ها در مدل کاهش می یابد. یک راه حل امیدوارکننده این است که جلوه
۲۸۵۱ های ویژگی را مجبور کنیم که یکنواخت باشند، به این معنی که افزایش یک ویژگی باید منجر به افزایش پیش
۲۸۵۲ بینی شود.

یک اشکال حکایتی: مقالات مدعی عملکرد خوب RuleFit هستند - اغلب نزدیک به عملکرد پیش‌بینی جنگل‌های تصادفی! - اما در موارد محدودی که من شخصاً آن را امتحان کردم، عملکرد نالمیدکننده بود. فقط آن را برای مشکل خود امتحان کنید و ببینید چگونه کار می‌کند.

محصولنهایی رویه RuleFit یک مدل خطی با ویژگی‌های فانتزی اضافی (قوانين تصمیم‌گیری) است. اما از آنجایی که این یک مدل خطی است، تفسیر وزن هنوز غیر قابل درک است. با همان «پاورقی» مدل رگرسیون خطی معمولی ارائه می‌شود: «... با توجه به اینکه همه ویژگی‌ها ثابت هستند». وقتی قوانینی با هم تداخل دارند کمی دشوارتر می‌شود. به عنوان مثال، یک قانون تصمیم (ویژگی) برای پیش‌بینی دوچرخه می‌تواند این باشد: «دما < ۱۰» و قانون دیگری می‌تواند «دما > ۱۵ و هوای خوب» باشد. اگر هوا خوب باشد و دما بالای ۱۵ درجه باشد، دما به طور خودکار از ۱۰ بیشتر می‌شود. در مواردی که قانون دوم اعمال می‌شود، قانون اول نیز اعمال می‌شود. تفسیر وزن تخمینی برای قانون دوم این است: «با فرض ثابت ماندن تمام ویژگی‌های دیگر، تعداد پیش‌بینی‌شده دوچرخه‌ها افزایش می‌یابد β . زمانی که هوا خوب و دما بالای ۱۵ درجه باشد. اما، اکنون واقعاً روش می‌شود که «همه ویژگی‌های دیگر ثابت شده‌اند» مشکل‌ساز است، زیرا اگر قانون ۲ اعمال شود، قانون ۱ نیز اعمال می‌شود و تفسیر بی‌معنی است.

۵,۵ نرم افزار و جایگزین

الگوریتم RuleFit در R توسط Fokkema and Christoffersen (2017) پیاده‌سازی شده است و می‌توانید نسخه GitHub را در Python پیدا کنید.

یک چارچوب بسیار مشابه skope-rules است، یک ماژول پایتون که قوانین را نیز از مجموعه‌ها استخراج می‌کند. در نحوه یادگیری قوانین نهایی متفاوت است: اول، قوانین skope قوانین با عملکرد پایین را بر اساس فراخوانی و آستانه‌های دقیق حذف می‌کنند. سپس قوانین تکراری و مشابه با انجام یک انتخاب بر اساس تنوع اصطلاحات منطقی (متغیر + عملکرد بزرگتر/کوچکتر) و عملکرد (امتیاز F1) قوانین حذف می‌شوند. این مرحله نهایی به استفاده از کمند متکی نیست، بلکه فقط امتیاز F1 خارج از کیف و شرایط منطقی را که قوانین را تشکیل می‌دهند در نظر می‌گیرد.

بسته imodels همچنین شامل اجرای مجموعه قوانین دیگر، مانند مجموعه قوانین بیزی، مجموعه قوانین تقویت شده، و مجموعه قوانین SLIPPER به عنوان یک بسته پایتون با یک رابط scikit-learn یکپارچه است.

۵,۷ سایر مدل های قابل تفسیر

۲۸۷۷ فهرست مدل های قابل تفسیر دائماً در حال افزایش است و اندازه آن نامشخص است. این شامل مدل های
۲۸۷۸ ساده ای مانند مدل های خطی، درخت های تصمیم گیری و Bayes ساده لوح است، اما همچنین مدل های
۲۸۷۹ پیچیده تری که مدل های یادگیری ماشین غیر قابل تفسیر را ترکیب یا اصلاح می کنند تا آنها را قابل تفسیر تر
۲۸۸۰ کنند. به ویژه نشریات مربوط به مدل های نوع دوم در حال حاضر با فرکانس بالا تولید می شوند و به سختی
۲۸۸۱ می توان با پیشرفت ها همراهی کرد. این کتاب فقط طبقه بندی کننده ساده بیز و k -نزدیک ترین همسایه ها را در
۲۸۸۲ این فصل آزار می دهد.
۲۸۸۳

۵,۷,۱ طبقه بندی کننده ساده بیز

۲۸۸۴ طبقه بندی کننده ساده بیز از قضیه احتمالات شرطی بیز استفاده می کند. برای هر ویژگی، احتمال یک کلاس
۲۸۸۵ را بسته به مقدار ویژگی محاسبه می کند. طبقه بندی کننده *Naive Bayes* احتمالات کلاس را برای هر
۲۸۸۶ ویژگی به طور مستقل محاسبه می کند، که معادل یک فرض قوی (= ساده لوحانه) استقلال شرطی ویژگی ها
۲۸۸۷ است *Naive Bayes*. یک مدل احتمال شرطی است و احتمال یک کلاس را C_k مدل می کند به شرح زیر
۲۸۸۸ است:
۲۸۸۹

$$P(C_k|x) = \frac{1}{Z} P(C_k) \prod_{i=1}^n P(x_i|C_k)$$

۲۸۹۰ عبارت Z یک پارامتر مقیاس پذیری است که تضمین می کند که مجموع احتمالات برای همه کلاس ها ۱ است
۲۸۹۱ (در غیر این صورت احتمالات نیستند). احتمال شرطی یک کلاس، احتمال کلاس ضریب احتمال هر ویژگی با
۲۸۹۲ توجه به کلاس است که با Z نormal شده است. این فرمول را می توان با استفاده از قضیه بیز به دست آورد.
۲۸۹۳

۲۸۹۴ ساده لوح بیز به دلیل فرض استقلال یک مدل قابل تفسیر است. می توان آن را در سطح مدولار تفسیر کرد.
۲۸۹۵ برای هر ویژگی بسیار واضح است که چقدر به پیش بینی کلاس خاصی کمک می کند، زیرا می توانیم احتمال
۲۸۹۶ شرطی را تفسیر کنیم.

۵,۷,۲ k -نزدیک ترین همسایه ها

۲۸۹۷ روش k -نزدیک ترین همسایه را می توان برای رگرسیون و طبقه بندی استفاده کرد و از نزدیک ترین همسایگان
۲۸۹۸ یک نقطه داده برای پیش بینی استفاده می کند. برای طبقه بندی، روش k -نزدیک ترین همسایه رایج ترین کلاس
۲۸۹۹ نزدیک ترین همسایه های یک نمونه را اختصاص می دهد. برای رگرسیون، میانگین نتیجه همسایگان را می گیرد.
۲۹۰۰ بخش های دشوار یافتن k مناسب و تصمیم گیری درباره نحوه اندازه گیری فاصله بین نمونه ها است که در نهایت
۲۹۰۱ همسایگی را مشخص می کند.
۲۹۰۲

۲۹۰۳ مدل -k نزدیکترین همسایه با سایر مدل‌های قابل تفسیر ارائه شده در این کتاب متفاوت است زیرا یک الگوریتم
۲۹۰۴ یادگیری مبتنی بر نمونه است. چگونه می‌توان -k نزدیک ترین همسایه‌ها را تفسیر کرد؟ اول از همه، هیچ
۲۹۰۵ پارامتری برای یادگیری وجود ندارد، بنابراین هیچ تفسیرپذیری در سطح مدولار وجود ندارد. علاوه بر این، عدم
۲۹۰۶ تفسیرپذیری مدل جهانی وجود دارد، زیرا مدل ذاتاً محلی است و هیچ وزن یا ساختار جهانی به صراحت آموخته
۲۹۰۷ نشده است. شاید در سطح محلی قابل تفسیر باشد؟ برای توضیح یک پیش‌بینی، همیشه می‌توانید k
۲۹۰۸ همسایه‌هایی را که برای پیش‌بینی استفاده شده‌اند، بازیابی کنید. اینکه آیا مدل قابل تفسیر است یا نه، تنها به
۲۹۰۹ این سؤال بستگی دارد که آیا می‌توانید یک نمونه واحد را در مجموعه داده «تفسیر» کنید. اگر یک نمونه از
۲۹۱۰ صدها یا هزاران ویژگی تشکیل شده باشد، من استدلال می‌کنم که قابل تفسیر نیست

۲۹۱۱

فصل عروش های مدل-آگنوتیک

۲۹۱۲

۲۹۱۳

۲۹۱۴ جداسازی توضیحات از مدل یادگیری ماشین (= روش‌های تفسیر مدل-آگنوتیک) مزایایی دارد, Ribeiro, ۲۹۱۵ Maitzit Bzrg Singh, and Guestrin 2016 27).

۲۹۱۶ خاص مدل، انعطاف پذیری آنهاست. توسعه دهنده‌گان یادگیری ماشینی آزادند که از هر مدل یادگیری ماشینی ۲۹۱۷ که دوست دارند استفاده کنند، زمانی که روش‌های تفسیر را می‌توان برای هر مدلی اعمال کرد. هر چیزی که ۲۹۱۸ مبتنی بر تفسیر یک مدل یادگیری ماشینی باشد، مانند گرافیک یا رابط کاربری، از مدل یادگیری ماشین ۲۹۱۹ اساسی نیز مستقل می‌شود. به طور معمول، نه تنها یک، بلکه بسیاری از انواع مدل‌های یادگیری ماشین برای ۲۹۲۰ حل یک کار ارزیابی می‌شوند، و هنگام مقایسه مدل‌ها از نظر تفسیرپذیری، کار با توضیحات مدل-آگنوتیک ۲۹۲۱ آسان‌تر است، زیرا از همان روش می‌توان برای هر نوع استفاده کرد. از مدل

۲۹۲۲ یک جایگزین برای روش‌های تفسیر مدل-آگنوتیک، استفاده از مدل‌های قابل تفسیر است، که اغلب دارای ۲۹۲۳ این عیب بزرگ است که عملکرد پیش‌بینی‌کننده در مقایسه با سایر مدل‌های یادگیری ماشین از بین می‌رود و ۲۹۲۴ شما خود را به یک نوع مدل محدود می‌کنید. جایگزین دیگر استفاده از روش‌های تفسیر مدل خاص است. عیب ۲۹۲۵ این کار این است که شما را به یک نوع مدل متصل می‌کند و تغییر به مدل دیگری دشوار خواهد بود.

۲۹۲۶ جنبه‌های مطلوب یک سیستم توضیح مدل-آگنوتیک عبارتند از Ribeiro, Singh, and Guestrin ۲۹۲۷ (2016):

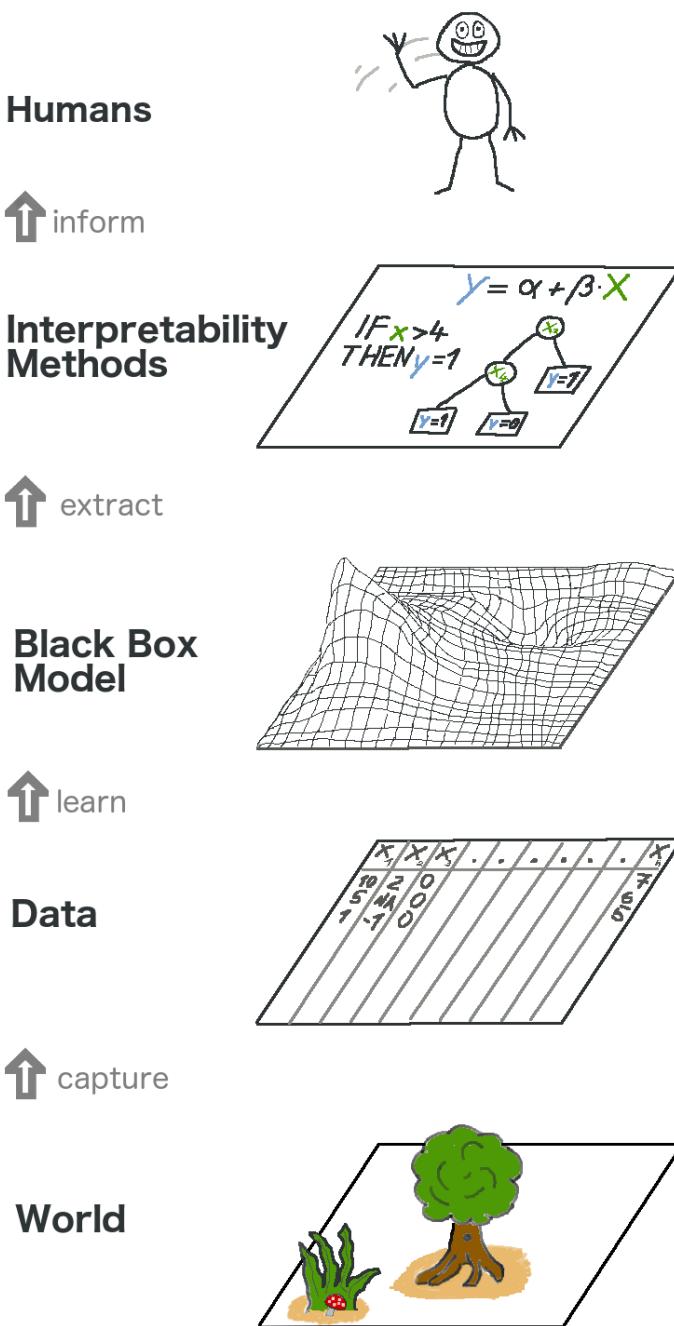
۲۹۲۸ انعطاف‌پذیری مدل: روش تفسیر می‌تواند با هر مدل یادگیری ماشینی مانند جنگل‌های تصادفی و شبکه‌های ۲۹۲۹ عصبی عمیق کار کند.

۲۹۳۰ انعطاف‌پذیری توضیح: شما محدود به شکل خاصی از توضیح نیستید. در برخی موارد ممکن است داشتن یک ۲۹۳۱ فرمول خطی مفید باشد، در موارد دیگر یک گرافیک با اهمیت ویژگی.

۲۹۳۲ انعطاف‌پذیری نمایش: سیستم توضیح باید بتواند از یک نمایش ویژگی متفاوت به عنوان مدل توضیح داده شده ۲۹۳۳ استفاده کند. برای طبقه‌بندی متنی که از بردارهای جاسازی کلمه انتزاعی استفاده می‌کند، ممکن است ترجیح ۲۹۳۴ داده شود که از وجود کلمات جداگانه برای توضیح استفاده شود.

۲۹۳۵ تصویر بزرگتر

۲۹۳۶ اجازه دهد نگاهی در سطح بالا به تفسیرپذیری مدل-اگنومتیک بیندازیم. ما با جمعآوری داده‌ها، جهان را به تصویر می‌کشیم و با یادگیری پیش‌بینی داده‌ها (برای کار) با یک مدل یادگیری ماشین، آن را بیشتر انتزاع می‌کنیم. تفسیرپذیری تنها لایه دیگری در بالای صفحه است که به درک انسان کمک می‌کند.



۲۹۳۹

۲۹۴۰ شکل ۱، تصویر بزرگ یادگیری ماشینی قابل توضیح. دنیای واقعی قبل از اینکه در قالب تبیین به دست انسان
۲۹۴۱ بررسد از لایه های زیادی عبور می کند.

۲۹۴۲

۲۹۴۳ پایین ترین لایه جهان است . این می تواند به معنای واقعی کلمه خود طبیعت باشد، مانند بیولوژی بدن انسان و
۲۹۴۴ نحوه واکنش آن به دارو، اما همچنین چیزهای انتزاعی تر مانند بازار املاک و مستغلات. لایه World شامل هر
۲۹۴۵ چیزی است که می توان مشاهده کرد و مورد توجه است. در نهایت، ما می خواهیم چیزی در مورد جهان
۲۹۴۶ بیاموزیم و با آن تعامل داشته باشیم.

۲۹۴۷ لایه دوم لایه داده است . ما باید جهان را دیجیتالی کنیم تا بتوانیم آن را برای کامپیوترها پردازش کنیم و
۲۹۴۸ همچنین اطلاعات را ذخیره کنیم. لایه داده شامل هر چیزی از تصاویر، متون، داده های جدولی و غیره است.

۲۹۴۹ با برآش مدل های یادگیری ماشین بر اساس لایه داده، لایه Black Box Model را دریافت می کنیم .
۲۹۵۰ الگوریتم های یادگیری ماشین با داده های دنیای واقعی یاد می گیرند تا پیش بینی کنند یا ساختارها را بیابند.

۲۹۵۱ در بالای لایه Black Box Model لایه روش های تفسیرپذیری قرار دارد که به ما کمک می کند تا با کدورت
۲۹۵۲ مدل های یادگیری ماشین مقابله کنیم. مهمترین ویژگی برای یک تشخیص خاص چه بود؟ چرا یک تراکنش
۲۹۵۳ مالی به عنوان تقلب طبقه بنده شد؟

۲۹۵۴ آخرین لایه توسط یک انسان اشغال شده است . نگاه کن این یکی برای شما تکان می دهد زیرا در حال خواندن
۲۹۵۵ این کتاب هستید و به ارائه توضیحات بهتر برای مدل های جعبه سیاه کمک می کنید! انسان ها در نهایت
۲۹۵۶ مصرف کننده توضیحات هستند.

۲۹۵۷ این انتزاع چند لایه همچنین به درک تفاوت های رویکردهای بین آماردانان و متخصصان یادگیری ماشین کمک
۲۹۵۸ می کند. آماردانان با لایه داده سروکار دارند، مانند برنامه ریزی آزمایشات بالینی یا طراحی نظرسنجی. آنها لایه
۲۹۵۹ Black Box Model را رد می کنند و مستقیماً به لایه روش های تفسیرپذیری می روند. متخصصان یادگیری
۲۹۶۰ ماشین همچنین با لایه داده سروکار دارند، مانند جمع آوری نمونه های برچسب گذاری شده از تصاویر سرطان
۲۹۶۱ پوست یا خزیدن ویکی پدیا. سپس آنها یک مدل یادگیری ماشین جعبه سیاه را آموزش می دهند. لایه
۲۹۶۲ روش های تفسیرپذیری نادیده گرفته می شود و انسان ها مستقیماً با پیش بینی های مدل جعبه سیاه سروکار
۲۹۶۳ دارند. بسیار خوب است که یادگیری ماشینی قابل تفسیر، کار آماردانان و متخصصان یادگیری ماشین را ترکیب
۲۹۶۴ می کند.

- ۲۹۶۵ البته این گرافیک همه چیز را به تصویر نمی کشد: داده ها می توانند از شبیه سازی به دست آیند. مدل های جعبه
سیاه همچنین پیش بینی هایی را ارائه می دهند که حتی ممکن است به دست انسان هم نرسد، بلکه فقط
۲۹۶۶ ماشین های دیگر و غیره را تامین می کنند. اما به طور کلی، در ک اینکه چگونه تفسیر پذیری به این لایه جدید در
۲۹۶۷ بالای مدل های یادگیری ماشین تبدیل می شود، یک انتزاع مفید است.
۲۹۶۸
- ۲۹۶۹ روش های تفسیر مدل - آگنوستیک را می توان به روش های محلی و جهانی تمایز کرد. کتاب نیز بر اساس این
۲۹۷۰ تمایز تنظیم شده است. روش های جهانی چگونگی تأثیر ویژگی ها بر پیش بینی را به طور متوسط توصیف
۲۹۷۱ می کنند . در مقابل، هدف روش های محلی توضیح پیش بینی های فردی است .
۲۹۷۲

فصل ۷ توضیحات مبتنی بر مثال

- ۲۹۷۳
- ۲۹۷۴
- ۲۹۷۵ روش‌های توضیح مبتنی بر مثال، نمونه‌های خاصی از مجموعه داده را برای توضیح رفتار مدل‌های یادگیری
- ۲۹۷۶ ماشین یا توضیح توزیع داده‌های اساسی انتخاب می‌کنند.
- ۲۹۷۷ توضیح‌های مبتنی بر مثال عمدتاً مدل‌ساز هستند، زیرا هر مدل یادگیری ماشینی را قابل تفسیرتر می‌سازند.
- ۲۹۷۸ تفاوت روش‌های مدل‌آگنوستیک در این است که روش‌های مبتنی بر مثال، یک مدل را با انتخاب نمونه‌هایی از
- ۲۹۷۹ مجموعه داده توضیح می‌دهند و نه با ایجاد خلاصه‌ای از ویژگی‌ها (مانند اهمیت ویژگی یا وابستگی جزئی).
- ۲۹۸۰ توضیحات مبتنی بر مثال تنها زمانی معنا پیدا می‌کنند که بتوانیم نمونه‌ای از داده‌ها را به روشی قابل فهم برای
- ۲۹۸۱ انسان نمایش دهیم. این به خوبی برای تصاویر کار می‌کند، زیرا می‌توانیم آنها را مستقیماً مشاهده کنیم. به
- ۲۹۸۲ طور کلی، روش‌های مبتنی بر مثال اگر مقادیر ویژگی‌های یک نمونه دارای زمینه بیشتری باشند، به خوبی کار
- ۲۹۸۳ می‌کنند، به این معنی که داده‌ها ساختاری دارند، مانند تصاویر یا متون. نمایش داده‌های جدولی به روشی
- ۲۹۸۴ معنadar چالش برانگیزتر است، زیرا یک نمونه می‌تواند از صدھا یا هزاران ویژگی (کمتر ساختارمند) تشکیل شده
- ۲۹۸۵ باشد. فهرست کردن تمام مقادیر ویژگی برای توصیف یک نمونه معمولاً مفید نیست. اگر فقط تعداد انگشت
- ۲۹۸۶ شماری از ویژگی‌ها وجود داشته باشد یا راهی برای خلاصه کردن یک نمونه داشته باشیم، به خوبی کار می‌
- ۲۹۸۷ کند.
- ۲۹۸۸ توضیحات مبتنی بر مثال به انسان کمک می‌کند تا مدل‌های ذهنی مدل یادگیری ماشین و داده‌هایی را که مدل
- ۲۹۸۹ یادگیری ماشینی روی آنها آموزش دیده است، بسازند. این به ویژه به درک توزیع داده‌های پیچیده کمک می‌
- ۲۹۹۰ کند. اما منظور من از توضیحات مبتنی بر مثال چیست؟ ما اغلب از آنها در شغل و زندگی روزمره خود استفاده
- ۲۹۹۱ می‌کنیم. اجازه دهید با چند مثال ۲۸ شروع کنیم.
- ۲۹۹۲ یک پزشک بیمار را با سرفه غیرمعمول و تب خفیف می‌بیند. علائم بیمار او را به یاد بیمار دیگری می‌اندازد که
- ۲۹۹۳ سال‌ها پیش با علائم مشابه داشت. او مشکوک است که بیمار فعلی اش ممکن است به همین بیماری مبتلا
- ۲۹۹۴ باشد و برای آزمایش این بیماری خاص نمونه خون می‌گیرد.
- ۲۹۹۵ یک دانشمند داده روی پروژه جدیدی برای یکی از مشتریان خود کار می‌کند: تجزیه و تحلیل عوامل خطر که
- ۲۹۹۶ منجر به خرابی ماشین‌های تولید صفحه کلید می‌شود. دانشمند داده پروژه مشابهی را که روی آن کار کرده
- ۲۹۹۷ بود به خاطر می‌آورد و از بخش‌هایی از کد پروژه قدیمی دوباره استفاده می‌کند زیرا فکر می‌کند مشتری
- ۲۹۹۸ همان تحلیل را می‌خواهد.

۲۹۹۹ بچه گربه ای روی طاقچه پنجره خانه ای در حال سوختن و خالی از سکنه نشسته است. آتش نشانی در حال حاضر آمده است و یکی از آتش نشان ها برای لحظه ای فکر می کند که آیا می تواند برای نجات بچه گربه به داخل ساختمان برود یا خیر. او موارد مشابهی را در زندگی خود به عنوان آتش نشان به یاد می آورد: خانه های چوبی قدیمی که مدتی است به آرامی می سوختند، اغلب ناپایدار بودند و در نهایت فرو ریختند. به دلیل شباht این مورد، تصمیم می گیرد وارد نشود، زیرا خطر ریزش خانه خیلی زیاد است. خوشبختانه، بچه گربه از پنجره به بیرون می پرد، به سلامت فرود می آید و هیچ کس در آتش آسیبی نمی بیند. پایان خوش.

۳۰۰۵

این داستان ها نحوه تفکر ما انسان ها را در مثال ها یا قیاس ها نشان می دهد. طرح اولیه توضیحات مبتنی بر مثال این است: چیز B شبیه چیز A است و A باعث ۷ است، بنابراین من پیش بینی می کنم که B نیز باعث ۷ شود. به طور ضمنی، برخی از رویکردهای یادگیری ماشین مبتنی بر مثال کار می کنند. درختان تصمیمداده ها را بر اساس شباht نقاط داده در ویژگی هایی که برای پیش بینی هدف مهم هستند، به گره ها تقسیم کنید. یک درخت تصمیم، پیش بینی یک نمونه داده جدید را با یافتن نمونه های مشابه (= در همان گره پایانی) و برگرداندن میانگین نتایج آن نمونه ها به عنوان پیش بینی، دریافت می کند. روش k-نزدیکترین همسایه (knn) به صراحت با پیش بینی های مبتنی بر مثال کار می کند. برای مثال جدید، یک مدل knn نزدیکترین همسایگان (مثلاً $k=3$ نزدیکترین نمونه) را تعیین می کند و میانگین نتایج آن همسایگان را به عنوان یک پیش بینی بر می گرداند. پیش بینی یک knn را می توان با برگرداندن همسایگان k توضیح داد، که - دوباره - تنها در صورتی معنی دار است که راه خوبی برای نمایش یک نمونه واحد داشته باشیم.

۳۰۱۵ روش های تفسیر زیر همگی مبتنی بر مثال هستند:

۳۰۱۶ توضیحات خلاف واقع به ما می گوید که چگونه یک نمونه باید تغییر کند تا پیش بینی آن به طور قابل توجهی تغییر کند. با ایجاد نمونه های خلاف واقع، در مورد اینکه مدل چگونه پیش بینی های خود را انجام می دهد و می تواند پیش بینی های فردی را توضیح دهد، می آموزیم.

۳۰۱۷

۳۰۱۸

۳۰۱۹ مثال های متخاصل ، خلاف واقع هایی هستند که برای فریب دادن مدل های یادگیری ماشین استفاده می شوند. ۳۰۲۰ تاکید بر ورق زدن پیش بینی و توضیح ندادن آن است.

۳۰۲۱ نمونه های اولیه گزیده ای از نمونه های نماینده از داده ها هستند و انتقادات مواردی هستند که به خوبی توسط آن نمونه های اولیه نشان داده نمی شوند. ۲۹

۳۰۲۲

- نمونه‌های تأثیرگذار، نقاط داده آموزشی هستند که بیشترین تأثیر را برای پارامترهای یک مدل پیش‌بینی یا خود پیش‌بینی‌ها داشتند. شناسایی و تجزیه و تحلیل نمونه‌های تأثیرگذار به یافتن مشکلات داده‌ها، اشکال زدایی مدل و درک بهتر رفتار مدل کمک می‌کند.
- مدل- k -نزدیکترین همسایگان : یک مدل یادگیری ماشین (قابل تفسیر) بر اساس مثالها.

۳۰۲۸

۳۰۲۹

فصل ۸ مدل جهانی-روشهای آگنوستیک

روش های جهانی رفتار متوسط یک مدل یادگیری ماشین را توصیف می کنند. همتای روش های جهانی روش

های محلی هستند. روش های جهانی اغلب به صورت مقادیر مورد انتظار بر اساس توزیع داده ها بیان می شوند.

برای مثال، نمودار وابستگی جزئی ، نمودار اثر ویژگی، پیش‌بینی مورد انتظار زمانی است که همه ویژگی‌های

دیگر به حاشیه رانده شوند. از آنجایی که روش‌های تفسیر کلی رفتار متوسط را توصیف می‌کنند، زمانی که

مدل‌ساز می‌خواهد مکانیسم‌های کلی در داده‌ها را بفهمد یا یک مدل را اشکال‌زدایی کند، بسیار مفید هستند.

در این کتاب، با تکنیک‌های تفسیر جهانی مدل-آگنوستیک زیر آشنا خواهید شد:

نمودار وابستگی جزئی یک روش اثر ویژگی است.

نمودار جلوه‌های محلی انباسته یکی دیگر از روش‌های اثر ویژگی است که در صورت وابسته بودن ویژگی‌ها

کار می‌کند.

تعامل ویژگی (آمار H) کمیت می‌کند که پیش‌بینی تا چه حد نتیجه اثرات مشترک ویژگی‌ها است.

تجزیه تابعی یک ایده مرکزی از تفسیرپذیری و تکنیکی است که تابع پیش‌بینی پیچیده را به بخش‌های

کوچکتر تجزیه می‌کند.

اهمیت ویژگی جایگشتی اهمیت یک ویژگی را به عنوان افزایش از دست دادن زمانی که ویژگی تغییر می‌کند

اندازه‌گیری می‌کند.

مدل‌های جانشین جهانی مدل اصلی را با مدلی ساده‌تر برای تفسیر جایگزین می‌کند.

نمونه‌های اولیه و انتقادات نشان‌دهنده نقطه داده یک توزیع هستند و می‌توانند برای افزایش تفسیرپذیری

استفاده شوند

۳۰۴۸

۱.۸ طرح وابستگی جزئی (PDP)

نmodar وابستگی جزئی (نقشه کوتاه PDP) یا (PD) اثر حاشیه ای یک یا دو ویژگی را بر نتیجه پیش بینی شده یک مدل یادگیری ماشین نشان می دهد . (Friedman 2001 JH) نmodar وابستگی جزئی می تواند نشان دهد که آیا رابطه بین هدف و یک ویژگی خطی، یکنواخت یا پیچیده تر است. به عنوان مثال، هنگامی که به یک مدل رگرسیون خطی اعمال می شود، نmodarهای وابستگی جزئی همیشه یک رابطه خطی را نشان می دهند.

تابع وابستگی جزئی برای رگرسیون به صورت زیر تعریف می شود:

$$\hat{f}_S(x_S) = E_{X_C} [\hat{f}(x_S, X_C)] = \int \hat{f}(x_S, X_C) d\mathbb{P}(X_C)$$

ویژگی هایی هستند که تابع وابستگی جزئی باید برای آنها ترسیم شود و X_C از دیگر ویژگی های مورد استفاده در مدل یادگیری ماشینی هستند ، که در اینجا به عنوان متغیرهای تصادفی در نظر گرفته می شوند. عموماً فقط یک یا دو ویژگی در مجموعه S وجود دارد. ویژگی(های) در S آنها بی هستند که می خواهیم تأثیر آنها را بر پیش بینی بدانیم. بردارهای ویژگی XS و XC ترکیبی فضای کلی ویژگی X را تشکیل می دهد. وابستگی جزئی با به حاشیه راندن خروجی مدل یادگیری ماشین بر روی توزیع ویژگی های مجموعه C عمل می کند، به طوری که تابع رابطه بین ویژگی های مجموعه S مورد علاقه ما و نتیجه پیش بینی شده را نشان می دهد. با به حاشیه راندن سایر ویژگی ها، تابعی دریافت می کنیم که فقط به ویژگی های S بستگی دارد، تعامل با سایر ویژگی ها نیز گنجانده شده است.

تابع جزئی f با محاسبه میانگین ها در داده های آموزشی که به روش مونت کارلو نیز معروف است، تخمین زده می شود:

$$\hat{f}_S(x_S) = \frac{1}{n} \sum_{i=1}^n \hat{f}(x_S, x_C^{(i)})$$

تابع جزئی به ما می گوید که برای مقدار(های) داده شده ویژگی S ، میانگین اثر حاشیه ای در پیش بینی چقدر است. در این فرمول، $x^{(i)}_C$ مقادیر واقعی ویژگی از مجموعه داده برای ویژگی هایی هستند که ما به آنها علاقه نداریم و n تعداد نمونه های موجود در مجموعه داده است. یک فرض PDP این است که ویژگی های موجود در C با ویژگی های S همبستگی ندارند. اگر این فرض نقض شود، میانگین های محاسبه شده برای نmodar وابستگی جزئی شامل نقاط داده ای خواهد بود که بسیار بعید یا حتی غیرممکن هستند (معایب را ببینید).

برای طبقه‌بندی که در آن مدل یادگیری ماشین احتمالات را خروجی می‌دهد، نمودار وابستگی جزئی احتمال را برای یک کلاس خاص با توجه به مقادیر مختلف برای ویژگی(ها) در S نشان می‌دهد. یک راه آسان برای مقابله با چندین کلاس ترسیم یک خط یا نمودار برای هر کلاس است.

نمودار وابستگی جزئی یک روش جهانی است: این روش همه نمونه‌ها را در نظر می‌گیرد و بیانیه ای در مورد رابطه کلی یک ویژگی با نتیجه پیش‌بینی شده ارائه می‌دهد.

ویژگی‌های طبقه‌بندی شده

تا اینجا فقط ویژگی‌های عددی را در نظر گرفته ایم. برای ویژگی‌های طبقه‌بندی، محاسبه وابستگی جزئی بسیار آسان است. برای هر یک از دسته‌ها، با وادار کردن همه نمونه‌های داده به داشتن دسته یکسان، یک تخمین PDP دریافت می‌کنیم. به عنوان مثال، اگر به مجموعه داده‌های اجاره دوچرخه نگاه کنیم و به نمودار وابستگی جزئی برای فصل علاقه مند باشیم، چهار عدد به دست می‌آوریم، یکی برای هر فصل. برای محاسبه مقدار "تابستان"، فصل تمام نمونه‌های داده را با "تابستان" جایگزین می‌کنیم و پیش‌بینی‌ها را میانگین می‌کنیم.

۳۰.۸۳ ۱.۱ اهمیت ویژگی مبتنی بر PDP
۳۰.۸۴ گرین ول و همکاران (۲۰۱۸) ۳۱ یک معیار اهمیت ویژگی مبتنی بر وابستگی جزئی را پیشنهاد کرد. انگیزه اصلی این است که یک PDP مسطح نشان می‌دهد که ویژگی مهم نیست، و هر چه PDP بیشتر تغییر کند، ویژگی مهم‌تر است. برای ویژگی‌های عددی، اهمیت به عنوان انحراف هر مقدار ویژگی منحصر به فرد از منحنی میانگین تعریف می‌شود:

$$I(x_S) = \sqrt{\frac{1}{K-1} \sum_{k=1}^K (\hat{f}_S(x_S^{(k)}) - \frac{1}{K} \sum_{k=1}^K \hat{f}_S(x_S^{(k)}))^2}$$

توجه داشته باشید که در اینجا $K^{(k)}$ مقادیر منحصر به فرد ویژگی the X_S هستند. برای ویژگی‌های دسته بندی ما داریم:

۳۰.۹۱ $I(x_S) = (max_k(\hat{f}_S(x_S^{(k)})) - min_k(\hat{f}_S(x_S^{(k)})))/4$

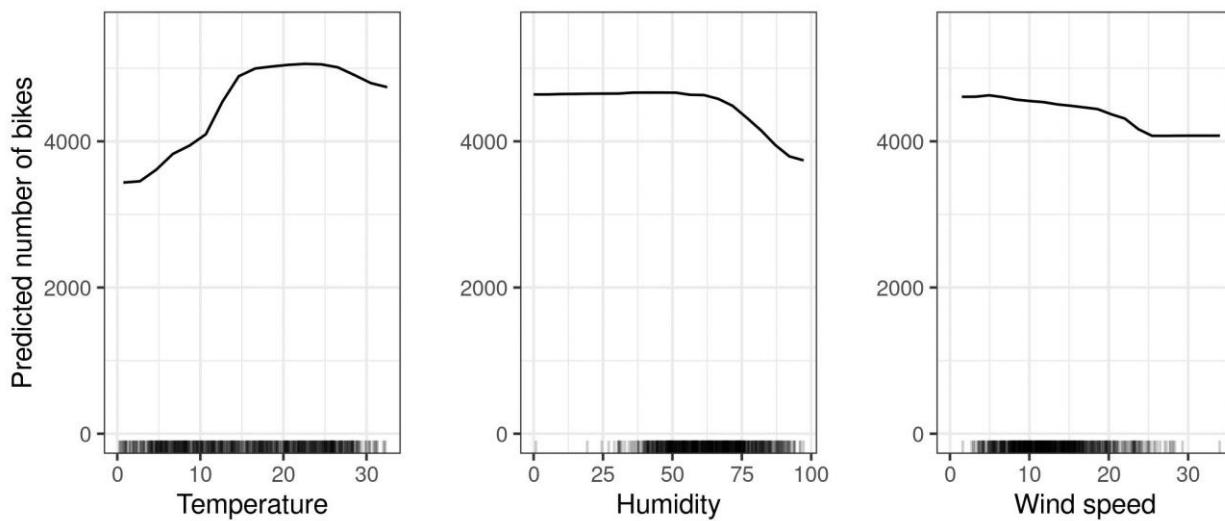
۳۰.۹۲ این محدوده مقادیر PDP برای دسته‌های منحصر به فرد تقسیم بر چهار است. این روش عجیب و غریب برای محاسبه انحراف، قانون محدوده نامیده می‌شود. هنگامی که شما فقط محدوده را می‌دانید، به دست آوردن یک تخمین تقریبی برای انحراف کمک می‌کند. و مخرج چهار از توزیع نرمال استاندارد می‌آید: در توزیع نرمال، ۹۵ درصد داده‌ها منهای دو و به علاوه دو انحراف معیار حول میانگین هستند. بنابراین محدوده تقسیم بر چهار تخمین تقریبی را به دست می‌دهد که احتمالاً واریانس واقعی را دست کم می‌گیرد.

۳۰۹۷ این اهمیت ویژگی مبتنی بر PDP باید با دقت تفسیر شود. فقط اثر اصلی ویژگی را می‌گیرد و تعاملات احتمالی
 ۳۰۹۸ ویژگی را نادیده می‌گیرد. یک ویژگی می‌تواند بر اساس روش‌های دیگر مانند اهمیت ویژگی جایگشت بسیار
 ۳۰۹۹ مهم باشد، اما PDP می‌تواند مسطح باشد زیرا این ویژگی عمده‌اً از طریق تعامل با سایر ویژگی‌ها بر پیش‌بینی
 ۳۱۰۰ تأثیر می‌گذارد. یکی دیگر از اشکالات این معیار این است که بر روی مقادیر منحصر به فرد تعریف شده است.
 ۳۱۰۱ یک مقدار ویژگی منحصر به فرد تنها با یک نمونه، در محاسبه اهمیت به همان وزنی داده می‌شود که دارای
 ۳۱۰۲ چندین نمونه است.

۳۱۰۳ ۸,۱,۲ مثال

۳۱۰۴ در عمل، مجموعه ویژگی‌های S معمولاً فقط شامل یک ویژگی یا حداکثر دو ویژگی است، زیرا یک ویژگی
 ۳۱۰۵ نمودارهای دو بعدی و دو ویژگی نمودارهای سه بعدی تولید می‌کنند. همه چیز فراتر از آن بسیار مشکل است.
 ۳۱۰۶ حتی سه بعدی روی کاغذ یا مانیتور دو بعدی چالش برانگیز است.

۳۱۰۷ اجازه دهید به مثال رگرسیون برگردیم، که در آن تعداد دوچرخه‌هایی را که در یک روز معین اجاره می‌شوند،
 ۳۱۰۸ پیش‌بینی می‌کنیم. ابتدا مدل یادگیری ماشینی را برازش می‌کنیم، سپس وابستگی‌های جزئی را تجزیه و
 ۳۱۰۹ تحلیل می‌کنیم. در این مورد، ما یک جنگل تصادفی برای پیش‌بینی تعداد دوچرخه‌ها و استفاده از نمودار
 ۳۱۱۰ وابستگی جزئی برای تجسم روابطی که مدل آموخته است، نصب کرده‌ایم. تأثیر ویژگی‌های آب و هوا بر تعداد
 ۳۱۱۱ دوچرخه‌های پیش‌بینی شده در شکل زیر نشان داده شده است.

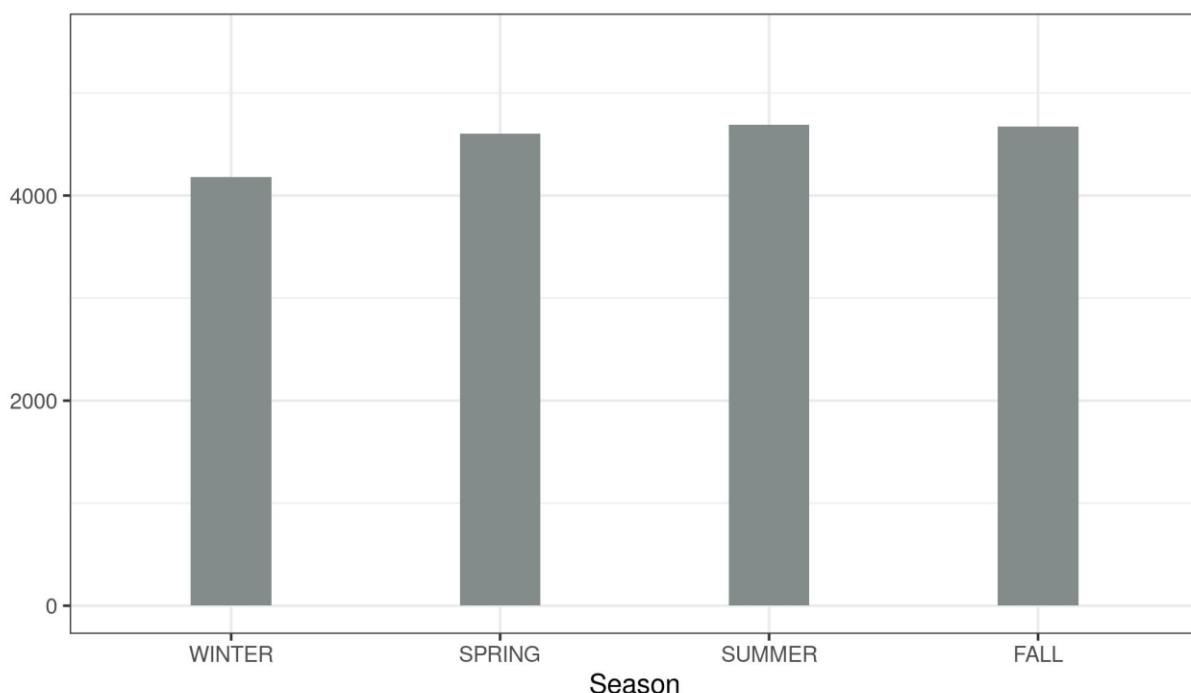


۳۱۱۲ شکل ۸,۱ PDP نیز مدل پیش‌بینی تعداد دوچرخه و دما، رطوبت و سرعت باد. بیشترین تفاوت را می‌توان در
 ۳۱۱۳ دما مشاهده کرد. هر چه گرمتر باشد، دوچرخه‌های بیشتری اجاره می‌شود. این روند تا ۲۰ درجه سانتیگراد بالا
 ۳۱۱۴

۳۱۱۵ می رود، سپس صاف می شود و در ۳۰ کمی کاهش می یابد. علائم روی محور X توزیع داده ها را نشان می
۳۱۱۶ دهد.

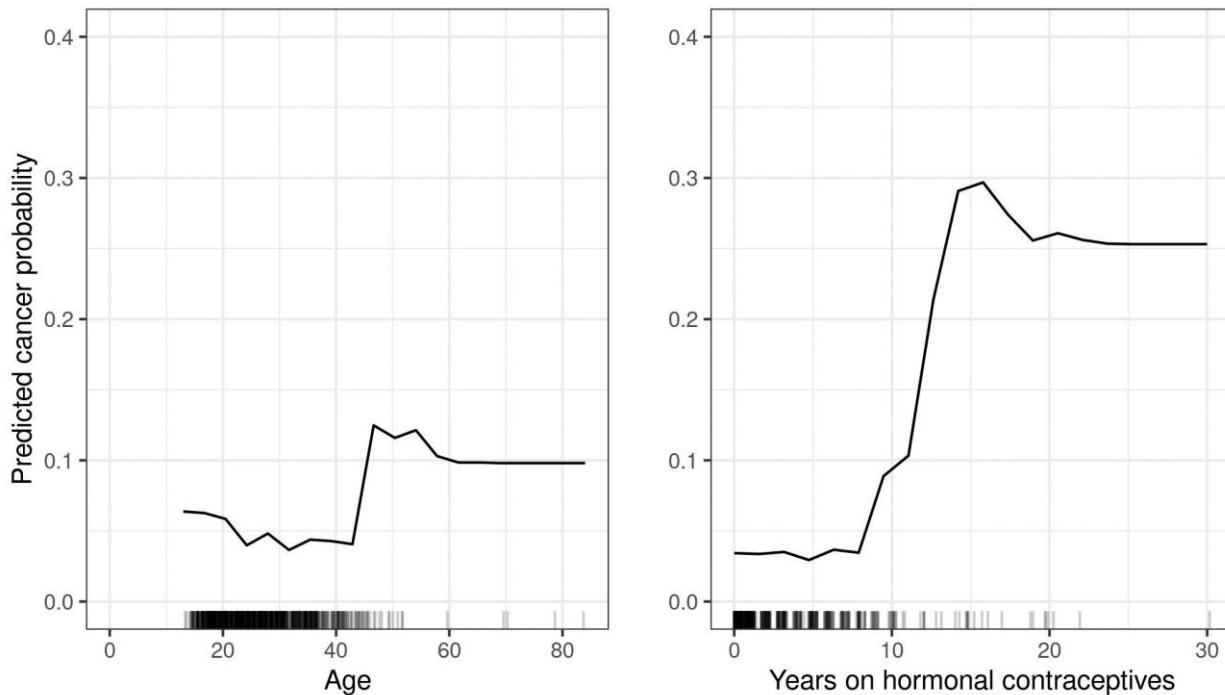
۳۱۱۷ برای آب و هوای گرم اما نه خیلی گرم، این مدل به طور متوسط تعداد زیادی دوچرخه کرایه ای را پیش بینی
۳۱۱۸ می کند. دوچرخه سواران بالقوه به طور فرایندهای از اجاره دوچرخه در زمانی که رطوبت از ۶۰٪ فراتر می رود، منع
۳۱۱۹ می شوند. علاوه بر این، هر چه باد بیشتر باشد، افراد کمتری دوست دارند دوچرخه سواری کنند، که منطقی
۳۱۲۰ است. جالب اینجاست که وقتی سرعت باد از ۲۵ به ۳۵ کیلومتر در ساعت می رسد، تعداد پیش بینی شده اجاره
۳۱۲۱ دوچرخه کاهش نمی یابد، اما داده های آموزشی زیادی وجود ندارد، بنابراین مدل یادگیری ماشینی احتمالاً
۳۱۲۲ نمی تواند پیش بینی معناداری برای این محدوده بیاموزد. حداقل به طور شهودی، من انتظار دارم با افزایش
۳۱۲۳ سرعت باد، تعداد دوچرخه ها کاهش یابد، به خصوص زمانی که سرعت باد بسیار زیاد است.

۳۱۲۴ برای نشان دادن یک نمودار وابستگی جزئی با یک ویژگی طبقه بندی شده، ما تأثیر ویژگی فصل را بر اجاره
۳۱۲۵ دوچرخه پیش بینی شده بررسی می کنیم.



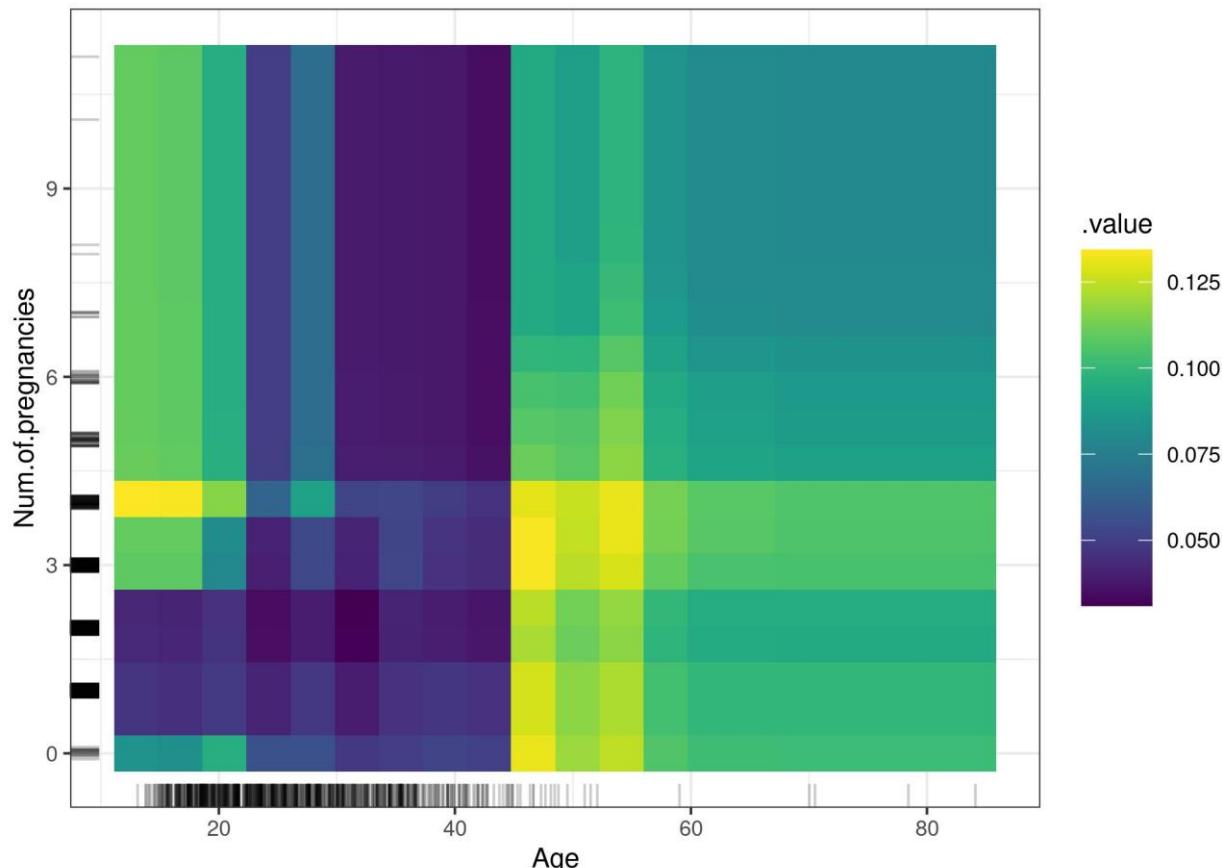
۳۱۲۶ شکل ۸,۲ PDP: برای مدل پیش بینی تعداد دوچرخه و فصل. بطور غیرمنتظره تمام فصول اثر مشابهی بر
۳۱۲۷ پیش بینی های مدل نشان می دهند، فقط برای زمستان مدل اجاره دوچرخه کمتری را پیش بینی می کند.
۳۱۲۸

۳۱۲۹ ما همچنین وابستگی جزئی را برای طبقه بندی سرطان دهانه رحم محاسبه می کنیم . این بار ما یک جنگل
۳۱۳۰ تصادفی را برای پیش‌بینی اینکه آیا یک زن ممکن است بر اساس عوامل خطر به سرطان دهانه رحم مبتلا شود،
۳۱۳۱ قرار می‌دهیم. ما وابستگی جزئی احتمال سرطان به ویژگی‌های مختلف را برای جنگل تصادفی محاسبه و تجسم
۳۱۳۲ می‌کنیم:



۳۱۳۳ ۳۱۳۴ شکل ۸.۳: های احتمال سرطان بر اساس سن و سال با داروهای ضد بارداری هورمونی. برای سن، PDP
۳۱۳۵ نشان می دهد که احتمال تا ۴۰ سالگی کم است و بعد از آن افزایش می یابد. هر چه سال‌ها بیشتر از داروهای
۳۱۳۶ ضد بارداری هورمونی استفاده کنید، خطر سرطان پیش‌بینی شده بیشتر می‌شود، مخصوصاً بعد از ۱۰ سال. برای
۳۱۳۷ هر دو ویژگی، نقاط داده زیادی با مقادیر زیاد در دسترس نبود، بنابراین برآوردهای PD در آن مناطق کمتر قابل
۳۱۳۸ اعتماد هستند.

۳۱۳۹ ما همچنین می توانیم وابستگی جزئی دو ویژگی را در یک زمان تجسم کنیم:



۳۱۴۰ شکل ۸,۴ PDP : احتمال سرطان و اثر متقابل سن و تعداد حاملگی. این نمودار افزایش احتمال ابتلا به سرطان را در ۴۵ سالگی نشان می دهد. برای سینین زیر ۲۵ سال، زنانی که ۱ یا ۲ بارداری داشتند، در مقایسه با زنانی که ۰ یا بیشتر از ۲ بارداری داشتند، خطر سرطان پیش بینی شده کمتری داشتند. اما هنگام نتیجه گیری مراقب باشید: این ممکن است فقط یک همبستگی باشد و نه علی!

۳۱۴۵ مزايا ۸,۱,۳
۳۱۴۶ محاسبه نمودارهای وابستگی جزئی بصری است : تابع وابستگی جزئی در یک مقدار مشخصه خاص، میانگین پیش بینی را نشان می دهد اگر همه نقاط داده را مجبور کنیم آن مقدار ویژگی را فرض کنند. بر اساس تجربه ۳۱۴۷ من، افراد غیر روحانی معمولاً ایده PDP ها را به سرعت درک می کنند.

۳۱۴۹ اگر مشخصه ای که PDP را برای آن محاسبه کرده اید با ویژگی های دیگر همبستگی نداشته باشد، PDP ها به ۳۱۵۰ خوبی نشان می دهند که چگونه ویژگی به طور متوسط بر پیش بینی تأثیر می گذارد. در مورد غیر همبسته، ۳۱۵۱ تفسیر واضح است : نمودار وابستگی جزئی نشان می دهد که چگونه میانگین پیش بینی در مجموعه داده شما با ۳۱۵۲ تغییر ویژگی-زم تغییر می کند. وقتی ویژگی ها با هم مرتبط باشند، پیچیده تر است، معایب را نیز ببینید.

۳۱۵۳ طرح های وابستگی جزئی به راحتی قابل پیاده سازی هستند.

۳۱۵۴ محاسبه برای نمودارهای وابستگی جزئی یک تفسیر علی دارد . ما روی یک ویژگی مداخله می کنیم و تغییرات
۳۱۵۵ پیش بینی ها را اندازه می گیریم. در انجام این کار، ما رابطه علی بین ویژگی و پیش بینی را تحلیل می کنیم. ۳۲
۳۱۵۶ این رابطه برای مدل علی است - زیرا ما به صراحت نتیجه را به عنوان تابعی از ویژگی ها مدل می کنیم - اما نه
۳۱۵۷ لزوماً برای دنیای واقعی !

۳۱۵۸ ۸,۱,۴ معایب
۳۱۵۹ حداکثر واقعی تعداد ویژگی ها در یک تابع وابستگی جزئی دو است. این تقصیر PDP ها نیست، بلکه از نمایش
۳۱۶۰ دو بعدی (کاغذ یا صفحه نمایش) و همچنین ناتوانی ما در تصور بیش از ۳ بعد است.

۳۱۶۱ برخی از نمودارهای PD توزیع ویژگی را نشان نمی دهند . حذف توزیع ممکن است گمراه کننده باشد، زیرا
۳۱۶۲ ممکن است مناطقی را که تقریباً هیچ داده ای ندارند، بیش از حد تفسیر کنید. این مشکل با نشان دادن یک
۳۱۶۳ فرش (شاخص نقاط داده در محور X یا هیستوگرام به راحتی حل می شود.

۳۱۶۴ فرض استقلال بزرگترین مشکل توطئه های PD است. فرض بر این است که ویژگی (هایی) که وابستگی جزئی
۳۱۶۵ برای آنها محاسبه می شود با سایر ویژگی ها همبستگی ندارند. برای مثال، فرض کنید با توجه به وزن و قد فرد
۳۱۶۶ می خواهید سرعت راه رفتن یک فرد را پیش بینی کنید. برای وابستگی جزئی یکی از ویژگی ها، مثلًا قد، فرض
۳۱۶۷ می کنیم که سایر ویژگی ها (وزن) با قد همبستگی ندارند، که بدیهی است یک فرض نادرست است. برای محاسبه
۳۱۶۸ PDP در ارتفاع معین (مثلث ۲۰۰ سانتی متر)، توزیع حاشیه ای وزن را میانگین می گیریم، که ممکن است شامل
۳۱۶۹ وزن کمتر از ۵۰ کیلوگرم باشد، که برای یک فرد ۲ متری غیر واقعی است. به عبارت دیگر: وقتی ویژگی ها با
۳۱۷۰ هم مرتبط هستند، ما نقاط داده جدیدی را در مناطقی از توزیع ویژگی ایجاد می کنیم که احتمال واقعی آن
۳۱۷۱ بسیار کم است (به عنوان مثال بعید است که فردی ۲ متر قد داشته باشد اما وزن آن کمتر از ۵۰ کیلوگرم
۳۱۷۲ باشد). یکی از راه حل های این مشکل است نمودارهای اثر محلی انباسته یا نمودارهای کوتاه ALE که با توزیع
۳۱۷۳ شرطی به جای توزیع حاشیه ای کار می کنند.

۳۱۷۴ اثرات ناهمگن ممکن است پنهان باشد زیرا نمودارهای PD فقط اثرات حاشیه ای متوسط را نشان می دهند.
۳۱۷۵ فرض کنید برای یک ویژگی، نیمی از نقاط داده شما ارتباط مثبتی با پیش بینی داشته باشند - هر چه مقدار
۳۱۷۶ ویژگی بزرگتر باشد، پیش بینی بزرگتر است - و نیمی دیگر دارای ارتباط منفی باشد - هر چه مقدار ویژگی
۳۱۷۷ کوچکتر باشد، پیش بینی بزرگتر است. منحنی PD می تواند یک خط افقی باشد، زیرا اثرات هر دو نیمه

۳۱۷۸ مجموعه داده می تواند یکدیگر را خنثی کند. سپس نتیجه می گیرید که این ویژگی تاثیری بر پیش بینی ندارد.
۳۱۷۹ با ترسیم منحنی های انتظار شرطی فردی به جای خط تجمعی، می توانیم اثرات ناهمگن را کشف کنیم.

۳۱۸۰ ۸,۱,۵ نرم افزار و جایگزین

۳۱۸۱ تعدادی بسته R وجود دارد که PDP ها را پیاده سازی می کنند. من از iml بسته برای نمونه ها استفاده کردم،
۳۱۸۲ اما pdp یا وجود دارد DALEX. در پایتون، نمودارهای وابستگی جزئی تعییه شده اند scikit-learn و می توانید
۳۱۸۳ از PDPBox.

۳۱۸۴ جایگزین های PDP ارائه شده در این کتاب نمودارهای ALE و منحنی های ICE هستند.

۳۱۸۵

۸.۲ طرح جلوه های محلی انباشته (ALE)

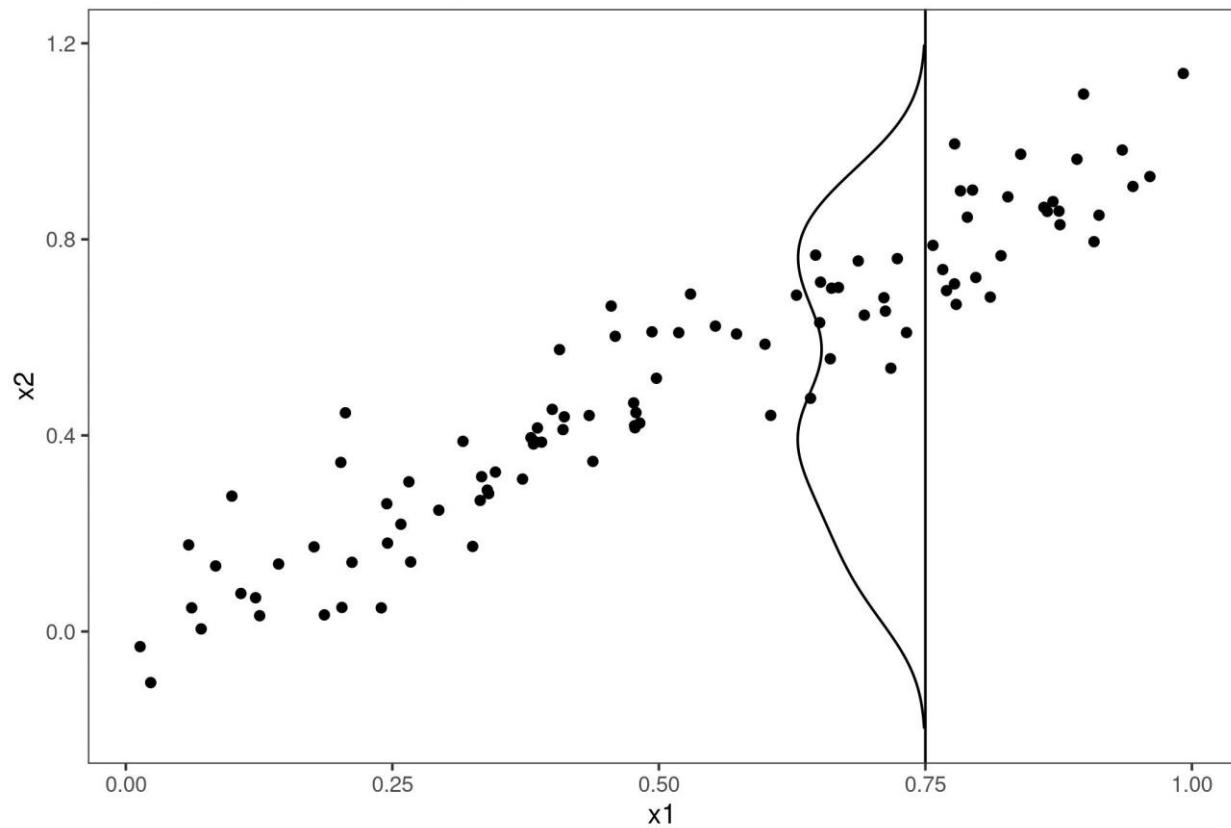
اثرات محلی انباشته شده ۳۳ توضیح می دهد که چگونه ویژگی ها به طور متوسط بر پیش بینی یک مدل یادگیری ماشین تأثیر می گذارند. نمودارهای ALE یک جایگزین سریعتر و بی طرفانه برای نمودارهای وابستگی جزئی (PDP) هستند.

توصیه می کنیم ابتدا فصل مربوط به نمودارهای وابستگی جزئی را مطالعه کنید، زیرا درک آنها آسان تر است و هر دو روش هدف یکسانی دارند: هر دو توضیح می دهند که چگونه یک ویژگی به طور متوسط بر پیش بینی تأثیر می گذارد. در بخش بعدی، می خواهم شما را متقدعاً کنم که نمودارهای وابستگی جزئی زمانی که ویژگی ها همبستگی دارند، یک مشکل جدی دارند.

۸.۱ انگیزه و شهود

اگر ویژگی های یک مدل یادگیری ماشینی با هم مرتبط باشند، نمی توان به نمودار وابستگی جزئی اعتماد کرد. محاسبه یک نمودار وابستگی جزئی برای یک ویژگی که به شدت با سایر ویژگی ها همبستگی دارد، شامل میانگین گیری پیش بینی های نمونه های داده مصنوعی است که در واقعیت بعید است. این می تواند تا حد زیادی اثر ویژگی تخمین زده شده را سوگیری کند. تصور کنید که نمودارهای وابستگی جزئی را برای یک مدل یادگیری ماشینی محاسبه کنید که ارزش یک خانه را بسته به تعداد اتاق ها و اندازه منطقه نشیمن پیش بینی می کند. ما به تأثیر منطقه زندگی بر مقدار پیش بینی شده علاقه مندیم. برای یادآوری، دستور نمودارهای وابستگی جزئی به این صورت است: ۱) ویژگی را انتخاب کنید. ۲) شبکه را تعریف کنید. ۳) در هر مقدار شبکه: الف) ویژگی را با مقدار شبکه جایگزین کنید و ب) پیش بینی های میانگین. ۴) منحنی را رسم کنید. برای محاسبه اولین مقدار شبکه - PDP مثلا ۳۰ متر^۲- فضای نشیمن را برای همه موارد ۳۰ متر مربع جایگزین می کنیم ، حتی برای خانه هایی با ۱۰ اتاق. به نظر من یک خانه بسیار غیر معمول است. طرح وابستگی جزئی شامل این خانه های غیر واقعی در تخمین اثر ویژگی می شود و وانمود می کند که همه چیز خوب است. شکل زیر دو ویژگی مرتبط را نشان می دهد و اینکه چگونه روش نمودار وابستگی جزئی پیش بینی های نمونه های بعید را میانگین می دهد.

Marginal distribution $P(x_2)$



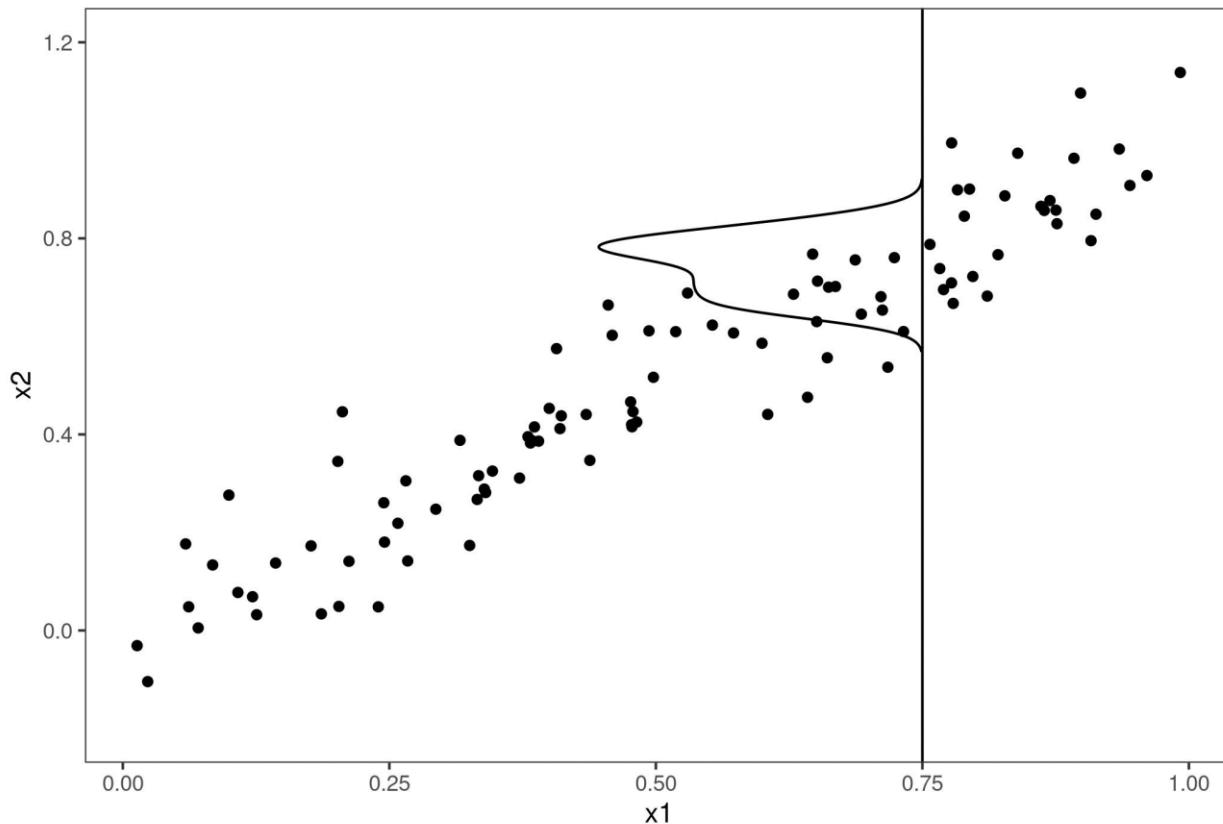
۳۲۰۹

شکل ۸.۵: ویژگی های x_1 و x_2 با همبستگی قوی. برای محاسبه اثر ویژگی x_1 در x_2 باز همه نمونه ها را با $x_1 = 0.75$ جایگزین می کند، به اشتباہ فرض می کنیم که توزیع x_2 در $x_1 = 0.75$ با توزیع حاشیه ای) x_2 خط عمودی) یکسان است. این منجر به ترکیبات بعید از x_1 و x_2 می شود (به عنوان مثال $x_2 = 0.2$ در $x_1 = 0.75$ ، که $P(x_2)$ برای محاسبه اثر میانگین استفاده می کند).

برای به دست آوردن تخمین اثر ویژگی که به همبستگی ویژگی ها احترام می گذارد، چه کاری می توانیم انجام دهیم؟ ما می توانیم میانگین توزیع شرطی ویژگی را داشته باشیم، به این معنی که در یک مقدار شبکه ای x_1 ، پیش‌بینی‌های نمونه‌هایی با مقدار x_1 مشابه را میانگین می‌کنیم. راه حل برای محاسبه اثرات ویژگی با استفاده از توزیع شرطی، نمودارهای حاشیه ای یا M-Plots نامیده می شود (نام گیج کننده، زیرا بر اساس توزیع شرطی است، نه توزیع حاشیه ای). صبر کن، آیا به شما قول ندادم که در مورد توطئه های ALE صحبت کنید؟ راه حلی نیست که ما به دنبال آن هستیم. چرا M-Plots مشکل ما را حل نمی کند؟ اگر میانگین M-Plots پیش‌بینی تمام خانه‌ها را در حدود ۳۰ مترمربع به دست آوریم، ترکیب آن را تخمین می‌زنیم تأثیر مساحت نشیمن و تعداد اتاق ها به دلیل همبستگی آنها. فرض کنید منطقه نشیمن هیچ تاثیری بر ارزش پیش‌بینی شده

یک خانه ندارد، فقط تعداد اتاق ها تاثیری دارد M-Plot. همچنان نشان می دهد که اندازه منطقه نشیمن مقدار پیش بینی شده را افزایش می دهد، زیرا تعداد اتاق ها با منطقه نشیمن افزایش می یابد. نمودار زیر برای دو ویژگی مرتبط نحوه کار M-Plots را نشان می دهد.

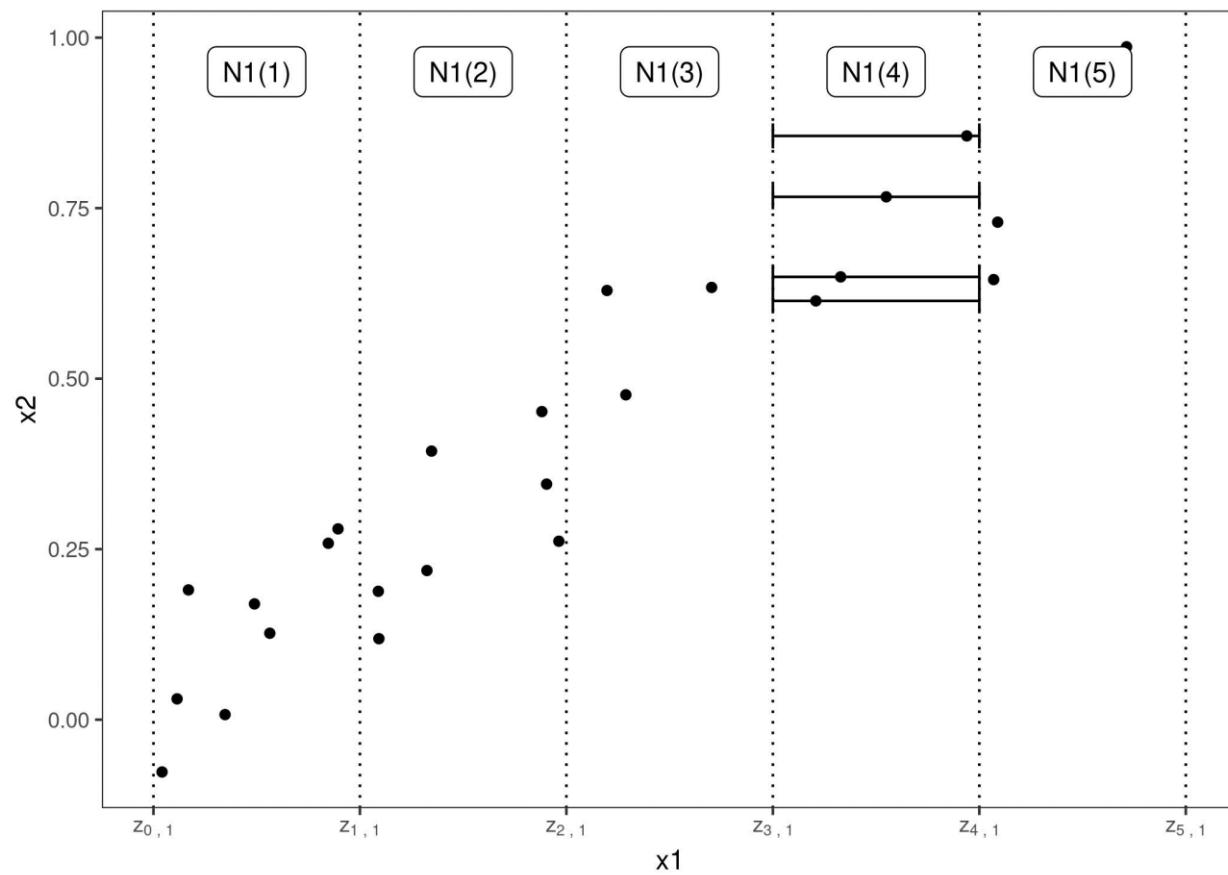
Conditional distribution $P(x_2|x_1=0.75)$



شکل ۸.۶: ویژگی های x_1 و x_2 با همبستگی قوی. میانگین M-Plots بیش از توزیع شرطی. در اینجا توزیع شرطی x_2 در $x_1 = 0.75$ است. میانگین گرفتن پیش‌بینی‌های محلی منجر به مخلوط کردن اثرات هر دو ویژگی می‌شود.

از میانگین پیش بینی نمونه های داده بعيد اجتناب می کند، اما آنها اثر یک ویژگی را با اثرات همه ویژگی های مرتبط مخلوط می کنند. نمودارهای ALE این مشکل را با محاسبه - همچنین بر اساس توزیع شرطی ویژگی ها - تفاوت در پیش بینی ها به جای میانگین ها حل می کند . برای تأثیر مساحت نشیمن در ۳۰ متر مربع ، روش ALE از تمام خانه های با مساحت ۳۰ متر مربع استفاده می کند ، پیش‌بینی‌های مدل را به دست می آورد که این خانه ها را ۳۱ متر مربع منهای پیش‌بینی تظاهر به ۲۹ متر مربع هستند. این به ما اثر

٣٢٣٤ خالص منطقه نشیمن می دهد و اثر را با اثرات ویژگی های مرتبط مخلوط نمی کند. استفاده از تفاوت ها اثر
 ٣٢٣٥ سایر ویژگی ها را مسدود می کند. نمودار زیر نحوه محاسبه نمودارهای ALE را نشان می دهد.



٣٢٣٦ شکل ٨.٧: محاسبه ALE برای ویژگی x_1 که با x_2 همبستگی دارد. ابتدا ویژگی را به فواصل (خطوط عمودی)
 ٣٢٣٧ تقسیم می کنیم. برای نمونه های داده (نقاط) در یک بازه، زمانی که ویژگی را با حد بالا و پایین بازه (خطوط
 ٣٢٣٨ افقی) جایگزین می کنیم، تفاوت پیش بینی را محاسبه می کنیم. این تفاوت ها بعداً انباشته و متمرکز می شوند و
 ٣٢٣٩ در نتیجه منحنی ALE ایجاد می شود.
 ٣٢٤٠

٣٢٤١ برای خلاصه کردن اینکه چگونه هر نوع نمودار ALE، M، PDP را در یک مقدار شبکه خاص ٧
 ٣٢٤٢ محاسبه می کند:

٣٢٤٣ نمودارهای وابستگی حزمی: «اجازه دهید به شما نشان دهم که وقتی هر نمونه داده مقدار ٧ را دارد، مدل به
 ٣٢٤٤ طور متوسط چه چیزی را پیش بینی می کند. برای آن ویژگی من نادیده می گیرم که آیا مقدار ٧ برای همه
 ٣٢٤٥ نمونه های داده معنا دارد یا خیر.

۳۲۴۶ اجازه دهید به شما نشان دهم که مدل به طور متوسط برای نمونه های داده ای که مقادیر نزدیک
 ۳۲۴۷ به ۷ را برای آن ویژگی دارند، چه چیزی را پیش بینی می کند. این تأثیر می تواند به دلیل آن ویژگی باشد، اما
 ۳۲۴۸ همچنین به دلیل ویژگی های مرتبط.»

۳۲۴۹ ALE رسم می کند : "اجازه دهید به شما نشان دهم که چگونه پیش بینی های مدل در "پنجره" کوچکی از
 ۳۲۵۰ ویژگی اطراف ۷ برای نمونه های داده در آن پنجره تغییر می کند".

۳۲۵۱ ۸,۲ نظریه

۳۲۵۲ نمودارهای PD ، ALE و چگونه از نظر ریاضی متفاوت هستند؟ مشترک هر سه روش این است که آنها تابع
 ۳۲۵۳ پیش بینی پیچیده f را به تابعی کاهش می دهند که فقط به یک (یا دو) ویژگی بستگی دارد. هر سه روش با
 ۳۲۵۴ میانگین گیری اثرات سایر ویژگی ها، تابع را کاهش می دهند، اما در محاسبه میانگین های پیش بینی ها یا تفاوت ها
 ۳۲۵۵ در پیش بینی ها و اینکه میانگین گیری بر روی توزیع حاشیه ای یا شرطی انجام می شود، تفاوت دارند.
 ۳۲۵۶ نمودارهای وابستگی جزئی میانگین پیش بینی ها را بر روی توزیع حاشیه ای نشان می دهند.

$$\hat{f}_{S,PDP}(x) = E_{X_C} \left[\hat{f}(x_S, X_C) \right] \\ = \int_{X_C} \hat{f}(x_S, X_C) d\mathbb{P}(X_C)$$

۳۲۵۸ این مقدار تابع پیش بینی f ، در مقدار(های) ویژگی $(S \setminus X_C)$ است که بر روی همه ویژگی ها در (X_C) میانگین گرفته شده است (در اینجا به عنوان متغیرهای تصادفی در نظر گرفته می شود). میانگین گیری به معنای
 ۳۲۵۹ محاسبه انتظار حاشیه ای E بر روی ویژگی های مجموعه C است، که انتگرال بیش از پیش بینی های وزن شده
 ۳۲۶۰ توسط توزیع احتمال است. جالب به نظر می رسد، اما برای محاسبه مقدار مورد انتظار بر روی توزیع حاشیه ای،
 ۳۲۶۱ ما به سادگی تمام نمونه های داده خود را می گیریم، آنها را مجبور می کنیم که یک مقدار شبکه مشخصی برای
 ۳۲۶۲ ویژگی های مجموعه S داشته باشند، و پیش بینی های این مجموعه داده دستکاری شده را به طور میانگین
 ۳۲۶۳ پیش بینی می کنیم. این روش تضمین می کند که ما از توزیع حاشیه ای ویژگی ها میانگین می گیریم.
 ۳۲۶۴

۳۲۶۵ نمودارهای M پیش بینی ها را بر روی توزیع شرطی میانگین می دهند.

$$\hat{f}_{S,M}(x_S) = E_{X_C|X_S} \left[\hat{f}(X_S, X_C) | X_S = x_S \right] \\ = \int_{X_C} \hat{f}(x_S, X_C) d\mathbb{P}(X_C | X_S = x_S)$$

۳۲۶۶ تنها چیزی که در مقایسه با PDP ها تغییر می کند این است که به جای اینکه توزیع حاشیه ای را در هر مقدار
 ۳۲۶۷ شبکه فرض کنیم، پیش بینی ها را مشروط به هر مقدار شبکه ای از ویژگی مورد نظر، میانگین می کنیم. در
 ۳۲۶۸ عمل، این بدان معنی است که ما باید یک محله تعریف کنیم، به عنوان مثال برای محاسبه اثر ۳۰ متر مربع بر
 ۳۲۶۹

۳۲۷۰ ارزش خانه پیش بینی شده، می توانیم میانگین پیش بینی همه خانه ها بین ۲۸ تا ۳۲ متر مربع را محاسبه کنیم.

۳۲۷۲ نمودارهای ALE میانگین تغییرات در پیش بینی ها را ترسیم می کنند و آنها را در شبکه جمع می کنند (در ادامه

۳۲۷۳ در مورد محاسبه بیشتر).

۳۲۷۴

$$f_{S,ALE}(x_S) = \int_{z_{0,S}}^{x_S} E_{X_C|X_S=x_S} [f^S(X_s, X_c)|X_S = z_S] dz_S - \text{constant}$$
$$= \int_{z_{0,S}}^{x_S} \left(\int_{x_C} f^S(z_s, X_c) d\mathbb{P}(X_C|X_S = z_S) \right) dz_S - \text{constant}$$

۳۲۷۵ این فرمول سه تفاوت را با M-Plots نشان می دهد. ابتدا، تغییرات پیش بینی ها را میانگین می گیریم، نه خود

۳۲۷۶ پیش بینی ها. تغییر به عنوان مشتق جزئی تعریف می شود (اما بعداً برای محاسبه واقعی، با تفاوت در پیش بینی

۳۲۷۷ ها در یک بازه زمانی جایگزین می شود).

۳۲۷۸

$$\hat{f}^S(x_s, x_c) = \frac{\partial f(x_s, x_c)}{\partial x_s}$$

۳۲۷۹ تفاوت دوم انتگرال اضافی روی Z است. ما مشتقات جزئی محلی را در محدوده ویژگی های مجموعه S جمع می

۳۲۸۰ کنیم، که به ما تأثیر ویژگی را بر پیش بینی می دهد. برای محاسبات واقعی، Z ها با شبکه ای از فواصل

۳۲۸۱ جایگزین می شوند که در آن تغییرات پیش بینی را محاسبه می کنیم. روش ALE به جای میانگین گیری

۳۲۸۲ مستقیم پیش بینی ها، تفاوت های پیش بینی را مشروط به ویژگی های S محاسبه می کند و مشتق را روی

۳۲۸۳ ویژگی های S برای تخمین اثر ادغام می کند. خوب، احتمانه به نظر می رسد. اشتراق و ادغام معمولاً یکدیگر را

۳۲۸۴ خنثی می کنند، مانند ابتدا تفریق و سپس جمع کردن همان عدد. چرا اینجا منطقی است؟ مشتق (یا تفاوت

۳۲۸۵ فاصله) اثر ویژگی مورد علاقه را جدا می کند و اثر ویژگی های همبسته را مسدود می کند.

۳۲۸۶ سومین تفاوت نمودارهای ALE با نمودارهای M این است که یک ثابت را از نتایج کم می کنیم. این مرحله

۳۲۸۷ نمودار ALE را در مرکز قرار می دهد تا اثر متوسط روی داده ها صفر شود.

۳۲۸۸ یک مشکل باقی می ماند: همه مدل ها دارای گرادیان نیستند، برای مثال جنگل های تصادفی گرادیان ندارند.

۳۲۸۹ اما همانطور که خواهید دید، محاسبات واقعی بدون گرادیان کار می کند و از فواصل استفاده می کند. اجازه

۳۲۹۰ دهید کمی عمیق تر به تخمین نمودارهای ALE بپردازیم.

۳۲۹۱

۳۲۹۲ ابتدا توضیح خواهم داد که چگونه نمودارهای ALE برای یک ویژگی عددی واحد، بعداً برای دو ویژگی عددی و

۳۲۹۳ برای یک ویژگی طبقه بندی واحد تخمین زده می شوند. برای تخمین اثرات محلی، ویژگی را به فواصل زیادی

۸,۲,۳ برآورد

۳۲۹۴ تقسیم می کنیم و تفاوت های پیش بینی ها را محاسبه می کنیم. این روش مشتقات را تقریب می کند و همچنین
۳۲۹۵ برای مدل های بدون مشتق کار می کند.

۳۲۹۶ ابتدا اثر بدون مرکز را تخمین می زنیم:

$$\hat{f}_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i: x_j^{(i)} \in N_j(k)} [\hat{f}(z_{k,j}, x_{-j}^{(i)}) - \hat{f}(z_{k-1,j}, x_{-j}^{(i)})]$$

۳۲۹۸ اجازه دهد این فرمول را از سمت راست شروع کنیم. نام Accumulated Local Effects به خوبی تمام
۳۲۹۹ اجزای منفرد این فرمول را منعکس می کند. در هسته خود، روش ALE تفاوت های پیش بینی ها را محاسبه
۳۳۰۰ می کند، که به موجب آن ویژگی مورد نظر را با مقادیر شبکه Z جایگزین می کنیم. تفاوت در پیش بینی اثری
۳۳۰۱ است که ویژگی برای یک نمونه فردی در یک بازه زمانی مشخص دارد. مجموع سمت راست اثرات تمام نمونه ها
۳۳۰۲ را در یک بازه جمع می کند که در فرمول به صورت همسایگی $(\lambda_j(N))$ ظاهر می شود. ما این مجموع را بر
۳۳۰۳ تعداد نمونه های این بازه تقسیم می کنیم تا میانگین اختلاف پیش بینی های این بازه را بدست آوریم. این میانگین
۳۳۰۴ در بازه با عبارت Local پوشش داده می شود به نام ALE. نماد مجموع سمت چپ به این معنی است که ما
۳۳۰۵ اثرات میانگین را در تمام بازه ها جمع می کنیم ALE (غیر مرکزی) یک مقدار ویژگی که مثلاً در بازه سوم قرار
۳۳۰۶ دارد، مجموع تأثیرات بازه های اول، دوم و سوم است. کلمه انباشته در ALE این را نشان می دهد.

۳۳۰۷ این اثر در مرکز قرار می گیرد تا اثر میانگین صفر شود.

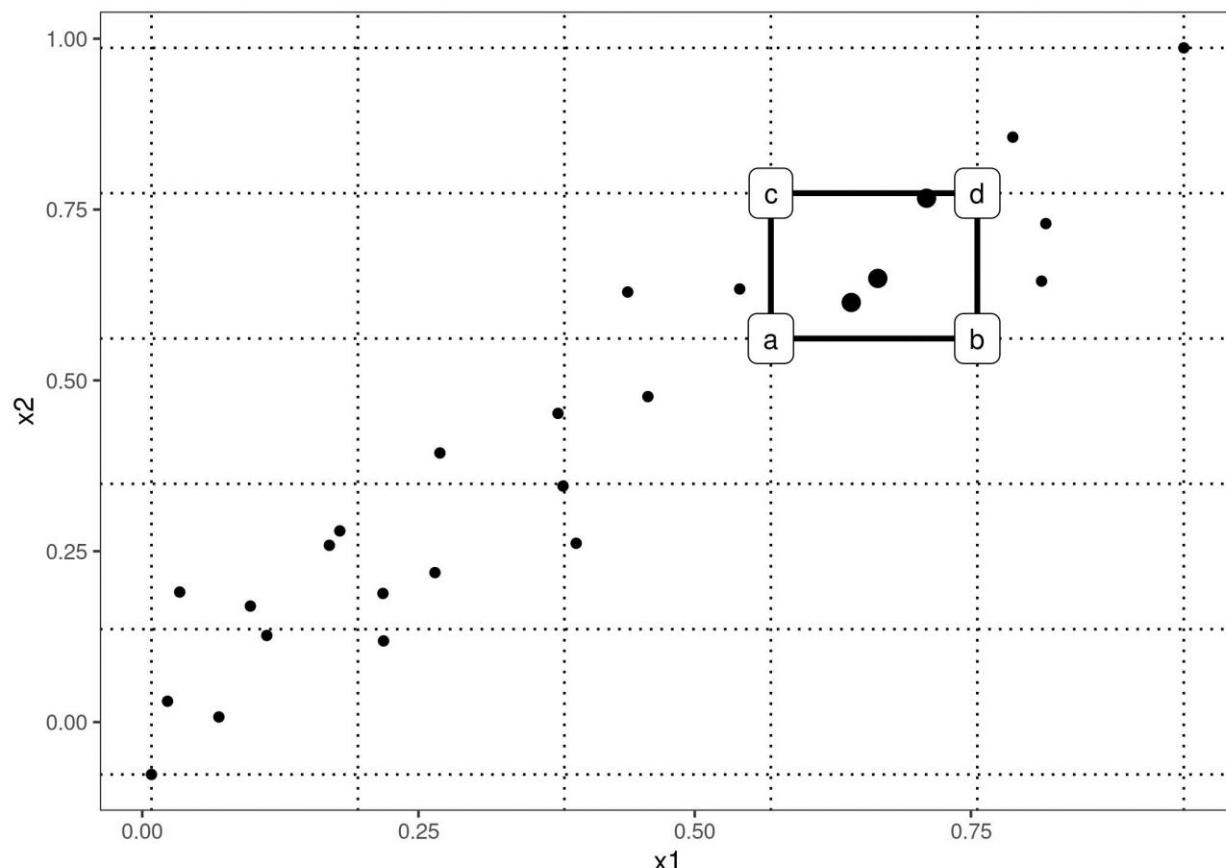
۳۳۰۸ $\hat{f}_{j,ALE}(x) = \hat{f}_{j,ALE}(x) - \frac{1}{n} \sum_{i=1}^n \hat{f}_{j,ALE}(x_j^{(i)})$

۳۳۰۹ مقدار ALE را می توان به عنوان تأثیر اصلی ویژگی در یک مقدار مشخص در مقایسه با میانگین پیش بینی داده
۳۳۱۰ ها تفسیر کرد. به عنوان مثال، تخمین ALE از $\lambda_j(x) = 3$ به این معنی است که وقتی ویژگی Z دارای
۳۳۱۱ مقدار 3 باشد، پیش بینی در مقایسه با پیش بینی متوسط 2 کمتر است.

۳۳۱۲ چندک های توزیع ویژگی به عنوان شبکه ای که فواصل را تعریف می شود. استفاده از چندک
۳۳۱۳ ها تضمین می کند که تعداد نمونه های داده ای یکسان در هر یک از بازه ها وجود دارد. کوانتیل ها این عیب را
۳۳۱۴ دارند که فاصله ها می توانند طول های بسیار متفاوتی داشته باشند. اگر ویژگی مورد نظر بسیار منحرف باشد، به
۳۳۱۵ عنوان مثال مقادیر بسیار کم و تنها چند مقدار بسیار بالا، این می تواند منجر به نمودارهای ALE عجیب و غریب
۳۳۱۶ شود.

۳۳۱۷ برای تعامل دو ویژگی ترسیم می کند ALE

۳۳۱۸ نمودارهای ALE همچنین می توانند اثر متقابل دو ویژگی را نشان دهند. اصول محاسبه مانند یک ویژگی است،
 ۳۳۱۹ اما ما به جای فواصل، با سلول های مستطیلی کار می کنیم، زیرا باید اثرات را در دو بعد جمع کنیم. علاوه بر
 ۳۳۲۰ تنظیم برای اثر میانگین کلی، ما همچنین اثرات اصلی هر دو ویژگی را تنظیم می کنیم. این بدان معنی است که
 ۳۳۲۱ برای دو ویژگی، اثر مرتبه دوم را تخمین می زند، که شامل اثرات اصلی ویژگی ها نمی شود. به عبارت
 ۳۳۲۲ ALE برای دو ویژگی فقط اثر متقابل اضافی دو ویژگی را نشان می دهد. من از فرمول های نمودارهای
 ۳۳۲۳ ALE دو بعدی صرف نظر می کنم زیرا خواندن آنها طولانی و ناخوشایند است. اگر به محاسبه علاقه دارید، شما
 ۳۳۲۴ را به مقاله، فرمول (۱۳) - (۱۶) ارجاع می دهم. من برای ایجاد شهود در مورد محاسبه ALE مرتبه دوم به
 ۳۳۲۵ تجسم ها تکیه خواهم کرد.



۳۳۲۶ شکل ۸: محاسبه D-ALE. ۳۳۲۷ روی دو ویژگی یک شبکه قرار می دهیم. در هر سلول شبکه تفاوت های مرتبه
 ۳۳۲۸ دوم را برای همه نمونه های درون محاسبه می کنیم. ابتدا مقادیر x_1 و x_2 را با مقادیر گوشه سلول جایگزین
 ۳۳۲۹ می کنیم. اگر a، b، c و d پیش‌بینی‌های "گوشه" یک نمونه دستکاری شده را نشان دهند (همانطور که در

نمودار نشان داده شده است)، تفاوت مرتبه دوم (b - a) - (c - d) است. میانگین اختلاف مرتبه دوم در هر سلول روی شبکه جمع شده و در مرکز قرار می گیرد.

در شکل قبل، بسیاری از سلول ها به دلیل همبستگی خالی هستند. در طرح ALE این را می توان با یک کادر خاکستری یا تیره تجسم کرد. یا می توانید تخمین ALE از دست رفته یک سلول خالی را با تخمین نزدیکترین سلول غیر خالی جایگزین کنید.

از آنجایی که تخمین های ALE برای دو ویژگی فقط اثر مرتبه دوم ویژگی ها را نشان می دهد، تفسیر نیاز به توجه ویژه دارد. اثر مرتبه دوم، اثر متقابل اضافی ویژگی ها است، پس از اینکه اثرات اصلی ویژگی ها را در نظر گرفتیم. فرض کنید دو ویژگی با هم تعامل ندارند، اما هر کدام یک اثر خطی بر نتیجه پیش بینی شده دارند. در نمودار ALE 1 بعدی برای هر ویژگی، ما یک خط مستقیم را به عنوان منحنی ALE تخمین زده می بینیم. اما وقتی تخمین های ALE دو بعدی را رسم می کنیم، باید نزدیک به صفر باشند، زیرا اثر مرتبه دوم فقط اثر اضافی تعامل است. نمودارهای ALE و نمودارهای PD از این نظر متفاوت هستند PD:ها همیشه اثر کل را نشان می دهند، نمودارهای ALE اثر مرتبه اول یا دوم را نشان می دهند. اینها تصمیمات طراحی هستند که به ریاضیات اساسی بستگی ندارند.

اثرات محلی انباشته شده را نیز می توان برای مرتبه های خودسرانه بالاتر (تعامل سه یا چند ویژگی) محاسبه کرد، اما همانطور که در فصل PDP بحث شد ، فقط تا دو ویژگی منطقی است، زیرا تعاملات بالاتر را نمی توان تجسم کرد یا حتی معنی دار تفسیر کرد.

ALE برای ویژگی های طبقه بندی شده

روش اثرات محلی انباشته - طبق تعریف - به مقادیر ویژگی نیاز دارد تا نظم داشته باشد، زیرا این روش اثرات را در جهت خاصی جمع می کند. ویژگی های طبقه بندی هیچ نظم طبیعی ندارند. برای محاسبه نمودار ALE برای یک ویژگی طبقه بندی، باید به نحوی یک سفارش ایجاد یا پیدا کنیم. ترتیب دسته ها بر محاسبه و تفسیر اثرات محلی انباشته شده تأثیر می گذارد.

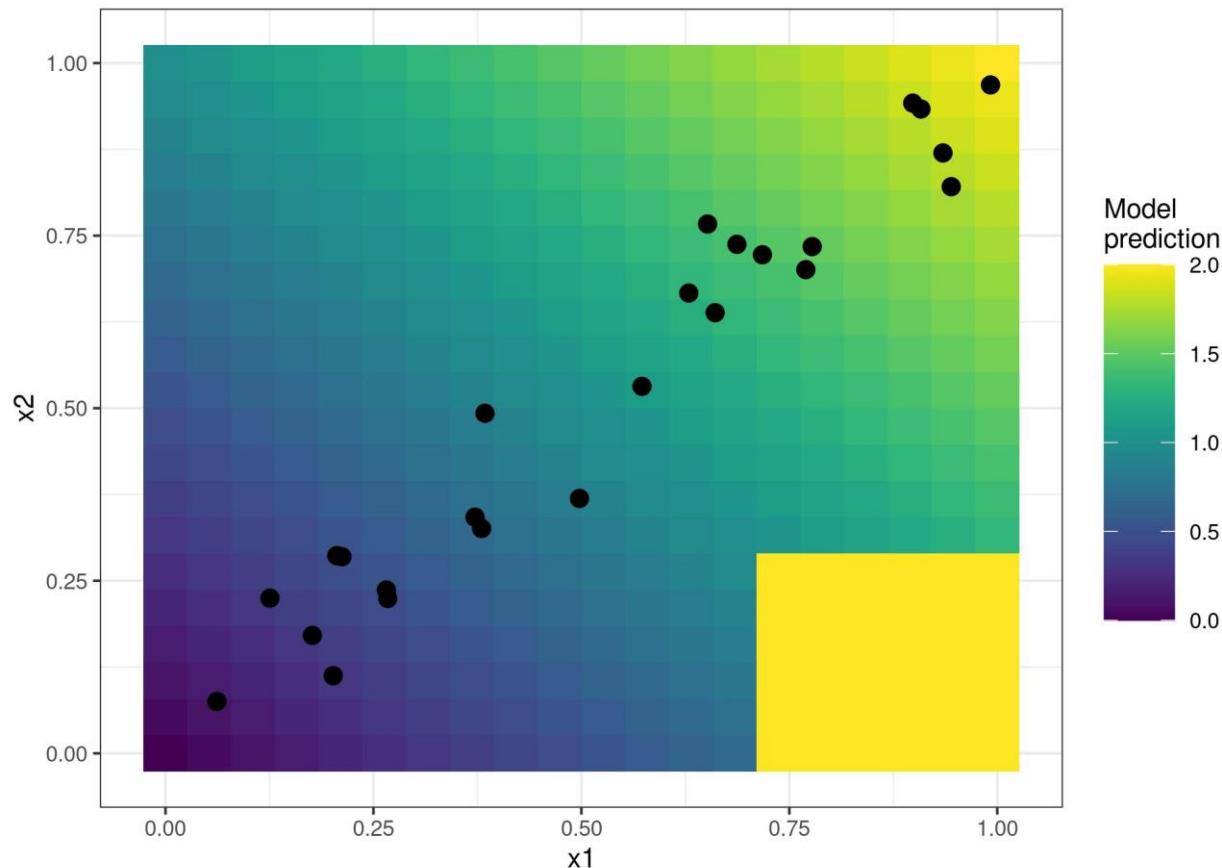
یک راه حل این است که دسته ها را بر اساس شباهت آنها بر اساس سایر ویژگی ها مرتب کنید. فاصله بین دو دسته مجموع فواصل هر ویژگی است. فاصله از نظر ویژگی، توزیع تجمعی را در هر دو دسته، که فاصله کولموگروف- اسمیرنوف نیز نامیده می شود (برای ویژگی های عددی) یا جداول فراوانی نسبی (برای ویژگی های طبقه بندی) مقایسه می کند. هنگامی که فواصل بین همه دسته ها را داریم، از مقیاس بندی چند بعدی برای

۳۳۵۵ کاهش ماتریس فاصله به اندازه فاصله یک بعدی استفاده می کنیم. این یک ترتیب مبتنی بر شباهت از دسته ها
۳۳۵۶ را به ما می دهد.

۳۳۵۷ برای اینکه این موضوع کمی واضح تر شود، در اینجا یک مثال آورده شده است: فرض کنید دو ویژگی طبقه
۳۳۵۸ بندی "فصل" و "آب و هوا" و یک ویژگی عددی "دما" را داریم. برای اولین ویژگی دسته بندی (فصل) می
۳۳۵۹ خواهیم ALE ها را محاسبه کنیم. این ویژگی دارای دسته های «بهار»، «تابستان»، «پاییز»، «زمستان» است. ما
۳۳۶۰ شروع به محاسبه فاصله بین دسته های "بهار" و "تابستان" می کنیم. فاصله مجموع فواصل بیش از ویژگی های
۳۳۶۱ دما و آب و هوا است. برای دما، همه نمونه ها را با فصل «بهار» می گیریم،تابع توزیع تجمعی تجربی را محاسبه
۳۳۶۲ می کنیم و همین کار را برای نمونه هایی با فصل «تابستان» انجام می دهیم و فاصله آنها را با آماره کولموگروف-
۳۳۶۳ اسمیرنوف اندازه گیری می کنیم. برای ویژگی آب و هوا، ما برای همه موارد "بهار" احتمالات را برای هر نوع آب و
۳۳۶۴ هوا محاسبه می کنیم. همین کار را برای نمونه های "تابستانی" انجام دهید و فواصل مطلق را در توزیع احتمال
۳۳۶۵ جمع کنید. اگر «بهار» و «تابستان» دما و آب و هوای بسیار متفاوتی داشته باشند، کل فاصله دسته بزرگ است.
۳۳۶۶ ما این روش را با سایر جفت های فصلی تکرار می کنیم و ماتریس فاصله حاصل را با مقیاس بندی چند بعدی به
۳۳۶۷ یک بعد کاهش می دهیم.

۸,۲,۴ مثالها

۳۳۶۸ اجازه دهید توطئه های ALE را در عمل ببینیم. من سناریویی ساخته ام که در آن توطئه های وابستگی جزئی
۳۳۶۹ شکست می خورند. این سناریو از یک مدل پیش بینی و دو ویژگی قویاً همبسته تشکیل شده است. مدل
۳۳۷۰ پیش بینی عمدتاً یک مدل رگرسیون خطی است، اما با ترکیبی از دو ویژگی که ما هرگز نمونه هایی را برای آن
۳۳۷۱ مشاهده نکرده ایم، کار عجیبی انجام می دهد.
۳۳۷۲



۳۳۷۳

۳۳۷۴

۳۳۷۵

۳۳۷۶

۳۳۷۷

۳۳۷۸

۳۳۷۹

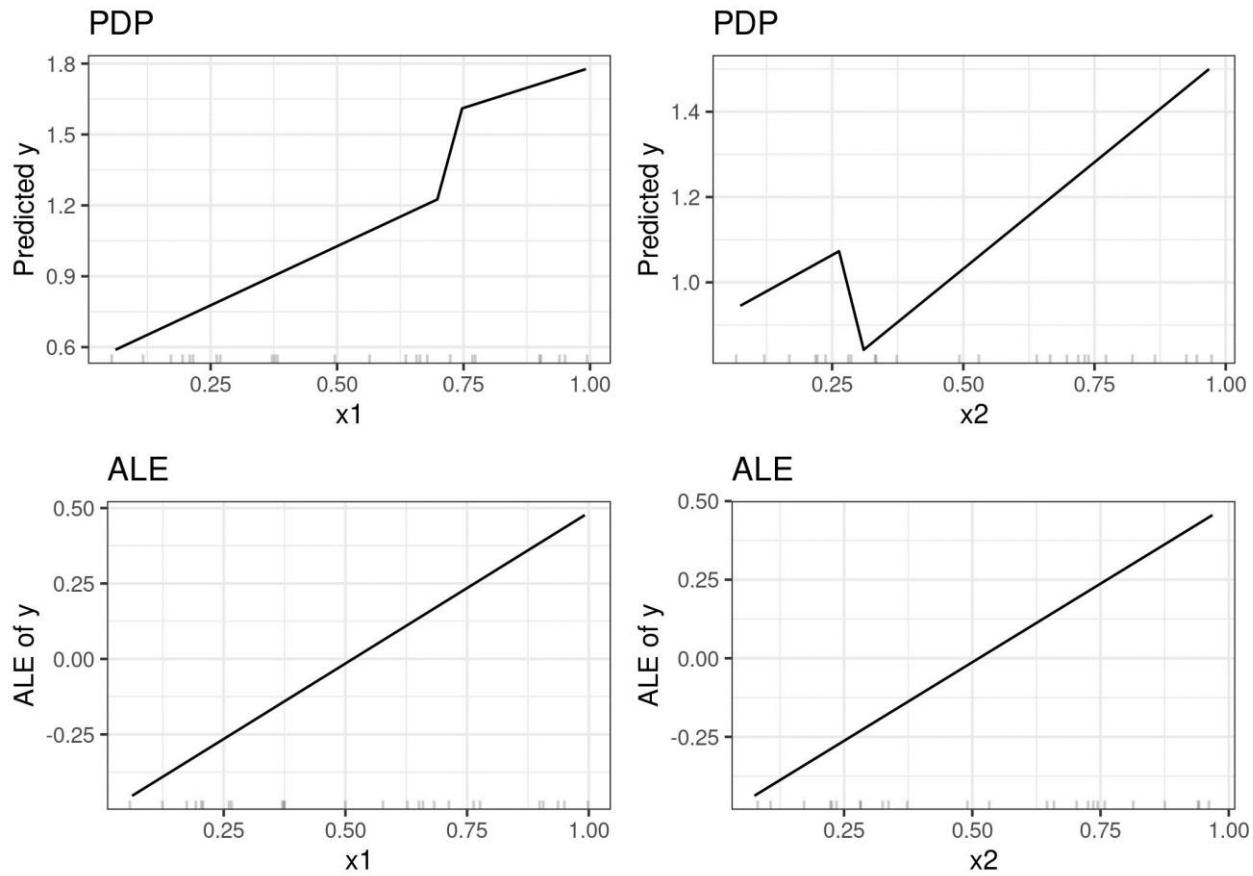
۳۳۸۰

۳۳۸۱

۳۳۸۲

۳۳۸۳

شکل ۸,۹: دو ویژگی و نتیجه پیش بینی شده. مدل مجموع دو ویژگی (پس زمینه سایه دار) را پیش بینی می کند، با این تفاوت که اگر x_1 بزرگتر از ۰,۳ و x_2 کمتر از ۰,۷ باشد، مدل همیشه ۲ را پیش بینی می کند. بر عملکرد مدل تأثیر نمی گذارد و همچنین نباید بر تفسیر آن تأثیر بگذارد.

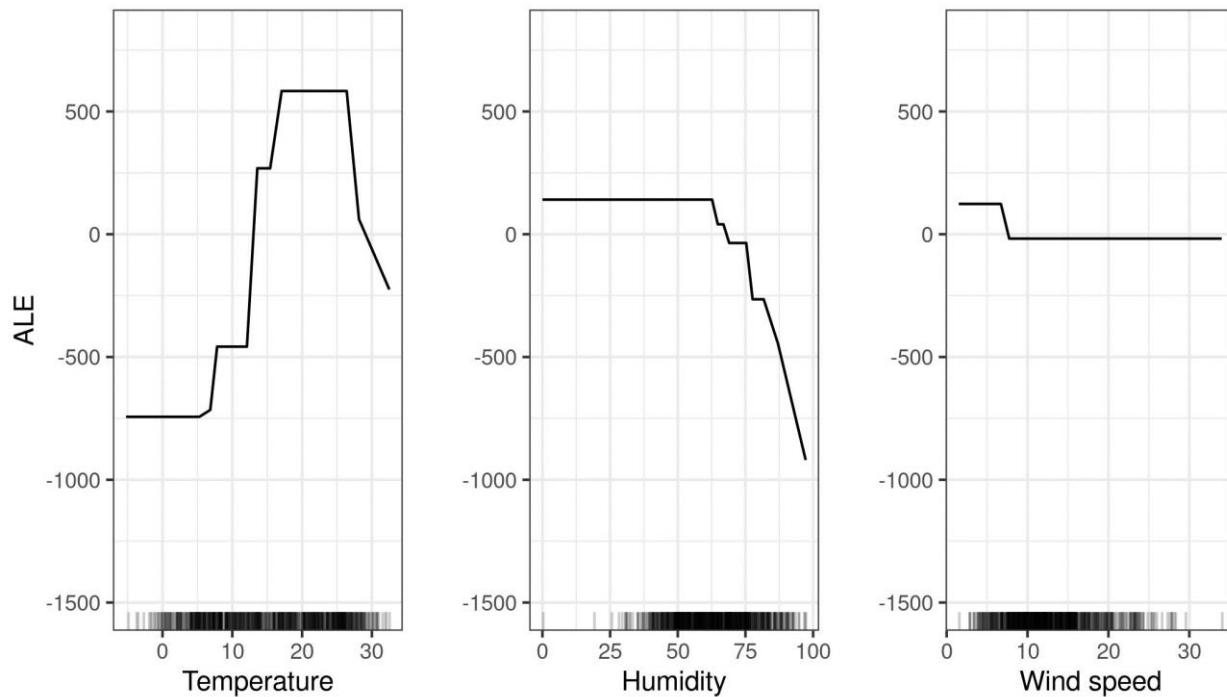


۳۳۸۴

شکل ۸,۱۰: مقایسه اثرات ویژگی محاسبه شده با (PDP ردیف بالا) و (ALE ردیف پایین). تخمین‌های تحت تاثیر رفتار عجیب مدل در خارج از توزیع داده‌ها (پرش‌های تند در نمودارها) قرار دارند. نمودارهای ALE به درستی مشخص می‌کنند که مدل یادگیری ماشین رابطه‌ای خطی بین ویژگی‌ها و پیش‌بینی دارد و مناطق بدون داده را نادیده می‌گیرد.

اما آیا جالب نیست که ببینیم مدل ما در $x_1 < 0.3$ و $x_2 > 0.7$ رفتار عجیبی دارد؟ خوب، بله و نه. از آنجایی که اینها نمونه‌های داده ای هستند که ممکن است از نظر فیزیکی غیرممکن یا حداقل بسیار بعید باشند، معمولاً بررسی این نمونه‌ها بی‌ربط است. اما اگر مشکوک هستید که توزیع آزمایشی شما ممکن است کمی متفاوت باشد و برخی از نمونه‌ها در واقع در آن محدوده هستند، جالب است که این ناحیه را در محاسبه اثرات ویژگی‌ها لحاظ کنید. اما این باید یک تصمیم آگاهانه باشد که شامل مناطقی شود که هنوز داده‌ها را مشاهده نکرده‌ایم و نباید یک اثر جانبی روش انتخابی مانند PDP باشد. اگر مشکوک هستید که مدل بعداً با داده‌های توزیع شده متفاوت استفاده می‌شود، توصیه می‌کنم از نمودارهای ALE استفاده کنید و توزیع داده‌هایی را که انتظار دارید شبیه سازی کنید.

۳۳۹۷ با عطف به یک مجموعه داده واقعی، اجازه دهید تعداد دوچرخه‌های اجاره‌ای را بر اساس آب و هوا و روز
 ۳۳۹۸ پیش‌بینی کنیم و بررسی کنیم که آیا نقشه‌های ALE واقعاً به همان اندازه که وعده داده شده کار می‌کنند یا
 ۳۳۹۹ خیر. ما یک درخت رگرسیون را برای پیش‌بینی تعداد دوچرخه‌های اجاره‌ای در یک روز معین آموزش می‌دهیم
 ۳۴۰۰ و از نمودارهای ALE برای تجزیه و تحلیل چگونگی تأثیر دما، رطوبت نسبی و سرعت باد بر پیش‌بینی‌ها استفاده
 ۳۴۰۱ می‌کنیم. باید ببینیم توطئه‌های ALE چه می‌گویند:



۳۴۰۲ شکل ۸,۱۱ ALE برای مدل پیش‌بینی دوچرخه بر اساس دما، رطوبت و سرعت باد ترسیم می‌کند. دما تأثیر
 ۳۴۰۳ زیادی در پیش‌بینی دارد. میانگین پیش‌بینی با افزایش دما افزایش می‌یابد، اما دوباره به بالای ۲۵ درجه
 ۳۴۰۴ سانتیگراد می‌رسد. رطوبت اثر منفی دارد: وقتی بالای ۶۰ درصد باشد، هر چه رطوبت نسبی بیشتر باشد، پیش
 ۳۴۰۵ بینی کمتر است. سرعت باد روی پیش‌بینی‌ها تأثیر زیادی ندارد.
 ۳۴۰۶

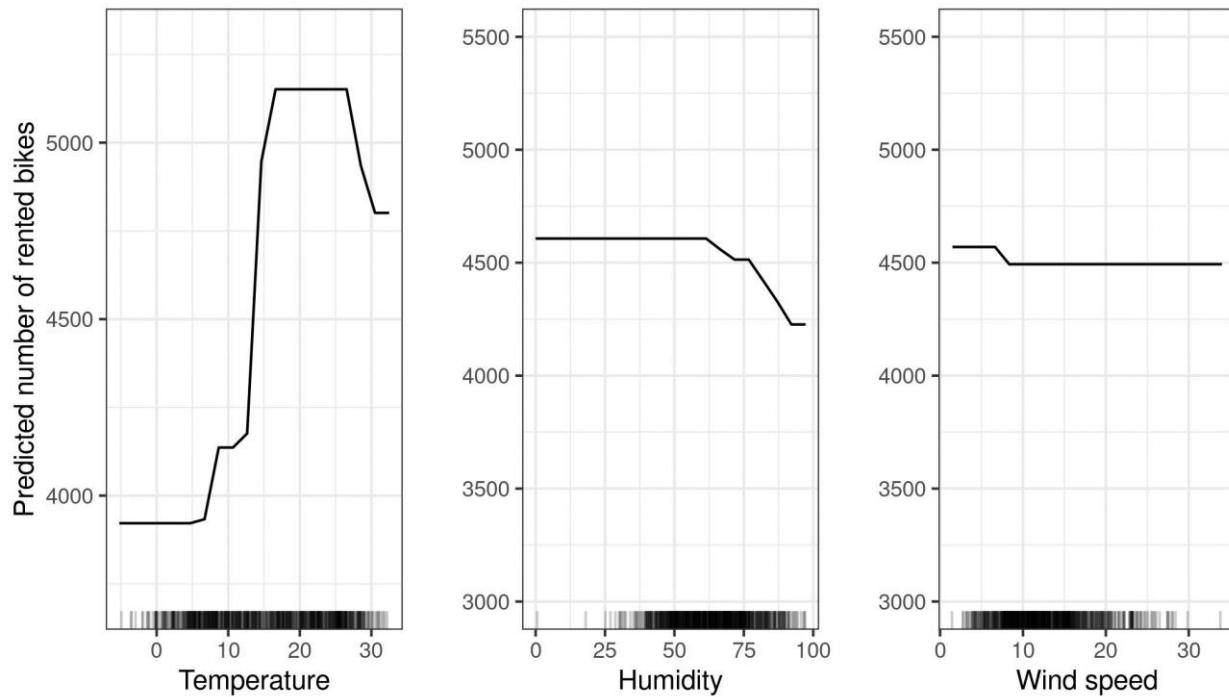
۳۴۰۷ اجازه دهید به همبستگی بین دما، رطوبت و سرعت باد و سایر ویژگی‌ها نگاه کنیم. از آنجایی که داده‌ها دارای
 ۳۴۰۸ ویژگی‌های طبقه‌بندی هستند، ما نمی‌توانیم فقط از ضریب همبستگی پیرسون استفاده کنیم که فقط در
 ۳۴۰۹ صورتی کار می‌کند که هر دو ویژگی عددی باشند. در عوض، من یک مدل خطی را آموزش می‌دهم تا مثلاً دما
 ۳۴۱۰ را بر اساس یکی از ویژگی‌های دیگر به عنوان ورودی پیش‌بینی کنم. سپس اندازه واریانس دیگری را در مدل
 ۳۴۱۱ خطی توضیح می‌دهم و جذر می‌گیرم. اگر ویژگی دیگر عددی بود، نتیجه برابر با قدر مطلق ضریب همبستگی
 ۳۴۱۲ پیرسون استاندارد است. اما این رویکرد مبتنی بر مدل واریانس توضیح داده شده (که ANOVA نیز مخفف

۳۴۱۳ نامیده می‌شود) حتی اگر ویژگی دیگر طبقه‌بندی شده باشد، کار می‌کند. اندازه
 ۳۴۱۴ گیری "توضیح واریانس" همیشه بین ۰ (بدون ارتباط) و ۱ قرار دارد (دما را می‌توان کاملاً از ویژگی دیگر پیش
 ۳۴۱۵ بینی کرد). ما واریانس توضیح داده شده دما، رطوبت و سرعت باد را با تمام ویژگی‌های دیگر محاسبه می‌کنیم.
 ۳۴۱۶ هر چه واریانس توضیح داده شده (همبستگی) بیشتر باشد، مشکلات (بالقوه) بیشتری در نمودارهای PD وجود
 ۳۴۱۷ دارد. شکل زیر نشان می‌دهد که چقدر ویژگی‌های آب و هوا با سایر ویژگی‌ها ارتباط دارد.



۳۴۱۸ شکل ۸,۱۲: قدرت همبستگی بین دما، رطوبت و سرعت باد با همه ویژگی‌ها، اندازه گیری شده به عنوان مقدار
 ۳۴۱۹ واریانس توضیح داده شده، زمانی که ما یک مدل خطی را با دما برای پیش‌بینی و فصل به عنوان ویژگی آموزش
 ۳۴۲۰ می‌دهیم. برای دما، ما - تعجب آور نیست - همبستگی بالایی با فصل و ماه مشاهده می‌کنیم. رطوبت با
 ۳۴۲۱ وضعیت آب و هوا ارتباط دارد.
 ۳۴۲۲

۳۴۲۳ این تجزیه و تحلیل همبستگی نشان می‌دهد که ما ممکن است با مشکلاتی با نمودارهای وابستگی جزئی مواجه
 ۳۴۲۴ شویم، به ویژه برای ویژگی دما. خب خودتون ببینید:



۳۴۲۵

۳۴۲۶

۳۴۲۷

۳۴۲۸

۳۴۲۹

شکل ۸,۱۳: PDP ها برای دما، رطوبت و سرعت باد. در مقایسه با نمودارهای ALE، PDP ها کاهش کمتری در تعداد پیش‌بینی شده دوچرخه‌ها برای دمای بالا یا رطوبت بالا نشان می‌دهند. از تمام نمونه‌های داده برای محاسبه اثر دماهای بالا استفاده می‌کند، حتی اگر مثلاً نمونه‌هایی با فصل «زمستان» باشند. نمودارهای ALE قابل اعتمادتر هستند.

۳۴۳۰

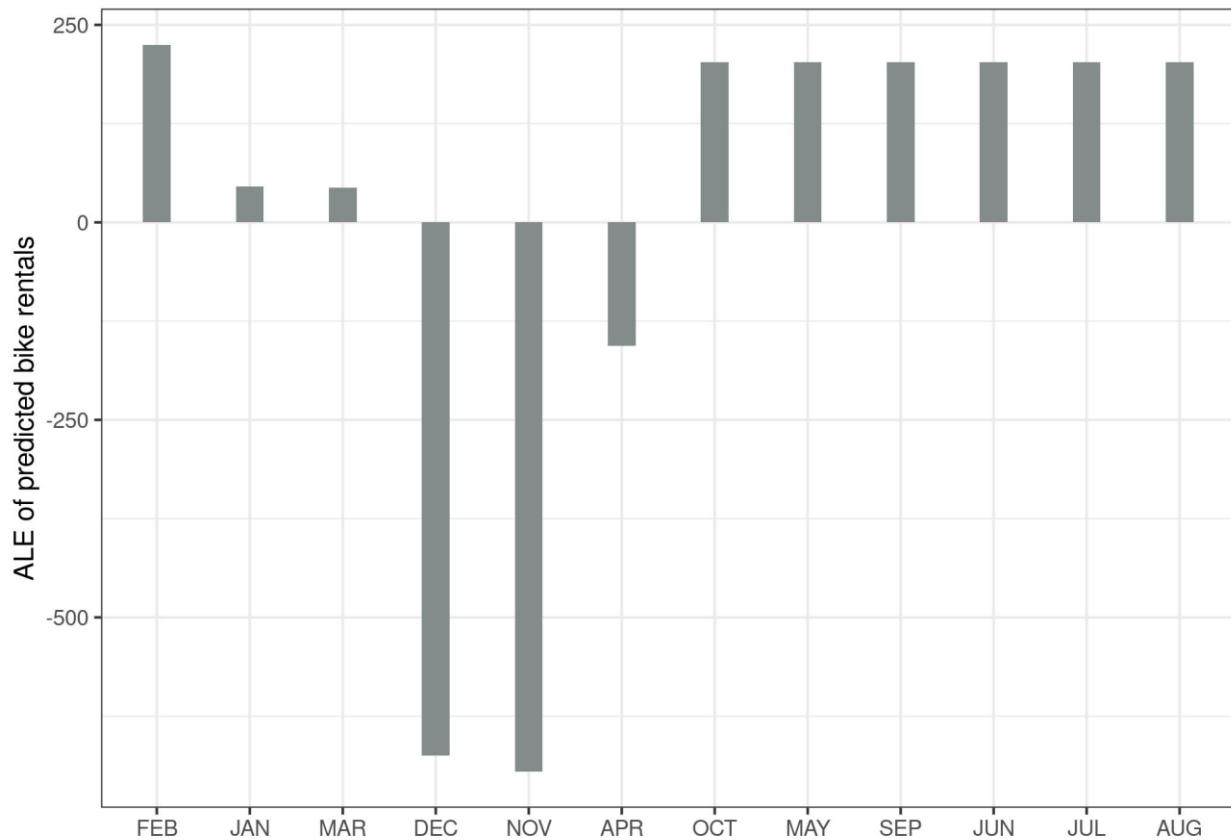
۳۴۳۱

۳۴۳۲

۳۴۳۳

۳۴۳۴

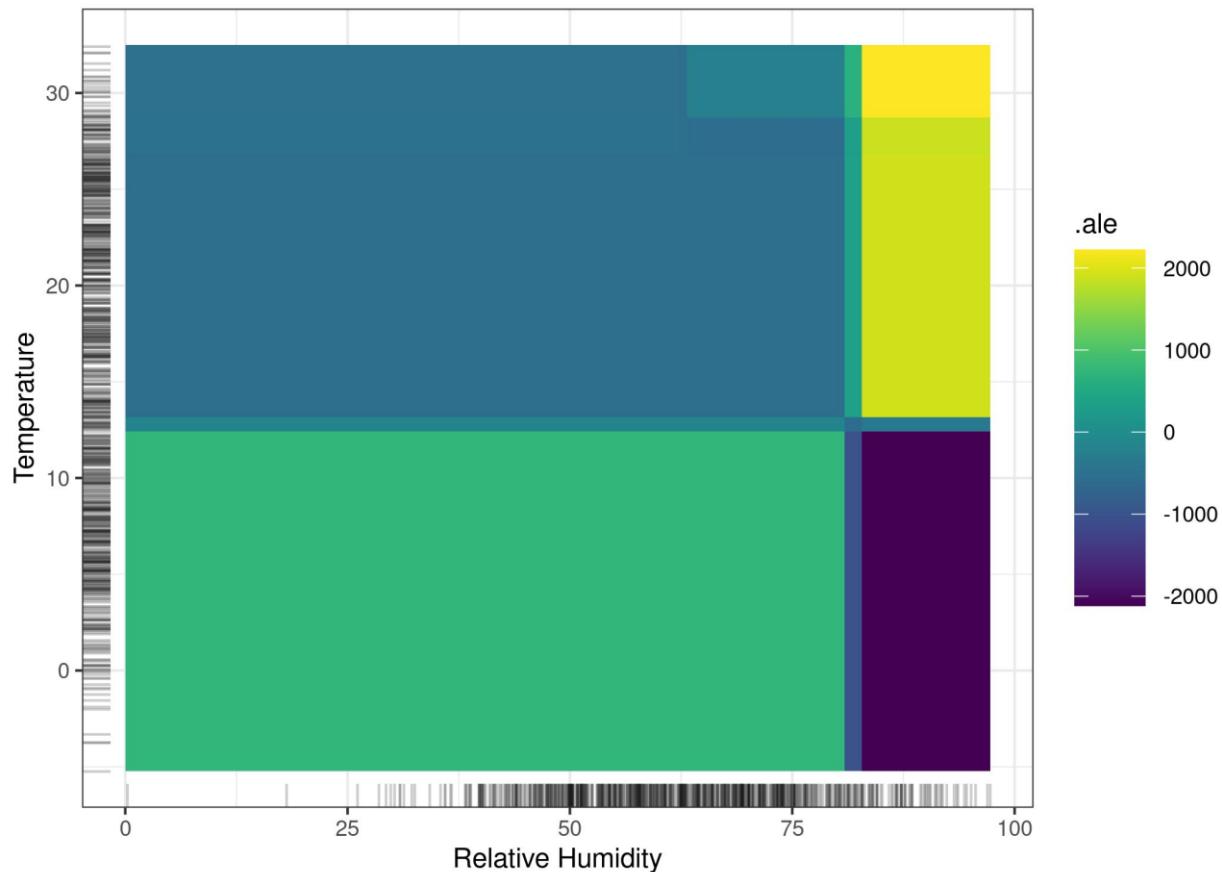
بعد، اجازه دهید نمودارهای ALE را در عمل برای یک ویژگی طبقه بندی ببینیم. ماه یک ویژگی طبقه بندی است که می‌خواهیم تأثیر آن بر تعداد پیش‌بینی شده دوچرخه‌ها را تجزیه و تحلیل کنیم. مسلماً، ماهها از قبل نظم خاصی دارند (زانویه تا دسامبر)، اما اجازه دهید ببینیم اگر ابتدا دسته‌ها را بر اساس شباهت دوباره ترتیب دهیم و سپس اثرات را محاسبه کنیم، چه اتفاقی می‌افتد. ماه‌ها بر اساس شباهت روزهای هر ماه بر اساس ویژگی‌های دیگر مانند دما یا تعطیلی آن مرتب می‌شوند.



۳۴۳۵ شکل ۱۴: نمودار ALE برای ماه ویژگی طبقه بندی شده. ماه ها بر اساس شباهت آنها به یکدیگر، بر اساس
۳۴۳۶ توزیع ویژگی های دیگر بر اساس ماه، مرتب می شوند. مشاهده می کنیم که ژانویه، مارس و آوریل، به ویژه آذر
۳۴۳۷ و آبان، در مقایسه با ماه های دیگر تأثیر کمتری بر تعداد دوچرخه های اجاره ای پیش بینی شده دارند.
۳۴۳۸

۳۴۳۹ از آنجایی که بسیاری از ویژگی ها مربوط به آب و هوای است، ترتیب ماه ها به شدت نشان دهنده شباهت آب و
۳۴۴۰ هوای بین ماه ها است. تمام ماه های سردتر در سمت چپ (فوریه تا آوریل) و ماه های گرمتر در سمت راست
۳۴۴۱ (اکتبر تا آگوست) قرار دارند. به خاطر داشته باشید که ویژگی های غیر آب و هوایی نیز در محاسبه شباهت
۳۴۴۲ لحاظ شده است، به عنوان مثال فراوانی نسبی تعطیلات وزنی برابر با دمای محاسبه شباهت بین ماه ها دارد.

۳۴۴۳ در مرحله بعد، تأثیر درجه دوم رطوبت و دما را بر تعداد پیش بینی شده دوچرخه ها در نظر می گیریم. به یاد
۳۴۴۴ داشته باشید که افکت مرتبه دوم اثر متقابل اضافی دو ویژگی است و افکت های اصلی را شامل نمی شود. این
۳۴۴۵ بدان معنی است که، برای مثال، شما اثر اصلی را نخواهید دید که رطوبت بالا منجر به تعداد کمتری از
۳۴۴۶ دوچرخه های پیش بینی شده به طور متوسط در نمودار ALE درجه دوم می شود.



۳۴۴۷

۳۴۴۸

۳۴۴۹

۳۴۵۰

۳۴۵۱

۳۴۵۲

۳۴۵۳

۳۴۵۴

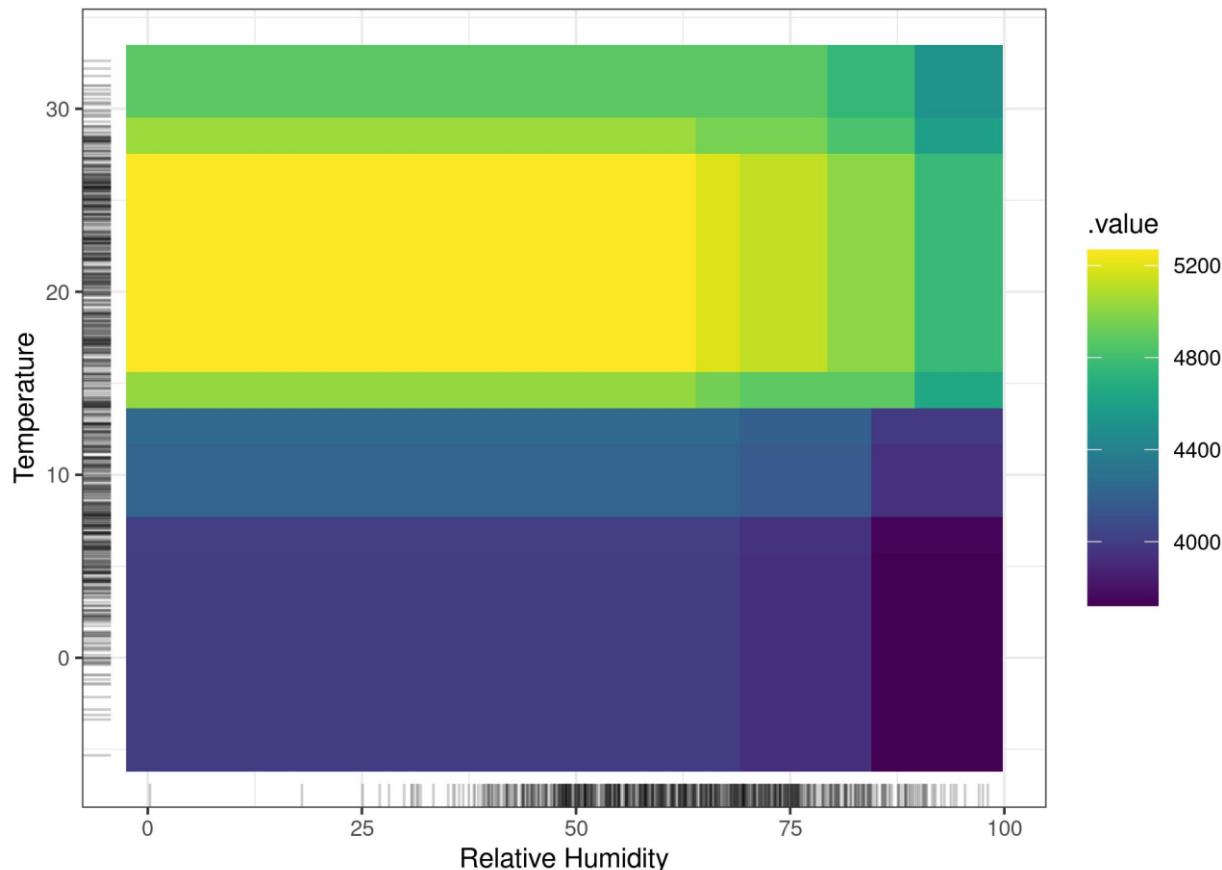
۳۴۵۵

۳۴۵۶

۳۴۵۷

شکل ۸,۱۵: نمودار ALE برای اثر مرتبه دوم رطوبت و دما بر تعداد پیش‌بینی شده دوچرخه‌های اجاره‌ای. سایه روشان‌تر نشان‌دهنده یک پیش‌بینی بالاتر از حد متوسط و سایه تیره‌تر یک پیش‌بینی کمتر از میانگین زمانی است که اثرات اصلی قبلًا در نظر گرفته شده‌اند. این طرح یک تعامل بین دما و رطوبت را نشان می‌دهد: هوای گرم و مرطوب پیش‌بینی را افزایش می‌دهد. در هوای سرد و مرطوب یک اثر منفی اضافی بر تعداد دوچرخه‌های پیش‌بینی شده نشان داده است.

به خاطر داشته باشید که هر دو اثر اصلی رطوبت و دما می‌گویند که تعداد پیش‌بینی شده دوچرخه‌ها در هوای بسیار گرم و مرطوب کاهش می‌یابد. بنابراین در هوای گرم و مرطوب، اثر ترکیبی دما و رطوبت مجموع اثرات اصلی نیست، بلکه بزرگتر از مجموع آن است. برای تأکید بر تفاوت بین افکت مرتبه دوم خالص (نقشه 2 ALE بعدی که همین الان دیدید) و اثر کل، اجازه دهید به طرح وابستگی جزئی نگاه کنیم PDP. اثر کل را نشان می‌دهد که پیش‌بینی میانگین، دو اثر اصلی و اثر مرتبه دوم (تقابل) را ترکیب می‌کند.

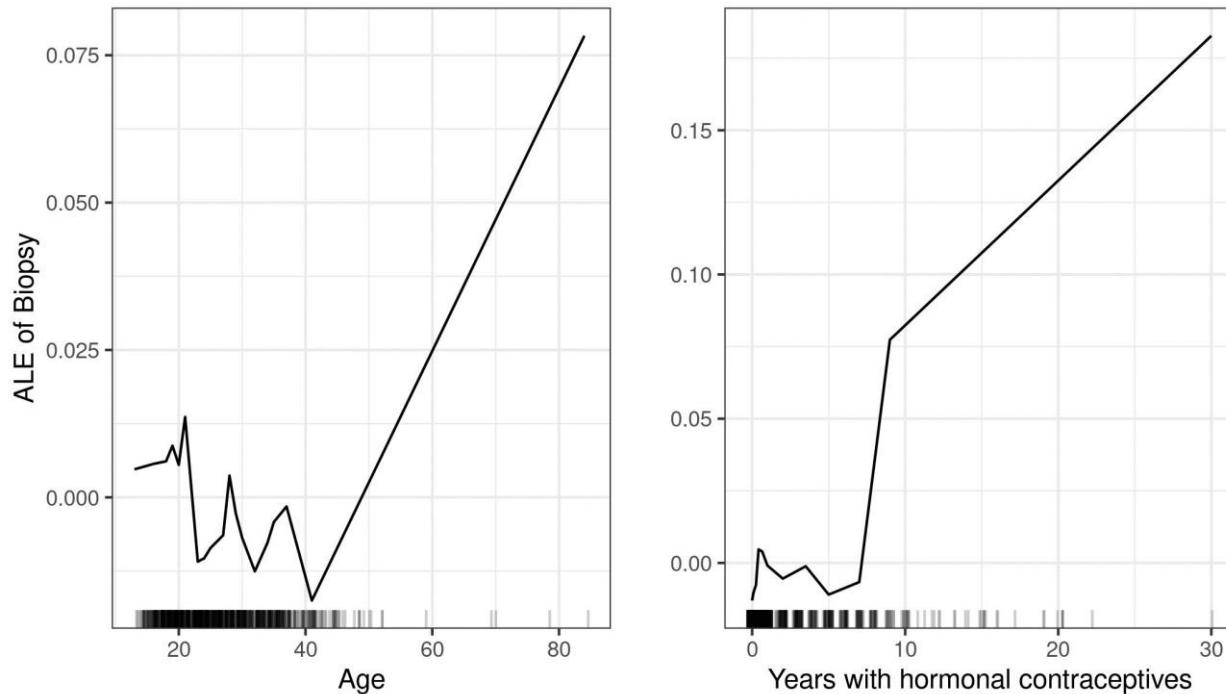


۳۴۵۸

شکل ۸,۱۶ PDP :اثر کل دما و رطوبت بر تعداد پیش‌بینی شده دوچرخه‌ها. طرح ترکیبی از اثر اصلی هر یک از
ویژگی‌ها و اثر متقابل آنها است، برخلاف طرح ۲ D-ALE که فقط تعامل را نشان می‌دهد.

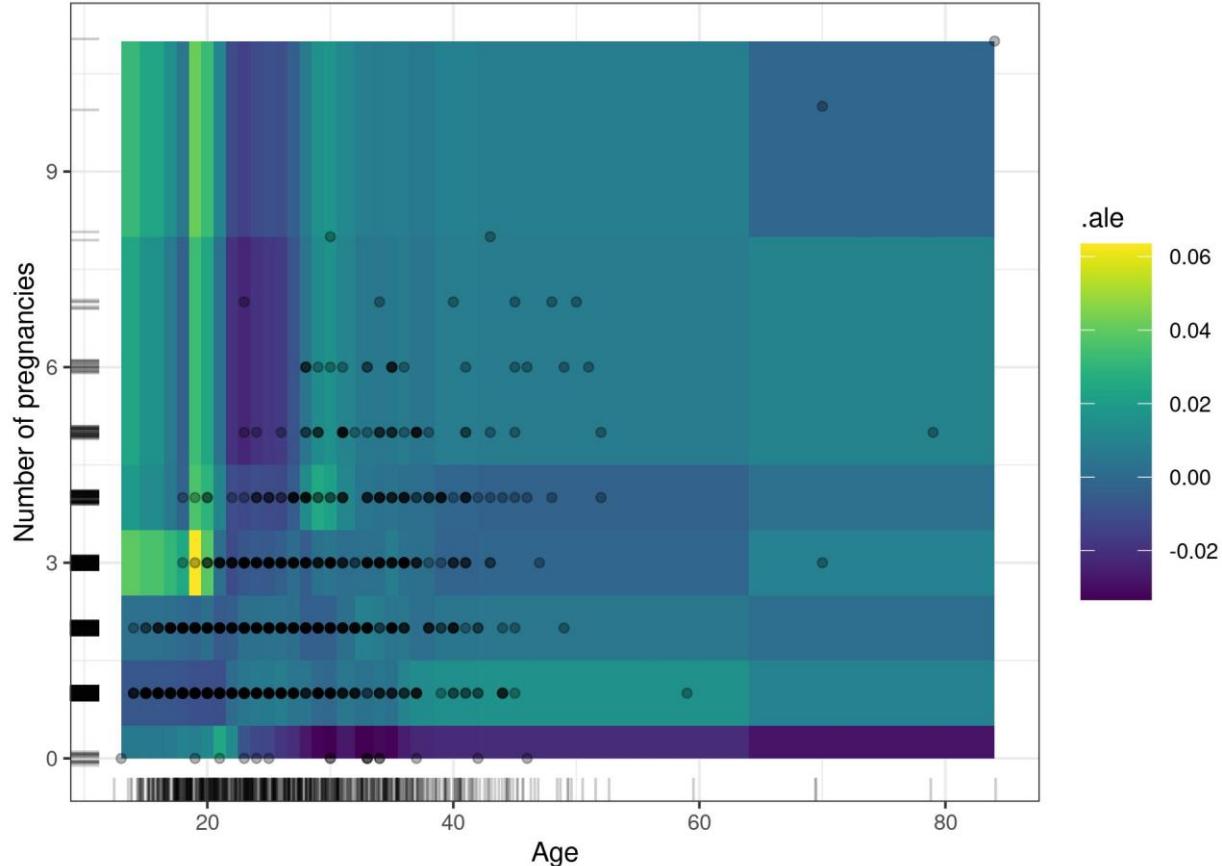
اگر فقط به تعامل علاقه دارید، باید به افکت‌های درجه دوم نگاه کنید، زیرا اثر کل، جلوه‌های اصلی را در طرح
ترکیب می‌کند. اما اگر می‌خواهید اثر ترکیبی ویژگی‌ها را بدانید، باید به اثر کل (که PDP نشان می‌دهد)
نگاه کنید. برای مثال، اگر می‌خواهید تعداد دوچرخه‌های مورد انتظار در دمای ۳۰ درجه سانتی‌گراد و رطوبت
درصد را بدانید، می‌توانید آن را مستقیماً از PDP دو بعدی بخوانید. اگر می‌خواهید همان را از نمودارهای
ALE بخوانید، باید به سه نمودار نگاه کنید: نمودار ALE برای دما، برای رطوبت و برای دما + رطوبت و همچنین
باید پیش‌بینی میانگین کلی را بدانید. در سناریویی که دو ویژگی هیچ تعاملی با هم ندارند، طرح کلی اثر دو
ویژگی می‌تواند گمراه‌کننده باشد، زیرا احتمالاً یک چشم‌انداز پیچیده را نشان می‌دهد، که نشان‌دهنده تعامل
است. اما صرفاً محصول دو اثر اصلی است. اثر مرتبه دوم بلافصله نشان می‌دهد که هیچ تعاملی وجود ندارد.

۳۴۶۹ در حال حاضر دوچرخه کافی است، اجازه دهدید به یک کار طبقه بندی بپردازیم. ما یک جنگل تصادفی برای
۳۴۷۰ پیش‌بینی احتمال سرطان دهانه رحم بر اساس عوامل خطر آموخته می‌دهیم. ما جلوه‌های محلی انباشته شده را
۳۴۷۱ برای دو ویژگی تجسم می‌کنیم:



۳۴۷۲ شکل ۸,۱۷ ALE: اثر سن و سال با داروهای ضد بارداری هورمونی بر احتمال پیش‌بینی شده سرطان دهانه رحم
۳۴۷۳ را ترسیم می‌کند. برای ویژگی سن، نمودار ALE نشان می‌دهد که احتمال سرطان پیش‌بینی شده به طور
۳۴۷۴ متوسط تا سن ۴۰ سالگی کم است و پس از آن افزایش می‌یابد. تعداد سالهای استفاده از داروهای ضد بارداری
۳۴۷۵ هورمونی با افزایش خطر سرطان پیش‌بینی شده بعد از ۸ سال مرتبط است.
۳۴۷۶

۳۴۷۷ در ادامه، به تعامل بین تعداد بارداری و سن نگاه می‌کنیم.



۳۴۷۸

شکل ۸,۱۸: نمودار ALE اثر مرتبه دوم تعداد بارداری و سن. تفسیر طرح کمی غیرقطعی است و نشان می دهد که چه چیزی بیش از حد مناسب است. به عنوان مثال، نمودار یک رفتار مدل عجیب و غریب را در سالین ۱۸ تا ۲۰ سالگی و بیش از ۳ بارداری نشان می دهد (تا ۵ درصد افزایش در احتمال سلطان). تعداد زیادی زن در داده ها با این صورت فلکی سن و تعداد حاملگی وجود ندارد (داده های واقعی به عنوان نقاط نمایش داده می شود)، بنابراین مدل در طول آموزش به دلیل اشتباه برای آن زنان جریمه جدی نمی شود.

۳۴۸۴

۸,۲,۵ مزايا

نمودارهای ALE بی طرفانه هستند ، به این معنی که وقتی ویژگی ها همبستگی دارند، همچنان کار می کنند. نمودارهای وابستگی جزئی در این سناریو شکست می خورند، زیرا ترکیب های بعید یا حتی فیزیکی غیرممکن از مقادیر ویژگی را به حاشیه می برنند.

نمودارهای ALE سریعتر از PDP ها محاسبه می شوند و با $O(n)$ مقیاس می شوند، زیرا بزرگترین تعداد بازه های ممکن تعداد نمونه هایی با یک بازه در هر نمونه است PDP. به n برابر تعداد تخمین نقاط شبکه نیاز دارد.

برای ۲۰ نقطه شبکه، PDP ها ۲۰ برابر بیشتر از نمودار ALE که در بدترین حالت ALE پیش بینی می کنند،
نیاز دارند که در آن فواصل به اندازه نمونه ها استفاده می شود.

تفسیر نمودارهای ALE واضح است : مشروط بر یک مقدار معین، اثر نسبی تغییر ویژگی در پیش بینی را می توان
از نمودار ALE خواند. نمودارهای ALE در مرکز صفر قرار دارند . این باعث می شود تفسیر آنها خوب باشد، زیرا
مقدار در هر نقطه از منحنی ALE تفاوت پیش بینی میانگین است. نمودار 2D ALE فقط تعامل را نشان می
دهد : اگر دو ویژگی با هم تعامل نداشته باشند، نمودار هیچ چیزی را نشان نمی دهد.

کل تابع پیش بینی را می توان به مجموع توابع ALE با ابعاد پایین تر تجزیه کرد ، همانطور که در فصل تجزیه
تابعی توضیح داده شد.

در مجموع، در بیشتر موقعیت ها، نمودارهای ALE را به PDP ترجیح می دهم ، زیرا ویژگی ها معمولاً تا حدی با
هم مرتبط هستند.

۸.۲.۶ معایب

اگر ویژگی ها به شدت همبستگی داشته باشند، تفسیر اثر در فواصل زمانی مجاز نیست . موردی را در نظر
بگیرید که در آن ویژگی های شما بسیار همبسته هستند، و شما به انتهای سمت چپ یک نمودار ۱D-ALE
می کنید. منحنی ALE ممکن است باعث تعبیر نادرست زیر شود: «منحنی ALE نشان می دهد که چگونه
پیش بینی به طور متوسط، زمانی که به تدریج مقدار ویژگی مربوطه را برای یک نمونه داده تغییر می دهیم، و
مقادیر دیگر ویژگی ها را ثابت نگه می داریم، تغییر می کند». اثرات در هر بازه (محلی) محاسبه می شوند و
بنابراین تفسیر اثر فقط می تواند محلی باشد. برای راحتی، جلوه های بازه ای برای نشان دادن یک منحنی صاف
جمع آوری می شوند، اما به خاطر داشته باشید که هر بازه با نمونه های داده متفاوتی ایجاد می شود.

اثرات ALE ممکن است با ضرایب مشخص شده در یک مدل رگرسیون خطی زمانی که ویژگی ها با هم تعامل
دارند و همبستگی دارند، متفاوت باشد. گرومپینگ (۲۰۲۰) ۳۴ نشان داد که در یک مدل خطی با دو ویژگی
همبسته و یک عبارت تعامل اضافی $\hat{f}(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ ()
($\beta_3 x_1 x_2$)، نمودارهای ALE مرتبه اول یک خط مستقیم را نشان نمی دهند. در عوض، آنها
کمی خمیده هستند زیرا بخش هایی از تعامل ضریبی ویژگی ها را در خود جای می دهند. برای درک آنچه در
اینجا اتفاق می افتد، توصیه می کنم فصل تجزیه تابع را بخوانید. به طور خلاصه، اثرات مرتبه اول (یا ۱D)
را متفاوت با فرمول خطی تعریف می کند. این لزوماً اشتباه نیست، زیرا وقتی ویژگی ها همبستگی دارند، نسبت
دادن تعاملات به آن روشن نیست. اما مطمئناً غیر قابل درک است که ALE و ضریب خطی مطابقت ندارند.

نمودارهای ALE می توانند کمی متزلزل شوند (بسیاری از فراز و نشیب های کوچک) با تعداد فواصل زیاد. در ۳۵۱۶ این مورد، کاهش تعداد فواصل، تخمین ها را پایدارتر می کند، اما برخی از پیچیدگی واقعی مدل پیش‌بینی را ۳۵۱۷ هموارتر و پنهان می کند. هیچ راه حل کاملی برای تنظیم تعداد فواصل وجود ندارد . اگر عدد خیلی کوچک ۳۵۱۸ باشد، نمودارهای ALE ممکن است خیلی دقیق نباشند. اگر عدد خیلی زیاد باشد، منحنی می تواند متزلزل شود. ۳۵۱۹

برخلاف PDP ها، نمودارهای ALE با منحنی های ICE همراه نیستند . برای PDP ها، منحنی های ICE عالی ۳۵۲۰ هستند زیرا می توانند ناهمگونی را در اثر ویژگی نشان دهند، به این معنی که اثر یک ویژگی برای زیر مجموعه ۳۵۲۱ های داده متفاوت به نظر می رسد. برای نمودارهای ALE فقط می توانید در هر بازه بررسی کنید که آیا اثر بین ۳۵۲۲ نمونه‌ها متفاوت است یا خیر، اما هر بازه دارای نمونه‌های متفاوتی است بنابراین با منحنی های ICE یکسان ۳۵۲۳ نیست. ۳۵۲۴

تخمین های مرتبه دوم ALE دارای ثبات متفاوتی در فضای ویژگی هستند که به هیچ وجه تجسم نمی شود. ۳۵۲۵ دلیل این امر این است که هر تخمین یک اثر محلی در یک سلول از تعداد متفاوتی از نمونه های داده استفاده ۳۵۲۶ می کند. در نتیجه، تمام تخمین ها دقت متفاوتی دارند (اما هنوز بهترین تخمین های ممکن هستند). مشکل در ۳۵۲۷ نسخه کمتر شدید برای نمودارهای ALE اثر اصلی وجود دارد. به لطف استفاده از چندک ها به عنوان شبکه، ۳۵۲۸ تعداد نمونه ها در همه فواصل یکسان است، اما در برخی مناطق فواصل کوتاه زیادی وجود دارد و منحنی ALE ۳۵۲۹ از تخمین های بسیار بیشتری تشکیل می شود. اما برای فواصل طولانی، که می تواند بخش بزرگی از کل ۳۵۳۰ منحنی را تشکیل دهد، موارد نسبتاً کمتری وجود دارد. این اتفاق در طرح ALE پیش‌بینی سرطان دهانه رحم ۳۵۳۱ برای سن بالا رخ داد. ۳۵۳۲

تفسیر طرح های افکت درجه دوم می تواند کمی آزاردهنده باشد ، زیرا همیشه باید جلوه های اصلی را در ذهن ۳۵۳۳ داشته باشید. خواندن نقشه های حرارتی به عنوان اثر کلی دو ویژگی وسوسه انگیز است، اما این فقط اثر اضافی ۳۵۳۴ تعامل است. افکت مرتبه دوم خالص برای کشف و کاوش فعل و انفعالات جالب است، اما برای تفسیر این که اثر ۳۵۳۵ چگونه به نظر می رسد، فکر می کنم ادغام جلوه های اصلی در طرح منطقی تر است. ۳۵۳۶

اجرای نمودارهای ALE در مقایسه با نمودارهای وابستگی جزئی بسیار پیچیده تر و کمتر بصری است. ۳۵۳۷

حتی اگر نمودارهای ALE در مورد ویژگی های همبسته مغرضانه نیستند، وقتی ویژگی ها به شدت همبسته ۳۵۳۸ باشند، تفسیر مشکل باقی می ماند . زیرا اگر همبستگی بسیار قوی داشته باشند، تنها تحلیل اثر تغییر هر دو ۳۵۳۹ ویژگی با هم و نه به صورت مجزا منطقی است. این نقطه ضعف مختص نمودارهای ALE نیست، بلکه یک مشکل ۳۵۴۰ کلی از ویژگی های همبستگی قوی است. ۳۵۴۱

۳۵۴۲ اگر ویژگی ها همبستگی ندارند و زمان محاسبه مشکلی ندارد، PDP ها کمی ترجیح داده می شوند زیرا در ک
۳۵۴۳ آنها آسان تر است و می توان همراه با منحنی های ICE رسم کرد.

۳۵۴۴ لیست معايب بسیار طولانی شده است، اما فریب تعداد کلماتی را که من به کار می برم نخورید: به عنوان یک
۳۵۴۵ قانون سرانگشتی: به جای PDP از ALE استفاده کنید.

۸,۲,۷ اجرا و جایگزین

۳۵۴۷ آیا اشاره کردم که نمودارهای وابستگی جزئی و منحنی های انتظار شرطی فردی جایگزین هستند؟(=)

۳۵۴۸ نمودارهای ALE در R در بسته ALEPlot توسعه خود مخترع و یک بار در بسته iml پیاده سازی می شوند .
۳۵۴۹ همچنین حداقل دو پیاده سازی پایتون با بسته ALEPython و در Alibi ALE دارد.

۸,۳ تعامل با ویژگی ها

۳۵۵۱ هنگامی که ویژگی ها در یک مدل پیش بینی با یکدیگر تعامل دارند، پیش بینی را نمی توان به عنوان مجموع اثرات ویژگی بیان کرد، زیرا تأثیر یک ویژگی به ارزش ویژگی دیگر بستگی دارد. قید ارسسطو «کل بزرگتر از مجموع اجزای آن است» در حضور فعل و انفعالات اعمال می شود.

۸,۳,۱ تعامل ویژگی؟

۳۵۵۲ اگر یک مدل یادگیری ماشینی بر اساس دو ویژگی پیش بینی کند، می توانیم پیش بینی را به چهار جمله تجزیه کنیم: یک جمله ثابت، یک جمله برای ویژگی اول، یک عبارت برای ویژگی دوم و یک عبارت برای تعامل بین دو ویژگی.

۳۵۵۳ تعامل بین دو ویژگی، تغییر در پیش بینی است که با تغییر ویژگی ها پس از در نظر گرفتن اثرات ویژگی های فردی رخ می دهد.

۳۵۵۴ به عنوان مثال، یک مدل ارزش یک خانه را با استفاده از اندازه خانه (بزرگ یا کوچک) و مکان (خوب یا بد) به عنوان ویژگی ها پیش بینی می کند که چهار پیش بینی ممکن را به دست می دهد:

Location	Size	Prediction
good	big	300,000
good	small	200,000
bad	big	250,000
bad	small	150,000

۳۵۶۳ ما پیش بینی مدل را به بخش های زیر تجزیه می کنیم: یک جمله ثابت (۱۵۰,۰۰۰)، یک اثر برای ویژگی اندازه (۱۰۰,۰۰۰+) اگر بزرگ، + اگر کوچک) و یک اثر برای مکان (+۵۰,۰۰۰ اگر خوب، +۰ اگر بد است). این تجزیه به طور کامل پیش بینی های مدل را توضیح می دهد. هیچ اثر متقابله وجود ندارد، زیرا پیش بینی مدل مجموع اثرات تک ویژگی برای اندازه و مکان است. وقتی یک خانه کوچک را بزرگ می کنید، بدون توجه به موقعیت مکانی، پیش بینی همیشه ۱۰۰,۰۰۰ افزایش می یابد. همچنین تفاوت پیش بینی موقعیت مکانی خوب و بد بدون توجه به اندازه ۵۰,۰۰۰ است.

بیایید اکنون به یک مثال با تعامل نگاه کنیم:

Location	Size	Prediction
good	big	400,000
good	small	200,000
bad	big	250,000
bad	small	150,000

جدول پیش‌بینی را به بخش‌های زیر تجزیه می‌کنیم: یک جمله ثابت (۱۵۰۰۰۰)، یک اثر برای ویژگی اندازه (۱۰۰۰۰۰+) اگر بزرگ، +۰ اگر کوچک) و یک اثر برای مکان (۵۰۰۰۰+) اگر خوب، +۰ اگر بد است.). برای این جدول ما به یک عبارت اضافی برای تعامل نیاز داریم: ۱۰۰۰۰۰+ اگر خانه بزرگ و در موقعیت خوبی باشد. این یک تعامل بین اندازه و مکان است، زیرا در این مورد تفاوت در پیش‌بینی بین یک خانه بزرگ و یک خانه کوچک به مکان بستگی دارد.

یکی از راه‌های تخمین قدرت تعامل این است که اندازه‌گیری شود که چقدر از تغییرات پیش‌بینی به تعامل ویژگی‌ها بستگی دارد. این اندازه گیری آماره H نامیده می‌شود که توسط فریدمن و پوپسکو (۲۰۰۸) معرفی شده است.

۸,۳,۲ نظریه: آماره H فریدمن

قصد داریم به دو مورد بپردازیم: اول، یک معیار تعامل دو طرفه که به ما می‌گوید آیا و تا چه حد دو ویژگی در مدل با یکدیگر تعامل دارند یا خیر. دوم، یک معیار تعامل کلی که به ما می‌گوید آیا و تا چه حد یک ویژگی در مدل با همه ویژگی‌های دیگر تعامل دارد یا خیر. در تئوری، تعاملات دلخواه بین هر تعداد ویژگی قابل اندازه گیری است، اما این دو مورد جالب ترین موارد هستند.

اگر دو ویژگی با هم تعامل نداشته باشند، می‌توانیم تابع وابستگی جزئی را به صورت زیر تجزیه کنیم (با فرض اینکه توابع وابستگی جزئی در مرکز صفر هستند):

$$PD_{jk}(x_j, x_k) = PD_j(x_j) + PD_k(x_k)$$

تابع وابستگی جزئی دو طرفه هر دو ویژگی و (x_j, x_k) توابع وابستگی جزئی ویژگی‌های منفرد.

به همین ترتیب، اگر یک ویژگی با هیچ یک از ویژگی‌های دیگر تعامل نداشته باشد، می‌توانیم تابع پیش‌بینی را بیان کنیم $f(x)$. به عنوان مجموع توابع وابستگی جزئی، که در آن جمع اول فقط به j و دومی به تمام ویژگی‌های دیگر به جزء بستگی دارد:

$$\hat{f}(x) = PD_j(x_j) + PD_{-j}(x_{-j})$$

۳۵۹۲

۳۵۹۳

جایی که (x_{-j}) تابع وابستگی جزئی است که به همه ویژگی‌ها به جز ویژگی- j ام بستگی دارد.

۳۵۹۴

این تجزیه تابع وابستگی جزئی (یا پیش‌بینی کامل) را بدون برهمکنش (بین ویژگی‌های j و k) یا به ترتیب j و همه ویژگی‌های دیگر) بیان می‌کند. در مرحله بعد، تفاوت بین تابع وابستگی جزئی مشاهده شده و تابع تجزیه شده را بدون برهمکنش اندازه‌گیری می‌کنیم. ما واریانس خروجی وابستگی جزئی (برای اندازه‌گیری تعامل بین دو ویژگی) یا کل تابع (برای اندازه‌گیری تعامل بین یک ویژگی و همه ویژگی‌های دیگر) را محاسبه می‌کنیم. مقدار واریانس توضیح داده شده توسط برهمکنش (تفاوت بین PD مشاهده شده و بدون تعامل) به عنوان آماره قدرت اندرکنش استفاده می‌شود. در صورتی که هیچ برهمکنشی وجود نداشته باشد، آماره \cdot و اگر تمام واریانس آن وجود نداشته باشد، ۱ است. PD_{jk} با مجموع توابع وابستگی جزئی توضیح داده می‌شود. آمار تعامل ۱ بین دو ویژگی به این معنی است که هر تابع PD ثابت است و تأثیر بر پیش‌بینی فقط از طریق تعامل حاصل می‌شود. آماره H همچنین می‌تواند بزرگتر از ۱ باشد که تفسیر آن دشوارتر است. این می‌تواند زمانی اتفاق بیفتد که واریانس تعامل دو طرفه بزرگتر از واریانس نمودار وابستگی جزئی دو بعدی باشد.

۳۶۰۴

از نظر ریاضی، آماره H ارائه شده توسط فریدمن و پوپسکو برای تعامل بین ویژگی j و k به صورت زیر است:

۳۶۰۵

۳۶۰۶

همین امر در مورد اندازه‌گیری اینکه آیا یک ویژگی j با هر ویژگی دیگری تعامل دارد یا خیر، صدق می‌کند:

۳۶۰۷

۳۶۰۸

ارزیابی آماره H گران است، زیرا در تمام نقاط داده تکرار می‌شود و در هر نقطه باید وابستگی جزئی ارزیابی شود که به نوبه خود با تمام n نقطه داده انجام می‌شود. در بدترین حالت، برای محاسبه آماره H دو طرفه j در مقابل k و $3n^2$ برای کل آماره j H در مقابل همه) به $2n^2$ فراخوانی به مدل‌های یادگیری ماشینی پیش‌بینی تابع نیاز داریم . برای سرعت بخشیدن به محاسبات، می‌توانیم از n نقطه داده نمونه برداری کنیم. این نقطه ضعف افزایش واریانس تخمین‌های وابستگی جزئی را دارد که باعث می‌شود آماره H ناپایدار باشد. بنابراین اگر از نمونه برداری برای کاهش بار محاسباتی استفاده می‌کنید، مطمئن شوید که از نقاط داده به اندازه کافی نمونه برداری کرده اید.

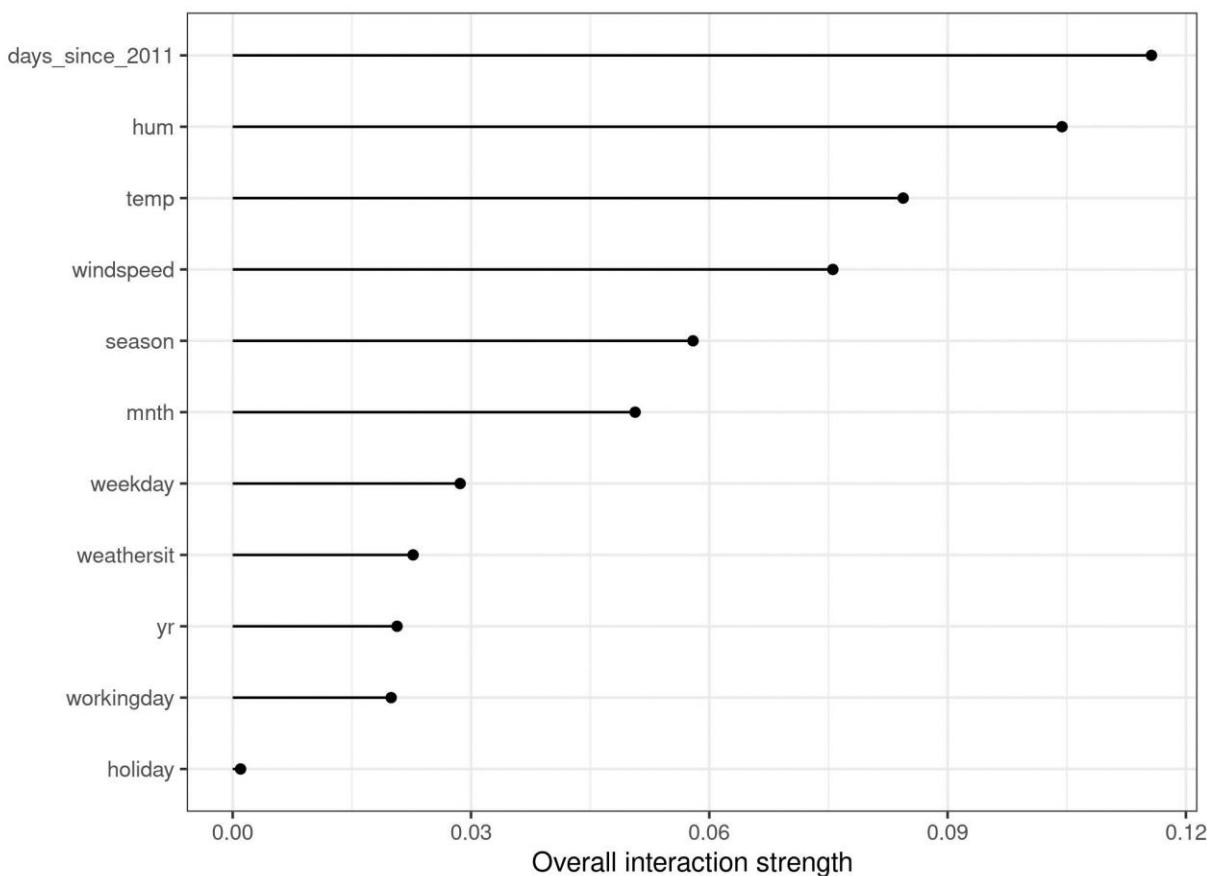
۳۶۱۴

۳۶۱۵ فریدمن و پوپسکو همچنین یک آمار آزمایشی برای ارزیابی اینکه آیا آماره H به طور قابل توجهی با صفر متفاوت است پیشنهاد می کنند. فرضیه صفر عدم وجود تعامل است. برای ایجاد آمار تعامل تحت فرضیه صفر، باید
 ۳۶۱۶ بتوانید مدل را طوری تنظیم کنید که هیچ تعاملی بین ویژگی Z و K یا همه موارد دیگر نداشته باشد. این امکان
 ۳۶۱۷ برای همه مدل ها وجود ندارد. بنابراین این آزمون مختص مدل است، نه مدل آگنوتیک، و به این ترتیب در
 ۳۶۱۸ اینجا پوشش داده نشده است.
 ۳۶۱۹

۳۶۲۰ اگر پیش‌بینی یک احتمال باشد، آماره قدرت تعامل نیز می‌تواند در یک تنظیم طبقه‌بندی اعمال شود.
 ۳۶۲۱

۸.۳.۳ مثالها

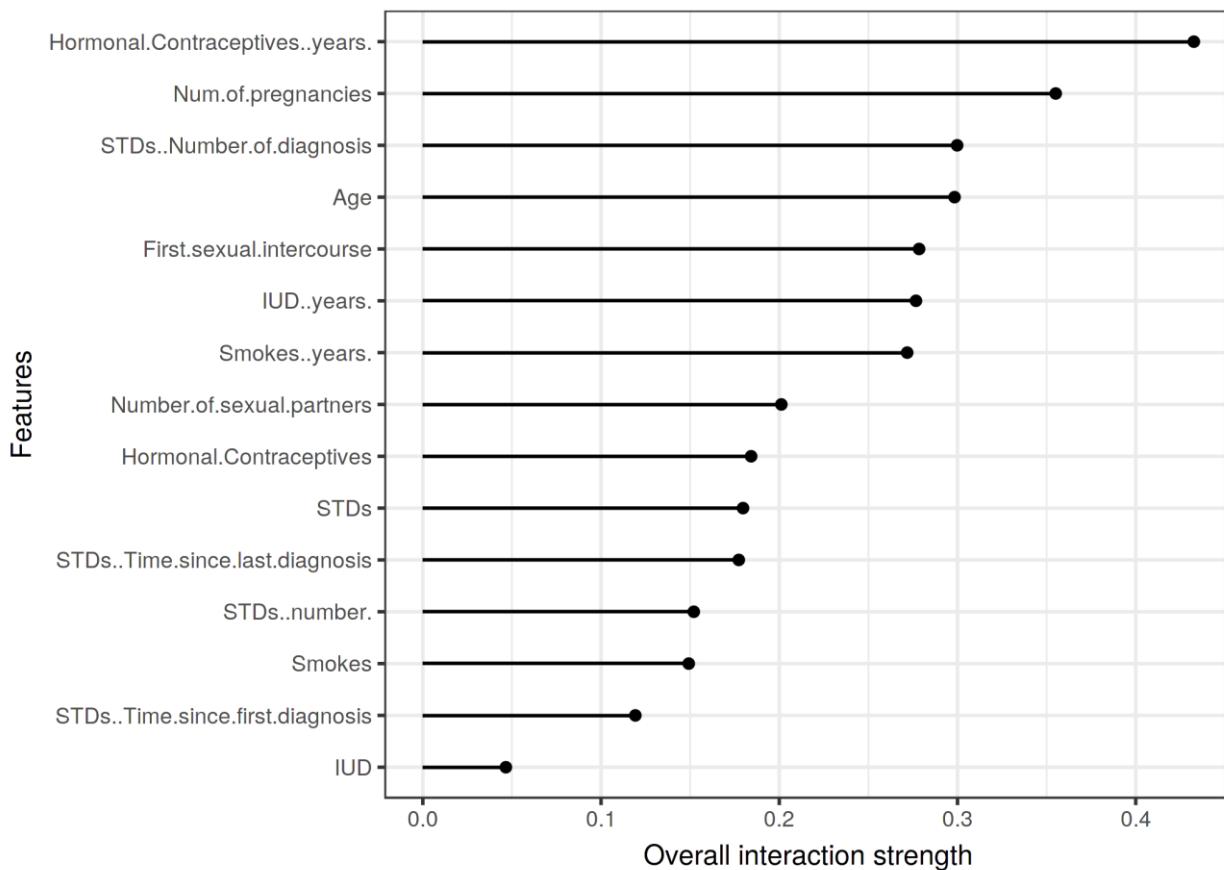
۳۶۲۲ بیایید بینیم تعاملات ویژگی در عمل چگونه است! ما قدرت تعامل ویژگی‌ها را در یک ماشین بردار پشتیبان
 ۳۶۲۳ اندازه‌گیری می‌کنیم که تعداد دوچرخه‌های اجاره‌ای را بر اساس ویژگی‌های آب و هوا و تقویم پیش‌بینی می‌کند.
 ۳۶۲۴ نمودار زیر آمار تعامل ویژگی H را نشان می‌دهد:



۳۶۲۵

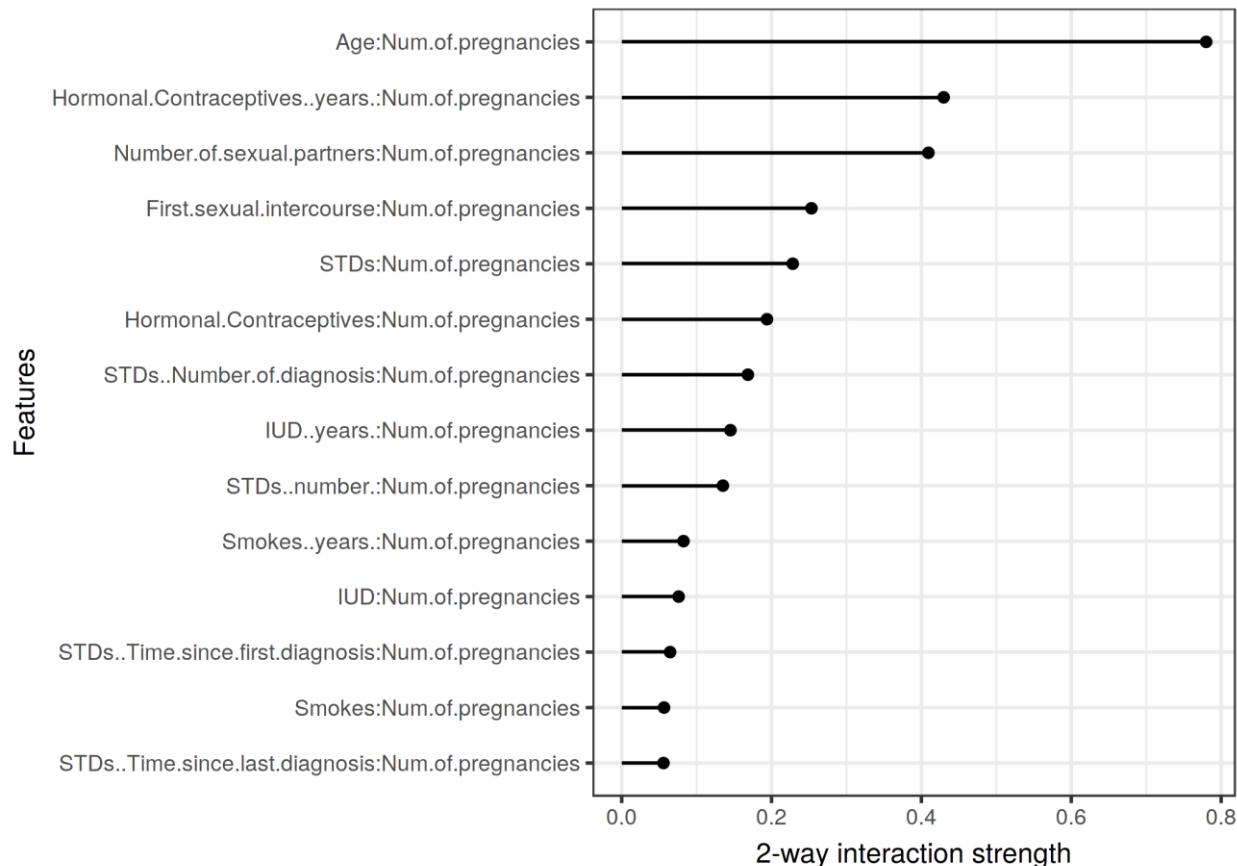
شکل ۸,۱۹: قدرت تعامل (آمار H) برای هر ویژگی با تمام ویژگی های دیگر برای ماشین بردار پشتیبان که اجاره دوچرخه را پیش بینی می کند. به طور کلی، اثرات متقابل بین ویژگی ها بسیار ضعیف است (زیر ۱۰ درصد واریانس توضیح داده شده در هر ویژگی).)

در مثال بعدی، ما آمار تعامل را برای یک مسئله طبقه بندی محاسبه می کنیم. ما تعاملات بین ویژگی ها را در یک جنگل تصادفی که برای پیش بینی سرطان دهانه رحم آموزش دیده است، با توجه به برخی عوامل خطر تجزیه و تحلیل می کنیم.



شکل ۸,۲۰: قدرت تعامل (آمار H) برای هر ویژگی با تمام ویژگی های دیگر برای یک جنگل تصادفی که احتمال سرطان دهانه رحم را پیش بینی می کند. سالهای استفاده از داروهای ضدبارداری هورمونی بالاترین اثر متقابل نسبی را با سایر ویژگی ها دارد و به دنبال آن تعداد حاملگی ها قرار دارد.

پس از بررسی تعاملات هر ویژگی با سایر ویژگی ها، می توانیم یکی از ویژگی ها را انتخاب کنیم و عمیق تر در تمام تعاملات دو طرفه بین ویژگی انتخاب شده و سایر ویژگی ها غوطه ور شویم.



۳۶۳۸

۳۶۳۹

۳۶۴۰

۳۶۴۱

۳۶۴۲

۳۶۴۳

۳۶۴۴

۳۶۴۵

۳۶۴۶

۳۶۴۷

۳۶۴۸

۸,۳,۴ مزايا

شکل ۲۱: نقاط قوت تعامل دو طرفه (آمار H) بین تعداد حاملگی ها و ویژگی های دیگر. یک تعامل قوی بین تعداد بارداری و سن وجود دارد.

آماره تعامل H دارای یک نظریه اساسی از طریق تجزیه و استیگی جزئی است.

آماره H تعبیر معناداری دارد : تعامل به عنوان سهم واریانسی که توسط تعامل توضیح داده می شود، تعریف می شود.

از آنجایی که آمار بدون بعد است ، در بین ویژگی ها و حتی در بین مدل ها قابل مقایسه است. این آمار انواع تعاملات را بدون توجه به شکل خاص آنها تشخیص می دهد.

با آماره H همچنین امکان تحلیل برهمکنش های دلخواه بالاتر مانند قدرت برهمکنش بین ۳ یا چند ویژگی وجود دارد.

۳۶۴۹
۳۶۵۰ اولين چيزی که متوجه خواهيد شد: محاسبه آماره H تعامل زمان زيادي می برد، زيرا از نظر محاسباتي گران است.
۳۶۵۱

۳۶۵۲ محاسبات شامل تخمين توزيع های حاشيه ای است. اگر از تمام نقاط داده استفاده نکنيم، اين تخمين ها داراي
۳۶۵۳ واريанс خاصی هستند . اين بدان معناست که با نمونهبرداری از نقاط، تخمين ها نيز از اجرا به اجرا متفاوت است
۳۶۵۴ و نتایج می توانند ناپايدار باشند . من توصيه می کنم محاسبه آماره H را چند بار تكرار کنيد تا ببینيد آيا داده
۳۶۵۵ های کافی برای به دست آوردن يك نتيجه پايدار داريد يا خير.

۳۶۵۶ مشخص نیست که آيا يك تعامل به طور قابل توجهی بيشتر از α است يا خير. ما باید يك آزمایش آماری انجام
۳۶۵۷ دهیم، اما این تست (هنوز) در يك نسخه مدل-آگنوستيك موجود نیست.

۳۶۵۸ در مورد مسئله آزمون، دشوار است که بگوییم چه زمانی آماره H به اندازه کافی بزرگ است که بتوانیم يك
۳۶۵۹ تعامل را «قوی» در نظر بگیریم.

۳۶۶۰ همچنین، آماره H می تواند بزرگتر از ۱ باشد که تفسیر را دشوار می کند.

۳۶۶۱ زمانی که اثر کلی دو ويژگی ضعيف باشد، اما بيشتر از تعاملات تشکيل شده باشد، آماره H بسيار بزرگ خواهد
۳۶۶۲ بود. اين فعل و انفعالات کاذب به مخرج کوچکی از آماره H نياز دارند و زمانی که ويژگي ها همبستگي دارند
۳۶۶۳ بدتر می شوند. يك تعامل جعلی را می توان به راحتی به عنوان يك اثر متقابل قوي بيش از حد تفسير کرد، در
۳۶۶۴ حالی که در واقعیت هر دو ويژگی نقش کوچکی در مدل دارند. يك راه حل ممکن اين است که نسخه غير عادي
۳۶۶۵ آماره H را تجسم کنيد، که جذر شماره کننده آماره H است . اين آماره H را به همان سطح پاسخ، حداقل برای
۳۶۶۶ رگرسیون، مقیاس می دهد و تاکید کمتری بر تعاملات جعلی دارد.

$$H_{jk}^* = \sqrt{\sum_{i=1}^n \left[PD_{jk}(x_j^{(i)}, x_k^{(i)}) - PD_j(x_j^{(i)}) - PD_k(x_k^{(i)}) \right]^2}$$

۳۶۶۷ آماره H قدرت تعاملات را به ما می گويد، اما به ما نمی گويد که چگونه تعاملات به نظر می رسند. اين همان
۳۶۶۸ چيزی است که توطئه های وابستگی جزئی برای آن هستند. يك گرددش کار معنی دار اين است که نقاط قوت
۳۶۶۹ تعامل را اندازه گيري کنيد و سپس برای تعاملاتی که به آنها علاقه داريد، نموذارهای وابستگی جزئی دو بعدی
۳۶۷۰ ایجاد کنيد.
۳۶۷۱

۳۶۷۲ اگر ورودی ها پیکسل باشند، آماره H نمی تواند به طور معناداري استفاده شود. بنابراین این تکنيک برای طبقه
۳۶۷۳ بندی کننده تصویر مفید نیست.

آمار تعامل با این فرض کار می کند که ما می توانیم ویژگی ها را به طور مستقل به هم بزنیم. اگر ویژگی ها به شدت همبستگی داشته باشند، این فرض نقض می شود و ترکیب های ویژگی هایی را که در واقعیت بسیار بعید هستند، ادغام می کنیم . این همان مشکلی است که نمودار های وابستگی جزئی دارند. ویژگی های همبسته می توانند به مقادیر زیادی از آماره H منجر شوند.

گاهی اوقات نتایج عجیب هستند و برای شبیه سازی های کوچک نتایج مورد انتظار را به همراه نمی آورند . اما این بیشتر یک مشاهده حکایتی است.

۸,۳,۶ پیاده سازی ها

برای مثال های این کتاب، از بسته R استفاده کردم `iml` که در CRAN و نسخه توسعه یافته در GitHub موجود است . پیاده سازی های دیگری نیز وجود دارد که بر روی مدل های خاص تمرکز دارند: بسته R از قبل H-statistic و RuleFit را پیاده سازی می کند . بسته R `gbm` مدل های تقویت شده گرادیان و آماره H را پیاده سازی می کند.

۸,۳,۷ گزینه های جایگزین

آماره H تنها راه برای اندازه گیری تعاملات نیست:

شبکه های تعامل متغیر (VIN) توسط هوکر (2004) ۳۷ رویکردی است کهتابع پیش بینی را به اثرات اصلی و تعاملات ویژگی تجزیه می کند. سپس تعاملات بین ویژگی ها به عنوان یک شبکه تجسم می شود. مatasfanه هنوز نرم افزاری در دسترس نیست.

تعامل ویژگی مبتنی بر وابستگی جزئی توسط گرین ول و همکاران. (2018) ۳۸ تعامل بین دو ویژگی را اندازه گیری می کند. این رویکرد اهمیت ویژگی (تعریف شده به عنوان واریانس تابع وابستگی جزئی) یک ویژگی را مشروط به نقاط مختلف و ثابت ویژگی دیگر می سنجد. اگر واریانس بالا باشد، ویژگی ها با یکدیگر تعامل دارند، اگر صفر باشد، تعامل ندارند. بسته R `GitHubvip` در دسترس است . این بسته همچنین نمودار های وابستگی جزئی و اهمیت ویژگی را پوشش می دهد.

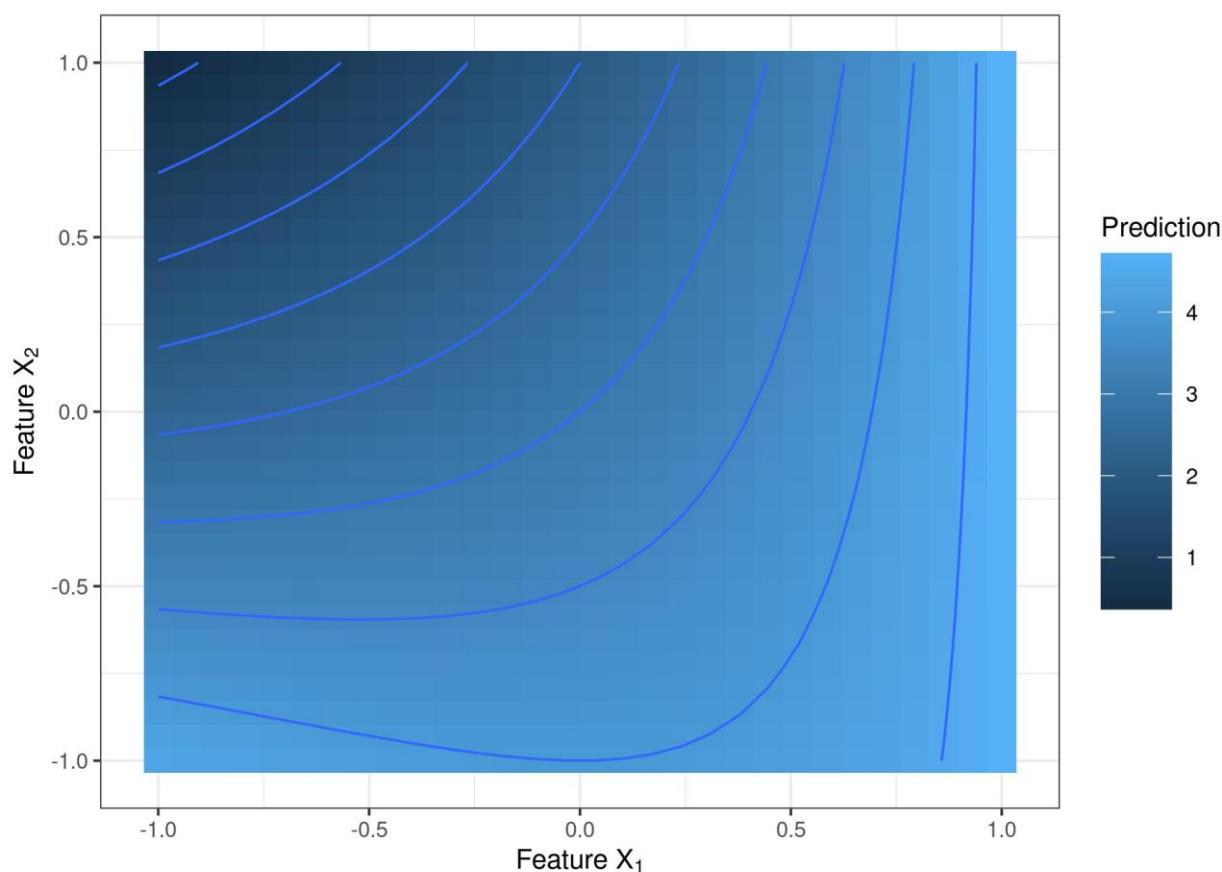
۸.۴ تجزیه عملکردی

۳۶۹۶ یک مدل یادگیری ماشین نظارت شده را می‌توان به عنوان تابع مشاهده کرد که یک بردار ویژگی با ابعاد بالا
۳۶۹۷ را به عنوان ورودی می‌گیرد و یک امتیاز پیش‌بینی یا طبقه‌بندی را به عنوان خروجی ایجاد می‌کند. تجزیه
۳۶۹۸ عملکردی یک تکنیک تفسیری است که عملکرد با ابعاد بالا را تجزیه می‌کند و آن را به صورت مجموع اثرات
۳۶۹۹ ویژگی‌های فردی و اثرات متقابل قابل تجسم بیان می‌کند. علاوه بر این، تجزیه عملکردی یک اصل اساسی است
۳۷۰۰ که زیربنای بسیاری از تکنیک‌های تفسیری است - به شما کمک می‌کند روش‌های تفسیری دیگر را بهتر درک
۳۷۰۱ کنید.
۳۷۰۲

۳۷۰۳ اجازه دهد مستقیماً وارد شویم و به یک تابع خاص نگاه کنیم. این تابع دو ویژگی را به عنوان ورودی می‌گیرد
۳۷۰۴ و یک خروجی یک بعدی تولید می‌کند:

$$y = \hat{f}(x_1, x_2) = 2 + e^{x_1} - x_2 + x_1 \cdot x_2$$

۳۷۰۵ عملکرد را به عنوان یک مدل یادگیری ماشین در نظر بگیرید. ما می‌توانیم تابع را با یک نمودار سه بعدی یا یک
۳۷۰۶ نقشه حرارتی با خطوط کانتور تجسم کنیم:
۳۷۰۷



۳۷۰۸

شکل ۸,۲۲: سطح پیش بینی یکتابع با دو ویژگی X_1 و X_2

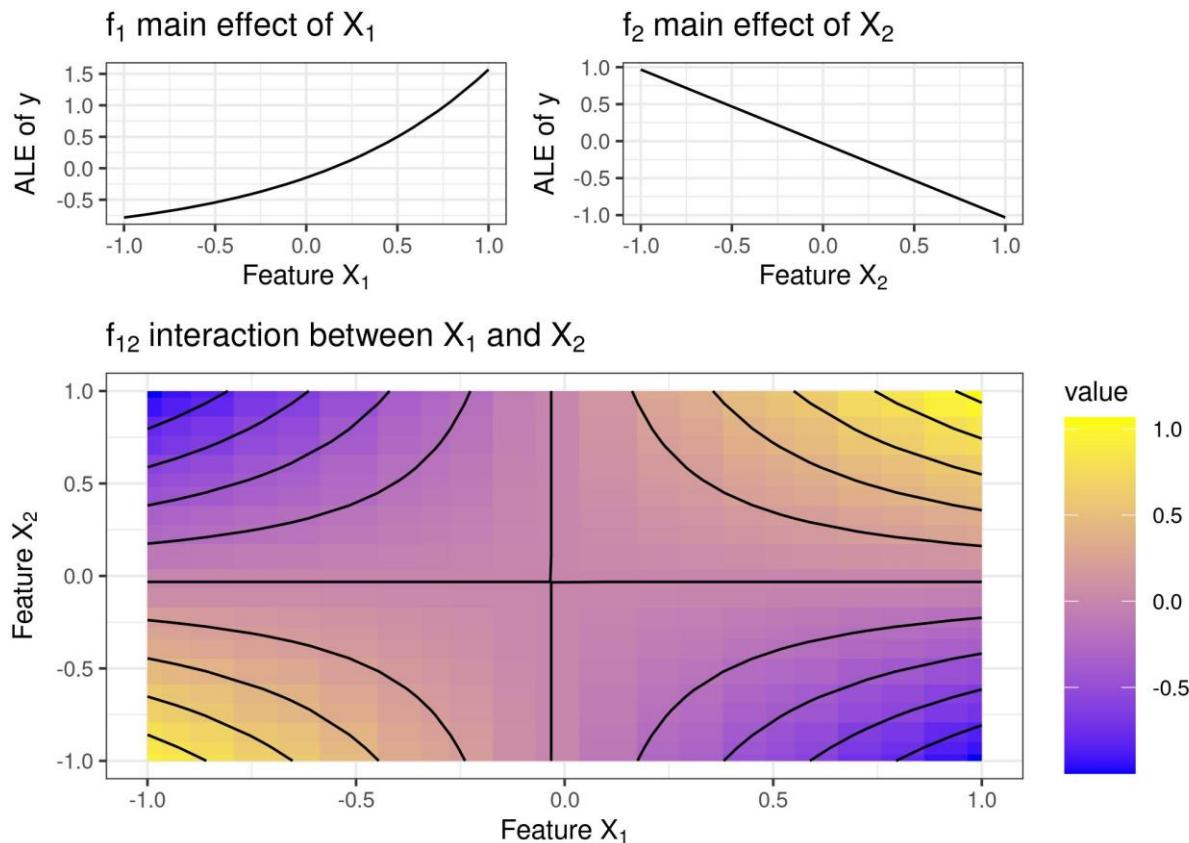
وقتی تابع مقادیر زیادی می گیرد X_1 بزرگ است و X_2 کوچک است و مقادیر کوچکی را برای بزرگ می گیرد
 ۳۷۱۰ کوچک . تابع پیش بینی صرفاً یک اثر افزایشی بین دو ویژگی نیست، بلکه یک تعامل بین این دو
 ۳۷۱۱ است. وجود یک تعامل را می توان در شکل مشاهده کرد - اثر تغییر مقادیر برای ویژگی X_1 بستگی به مقدار آن
 ۳۷۱۲ ویژگی دارد X_2 دارد.
 ۳۷۱۳

کار ما اکنون این است که این تابع را به اثرات اصلی ویژگی ها تجزیه کنیم X_1 و یک اصطلاح تعامل. برای
 ۳۷۱۴ عملکرد دو بعدی f که فقط به دو ویژگی ورودی بستگی دارد $X_1 X_2$ ، ما می خواهیم هر جزء یک اثر اصلی را
 ۳۷۱۵ نشان دهد f و (f) اثر متقابل (f) یا رهگیری) ۲
 ۳۷۱۶

$$\hat{f}(x_1, x_2) = \hat{f}_0 + \hat{f}_1(x_1) + \hat{f}_2(x_2) + \hat{f}_{1,2}(x_1, x_2)$$

اثرات اصلی نشان می دهد که چگونه هر ویژگی بر پیش بینی تأثیر می گذارد، مستقل از مقادیر ویژگی دیگر. اثر
 ۳۷۱۸ متقابل اثر مشترک ویژگی ها را نشان می دهد. رهگیری به سادگی به ما می گوید که وقتی همه اثرات ویژگی
 ۳۷۱۹ روی صفر تنظیم می شوند، پیش بینی چیست. توجه داشته باشید که اجزاء خود توابعی هستند (به جز
 ۳۷۲۰ رهگیری) با ابعاد ورودی متفاوت.
 ۳۷۲۱

من فقط کامپوننت ها را اکنون به شما می دهم و بعداً توضیح می دهم که از کجا آمده اند. رهگیری است
 ۳۷۲۲ ۳,۱۸f~ از آنجایی که سایر مؤلفه ها توابع هستند، می توانیم آنها را تجسم کنیم:
 ۳۷۲۳



شکل ۸.۲۳: تجزیه یک تابع.

آیا با توجه به فرمول واقعی بالا، با نادیده گرفتن این که مقدار رهگیری کمی تصادفی به نظر می‌رسد، آیا فکر می‌کنید اجزاء منطقی هستند؟ را ایکس ۱ ویژگی یک اثر اصلی نمایی را نشان می‌دهد و ایکس ۲ اثر خطی منفی را نشان می‌دهد. اصطلاح تعامل کمی شبیه تراشه Pringles است. در اصطلاحات کمتر ترد و ریاضی، یک سهمی هذلولی است، همانطور که ما انتظار داریم. ایکس ۱ ایکس ۲ . هشدار اسپویلر: تجزیه بر اساس نمودارهای اثر محلی انباشته شده است که بعداً در فصل به آن خواهیم پرداخت.

۸.۴.۱ چگونه کامپوننت‌ها را محاسبه نکنیم؟

اما چرا این همه هیجان؟ یک نگاه به فرمول قبل پاسخ تجزیه را به ما می‌دهد، بنابراین نیازی به روش‌های فانتزی نیست، درست است؟ برای ویژگی ایکس ۱ ، می‌توانیم تمام جمع‌هایی را که فقط شامل می‌شوند، بگیریم ایکس ۱ به عنوان جزء برای آن ویژگی. که خواهد بود $f_1 = \alpha_1 + \beta_1 X_1$ ایکس ۲ به عنوان جزء برای آن ویژگی. که خواهد بود $f_2 = \alpha_2 + \beta_2 X_2$ ایکس ۱ و ایکس ۲ . در حالی که این پاسخ صحیح برای این مثال است (تا ثابت‌ها)، دو مشکل در این رویکرد وجود دارد: مشکل ۱): در حالی که مثال با فرمول

شروع شد، واقعیت این است که تقریباً هیچ مدل یادگیری ماشینی را نمی توان با چنین چیزی توصیف کرد.
 فرمول تمیز مسئله ۲) بسیار پیچیده تر است و به چیستی یک تعامل مربوط می شود. یکتابع ساده را تصور
 کنید^{۸f}=ایکس۱،ایکس۲، که در آن هر دو ویژگی مقادیر بزرگتر از صفر می گیرند و مستقل
 از یکدیگر هستند. با استفاده از تاکتیک نگاه کردن به فرمول، به این نتیجه می رسیم که یک تعامل بین ویژگی
 ها وجود دارد ایکس۱ و ایکس۲ ، اما نه اثرات ویژگی های فردی. اما آیا واقعاً می توانیم این ویژگی را بگوییم
 ایکس۱ هیچ تأثیر فردی بر عملکرد پیش بینی ندارد؟ صرف نظر از اینکه ویژگی دیگر چه ارزشی دارد ایکس۲
 طول می کشد، پیش بینی با افزایش ما افزایش می یابد ایکس۱ . به عنوان مثال، برای ایکس۲=۱ ، اثر ایکس۱
 است^{۸f}=ایکس۱ ، و وقتی که ایکس۲=۱۰ است^{۸f}=ایکس۱۰=۱۰. بنابراین،
 مشخص است که ویژگی ایکس۱ تأثیر مثبتی بر پیش بینی دارد، مستقل از ایکس۲ ، و صفر نیست.

برای حل مشکل ۱) عدم دسترسی به یک فرمول منظم، به روشی نیاز داریم که فقط از تابع پیش بینی یا امتیاز
 طبقه بندی استفاده کند. برای حل مشکل ۲) عدم تعریف، به برخی بدیهیات نیاز داریم که به ما بگوید اجزاء
 چگونه باید باشند و چگونه با یکدیگر ارتباط دارند. اما ابتدا باید به طور دقیق تر تعریف کنیم که تجزیه
 عملکردی چیست.

۸.۴.۲ تجزیه عملکردی
 یک تابع پیش بینی طول می کشد پ ویژگی ها به عنوان ورودی، $\rightarrow \text{آرپ}^{\text{۸f}}$ آر و خروجی تولید می کند. این
 می تواند یک تابع رگرسیون باشد، اما همچنین می تواند احتمال طبقه بندی برای یک کلاس مشخص یا امتیاز
 برای یک خوشه معین (یادگیری ماشین بدون نظارت) باشد. به طور کامل تجزیه می شود، می توانیم تابع
 پیش بینی را به صورت مجموع مؤلفه های عملکردی نشان دهیم:

$$\begin{aligned}
 \hat{f}(x) = & \hat{f}_0 + \hat{f}_1(x_1) + \dots + \hat{f}_p(x_p) \\
 & + \hat{f}_{1,2}(x_1, x_2) + \dots + \hat{f}_{1,p}(x_1, x_p) + \dots + \hat{f}_{p-1,p}(x_{p-1}, x_p) \\
 & + \dots \\
 & + \hat{f}_{1,\dots,p}(x_1, \dots, x_p)
 \end{aligned}$$

می توانیم فرمول تجزیه را با فهرست کردن همه زیرمجموعه های ممکن ترکیب ویژگی ها کمی زیباتر کنیم :
 اس \subseteq ۱،۰۰،پ . {این مجموعه شامل رهگیری) اس $\subseteq\emptyset$ ، جلوه های اصلی | اس $=1$ و تمام تعاملات)
 اس ≥ 1 . با تعریف این زیر مجموعه، می توانیم تجزیه را به صورت زیر بنویسیم:

$$\hat{f}(x) = \sum_{S \subseteq \{1, \dots, p\}} \hat{f}_S(x_S)$$

در فرمول، ایکس اس بردار ویژگی های مجموعه شاخص است اس . و هر زیر مجموعه اس یک مؤلفه عملکردی را نشان می دهد، به عنوان مثال یک افکت اصلی اگر S فقط یک ویژگی داشته باشد یا یک تعامل اگر | اس > ۱ در فرمول بالا چند جزء وجود دارد؟ پاسخ به چند زیر مجموعه ممکن خلاصه می شود اس از ویژگی ها ۱،۰۰..؛پ می توانیم تشکیل دهیم. و اینها هستند $\sum p_{\text{من}}=0$ ؛پ زیر مجموعه های ممکن! به عنوان مثال، اگر یک تابع از ۱۰ ویژگی استفاده کند، می توانیم تابع را به ۱۰۴۲ جزء تجزیه کنیم: ۱ قطع، ۱۰ اثر اصلی، ۹۰ عبارت تعامل دو طرفه، ۷۲۰ عبارت تعامل سه طرفه، ... و با هر ویژگی اضافی، تعداد اجزا دو برابر می شود واضح است که برای اکثر توابع، محاسبه همه مؤلفه ها امکان پذیر نیست. دلیل دیگر برای محاسبه نکردن همه اجزا این است که اجزا با | اس > ۲ تجسم و تفسیر آنها دشوار است.

۸.۴.۳ چگونه کامپوننت ها را محاسبه نکنیم ||

تا کنون از صحبت در مورد چگونگی تعریف و محاسبه مؤلفه ها اجتناب کرده ام. تنها محدودیت هایی که به طور ضمنی در مورد آن صحبت کردیم، تعداد و ابعاد اجزا بود، و اینکه مجموع مؤلفه ها باید تابع اصلی را ایجاد کند. اما بدون محدودیت بیشتر در مورد اینکه چه اجزایی باید باشند، آنها منحصر به فرد نیستند. این بدان معناست که می توانیم جلوه ها را بین جلوه های اصلی و تعاملات، یا تعاملات مرتبه پایین (چند ویژگی) و تعاملات مرتبه بالاتر (ویژگی های بیشتر) تغییر دهیم. در مثال ابتدای فصل می توانیم هر دو افکت اصلی را صفر کنیم و جلوه های آنها را به اثر تعامل اضافه کنیم.

در اینجا یک مثال شدیدتر وجود دارد که نیاز به محدودیت در اجزا را نشان می دهد. فرض کنید یک تابع سه بعدی دارید. واقعاً مهم نیست که این تابع چگونه به نظر می رسد، اما تجزیه زیر همیشه کار می کند $f_1 = 12$. است $f_2 = 1.02x_1 + 1.02x_2 + 1.02x_3$. ایکس ۱ تعداد کفشه که دارید $f_3 = 1.38f_1 + 2.38f_2 + 3.8f_3$. همه صفر هستند و برای اینکه این ترفند کار کند، تعریف می کنم $f_4 = f_1 + f_2 + f_3$. ایکس ۲، ایکس ۳، ایکس ۴ = (ایکس ۱، ایکس ۲، ایکس ۳) ایکس ۴ = ایکس ۴. اثراً باقی مانده را می کشد، که طبق تعریف همیشه کار می کند، به این معنا که مجموع همه اجزاء تابع پیش بینی اصلی را به ما می دهد. اگر بخواهید این را به عنوان تفسیر مدل خود ارائه دهید، این تجزیه خیلی معنی دار و کاملاً گمراه کننده نخواهد بود.

ابهام را می توان با تعیین محدودیت های بیشتر یا روش های خاص برای محاسبه مؤلفه ها اجتناب کرد. در این فصل، سه روش را مورد بحث قرار خواهیم داد که به روش های مختلف به تجزیه عملکردی نزدیک می شوند:

ANOVA عملکردی (تعمیم شده).

جلوه های محلی اباحت شده

۳۷۸۷

۳۷۸۸

مدل های رگرسیون آماری

۳۷۸۹

ANOVA عملکردی ۱,۴,۴

۳۷۹۰

تابعی توسط هوکر (۲۰۰۴) ۳۹ پیشنهاد شد . لازمه این رویکرد این است که تابع پیش بینی مدل

۳۷۹۱

باشد f^* مربع قابل ادغام است. مانند هر تجزیه عملکردی، ANOVA عملکردی تابع را به اجزای زیر تجزیه می

۳۷۹۲

کند:

۳۷۹۳

$$\hat{f}(x) = \sum_{S \subseteq \{1, \dots, p\}} \hat{f}_S(x_S)$$

۳۷۹۴

هوکر (۲۰۰۴) هر جزء را با فرمول زیر تعریف می کند:

۳۷۹۵

$$\hat{f}_S(x) = \int_{X_{-S}} \left(\hat{f}(x) - \sum_{V \subset S} \hat{f}_V(x) \right) dX_{-S}$$

۳۷۹۶

خوب، اجازه دهید این چیز را از هم جدا کنیم. ما می توانیم جزء را به صورت زیر بازنویسی کنیم:

۳۷۹۷

$$\hat{f}_S(x) = \int_{X_{-S}} (\hat{f}(x)) dX_{-S} - \int_{X_{-S}} \left(\sum_{V \subset S} \hat{f}_V(x) \right) dX_{-S}$$

۳۷۹۸

در سمت چپ انتگرال بر روی تابع پیش بینی با توجه به ویژگی های حذف شده از مجموعه است اس ، نشان

۳۷۹۹

اده شده با- اس . به عنوان مثال، اگر مؤلفه تعامل دو طرفه را برای ویژگی های ۲ و ۳ محاسبه کنیم، روی

۳۸۰۰

ویژگی های ۱ ، ۴ ، ۵ ، ... انتگرال را می توان به عنوان مقدار مورد انتظار تابع پیش بینی با توجه به ایکس-اس ، با

۳۸۰۱

فرض اینکه همه ویژگی ها از یک توزیع یکنواخت از حداقل تا حداکثر خود پیروی می کنند. از این بازه، همه

۳۸۰۲

اجزا را با زیرمجموعه های از کم می کنیم اس . این تفرقی اثر تمام جلوه های مرتبه پایین را حذف می کند و

۳۸۰۳

اثر را در مرکز قرار می دهد. برای اس $\{1, 2\}$ ، اثرات اصلی هر دو ویژگی را کم می کنیم $f^*_{1,2}$ و $f^*_{2,1}$ و

۳۸۰۴

همچنین رهگیری f^* . وقوع این اثرات مرتبه پایین تر، فرمول را بازگشتی می کند: ما باید از سلسله مراتب زیر

۳۸۰۵

مجموعه ها عبور کرده و همه این اجرا را محاسبه کنیم. برای جزء رهگیری f^* ، زیر مجموعه مجموعه خالی

۳۸۰۶

است اس $\{\emptyset\}$ و بنابراین- اس شامل تمام ویژگی ها:

۳۸۰۷

$$\hat{f}_0(x) = \int_X \hat{f}(x) dX$$

۳۸۰۸ این به سادگی تابع پیش بینی است که روی همه ویژگی ها یکپارچه شده است، زمانی که فرض کنیم همه
 ۳۸۰۹ ویژگی ها به طور یکنواخت توزیع شده اند، رهگیری را می توان به عنوان انتظار تابع پیش بینی نیز تفسیر کرد.
 ۳۸۱۰ حالا که می دانیم f^0 ، می توانیم محاسبه کنیم f^1 و به طور معادل f^2 :

$$f_1(x) = \int_{X_{-1}} \left(f(x) - f_0 \right) dX_{-S}$$

۳۸۱۱

برای پایان دادن به محاسبه برای جزء f^1 ، ما می توانیم همه چیز را کنار هم بگذاریم:

۳۸۱۲

$$\hat{f}_{1,2}(x) = \int_{X_{3,4}} \left(\hat{f}(x) - (\hat{f}_0(x) + \hat{f}_1(x) - \hat{f}_0 + \hat{f}_2(x) - \hat{f}_0) \right) dX_3, X_4$$

۳۸۱۳

۳۸۱۴ این مثال نشان می دهد که چگونه هر افکت مرتبه بالاتر با ادغام کردن همه ویژگی های دیگر تعریف می شود، اما
 ۳۸۱۵ همچنان با حذف تمام جلوه های مرتبه پایین تر که زیرمجموعه ای از مجموعه ویژگی های مورد علاقه ما هستند.

۳۸۱۶ هوکر (۲۰۰۴) نشان داده است که این تعریف از اجزای عملکردی این بدیهیات مطلوب را برآورده می کند:

۳۸۱۷ صفر یعنی $\nabla^1 f$: اس)ایکس اس (دایکسیس = ۰ برای هر اس . $\neq \emptyset$

۳۸۱۸ متعامد بودن $\nabla^1 f$: اس)ایکس اس (ایکس $\nabla^1 f$: اس)ایکس اس (دایکس = ۰ برای اس $\neq \emptyset$

۳۸۱۹ تجزیه واریانس: اجازه دهید $\sigma_1^2 = \nabla^1 f(\text{ایکس}^1)$ ، $\sigma_2^2 = \nabla^1 f(\text{ایکس}^2)$ ، سپس $\sigma_0^2 = \sum_{i=1}^n \text{اس}^i$
 ۳۸۲۰) اس $\nabla^1 f$ (اس)

۳۸۲۱ صفر به این معنی است که همه اثرات یا تعاملات حول محور صفر هستند. در نتیجه، تفسیر در موقعیت X نسبت
 ۳۸۲۲ به پیش بینی مرکز است و نه پیش بینی مطلق.

۳۸۲۳ اصل متعامد نشان می دهد که اجزاء اطلاعات را به اشتراک نمی گذارند. به عنوان مثال، اثر مرتبه اول ویژگی
 ۳۸۲۴ ایکس ۱ و مدت تعامل از ایکس ۱ و ایکس ۲ همبستگی ندارند. به دلیل متعامد بودن، همه اجزا به این معنا که
 ۳۸۲۵ اثرات را با هم ترکیب نمی کنند، "خالص" هستند. بسیار منطقی است که کامپوننت برای مثلاً ویژگی باشد
 ۳۸۲۶ ایکس ۴ باید مستقل از اصطلاح تعامل بین ویژگی ها باشد ایکس ۱ و ایکس ۲ . پیامد جالب تر برای متعامد بودن
 ۳۸۲۷ مؤلفه های سلسله مراتبی به وجود می آید، جایی که یک مؤلفه حاوی ویژگی های دیگری است، برای مثال تعامل
 ۳۸۲۸ بین ایکس ۱ و ایکس ۲ ، و اثر اصلی ویژگی ایکس ۱ . در مقابل، یک نمودار وابستگی جزئی دو بعدی برای ایکس ۱
 ۳۸۲۹ و ایکس ۲ شامل چهار اثر است: رهگیری، دو اثر اصلی ایکس ۱ و ایکس ۲ و تعامل بین آنها جزء عملکردی
 ۳۸۳۰ برای ANOVA ایکس ۱، ایکس ۲ (فقط شامل تعامل خالص است).

۳۸۳۱ تجزیه واریانس به ما امکان می دهد واریانس تابع را تقسیم کنیم f در میان مولفه ها، و تضمین می کند که
۳۸۳۲ واریانس کل تابع را در پایان جمع می کند. خاصیت تجزیه واریانس همچنین می تواند دلیل فراخوانی روش را
۳۸۳۳ برای ما توضیح دهد

۳۸۳۴ $ANOVA$ عملکردی. در آمار، $ANOVA$ مخفف ANAlysis Of VAriance است. به مجموعه ای از
۳۸۳۵ روش ها اطلاق می شود که تفاوت ها را در میانگین یک متغیر هدف تحلیل می کنند $ANOVA$. با تقسیم واریانس
۳۸۳۶ و نسبت دادن آن به متغیرها کار می کند. بنابراین، $ANOVA$ عملکردی را می توان به عنوان بسط این مفهوم
۳۸۳۷ برای هر تابعی دید.

۳۸۳۸ مشکلات با $ANOVA$ عملکردی زمانی که ویژگی ها همبسته هستند به وجود می آیند. به عنوان راه حل،
۳۸۳۹ $ANOVA$ عملکردی تعمیم یافته پیشنهاد شده است.

۳۸۴۰ $ANOVA$ ۸,۴,۵ عملکردی تعمیم یافته برای ویژگی های وابسته
۳۸۴۱ مشابه بیشتر تکنیک های تفسیری مبتنی بر داده های نمونه گیری (مانند PDP، ANOVA عملکردی می تواند
۳۸۴۲ نتایج گمراه کننده ای را در صورت همبستگی ویژگی ها ایجاد کند. اگر روی توزیع یکنواخت ادغام کنیم، زمانی که
۳۸۴۳ در واقعیت ویژگی ها وابسته هستند، یک مجموعه داده جدید ایجاد می کنیم که از توزیع مشترک منحرف
۳۸۴۴ می شود و به ترکیبات غیر محتمل مقادیر ویژگی تعمیم می یابد.

۳۸۴۵ هوکر (۲۰۰۷) $ANOVA$ ۴۰ عملکردی تعمیم یافته را پیشنهاد کرد، تجزیه ای که برای ویژگی های وابسته کار
۳۸۴۶ می کند. این یک تعمیم $ANOVA$ عملکردی است که قبلاً با آن مواجه شدیم، به این معنی که $ANOVA$
۳۸۴۷ عملکردی یک مورد خاص از $ANOVA$ عملکردی تعمیم یافته است. مولفه ها به عنوان پیش بینی f بر روی
۳۸۴۸ فضای توابع افزایشی تعریف می شوند:

$$\hat{f}_S(x_S) = \underset{g_S \in L^2(\mathbb{R}^S)_{S \subseteq P}}{\operatorname{argmin}} \int \left(\hat{f}(x) - \sum_{S \subseteq P} g_S(x_S) \right)^2 w(x) dx.$$

۳۸۵۰ به جای متعامد، مولفه ها یک شرط قائم مقامی سلسله مراتبی را برآورده می کنند:

$$\forall \hat{f}_S(x_S) | S \subset U : \int \hat{f}_S(x_S) \hat{f}_U(x_U) w(x) dx = 0$$

۳۸۵۲ تعامد سلسله مراتبی با متعامد متفاوت است. برای دو مجموعه ویژگی S و U که هیچ کدام زیر مجموعه دیگری
۳۸۵۳ نیستند (مثلاً اس $= \{1, 2\}$ و $U = \{2, 3\}$)، اجزاء f_S اس و f_U لازم نیست متعامد باشد تا تجزیه به صورت سلسله
۳۸۵۴ مراتبی متعامد باشد. اما همه اجزا برای همه زیر مجموعه های اس باید متعامد به f_S باشند. در نتیجه، تفسیر به
۳۸۵۵ روش های مرتبط متفاوت است: مشابه ALE، مؤلفه های $ANOVA$ عملکردی تعمیم یافته

۳۸۵۶ می توانند اثرات (حاشیه‌ای) ویژگی‌های همبسته را در هم بینندن. اینکه آیا اجزاء با اثرات حاشیه‌ای درگیر می
۳۸۵۷ شوند یا خیر، به انتخاب تابع وزن نیز بستگی دارد) W ایکس . (اگر W را به عنوان اندازه یکنواخت در مکعب
۳۸۵۸ واحد انتخاب کنیم، عملکردی را از بخش بالا بدست می آوریم. یک انتخاب طبیعی برای W تابع
۳۸۵۹ توزیع احتمال مشترک است. با این حال، توزیع مشترک معمولاً ناشناخته است و تخمین آن دشوار است. یک
۳۸۶۰ ترفند می تواند این باشد که با اندازه گیری یکنواخت روی مکعب واحد شروع کنید و مناطقی را بدون داده
۳۸۶۱ بردارید.

۳۸۶۲ برآورد بر روی شبکه ای از نقاط در فضای ویژگی انجام می شود و به عنوان یک مسئله کمینه سازی بیان می
۳۸۶۳ شود که با استفاده از تکنیک های رگرسیون قابل حل است. با این حال، مولفه ها را نمی توان به صورت جداگانه
۳۸۶۴ یا به صورت سلسله مراتبی محاسبه کرد، اما یک سیستم پیچیده از معادلات شامل اجزای دیگر باید حل شود.
۳۸۶۵ بنابراین محاسبات بسیار پیچیده و محاسباتی فشرده است.

۴,۶ نمودارهای اثر محلی انباشته شده

۳۸۶۶ نمودارهای ALE (Apley و Zhu 2020 41) همچنین یک تجزیه عملکردی ارائه می دهند، به این معنی که
۳۸۶۷ با افروzen تمام نمودارهای ALE از قطع، نمودارهای 1 ALE بعدی، نمودارهای 2 ALE بعدی و غیره، تابع پیش
۳۸۶۸ بینی به دست می آید ALE با ANOVA عملکردی (تعمیم یافته) متفاوت است، زیرا اجزای آن متعامد نیستند،
۳۸۶۹ اما، همانطور که نویسندهای آن را نامیده اند، شبه متعامد هستند. برای درک شبه متعامد، باید عملگر را تعریف
۳۸۷۰ کنیم اج اس ، که یک تابع می گیرد f^8 و آن را به نمودار ALE خود برای زیر مجموعه ویژگی نگاشت می کند
۳۸۷۱

۳۸۷۲ اس . مثلاً اپراتور اج ۱,۲ یک مدل یادگیری ماشینی را به عنوان ورودی می گیرد و نمودار ALE دو بعدی را برای
۳۸۷۳ ویژگی های ۱ و ۲ تولید می کند : $\text{اج} = f_1, 2, \text{ALE}^{(8f)}$. اگر یک عملگر را دو بار اعمال کنیم، همان نمودار
۳۸۷۴ ALE را بدست می آوریم. پس از اعمال اپراتور اج ۱,۲ به f یک بار، نمودار 2D را دریافت می کنیم
۳۸۷۵ $\text{ALE}^{(8f)} = f_1, 2, \text{ALE}^{(8f)}$. سپس دوباره عملگر را اعمال می کنیم نه به f اما به $f_1, 2, \text{ALE}^{(8f)}$. این امکان پذیر است زیرا جزء D ۲
۳۸۷۶ ALE خود یک تابع است. نتیجه دوباره است $\text{ALE}^{(8f)} = f_1, 2, \text{ALE}^{(8f)}$ ، یعنی می توانیم یک عملگر را چندین بار اعمال کنیم و
۳۸۷۷ همیشه همان نمودار ALE را دریافت کنیم. این قسمت اول شبه متعامد است. اما اگر دو عملگر متفاوت را برای
۳۸۷۸ مجموعه ویژگی های مختلف اعمال کنیم، نتیجه چیست؟ مثلاً، اج ۱,۲ و اج ۱، یا اج ۱, ۲ و اج ۳, ۴, ۵ ؟ جواب صفر
۳۸۷۹ است. اگر ابتدا عملگر ALE را اعمال کنیم اج اس به یک تابع و سپس اعمال می شود اج L به نتیجه (با اس $L \neq$ ،
۳۸۸۰ سپس نتیجه صفر می شود. به عبارت دیگر، نمودار ALE یک نمودار ALE صفر است، مگر اینکه همان نمودار
۳۸۸۱ ALE را دو بار اعمال کنید. یا به عبارت دیگر، نمودار ALE برای مجموعه ویژگی S شامل هیچ نمودار

دیگری در آن نیست. یا در اصطلاح ریاضی، عملگر ALE توابع را به زیرفضاهای متعامد یک فضای مخصوص
داخلی نگاشت می کند.

Apley و Zhu (2020) اشاره می کنند، شبه متعامد ممکن است مطلوب تر از تعامد سلسله مراتبی باشد، زیرا اثرات حاشیه ای ویژگی ها را در هم نمی بندد. علاوه بر این، ALE نیازی به تخمین توزیع مشترک ندارد. مولفه ها را می توان به صورت سلسله مراتبی تخمین زد، به این معنی که محاسبه 2 ALE بعدی برای ویژگی های ۱ و ۲ فقط به محاسبات اجزای ALE منفرد ۱ و ۲ و اصطلاح رهگیری به علاوه نیاز دارد.

۸,۴,۷ مدل های رگرسیون آماری

این رویکرد با مدل های قابل تفسیر، به ویژه مدل های افزایشی تعمیم یافته پیوند دارد. به جای تجزیه یکتابع پیچیده، می توانیم محدودیت هایی را در فرآیند مدل سازی ایجاد کنیم تا بتوانیم به راحتی اجزای جداگانه را بخوانیم. در حالی که تجزیه را می توان به رو شی از بالا به پایین انجام داد، جایی که ما با یک تابع با ابعاد بالا شروع می کنیم و آن را تجزیه می کنیم، مدل های افزایشی تعمیم یافته یک رویکرد از پایین به بالا ارائه می دهد، جایی که ما مدل را از اجزای ساده می سازیم. هر دو رویکرد مشترک هستند که هدف آنها ارائه مؤلفه های فردی و قابل تفسیر است. در مدل های آماری، تعداد مؤلفه ها را محدود می کنیم تا نه همه ۲ پ اجزا باید نصب شوند ساده ترین نسخه رگرسیون خطی است:

$$\hat{f}(x) = \beta_0 + \beta_1 x_1 + \dots \beta_p x_p$$

فرمول بسیار شبیه به تجزیه عملکردی است، اما با دو تغییر عمده. اصلاح ۱: همه اثرات متقابل حذف می شوند و ما فقط جلوه های رهگیری و اصلی را نگه می داریم. اصلاح ۲: اثرات اصلی ممکن است فقط در ویژگی ها خطی باشند $\hat{f}(\beta, \alpha)$. زیا مشاهده مدل رگرسیون خطی از طریق لنز تجزیه عملکردی، می بینیم که خود مدل تجزیه عملکردی تابع واقعی را نشان می دهد که از ویژگی ها به هدف نگاشت می شود، اما تحت فرضیات قوی که اثرات اثرات خطی هستند و هیچ تعاملی وجود ندارد.

مدل افزودنی تعمیم یافته با اجازه دادن به عملکردهای انعطاف پذیرتر، فرض دوم را تسهیل می کند $\hat{f}(\beta, \alpha)$ از طریق استفاده از اسپلاین ها فعل و افعالات نیز می توانند اضافه شوند، اما این فرآیند نسبتاً دستی است. رویکردهایی مانند GAM تلاش می کنند تا تعاملات دو طرفه را به طور خودکار به یک GAM اضافه کنند.

تصور یک مدل رگرسیون خطی یا یک GAM به عنوان تجزیه عملکردی نیز می تواند منجر به سردرگمی شود. اگر رویکردهای تجزیه را در فصل قبل اعمال کنید ANOVA (عملکردی تعمیم یافته و اثرات محلی انباسته)، ممکن است مؤلفه هایی را دریافت کنید که با مؤلفه هایی که مستقیماً از GAM خوانده می شوند متفاوت

۳۹۰۸ باشند. این می تواند زمانی انفاق بیفتند که اثرات متقابل ویژگی های همبسته در GAM مدل شود. این اختلاف
۳۹۰۹ به این دلیل رخ می دهد که سایر رویکردهای تجزیه عملکردی تأثیرات را به طور متفاوتی بین برهمکنش ها و
۳۹۱۰ اثرات اصلی تقسیم می کنند.

۳۹۱۱ پس چه زمانی باید از GAM به جای مدل پیچیده + تجزیه استفاده کرد؟ زمانی که بیشتر تعاملات صفر است،
۳۹۱۲ باید به GAM ها پایبند باشید، به خصوص زمانی که هیچ تعاملی با سه یا چند ویژگی وجود ندارد. اگر بدانیم
۳۹۱۳ که حداقل تعداد ویژگی های درگیر در تعاملات دو است |) اس کا ۲ | سپس می توانیم از رویکردهایی مانند
۳۹۱۴ MARS یا GA2M استفاده کنیم. در نهایت، عملکرد مدل در داده های آزمایشی ممکن است نشان دهد که آیا
۳۹۱۵ یک GAM کافی است یا یک مدل پیچیده تر بسیار بهتر عمل می کند.

۳۹۱۶ ۸,۴,۸ پاداش: طرح وابستگی جزئی

آیا طرح وابستگی جزئی نیز تجزیه عملکردی را ارائه می دهد؟ پاسخ کوتاه: خیر. پاسخ طولانی تر: نمودار
۳۹۱۷ وابستگی جزئی برای مجموعه ویژگی اس همیشه شامل تمام اثرات سلسله مراتب PDP - برای {۱,۲} {نه تنها
۳۹۱۸ تعامل، بلکه اثرات ویژگی های فردی را نیز شامل می شود. در نتیجه، افزودن تمام PDP ها برای همه زیر
۳۹۱۹ مجموعه ها، تابع اصلی را به دست نمی آورد، و بنابراین تجزیه معتبری نیست. اما آیا می توانیم PDP را با حذف
۳۹۲۰ همه اثرات پایین تر تنظیم کنیم؟ بله، می توانیم، اما چیزی شبیه به ANOVA عملکردی دریافت می کنیم. با این
۳۹۲۱ حال، به جای ادغام بر روی یک توزیع یکنواخت، PDP روی توزیع حاشیه ای ادغام می شود ایکس-اس که با
۳۹۲۲ استفاده از نمونه گیری مونت کارلو برآورد شده است.
۳۹۲۳

۳۹۲۴ ۸,۴,۹ مزايا

۳۹۲۵ من تجزیه عملکردی را مفهوم اصلی تفسیرپذیری یادگیری ماشین می دانم.
۳۹۲۶

۳۹۲۷ تجزیه عملکردی یک توجیه نظری برای تجزیه مدل های یادگیری ماشینی با بعد بالا و پیچیده به اثرات و
۳۹۲۸ تعاملات فردی به ما می دهد - مرحله ای ضروری که به ما امکان می دهد تا اثرات فردی را تفسیر کنیم. تجزیه
۳۹۲۹ تابعی ایده اصلی تکنیک هایی مانند مدل های رگرسیون آماری، ALE، ANOVA، PDP (تممیم یافته)،
۳۹۳۰ آماره H و منحنی های ICE است.

۳۹۳۱ تجزیه عملکردی همچنین درک بهتری از روش های دیگر فراهم می کند . به عنوان مثال، اهمیت ویژگی
۳۹۳۲ جایگشت ارتباط بین یک ویژگی و هدف را می شکند. از طریق لنز تجزیه عملکردی مشاهده می کنیم، می توانیم
۳۹۳۳ ببینیم که جایگشت اثر تمام اجزایی را که ویژگی در آن دخیل است، «از بین می برد». این بر اثر اصلی ویژگی،

بلکه بر تمام تعاملات با سایر ویژگی ها تأثیر می گذارد. به عنوان مثال دیگری، مقادیر Shapley یک پیش‌بینی را به اثرات افزایشی ویژگی فردی تجزیه می‌کند. اما تجزیه عملکردی به ما می‌گوید که باید اثرات متقابلی نیز در تجزیه وجود داشته باشد، پس آنها کجا هستند؟ ارزش‌های Shapley نسبت دادن منصفانه اثرات به ویژگی‌های فردی را ارائه می‌دهند، به این معنی که همه تعاملات نیز به طور منصفانه به ویژگی‌ها نسبت داده می‌شوند و بنابراین بین مقادیر Shapley تقسیم می‌شوند.

هنگام در نظر گرفتن تجزیه عملکردی به عنوان یک ابزار، استفاده از نمودارهای ALE مزایای بسیاری را ارائه می‌دهد. نمودارهای ALE تجزیه عملکردی را ارائه می‌دهند که محاسبه آن سریع است، دارای پیاده سازی نرم افزاری است (به فصل ALE مراجعه کنید)، و ویژگی‌های شبه متعامد مطلوب.

۳۹۴۲ ۸,۴,۱۰ معايib

مفهوم تجزیه عملکردی به سرعت به مرزهای خود برای اجزای با ابعاد بالا فراتر از تعامل بین دو ویژگی می‌رسد. این انفجار نمایی در تعداد ویژگی‌ها نه تنها عملی بودن را محدود می‌کند، زیرا نمی‌توانیم به راحتی تعاملات مرتبه بالاتر را تجسم کنیم، بلکه اگر بخواهیم همه تعاملات را محاسبه کنیم، زمان محاسباتی دیوانه‌کننده است.

هر روش روش تجزیه عملکردی دارای معايib فردی خود است . رویکرد پايين به بالا - ساخت مدل‌های رگرسيون - يك فرآيند کاملاً دستی است و محدودیت‌های زيادي را بر مدل تحميل می‌کند که می‌تواند بر عملکرد پیش‌بینی تأثير بگذارد ANOVA. عملکردی به ویژگی‌های مستقل نياز دارد. تخمين ANOVA عملکردی تعديم يافته بسيار دشوار است. نمودارهای اثر محلی انباسته تجزیه واريанс را ارائه نمی‌دهند.

رویکرد تجزیه عملکردی برای تجزیه و تحلیل داده‌های جدولی مناسب تر از متن یا تصاویر است.

۸,۵ اهمیت ویژگی جایگشت

۳۹۵۲ اهمیت ویژگی جایگشت، افزایش خطای پیش‌بینی مدل را پس از تغییر مقادیر ویژگی اندازه‌گیری می‌کند، که
۳۹۵۳ رابطه بین ویژگی و نتیجه واقعی را قطع می‌کند.
۳۹۵۴

۸,۵,۱ نظریه

۳۹۵۵ مفهوم واقعاً ساده است: ما اهمیت یک ویژگی را با محاسبه افزایش خطای پیش‌بینی مدل پس از تغییر ویژگی
۳۹۵۶ اندازه می‌گیریم. یک ویژگی در صورتی "مهم" است که به هم زدن مقادیر آن خطای مدل را افزایش دهد، زیرا
۳۹۵۷ در این مورد مدل برای پیش‌بینی به ویژگی متکی است. یک ویژگی "بی اهمیت" است اگر به هم زدن مقادیر
۳۹۵۸ آن خطای مدل را بدون تغییر باقی بگذارد، زیرا در این مورد مدل ویژگی را برای پیش‌بینی نادیده می‌گیرد.
۳۹۵۹ اندازه گیری اهمیت ویژگی جایگشت توسط Breiman (2001) ۴۳ برای جنگل‌های تصادفی معرفی شد. بر
۳۹۶۰ اساس این ایده، فیشر، رودین و دومینیسی (۲۰۱۸) ۴۴ یک نسخه مدل-آگنوتیک از اهمیت ویژگی پیشنهاد
۳۹۶۱ کرد و آن را اتکای مدل نامید. آنها همچنین ایده‌های پیشرفته تری را در مورد اهمیت ویژگی معرفی کردند، به
۳۹۶۲ عنوان مثال یک نسخه (مختص مدل) که این را در نظر می‌گیرد که بسیاری از مدل‌های پیش‌بینی ممکن
۳۹۶۳ است داده‌ها را به خوبی پیش‌بینی کنند. مقاله آنها ارزش خواندن دارد.
۳۹۶۴

۳۹۶۵ الگوریتم اهمیت ویژگی جایگشت بر اساس فیشر، رودین و دومینیسی (۲۰۱۸):

۳۹۶۶ ورودی: مدل آموزش دیده^۸ ، ماتریس ویژگی ایکس ، بردار هدف^۷ ، اندازه گیری خطای^۶ (L^{۸f}) .

۳۹۶۷ ۱-خطای مدل اصلی را تخمین بزنید^۵ من^۴(y=g^{۸f},L^۵) ((مثالاً میانگین مربعات خطای)

۳۹۶۸ ۲-برای هر ویژگی^۱...^p { انجام دادن:

۳۹۶۹ -ایجاد ماتریس ویژگی ایکس پ^۵ متر با جابجایی ویژگی^۷ در داده X. این ارتباط بین ویژگی^۷ و نتیجه واقعی^۶
۳۹۷۰ را از بین می‌برد.

۳۹۷۱ -خطای برآورده^۵ پ^۶ متر(=L^{۸f}) ایکس پ^۵ متر ((بر اساس پیش‌بینی داده‌های تغییر یافته). اهمیت ویژگی^۷
۳۹۷۲ جایگشت را به عنوان ضریب محاسبه کنید اف من=ا^۵ پ^۶ متر/۵ من^۴ g با تفاوت اف من=ا^۵ پ^۶ متر-
۳۹۷۳ ۵ من^۴ g

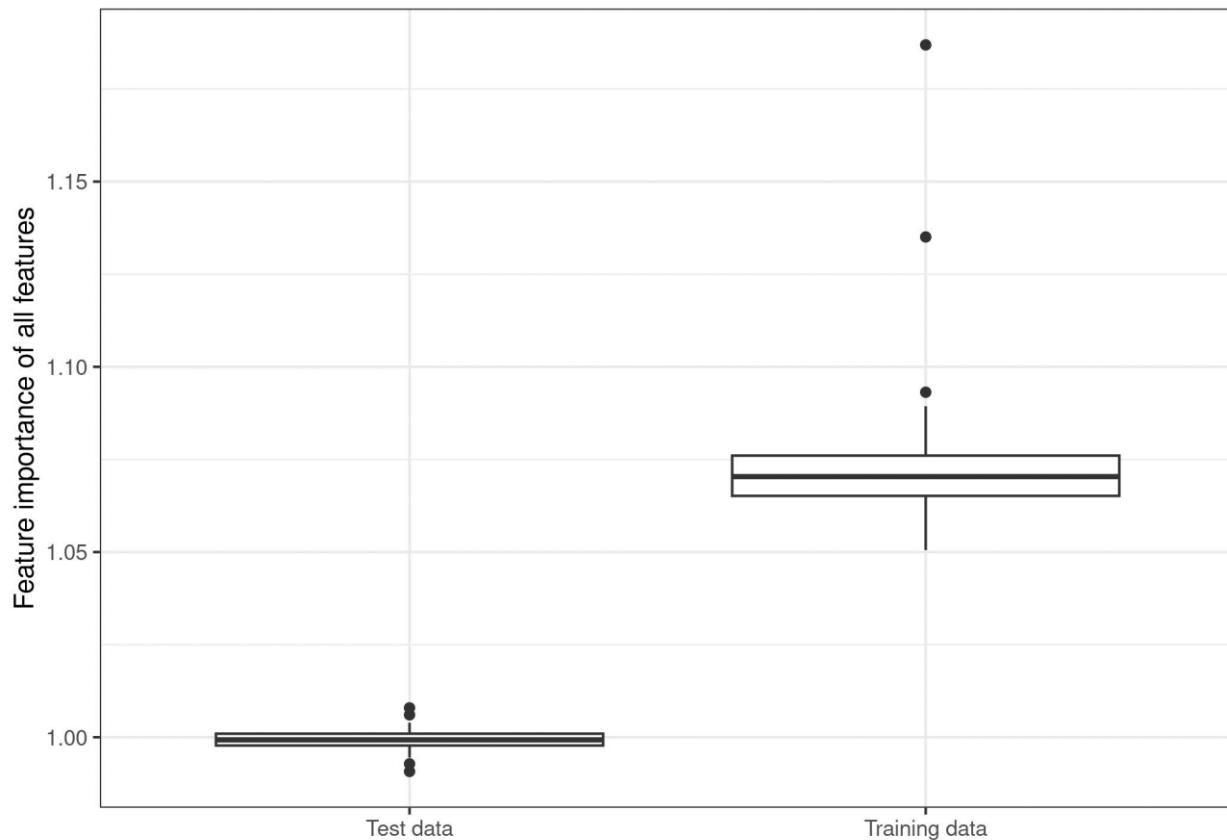
۳۹۷۴ ۳-مرتب سازی ویژگی‌ها با نزولی Fl.

۳۹۷۵ فیشر، رودین و دومینیسی (۲۰۱۸) در مقاله خود پیشنهاد می‌کنند که مجموعه داده را به نصف تقسیم کرده و
۳۹۷۶ مقادیر ویژگی^۷ از دو نیمه را به جای تغییر ویژگی^۷ تغییر دهید. اگر در مورد آن فکر کنید، این دقیقاً مشابه

۳۹۷۷ تغییر ویژگی \hat{z} است. اگر تخمین دقیق‌تری می‌خواهید، می‌توانید با جفت کردن هر نمونه با مقدار ویژگی \hat{z}
۳۹۷۸ نمونه‌های دیگر (به جز با خودش) خطای تغییر ویژگی \hat{z} را تخمین بزنید. این یک مجموعه داده با اندازه برای
۳۹۷۹ تخمین خطای جایگشت به شما می‌دهد ($n-1$) زمان محاسباتی زیادی را می‌طلبد. من فقط در صورتی می
۳۹۸۰ توانم استفاده از $(n-1)$ روش را توصیه کنم که در مورد تخمین‌های بسیار دقیق جدی هستید.

۳۹۸۱ آیا باید اهمیت داده‌های آموزش یا آزمون را محاسبه کنم؟
۳۹۸۲ احتمالاً باید از داده‌های آزمایشی استفاده کنید.

۳۹۸۳ پاسخ به سؤال در مورد داده‌های آزمایشی، این سؤال اساسی را در مورد اهمیت ویژگی نشان می
۳۹۸۴ دهد. بهترین راه برای درک تفاوت بین اهمیت ویژگی بر اساس آموزش در مقابل داده‌های آزمایشی، یک مثال
۳۹۸۵ «افراتی» است. من یک ماشین بردار پشتیبان را آموزش دادم تا یک نتیجه هدف پیوسته و تصادفی را با توجه
۳۹۸۶ به ۵۰ ویژگی تصادفی (۲۰۰ نمونه) پیش‌بینی کند. منظور من از "تصادفی" این است که نتیجه هدف مستقل
۳۹۸۷ از ۵۰ ویژگی است. این مانند پیش‌بینی دمای فردا با توجه به آخرین اعداد قرعه کشی است. اگر مدل هر رابطه
۳۹۸۸ ای را "یاد بگیرد"، بیش از حد برازش می‌کند. و در واقع، SVM بر روی داده‌های آموزشی بیش از حد تطبیق
۳۹۸۹ داد. میانگین خطای مطلق (کوتاه mae: برای داده‌های آموزشی ۰,۲۹ و برای داده‌های آزمون ۸۲,۰۰ است که
۳۹۹۰ همچنین خطای مطلق بهترین مدل ممکن است که همیشه میانگین نتیجه ۰ را پیش‌بینی می‌کند (mae از ۷۸,۰).
۳۹۹۱ به عبارت دیگر مدل SVM زباله است. چه مقادیری برای اهمیت ویژگی برای ۵۰ ویژگی این SVM بیش از حد
۳۹۹۲ برازش شده انتظار دارید؟ صفر است زیرا هیچ یک از ویژگی‌ها به بهبود عملکرد در داده‌های آزمایشی دیده
۳۹۹۳ نشده کمک نمی‌کند؟ یا اینکه آیا این اهمیت‌ها باید منعکس کنند که چقدر مدل به هر یک از ویژگی‌ها
۳۹۹۴ بستگی دارد، صرف نظر از اینکه آیا روابط آموخته شده به داده‌های دیده نشده تعمیم می‌یابد؟ اجازه دهید
۳۹۹۵ نگاهی بیندازیم که چگونه توزیع اهمیت ویژگی‌ها برای داده‌های آموزشی و آزمایشی متفاوت است. صرف نظر
۳۹۹۶ از اینکه آیا روابط آموخته شده به داده‌های دیده نشده تعمیم می‌یابد؟ اجازه دهید نگاهی بیندازیم که چگونه
۳۹۹۷ توزیع اهمیت ویژگی‌ها برای داده‌های آموزشی و آزمایشی متفاوت است. صرف نظر از اینکه آیا روابط آموخته
۳۹۹۸ شده به داده‌های دیده نشده تعمیم می‌یابد؟ اجازه دهید نگاهی بیندازیم که چگونه توزیع اهمیت ویژگی‌ها
۳۹۹۹ برای داده‌های آموزشی و آزمایشی متفاوت است.



شکل ۸,۲۴: توزیع مقداری اهمیت ویژگی بر اساس نوع داده. یک SVM بر روی یک مجموعه داده رگرسیون با ۵۰ ویژگی تصادفی و ۲۰۰ نمونه آموزش داده شد SVM. داده ها را بیش از حد برآذش می دهد: اهمیت ویژگی بر اساس داده های آموزشی بسیاری از ویژگی های مهم را نشان می دهد. با محاسبه داده های آزمایشی دیده نشده، اهمیت ویژگی ها نزدیک به نسبت یک ($=\text{بی اهمیت}$) است.

برای من مشخص نیست که کدام یک از این دو نتیجه مطلوب تر است. بنابراین من سعی خواهم کرد برای هر دو نسخه موردنی ایجاد کنم.

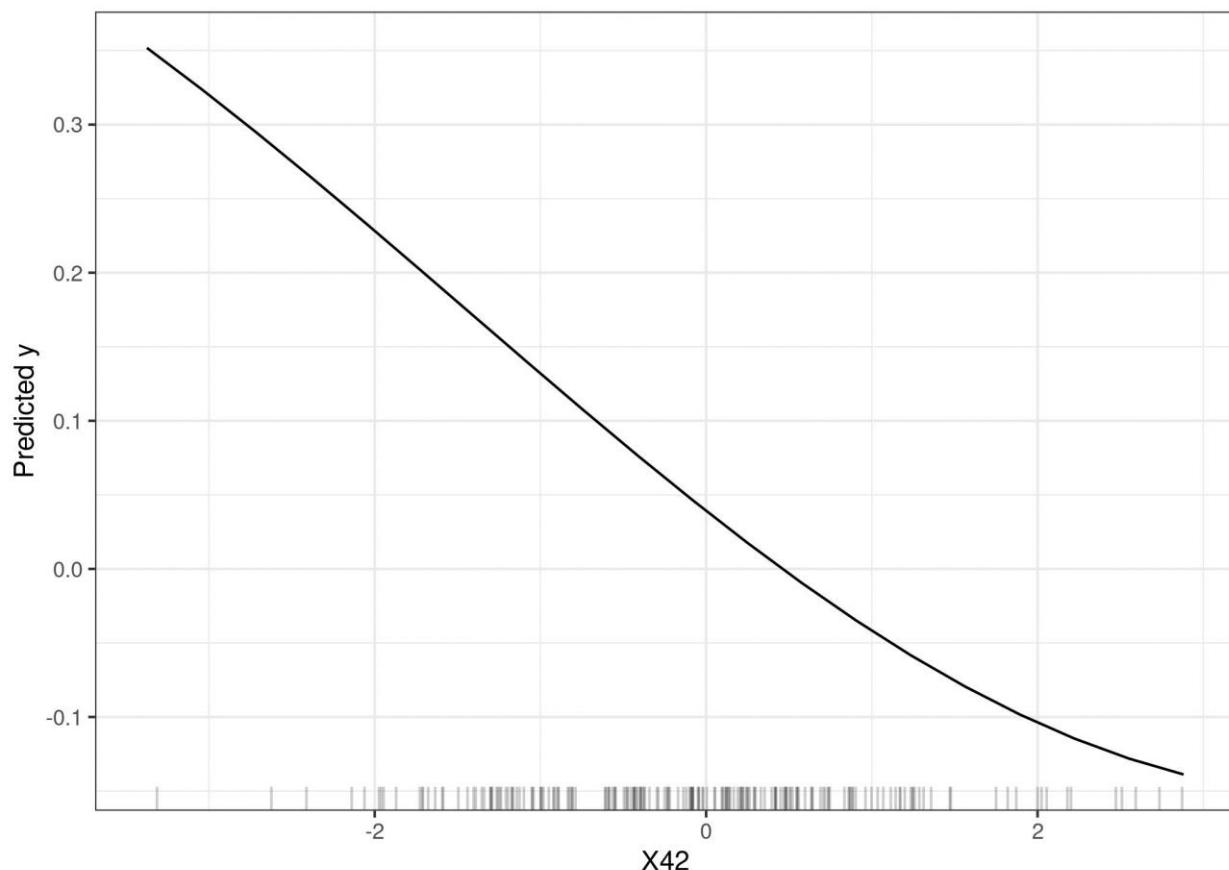
۴۰۰۷ مورد برای داده های تست

۴۰۰۸ این یک مورد ساده است: تخمین های خطای مدل بر اساس داده های آموزشی زباله هستند -> اهمیت ویژگی به تخمین خطای مدل متکی است -> اهمیت ویژگی بر اساس داده های آموزشی زباله است. در واقع، این یکی از ۴۰۰۹ اولین چیزهایی است که در یادگیری ماشین یاد می گیرید: اگر خطای مدل (یا عملکرد) را بر روی همان داده هایی که مدل آموزش داده شده است اندازه گیری کنید، اندازه گیری معمولاً خیلی خوش بینانه است، به این ۴۱۰ معنی که به نظر می رسد مدل خیلی بهتر از واقعیت کار می کند. و از آنجایی که اهمیت ویژگی جایگشت به ۴۱۱

۴۰۱۳ اندازه گیری خطای مدل بستگی دارد، باید از داده های آزمایشی دیده نشده استفاده کنیم. اهمیت ویژگی بر
۴۰۱۴ اساس داده های آموزشی باعث می شود ما به اشتباه فکر کنیم که ویژگی ها برای پیش بینی ها مهم هستند، در
۴۰۱۵ حالی که در واقعیت مدل بیش از حد برازش داشت و ویژگی ها اصلاً مهم نبودند.

۴۰۱۶ موردی برای داده های آموزشی

۴۰۱۷ فرمول بندی استدلال های استفاده از داده های آموزشی تا حدودی دشوارتر است، اما IMHO به اندازه
۴۰۱۸ استدلال های استفاده از داده های آزمون قانع کننده است. نگاهی دیگر به SVM زباله خود می اندازیم. بر اساس
۴۰۱۹ داده های آموزشی، مهم ترین ویژگی X42 بود. اجازه دهید به نمودار وابستگی جزئی ویژگی X42 نگاه کنیم.
۴۰۲۰ نمودار وابستگی جزئی نشان می دهد که چگونه خروجی مدل بر اساس تغییرات ویژگی تغییر می کند و بر
۴۰۲۱ خطای تعمیم تکیه نمی کند. فرقی نمی کند که PDP با داده های آموزشی یا آزمایشی محاسبه شود.



۴۰۲۲ شکل ۸,۲۵ PDP ویژگی X42، که با توجه به اهمیت ویژگی بر اساس داده های آموزشی، مهم ترین ویژگی
۴۰۲۳ است. نمودار نشان می دهد که چگونه SVM برای پیش بینی به این ویژگی وابسته است
۴۰۲۴

۴۰۲۵

نmodار به وضوح نشان می دهد که SVM یاد گرفته است برای پیش بینی های خود به ویژگی X42 تکیه کند، اما با توجه به اهمیت ویژگی بر اساس داده های آزمایشی (۱)، مهم نیست. بر اساس داده های آموزشی، اهمیت ۱,۱۹ است که نشان می دهد مدل یاد گرفته است از این ویژگی استفاده کند. اهمیت ویژگی بر اساس داده های آموزشی به ما می گوید که کدام ویژگی برای مدل مهم است به این معنا که برای پیش بینی به آنها بستگی دارد.

به عنوان بخشی از مورد استفاده از داده های آموزشی، می خواهیم استدلالی علیه داده های آزمایشی معرفی کنم. در عمل، شما می خواهید از تمام داده های خود برای آموزش مدل خود استفاده کنید تا در نهایت بهترین مدل ممکن را بدست آورید. این بدان معناست که هیچ داده آزمایشی استفاده نشده ای برای محاسبه اهمیت ویژگی باقی نمانده است. هنگامی که می خواهید خطای تعیین مدل خود را تخمین بزنید، همین مشکل را دارید. اگر از اعتبارسنجی متقطع (تودرتو) برای تخمین اهمیت ویژگی استفاده کنید، با این مشکل مواجه خواهید شد که اهمیت ویژگی در مدل نهایی با همه داده ها محاسبه نمی شود، بلکه در مدل هایی با زیرمجموعه هایی از داده ها که ممکن است رفتار متفاوتی داشته باشند، محاسبه می شود.

با این حال، در پایان توصیه می کنم از داده های آزمایشی برای اهمیت ویژگی جایگشت استفاده کنید. زیرا اگر علاقه مند هستید که پیش بینی های مدل چقدر تحت تأثیر یک ویژگی است، باید از معیارهای اهمیت دیگری مانند اهمیت SHAP استفاده کنید.

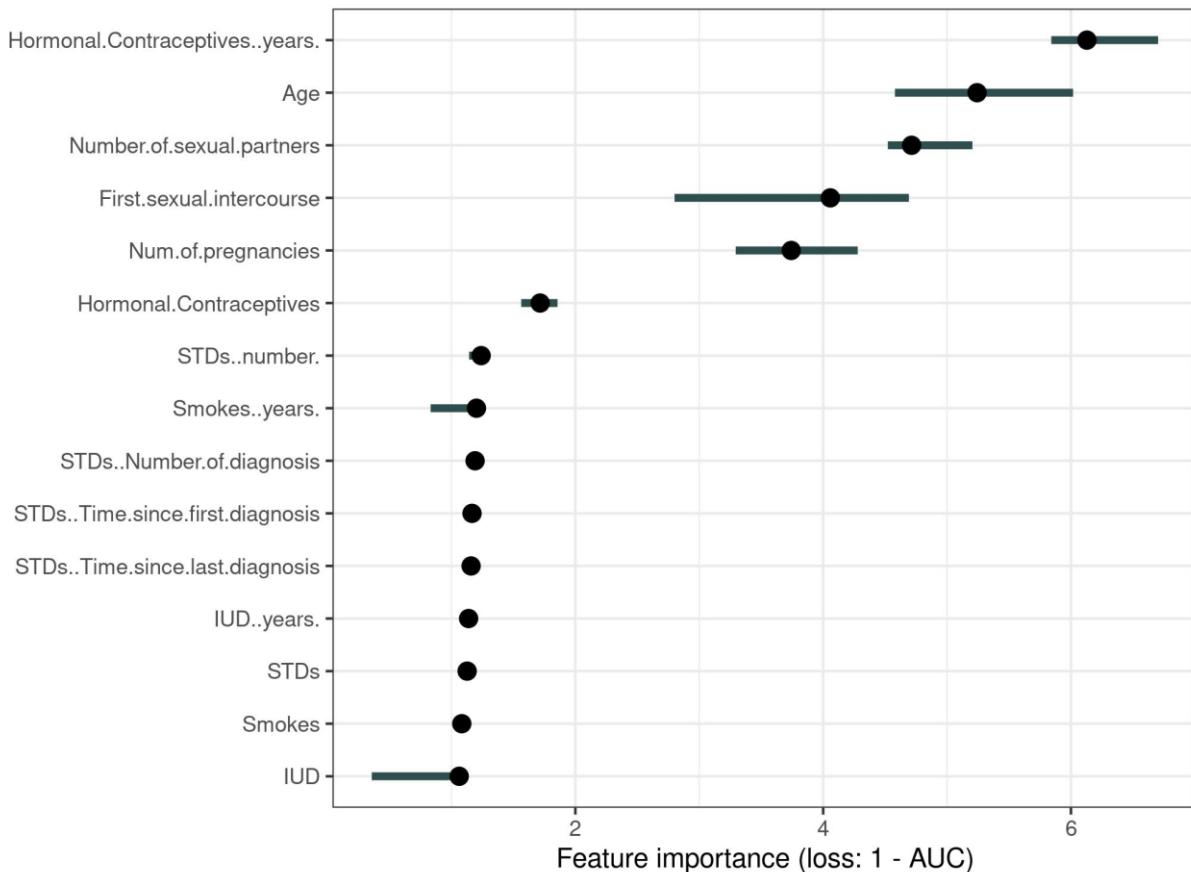
در ادامه به چند نمونه نگاه می کنیم. من محاسبه اهمیت را بر اساس داده های آموزشی استوار کردم، زیرا باید یکی را انتخاب می کردم و استفاده از داده های آموزشی به چند خط کد کمتر نیاز داشت.

۸,۵,۳ مثال و تفسیر

من نمونه هایی را برای طبقه بندی و رگرسیون نشان می دهم.

سرطان دهانه رحم (طبقه بندی)

ما یک مدل جنگل تصادفی را برای پیش بینی سرطان دهانه رحم برآش می کنیم . افزایش خطا را با ۱-AUC -ROC اندازه گیری می کنیم. ویژگی های مرتبط با افزایش خطای مدل با ضریب ۱ (= بدون تغییر) برای پیش بینی سرطان دهانه رحم مهم نبود.

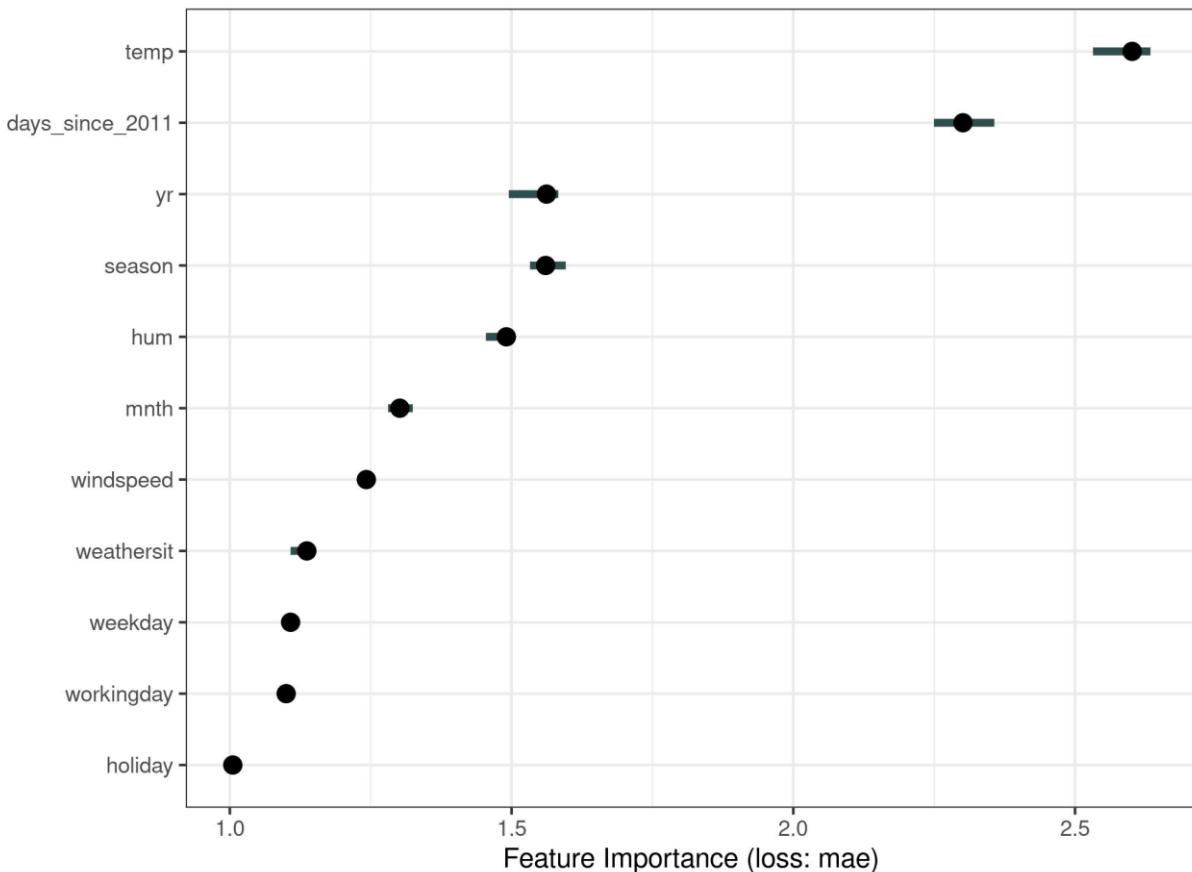


شکل ۸،۲۶: اهمیت هر یک از ویژگی ها برای پیش بینی سرطان دهانه رحم با یک جنگل تصادفی. مهم ترین ویژگی هورمونی بود. پیشگیری از بارداری..سال. تغییر هورمونی.پیشگیری از بارداری..سال. منجر به افزایش ۱-AUC با ضریب ۶,۱۳ شد

۴۰۵۳ مهمترین ویژگی هورمونی و ضد بارداری سالها بود. با افزایش خطای ۶,۱۳ پس از جایگشت همراه است.

۴۰۵۴ اشتراک دوچرخه (رگرسیون)

۴۰۵۵ ما یک مدل ماشین بردار پشتیبان را برای پیش بینی تعداد دوچرخه های اجاره ای ، با توجه به شرایط آب و هوایی
۴۰۵۶ و اطلاعات تقویم، برآذش می کنیم. به عنوان اندازه گیری خطا از میانگین خطای مطلق استفاده می کنیم.



شکل ۸,۲۷: اهمیت هر یک از ویژگی ها در پیش بینی شمارش دوچرخه با ماشین بردار پشتیبان. مهمترین ویژگی دما بود، کمترین اهمیت تعطیلات بود.

۸,۵,۴ مزايا

تفسیر خوب : اهمیت ویژگی افزایش خطای مدل زمانی است که اطلاعات ویژگی از بین می رود.
اهمیت ویژگی یک بینش بسیار فشرده و جهانی در مورد رفتار مدل ارائه می دهد.

یک جنبه مثبت استفاده از نسبت خطای تفاوت خطا به جای اندازه گیری اهمیت ویژگی در مسائل مختلف قابل مقایسه است.

اندازه گیری اهمیت به طور خودکار تمام تعاملات با سایر ویژگی ها را در نظر می گیرد . با جابجایی ویژگی، افکت های تعامل با سایر ویژگی ها را نیز از بین می برید. این بدان معنی است که اهمیت ویژگی جایگشت هم اثر ویژگی اصلی و هم اثرات متقابل بر عملکرد مدل را در نظر می گیرد. این نیز یک نقطه ضعف است زیرا اهمیت تعامل بین دو ویژگی در اندازه گیری اهمیت هر دو ویژگی گنجانده شده است. این بدان معناست که اهمیت

ویژگی‌ها به کاهش کل عملکرد اضافه نمی‌شود، اما مجموع آن بزرگ‌تر است. تنها در صورتی که هیچ تعاملی بین
۴۰۶۹
ویژگی‌ها وجود نداشته باشد، مانند یک مدل خطی، اهمیت تقریباً افزایش می‌یابد.

۴۰۷۰
۴۰۷۱ اهمیت ویژگی جایگشت نیازی به آموزش مجدد مدل ندارد. برخی از روش‌های دیگر حذف یک ویژگی، آموزش
۴۰۷۲ مجدد مدل و سپس مقایسه خطای مدل را پیشنهاد می‌کنند. از آنجایی که بازآموزی یک مدل یادگیری ماشینی
۴۰۷۳ می‌تواند زمان زیادی طول بکشد، « فقط » تغییر یک ویژگی می‌تواند در زمان زیادی صرفه‌جویی کند. روش‌های
۴۰۷۴ مهمی که مدل را با زیرمجموعه‌ای از ویژگی‌ها بازآموزی می‌کنند در نگاه اول بصری به نظر می‌رسند، اما مدل با
۴۰۷۵ داده‌های کاهش‌یافته برای اهمیت ویژگی بی‌معنی است. ما به اهمیت ویژگی یک مدل ثابت علاقه مندیم.
۴۰۷۶ بازآموزی با مجموعه داده کاهش‌یافته، مدلی متفاوت از مدل مورد علاقه ما ایجاد می‌کند. فرض کنید یک مدل
۴۰۷۷ خطی پراکنده (با کمند) با تعداد مشخصی از ویژگی‌ها با وزن غیر صفر را آموزش می‌دهید. مجموعه داده دارای
۴۰۷۸ ۱۰۰ ویژگی است، شما تعداد وزن‌های غیر صفر را ۵ می‌کنید. اهمیت یکی از ویژگی‌هایی که وزن غیر صفر
۴۰۷۹ دارند را تحلیل می‌کنید. شما ویژگی را حذف کرده و مدل را دوباره آموزش می‌دهید. عملکرد مدل ثابت می‌
۴۰۸۰ ماند زیرا یکی دیگر از ویژگی‌های به همان اندازه خوب وزن غیر صفر می‌گیرد و نتیجه شما این است که
۴۰۸۱ ویژگی مهم نبوده است. مثال دیگر: مدل یک درخت تصمیم است و ما اهمیت ویژگی را که به عنوان اولین
۴۰۸۲ تقسیم انتخاب شده است تجزیه و تحلیل می‌کنیم. شما ویژگی را حذف کرده و مدل را دوباره آموزش می‌
۴۰۸۳ دهید. از آنجایی که ویژگی دیگری به عنوان اولین تقسیم انتخاب شده است، کل درخت می‌تواند بسیار متفاوت
۴۰۸۴ باشد، به این معنی که نرخ خطای درختان (به طور بالقوه) کاملاً متفاوت را مقایسه می‌کنیم تا تصمیم بگیریم
۴۰۸۵ که این ویژگی برای یکی از درختان چقدر مهم است. مدل یک درخت تصمیم است و ما اهمیت ویژگی را که به
۴۰۸۶ عنوان اولین تقسیم انتخاب شده است تجزیه و تحلیل می‌کنیم. شما ویژگی را حذف کرده و مدل را دوباره
۴۰۸۷ آموزش می‌دهید. از آنجایی که ویژگی دیگری به عنوان اولین تقسیم انتخاب شده است، کل درخت می‌تواند
۴۰۸۸ بسیار متفاوت باشد، به این معنی که نرخ خطای درختان (به طور بالقوه) کاملاً متفاوت را مقایسه می‌کنیم تا
۴۰۸۹ تصمیم بگیریم که این ویژگی برای یکی از درختان چقدر مهم است. مدل یک درخت تصمیم است و ما اهمیت
۴۰۹۰ ویژگی را که به عنوان اولین تقسیم انتخاب شده است تجزیه و تحلیل می‌کنیم. شما ویژگی را حذف کرده و
۴۰۹۱ مدل را دوباره آموزش می‌دهید. از آنجایی که ویژگی دیگری به عنوان اولین تقسیم انتخاب شده است، کل
۴۰۹۲ درخت می‌تواند بسیار متفاوت باشد، به این معنی که نرخ خطای درختان (به طور بالقوه) کاملاً متفاوت را
۴۰۹۳ مقایسه می‌کنیم تا تصمیم بگیریم که این ویژگی برای یکی از درختان چقدر مهم است.

۴۰۹۵ اهمیت ویژگی جایگشت به خطای مدل مرتبط است . این ذاتا بد نیست، اما در برخی موارد آن چیزی نیست که
 ۴۰۹۶ شما نیاز دارید. در برخی موارد، ممکن است ترجیح دهید بدانید که خروجی مدل برای یک ویژگی چقدر
 ۴۰۹۷ متفاوت است بدون اینکه به معنای عملکرد آن توجه کنید. به عنوان مثال، شما می خواهید بفهمید که وقتی
 ۴۰۹۸ شخصی ویژگی ها را دستکاری می کند، خروجی مدل شما چقدر قوی است. در این مورد، شما علاقه ای به این
 ۴۰۹۹ نخواهید داشت که با تغییر یک ویژگی، عملکرد مدل چقدر کاهش می یابد، بلکه چقدر از واریانس خروجی مدل
 ۴۱۰۰ توسط هر ویژگی توضیح داده می شود. واریانس مدل (توضیح داده شده توسط ویژگی ها) و اهمیت ویژگی
 ۴۱۰۱ زمانی که مدل به خوبی تعمیم می یابد (یعنی بیش از حد برازنده نمی شود) به شدت با هم مرتبط هستند.

۴۱۰۲ شما نیاز به دسترسی به نتیجه واقعی دارید . اگر کسی فقط مدل و داده های بدون برچسب را در اختیار شما
 ۴۱۰۳ قرار دهد - اما نتیجه واقعی را نه - نمی توانید اهمیت ویژگی جایگشت را محاسبه کنید.

۴۱۰۴ اهمیت ویژگی جایگشت بستگی به به هم زدن ویژگی دارد که تصادفی بودن را به اندازه گیری اضافه می کند.
 ۴۱۰۵ هنگامی که جایگشت تکرار می شود، نتایج ممکن است بسیار متفاوت باشد . تکرار جایگشت و میانگین
 ۴۱۰۶ معیارهای اهمیت نسبت به تکرارها، اندازه گیری را ثابت می کند، اما زمان محاسبه را افزایش می دهد.

۴۱۰۷ اگر ویژگی ها با هم مرتبط باشند، اهمیت ویژگی جایگشت می تواند توسط نمونه های داده غیرواقعی سوگیری
 ۴۱۰۸ شود . مشکل همانند نمودارهای وابستگی جزئی است: جایگشت ویژگی ها در هنگام همبستگی دو یا چند
 ۴۱۰۹ ویژگی، نمونه های داده بعيد ایجاد می کند. وقتی همبستگی مثبت دارند (مانند قد و وزن یک فرد) و من یکی
 ۴۱۱۰ از ویژگی ها را به هم می زنم، نمونه های جدیدی ایجاد می کنم که بعيد یا حتی از نظر فیزیکی غیرممکن است
 ۴۱۱۱ (مثلًا فرد ۲ متری با وزن ۳۰ کیلوگرم)، اما از این نمونه های جدید استفاده می کنم. برای اندازه گیری اهمیت
 ۴۱۱۲ به عبارت دیگر، برای اهمیت ویژگی جایگشتی یک ویژگی همبسته، ما در نظر می گیریم که وقتی ویژگی را با
 ۴۱۱۳ مقادیری که هرگز در واقعیت مشاهده نمی کنیم مبادله می کنیم، عملکرد مدل چقدر کاهش می یابد. بررسی
 ۴۱۱۴ کنید که آیا ویژگی ها به شدت همبستگی دارند و در صورت وجود، در مورد تفسیر اهمیت ویژگی دقت کنید. با
 ۴۱۱۵ این حال، همبستگی های زوجی ممکن است برای آشکار کردن مشکل کافی نباشد.

۴۱۱۶ نکته دشوار دیگر: افزودن یک ویژگی مرتبط می تواند اهمیت ویژگی مرتبه را کاهش دهد با تقسیم اهمیت بین
 ۴۱۱۷ هر دو ویژگی. اجازه دهید مثالی از منظورم از «تقسیم کردن» اهمیت ویژگی به شما بگوییم: ما می خواهیم
 ۴۱۱۸ احتمال باران را پیش بینی کنیم و از دمای ساعت ۸ صبح روز قبل به عنوان یک ویژگی همراه با سایر ویژگی های
 ۴۱۱۹ نامرتبط استفاده کنیم. من یک جنگل تصادفی تمرین می کنم و معلوم می شود که دما مهمترین ویژگی است و
 ۴۱۲۰ همه چیز خوب است و شب بعد خوب می خوابیم. حال سناریوی دیگری را تصور کنید که در آن دمای ۹:۰۰

صبح را به عنوان یک ویژگی که به شدت با دمای ساعت ۸:۰۰ صبح مرتبط است، لحاظ کنم، دمای ساعت ۹:۰۰
صبح اگر از قبل دمای ساعت ۸:۰۰ صبح را بدانم، اطلاعات بیشتری به من نمی دهد. اما داشتن ویژگی های
بیشتر همیشه خوب است، درست است؟ من یک جنگل تصادفی را با دو ویژگی دما و ویژگی های نامرتبط
آموزش می دهم. برخی از درختان در جنگل تصادفی دمای ۸ صبح را می گیرند، برخی دیگر دمای ۹:۰۰ صبح،
دوباره برخی دیگر هر دو و برخی دیگر هیچ کدام. دو ویژگی دما در کنار هم کمی اهمیت بیشتری نسبت به
ویژگی دمای واحد قبلی دارند، اما به جای قرار گرفتن در بالای لیست ویژگی های مهم، هر دما اکنون جایی در
وسط است. با معرفی یک ویژگی همبسته، مهمترین ویژگی را از بالای نردهبان اهمیت به حد متوسط رساندم. از
یک طرف این خوب است، زیرا به سادگی رفتار مدل یادگیری ماشین زیربنایی، در اینجا جنگل تصادفی را
منعکس می کند. دمای ۸:۰۰ صبح به سادگی از اهمیت کمتری برخوردار شده است زیرا مدل اکنون می تواند
به اندازه گیری ۹:۰۰ صبح نیز تکیه کند. از سوی دیگر، تفسیر اهمیت ویژگی را به میزان قابل توجهی دشوارتر
می کند. تصور کنید می خواهید ویژگی ها را برای خطاهای اندازه گیری بررسی کنید. چک گران است و شما
تصمیم می گیرید فقط ۳ مورد از مهمترین ویژگی ها را بررسی کنید. در مورد اول دما را بررسی می کنید، در
مورد دوم هیچ ویژگی دما را فقط به این دلیل که آنها اکنون اهمیت را به اشتراک می گذارند درج نمی کنند.
حتی اگر مقادیر اهمیت ممکن است در سطح رفتار مدل معنا پیدا کند، اگر ویژگی های همبسته داشته باشید
گیج کننده است.

الگوریتمی به نام PIMP الگوریتم اهمیت ویژگی جایگشت را برای ارائه مقادیر p برای اهمیت ها تطبیق می
دهد. یکی دیگر از جایگزین های مبتنی بر ضرر، حذف ویژگی از داده های آموزشی، آموزش مجدد مدل و
اندازه گیری افزایش تلفات است. جایگشت یک ویژگی و اندازه گیری افزایش از دست دادن تنها راه برای اندازه
گیری اهمیت یک ویژگی نیست. معیارهای مختلف اهمیت را می توان به روش های مدل خاص و مدل-
آنالیتیک تقسیم کرد. اهمیت جینی برای جنگل های تصادفی یا ضرایب رگرسیون استاندارد برای مدل های
رگرسیون نمونه هایی از معیارهای اهمیت ویژه مدل هستند.

یک جایگزین مدل آنالیتیک برای اهمیت ویژگی جایگشت، معیارهای مبتنی بر واریانس هستند. معیارهای
اهمیت ویژگی مبتنی بر واریانس مانند شاخص های ANOVA یا Sobol عملکردی به ویژگی هایی که باعث
واریانس بالایی درتابع پیش بینی می شوند اهمیت بیشتری می دهند. همچنین اهمیت SHAP شباهت هایی به
اندازه گیری اهمیت مبتنی بر واریانس دارد. اگر تغییر یک ویژگی خروجی را تا حد زیادی تغییر می دهد، مهم
است. این تعریف اهمیت با تعریف مبتنی بر ضرر در مورد اهمیت ویژگی جایگشت متفاوت است. این در مواردی

مشهود است که یک مدل بیش از حد نصب شده باشد. اگر مدلی بیش از حد برآش می‌کند و از ویژگی غیرمرتب با خروجی استفاده می‌کند، اهمیت ویژگی جایگشتی اهمیتی برابر با صفر می‌دهد زیرا این ویژگی به تولید پیش‌بینی‌های صحیح کمک نمی‌کند. از سوی دیگر، اندازه‌گیری اهمیت مبنی بر واریانس، ممکن است به ویژگی اهمیت بالایی بددهد زیرا زمانی که ویژگی تغییر می‌کند، پیش‌بینی می‌تواند تغییرات زیادی داشته باشد.

نمای کلی خوبی از تکنیک‌های مختلف اهمیت در مقاله توسط Wei (2015) ارائه شده است.

نرم افزار ۸,۵,۷
بسته imlR برای نمونه‌ها استفاده شد. بسته‌های DALEX R و vip همچنین کتابخانه پایتون alibi و scikit-learn همچنین rfpimp اهمیت ویژگی جایگشت مدل-آگوستیک را پیاده‌سازی می‌کنند.

۸.۶ جانشین جهانی

یک مدل جایگزین جهانی یک مدل قابل تفسیر است که برای تقریب پیش‌بینی‌های یک مدل جعبه سیاه آموزش داده شده است. ما می‌توانیم با تفسیر مدل جایگزین در مورد مدل جعبه سیاه نتیجه گیری کنیم. حل تفسیر پذیری یادگیری ماشین با استفاده از یادگیری ماشین بیشتر!

۱.۶.۱ نظریه

مدل‌های جایگزین در مهندسی نیز استفاده می‌شوند: اگر یک نتیجه مورد علاقه گران، زمان‌بر یا اندازه‌گیری آن دشوار باشد (مثلاً به دلیل اینکه از یک شبیه‌سازی رایانه‌ای پیچیده می‌آید)، می‌توان به جای آن از یک مدل جایگزین ارزان و سریع نتیجه استفاده کرد. تفاوت بین مدل‌های جایگزین مورد استفاده در مهندسی و یادگیری ماشین قابل تفسیر در این است که مدل زیربنایی یک مدل یادگیری ماشینی است (نه شبیه‌سازی) و اینکه مدل جایگزین باید قابل تفسیر باشد. هدف از مدل‌های جایگزین (قابل تفسیر) تقریب پیش‌بینی‌های مدل زیربنایی تا حد امکان دقیق و در عین حال قابل تفسیر بودن است. ایده مدل‌های جایگزین را می‌توان با نام‌های مختلفی یافت: مدل تقریبی، متamodel، مدل سطح پاسخ، شبیه‌ساز، ...

درباره تئوری: در واقع برای درک مدل‌های جایگزین به نظریه زیادی نیاز نیست. ما می‌خواهیم تابع پیش‌بینی جعبه سیاه f را تا حد امکان با تابع پیش‌بینی مدل جایگزین g ، تحت این محدودیت که g قابل تفسیر است، تقریب بزنیم. برای تابع g می‌توان از هر مدل قابل تفسیر - به عنوان مثال از فصل مدل‌های قابل تفسیر - استفاده کرد.

به عنوان مثال یک مدل خطی:

$$g(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

یا درخت تصمیمی:

$$g(x) = \sum_{m=1}^M c_m I\{x \in R_m\}$$

آموزش یک مدل جایگزین یک روش مدل-آگنوستیک است، زیرا به هیچ اطلاعاتی در مورد عملکرد درونی مدل جعبه سیاه نیاز ندارد، فقط دسترسی به داده‌ها و عملکرد پیش‌بینی ضروری است. اگر مدل یادگیری ماشین زیربنایی با مدل دیگری جایگزین شد، همچنان می‌توانید از روش جایگزین استفاده کنید. انتخاب نوع مدل جعبه سیاه و نوع مدل جایگزین جدا شده است.

برای بدست آوردن مدل جایگزین مراحل زیر را انجام دهید:

۱- یک مجموعه داده X را انتخاب کنید. این می تواند همان مجموعه داده ای باشد که برای آموزش مدل جعبه سیاه استفاده شده است یا یک مجموعه داده جدید از همان توزیع. حتی می توانید بسته به برنامه خود زیر مجموعه ای از داده ها یا شبکه ای از نقاط را انتخاب کنید.

۲- برای مجموعه داده انتخابی X ، پیش بینی های مدل جعبه سیاه را دریافت کنید.

۳- یک نوع مدل قابل تفسیر (مدل خطی، درخت تصمیم، ...) را انتخاب کنید.

۴- مدل قابل تفسیر را بر روی مجموعه داده X و پیش بینی های آن آموزش دهید.

۵- تبریک می گویم! شما اکنون یک مدل جایگزین دارید.

۶- اندازه گیری کنید که مدل جایگزین چقدر پیش بینی های مدل جعبه سیاه را تکرار می کند.

۷- مدل جایگزین را تفسیر کنید.

ممکن است روش هایی برای مدل های جایگزین پیدا کنید که مراحل اضافی دارند یا کمی متفاوت هستند، اما ایده کلی معمولاً همانطور که در اینجا توضیح داده شده است.

یکی از راه های اندازه گیری اینکه جانشین چقدر مدل جعبه سیاه را تکرار می کند، اندازه گیری R-squared است:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^n (\hat{y}_*^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^n (\hat{y}^{(i)} - \bar{y})^2}$$

جایی که $\hat{y}^{(i)}$ من $*$ (پیش بینی نمونه $-a$) مدل جایگزین است، $\hat{y}_*^{(i)}$ من (پیش بینی مدل جعبه سیاه و \bar{y} میانگین پیش بینی های مدل جعبه سیاه SSE مخفف مجموع مربعات خطای SST مخفف مجموع مربعات کل است. معیار R-squared را می توان به عنوان درصد واریانسی که توسط مدل جانشین گرفته شده است تفسیر کرد. اگر R-squared نزدیک به $1 = SSE$ کم باشد، مدل قابل تفسیر رفتار مدل جعبه سیاه را به خوبی تقریب می کند. اگر مدل قابل تفسیر بسیار نزدیک است، ممکن است بخواهید مدل پیچیده را با مدل قابل تفسیر جایگزین کنید. اگر R-squared نزدیک به $0 = SSE$ بالا باشد، مدل قابل تفسیر نمی تواند مدل جعبه سیاه را توضیح دهد.

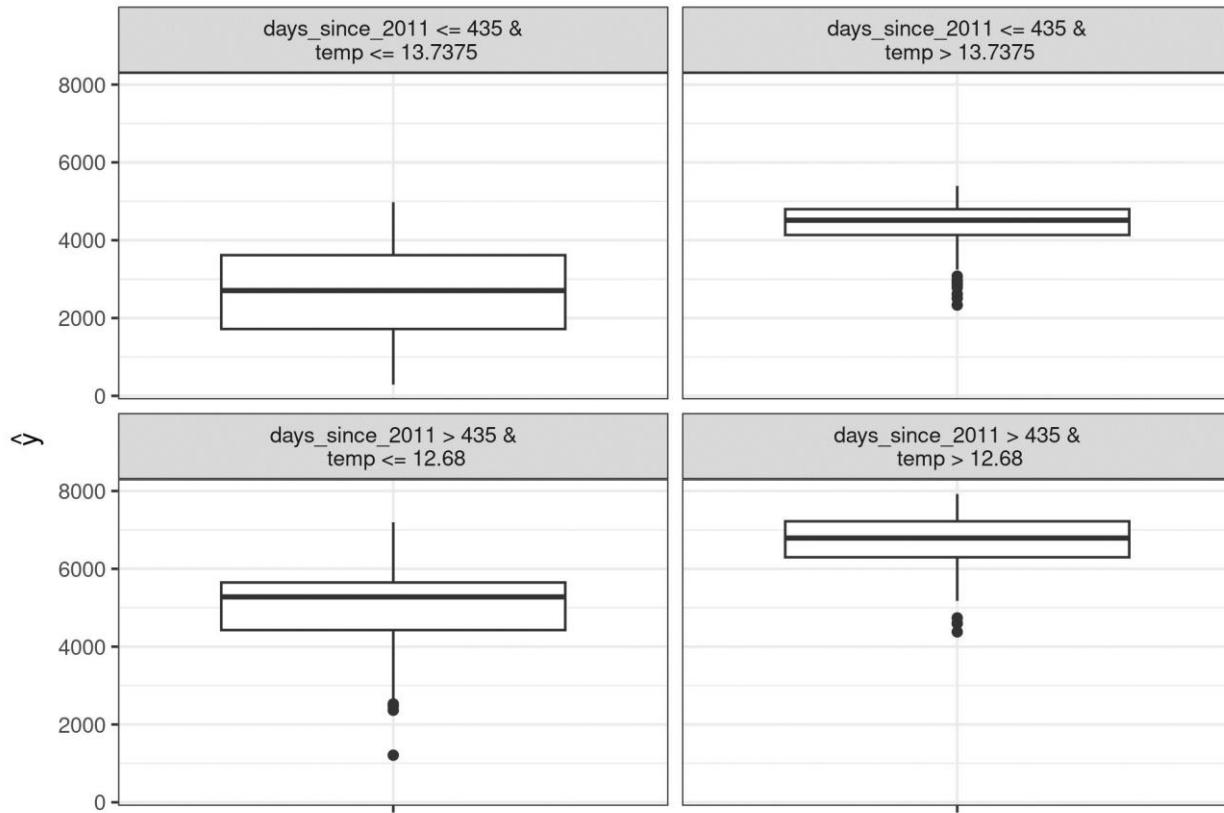
توجه داشته باشید که ما در مورد عملکرد مدل جعبه سیاه زیربنایی صحبت نکرده ایم، یعنی اینکه چقدر خوب یا بد در پیش بینی نتیجه واقعی عمل می کند. عملکرد مدل جعبه سیاه نقشی در آموزش مدل جانشین ندارد. تفسیر مدل جانشین همچنان معتبر است زیرا در مورد مدل اظهاراتی می کند نه در مورد دنیای واقعی. اما

۴۰۶ مسلماً اگر مدل جعبه سیاه بد باشد، تفسیر مدل جایگزین بی ربط می شود، زیرا در این صورت مدل جعبه سیاه
۴۰۷ خود بی ربط است.

۴۰۸ همچنین می‌توانیم یک مدل جایگزین براساس زیرمجموعه‌ای از داده‌های اصلی بسازیم یا نمونه‌ها را دوباره وزن
۴۰۹ کنیم. به این ترتیب، توزیع ورودی مدل جایگزین را تغییر می‌دهیم، که تمرکز تفسیر را تغییر می‌دهد (پس
۴۱۰ دیگر واقعاً جهانی نیست). اگر داده‌ها را به صورت محلی با یک نمونه خاص از داده‌ها وزن کنیم (هر چه نمونه
۴۱۱ ها به نمونه انتخاب شده نزدیکتر باشند، وزن آنها بیشتر است)، یک مدل جایگزین محلی دریافت می‌کنیم که
۴۱۲ می‌تواند پیش‌بینی فردی نمونه را توضیح دهد. در فصل بعدی در مورد مدل‌های محلی بیشتر بخوانید.

۴۱۳ **۸.۶.۲ مثال**
۴۱۴ برای نشان دادن مدل‌های جایگزین، یک رگرسیون و یک مثال طبقه‌بندی را در نظر می‌گیریم.

۴۱۵ ابتدا، ما یک ماشین بردار پشتیبان را آموزش می‌دهیم تا تعداد دوچرخه‌های اجاره شده روزانه را با توجه به
۴۱۶ اطلاعات آب و هوا و تقویم پیش‌بینی کند. ماشین بردار پشتیبان خیلی قابل تفسیر نیست، بنابراین ما یک
۴۱۷ جایگزین را با یک درخت تصمیم CART به عنوان مدل قابل تفسیر برای تقریب رفتار ماشین بردار پشتیبان
۴۱۸ آموزش می‌دهیم.

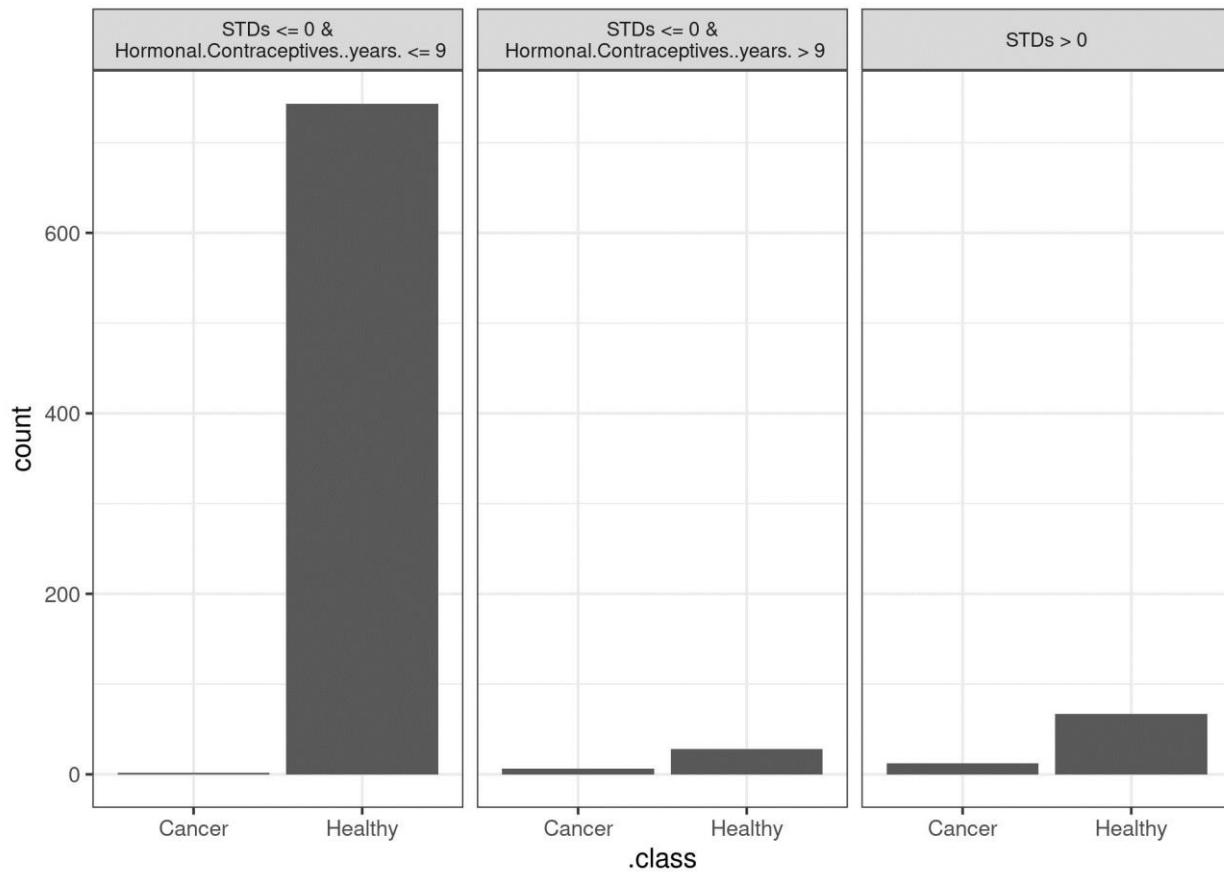


۴۲۱۹

شکل ۸،۲۸: گره های پایانی یک درخت جایگزین که پیش بینی های یک ماشین بردار پشتیبان آموزش دیده بر روی مجموعه داده های اجاره دوچرخه را تقریب می زند. توزیع ها در گره ها نشان می دهد که درخت جایگزین تعداد بیشتری از دوچرخه های اجاره ای را زمانی که دما بالای ۱۳ درجه سانتی گراد است و زمانی که روز بعد در دوره ۲ ساله بود (نقطه برش در ۴۳۵ روز) پیش بینی می کند.

مدل جایگزین دارای یک (R-squared) واریانس توضیح داده شده) ۰،۷۷ است که به این معنی است که رفتار جعبه سیاه زیرین را کاملاً خوب تقریب می کند، اما نه کاملاً. اگر تناسب کامل بود، می توانیم دستگاه بردار پشتیبانی را دور بیندازیم و به جای آن از درخت استفاده کنیم.

در مثال دوم، احتمال سرطان دهانه رحم را با یک جنگل تصادفی پیش بینی می کنیم. دوباره یک درخت تصمیم را با مجموعه داده اصلی آموزش می دهیم، اما با پیش بینی جنگل تصادفی به عنوان نتیجه، به جای کلاس های واقعی (سالم در مقابل سرطان) از داده ها.



شکل ۸،۲۹: گره های انتهایی یک درخت جایگزین که پیش بینی های یک جنگل تصادفی آموزش دیده بر روی مجموعه داده سرطان دهانه رحم را تقریب می زند. شمارش در گره ها فراوانی طبقه بندی مدل های جعبه سیاه را در گره ها نشان می دهد.

مدل جایگزین دارای یک) R-squared واریانس توضیح داده شده) ۰،۱۹ است، به این معنی که به خوبی جنگل تصادفی را تقریب نمی کند و ما نباید درخت را در هنگام نتیجه گیری در مورد مدل پیچیده تفسیر کنیم.

روش مدل جایگزین انعطاف پذیر است : هر مدلی از فصل مدل های قابل تفسیر قابل استفاده است. این همچنین به این معنی است که شما می توانید نه تنها مدل قابل تفسیر، بلکه مدل جعبه سیاه زیرین را نیز مبادله کنید. فرض کنید مدل پیچیده ای ایجاد کرده اید و آن را برای تیم های مختلف شرکت خود توضیح می دهید. یک تیم با مدل های خطی آشنا است، تیم دیگر می تواند درخت تصمیم را درک کند. شما می توانید دو مدل جایگزین (مدل خطی و درخت تصمیم) را برای مدل جعبه سیاه اصلی آموزش دهید و دو نوع توضیح ارائه دهید.

۴۲۴۲ اگر مدل جعبه سیاه با عملکرد بهتری پیدا کردید، لازم نیست روش تفسیر خود را تغییر دهید، زیرا می توانید از
۴۲۴۳ همان کلاس مدل های جایگزین استفاده کنید.

۴۲۴۴ من استدلال می کنم که رویکرد بسیار شهودی و سرراست است. این بدان معناست که پیادهسازی آن آسان
۴۲۴۵ است، اما توضیح آن برای افرادی که با علم داده یا یادگیری ماشین آشنایی ندارند نیز آسان است.

۴۲۴۶ با اندازه گیری مربع R^2 ، می توانیم به راحتی اندازه گیری کنیم که مدل های جایگزین ما در تقریب پیش بینی
۴۲۴۷ های جعبه سیاه چقدر خوب هستند.

۴۲۴۸ **۸.۶.۴ معايب**
۴۲۴۹ شما باید آگاه باشید که در مورد مدل نتیجه می گیرید نه در مورد داده ها ، زیرا مدل جایگزین هرگز نتیجه
۴۲۵۰ واقعی را نمی بیند.

۴۲۵۱ مشخص نیست بهترین برش برای R-squared چیست تا مطمئن شویم که مدل جایگزین به اندازه کافی به
۴۲۵۲ مدل جعبه سیاه نزدیک است. ۸۰ درصد واریانس توضیح داده شده است؟ ۵۰ درصد؟ ۹۹ درصد؟

۴۲۵۳ ما می توانیم اندازه گیری کنیم که مدل جانشین چقدر به مدل جعبه سیاه نزدیک است. بیایید فرض کنیم
خیلی نزدیک نیستیم، اما به اندازه کافی نزدیک هستیم. ممکن است این اتفاق بیفتد که مدل قابل تفسیر برای
۴۲۵۴ یک زیر مجموعه از مجموعه داده بسیار نزدیک باشد، اما برای زیرمجموعه دیگر کاملاً واگرا باشد . در این مورد،
۴۲۵۵ تفسیر مدل ساده برای همه نقاط داده به یک اندازه خوب نخواهد بود.
۴۲۵۶

۴۲۵۷ مدل قابل تفسیری که شما به عنوان جانشین انتخاب می کنید با تمام مزايا و معايب خود همراه است.

۴۲۵۸ برخی افراد استدلال می کنند که به طور کلی، هیچ مدل ذاتی قابل تفسیر (از جمله مدل های خطی و درخت
۴۲۵۹ های تصمیم گیری) وجود ندارد و حتی داشتن توهم تفسیرپذیری خطرناک است. اگر شما هم با این نظر موافق
۴۲۶۰ هستید، مطمئناً این روش برای شما مناسب نیست.

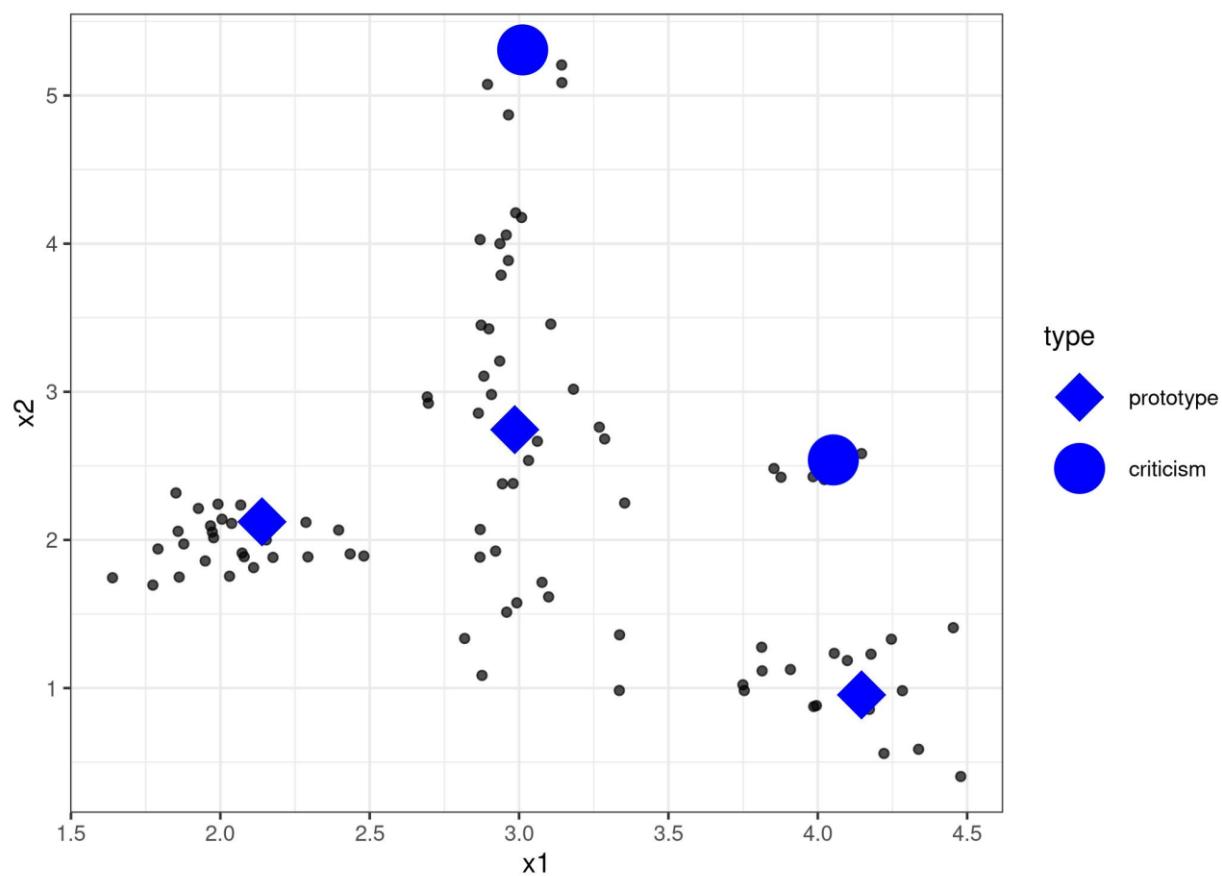
۴۲۶۱ **۸.۶.۵ نرم افزار**
۴۲۶۲ من از mlpack برای نمونه ها استفاده کردم. اگر می توانید یک مدل یادگیری ماشینی آموزش دهید، باید
۴۲۶۳ خودتان بتوانید مدل های جایگزین را پیادهسازی کنید. به سادگی یک مدل قابل تفسیر را برای پیش بینی پیش
۴۲۶۴ بینی های مدل جعبه سیاه آموزش دهید

۴۲۶۵

۸.۷ نمونه های اولیه و انتقادات

۴۲۶۶ نمونه اولیه یک نمونه داده است که نماینده همه داده ها است. انتقاد یک نمونه داده است که به خوبی توسط
 ۴۲۶۷ مجموعه نمونه های اولیه نمایش داده نمی شود. هدف از انتقاد، ارائه بینش همراه با نمونه های اولیه، به ویژه
 ۴۲۶۸ برای نقاط داده ای است که نمونه های اولیه به خوبی نشان نمی دهند. نمونه های اولیه و انتقادات را می توان
 ۴۲۶۹ به طور مستقل از یک مدل یادگیری ماشین برای توصیف داده ها استفاده کرد، اما همچنین می توان از آنها برای
 ۴۲۷۰ ایجاد یک مدل قابل تفسیر یا برای تفسیرپذیر ساختن مدل جعبه سیاه استفاده کرد.
 ۴۲۷۱

۴۲۷۲ در این فصل از عبارت «نقطه داده» برای اشاره به یک نمونه استفاده می کنم، تا بر این تفسیر تأکید کنم که یک
 ۴۲۷۳ نمونه همچنین نقطه ای در یک سیستم مختصات است که در آن هر ویژگی یک بعد است. شکل زیر توزیع داده
 ۴۲۷۴ های شبیه سازی شده را نشان می دهد که برخی از نمونه ها به عنوان نمونه اولیه و برخی به عنوان انتقاد
 ۴۲۷۵ انتخاب شده اند. نقاط کوچک داده ها، نقاط بزرگ انتقادات و مربع های بزرگ نمونه های اولیه هستند. نمونه
 ۴۲۷۶ های اولیه (به صورت دستی) برای پوشش مراکز توزیع داده ها انتخاب می شوند و انتقادات، نقاطی در یک
 ۴۲۷۷ خوش بدون نمونه اولیه هستند. نمونه های اولیه و انتقادات همیشه نمونه های واقعی از داده ها هستند.



۴۲۷۸

شکل ۸,۳۰: نمونه های اولیه و انتقادات برای توزیع داده با دو ویژگی X_1 و X_2

من نمونه های اولیه را به صورت دستی انتخاب کردم که مقیاس خوبی ندارد و احتمالاً منجر به نتایج ضعیف می شود. روش های زیادی برای یافتن نمونه های اولیه در داده ها وجود دارد. یکی از آنها k-medoids است، یک الگوریتم خوش بندی مرتبط با k-means. هر الگوریتم خوش بندی که نقاط داده واقعی را به عنوان مرکز خوش برمی گرداند، واجد شرایط انتخاب نمونه های اولیه است. اما اکثر این روش ها فقط نمونه های اولیه را پیدا می کنند، اما هیچ انتقادی ندارند. این فصل منتقد MMD توسط کیم و همکاران را ارائه می کند. (۲۰۱۶) ۴۶، رویکردی که نمونه های اولیه و انتقادات را در یک چارچوب واحد ترکیب می کند.

MMD-critic توزیع داده ها و توزیع نمونه های اولیه انتخاب شده را مقایسه می کند. این مفهوم اصلی برای درک روش انتقادی MMD-critic است. نمونه های اولیه ای را انتخاب می کند که اختلاف بین دو توزیع را به حداقل می رساند. نقاط داده در مناطق با تراکم بالا نمونه های اولیه خوبی هستند، به خصوص زمانی که نقاط از "خوش های داده" مختلف انتخاب می شوند. نقاط داده از مناطقی که به خوبی توسط نمونه های اولیه توضیح داده نشده اند به عنوان انتقاد انتخاب می شوند.

اجازه دهید عمیق تر به نظریه ببردازیم.

۸,۷,۱ نظریه

روش انتقادی MMD در سطح بالا را می توان به طور خلاصه خلاصه کرد:

- ۱- تعداد نمونه های اولیه و انتقاداتی را که می خواهید پیدا کنید انتخاب کنید.
- ۲- نمونه های اولیه را با جستجوی حریصانه پیدا کنید. نمونه های اولیه به گونه ای انتخاب می شوند که توزیع نمونه های اولیه به توزیع داده ها نزدیک باشد.
- ۳- انتقادات را با جستجوی حریصانه پیدا کنید. نقاطی به عنوان انتقاد انتخاب می شوند که در آن توزیع نمونه های اولیه با توزیع داده ها متفاوت است.

برای یافتن نمونه های اولیه و انتقادات برای مجموعه داده با MMD-critic، به چند عنصر نیاز داریم. به عنوان اساسی ترین عنصر، ما به یک تابع هسته برای تخمین چگالی داده ها نیاز داریم. هسته تابعی است که دو نقطه داده را با توجه به مجاورت آنها وزن می کند. بر اساس برآوردهای چگالی، ما به معیاری نیاز داریم که به ما بگوید دو توزیع چقدر متفاوت هستند تا بتوانیم تعیین کنیم که آیا توزیع نمونه های اولیه که انتخاب می کنیم به توزیع داده نزدیک است یا خیر. این با اندازه گیری حداکثر اختلاف میانگین (MMD) حل می شود.

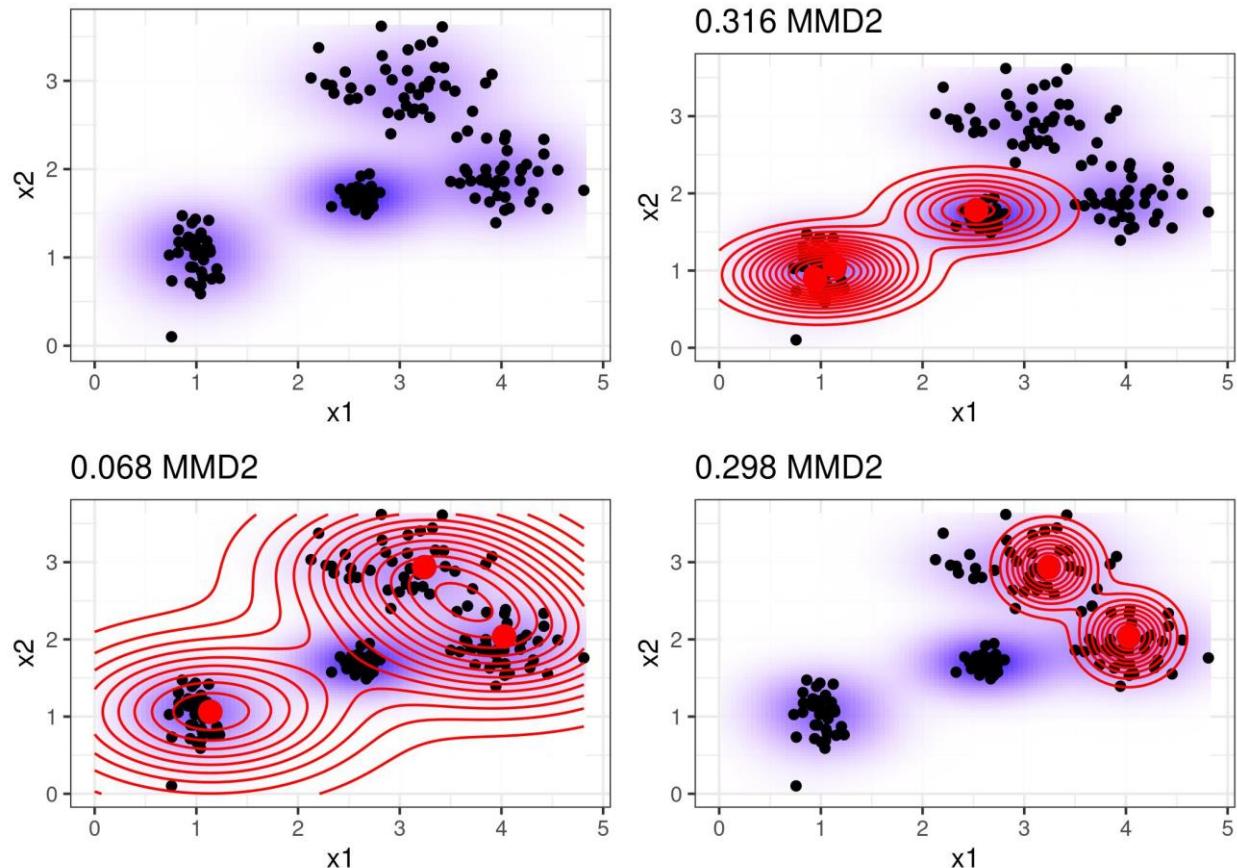
همچنین بر اساس تابع هسته، به تابع شاهد نیاز داریم تا به ما بگوید که دو توزیع در یک نقطه داده خاص چقدر متفاوت هستند. با تابع شاهد، می‌توانیم انتقادات را انتخاب کنیم، یعنی نقاط داده‌ای که در آن توزیع نمونه‌های اولیه و داده‌ها از هم جدا می‌شود و تابع شاهد مقادیر مطلق زیادی به خود می‌گیرد. آخرین عنصر یک استراتژی جستجو برای نمونه‌های اولیه و انتقادات خوب است که با یک جستجوی حریصانه ساده حل می‌شود.

اجازه دهید با حداکثر اختلاف میانگین (MMD) شروع کنیم ، که اختلاف بین دو توزیع را اندازه گیری می کند. انتخاب نمونه های اولیه، توزیع چگالی نمونه های اولیه را ایجاد می کند. ما می خواهیم ارزیابی کنیم که آیا توزیع نمونه اولیه با توزیع داده متفاوت است یا خیر. ما هر دو را با توابع چگالی هسته تخمین می زنیم. حداکثر اختلاف میانگین تفاوت بین دو توزیع را اندازه گیری می کند، که برتری بیش از یک فضای تابعی از تفاوت بین انتظارات بر اساس دو توزیع است. همه چیز روشن است؟ من شخصاً وقتی می بینم که چگونه چیزی با داده ها محاسبه می شود، این مفاهیم را خیلی بهتر درک می کنم. فرمول زیر نحوه محاسبه مجدد اندازه گیری (MMD) (MMD2) را نشان می دهد:

$$MMD^2 = \frac{1}{m^2} \sum_{i,j=1}^m k(z_i, z_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(z_i, x_j) + \frac{1}{n^2} \sum_{i,j=1}^n k(x_i, x_j)$$

یک تابع هسته است که شباهت دو نقطه را اندازه گیری می کند، اما در ادامه در مورد آن بیشتر توضیح خواهیم داد m . تعداد نمونه های اولیه Z و n تعداد نقاط داده X در مجموعه داده اصلی ما است. نمونه های اولیه Z مجموعه ای از نقاط داده X هستند. هر نقطه چند بعدی است، یعنی می تواند چندین ویژگی داشته باشد. هدف MMD-critic به حداقل رساندن MMD2 است. هرچه MMD2 به صفر نزدیکتر باشد، توزیع نمونه های اولیه بهتر با داده ها مطابقت دارد. کلید به صفر رساندن MMD2 عبارت وسطی است که میانگین نزدیکی بین نمونه های اولیه و سایر نقاط داده را محاسبه می کند (ضرب در ۲). اگر این عبارت با عبارت اول (میانگین نزدیکی نمونه های اولیه به یکدیگر) به اضافه جمله آخر جمع شود (میانگین نزدیکی داده ها به یکدیگر اشاره می کند)، نمونه های اولیه داده ها را کاملاً توضیح می دهند.

نمودار زیر اندازه گیری MMD2 را نشان می دهد. نمودار اول نقاط داده را با دو ویژگی نشان می دهد که به موجب آن تخمین چگالی داده ها با پس زمینه سایه دار نمایش داده می شود. هر یک از نمودارهای دیگر انتخاب های مختلفی از نمونه های اولیه را به همراه معیار MMD2 در عناوین طرح نشان می دهد. نمونه های اولیه نقاط بزرگ هستند و توزیع آنها به صورت خطوط کانتور نشان داده شده است. انتخاب نمونه های اولیه که به بهترین شکل داده ها را در این سناریوها پوشش می دهند (پایین سمت چپ) کمترین مقدار اختلاف را دارد.



شکل ۴.۳۱: مجدور اندازه گیری حداکثر میانگین اختلاف (MMD2) برای یک مجموعه داده با دو ویژگی و انتخاب های مختلف نمونه های اولیه.

یک انتخاب برای کرنل، هسته تابع پایه شعاعی است:

$$k(x, x') = \exp(-\gamma \|x - x'\|^2)$$

کجا $\|x - x'\|$ فاصله اقلیدسی بین دو نقطه و γ یک پارامتر مقیاس بندی است. ارزش هسته با فاصله بین دو نقطه کاهش می باید و بین صفر و یک قرار می گیرد: زمانی که دو نقطه بینهایت از هم فاصله دارند، صفر می شود. یکی زمانی که دو نقطه برابر باشند.

ما اندازه گیری MMD2، هسته و جستجوی حریصانه را در الگوریتمی برای یافتن نمونه های اولیه ترکیب می کنیم:

- با یک لیست خالی از نمونه های اولیه شروع کنید.

- در حالی که تعداد نمونه های اولیه کمتر از عدد انتخابی m است:

برای هر نقطه از مجموعه داده، بررسی کنید که با افزودن نقطه به لیست نمونه های اولیه، چقدر MMD2 کاهش می یابد. نقطه داده ای را که MMD2 را به حداقل می رساند به لیست اضافه کنید.

لیست نمونه های اولیه را برگردانید.

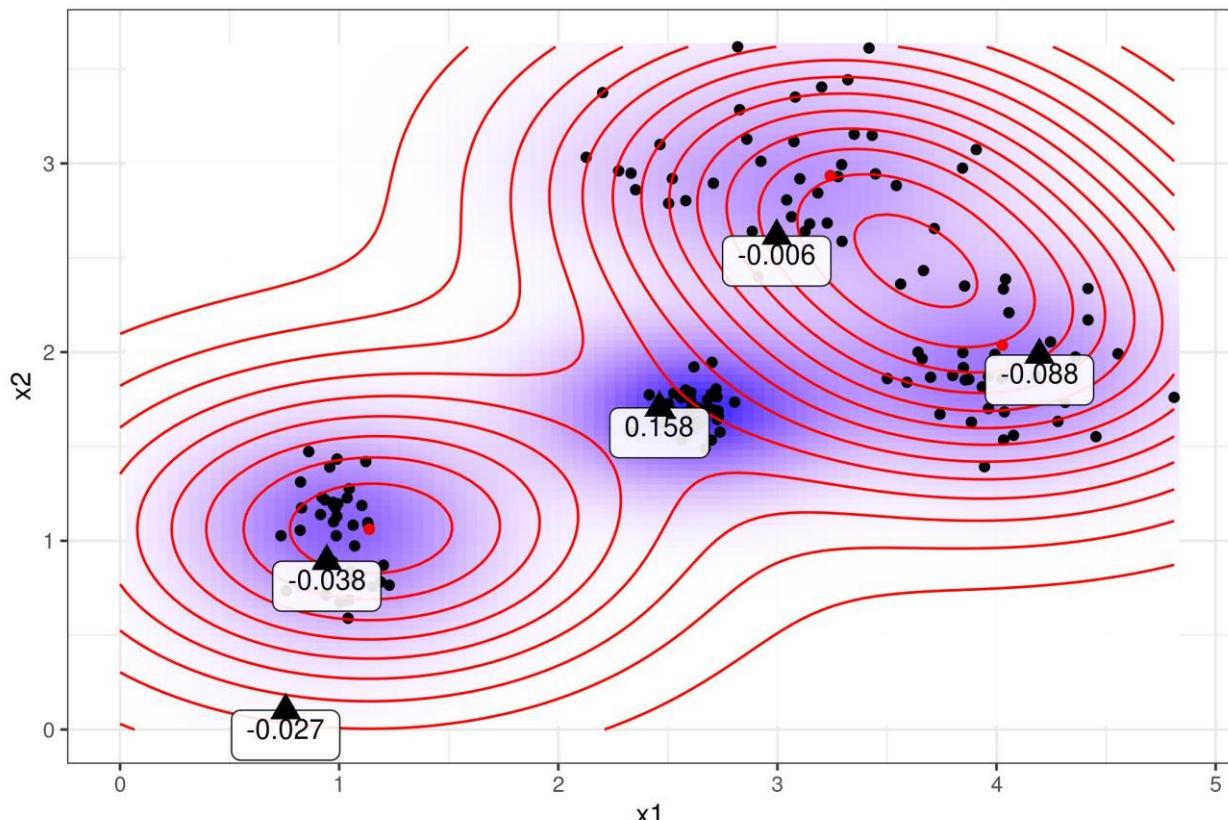
عنصر باقیمانده برای یافتن انتقادات تابع شاهد است که به ما می گوید چقدر دو تخمین چگالی در یک نقطه خاص تفاوت دارند. می توان با استفاده از:

$$witness(x) = \frac{1}{n} \sum_{i=1}^n k(x, x_i) - \frac{1}{m} \sum_{j=1}^m k(x, z_j)$$

برای دو مجموعه داده (با ویژگی های یکسان)، تابع شاهد به شما ابزاری برای ارزیابی اینکه در کدام توزیع تجربی نقطه X بهتر برازنده است، می دهد. برای یافتن انتقادات، ما به دنبال مقادیر افراطی عملکرد شاهد در دو جهت منفی و مثبت هستیم. عبارت اول در تابع شاهد میانگین نزدیکی بین نقطه X و داده ها و عبارت دوم به ترتیب میانگین نزدیکی بین نقطه X و نمونه های اولیه است. اگر تابع شاهد برای نقطه X نزدیک به صفر باشد، تابع چگالی داده ها و نمونه های اولیه به هم نزدیک هستند، به این معنی که توزیع نمونه های اولیه شبیه توزیع داده ها در نقطه X است. تابع شاهد منفی در نقطه X به این معنی است که توزیع نمونه اولیه توزیع داده ها را بیش از حد تخمین می زند (به عنوان مثال اگر نمونه اولیه را انتخاب کنیم اما تنها چند نقطه داده در این نزدیکی وجود دارد). تابع شاهد مثبت در نقطه X به این معنی است که توزیع نمونه اولیه توزیع داده را دست کم می گیرد (به عنوان مثال اگر نقاط داده زیادی در اطراف X وجود داشته باشد اما ما هیچ نمونه اولیه ای را در این نزدیکی انتخاب نکرده باشیم).

برای اینکه شهود بیشتری به شما بدهیم، اجازه دهید از نمونه های اولیه طرح با کمترین MMD2 مجدد استفاده کنیم و عملکرد شاهد را برای چند نقطه انتخاب شده به صورت دستی نمایش دهیم. برحسب های نمودار زیر مقدار تابع شاهد را برای نقاط مختلفی که به صورت مثلث مشخص شده اند نشان می دهد. فقط نقطه وسط ارزش مطلق بالایی دارد و بنابراین کاندیدای خوبی برای انتقاد است.

۰.۰۶۸ MMD2



۴۳۶۱

۴۳۶۲

شکل ۸,۳۲: ارزیابی عملکرد شاهد در نقاط مختلف.

تابع شاهد به ما این امکان را می دهد که به طور صریح نمونه های داده ای را جستجو کنیم که به خوبی توسط نمونه های اولیه نمایش داده نمی شوند. نقدها نکاتی با ارزش مطلق بالا در کارکرد شاهد هستند. مانند نمونه های اولیه، انتقادات نیز از طریق جستجوی حریصانه یافت می شود. اما به جای کاهش کلی MMD2 ، ما به دنبال نقاطی هستیم که یک تابع هزینه را که شامل تابع شاهد و یک اصطلاح تنظیم کننده است، به حداقل رساند. عبارت اضافی در تابع بهینه سازی، تنوع در نقاط را اعمال می کند، که لازم است تا نقاط از خوشه های مختلف آمده باشند.

این مرحله دوم مستقل از نحوه یافتن نمونه های اولیه است. همچنین می توانستم چند نمونه اولیه را انتخاب کنم و از روشی که در اینجا توضیح داده شد برای یادگیری انتقادات استفاده کنم. یا نمونه های اولیه می توانند از هر روش خوش بندی مانند k-medoids حاصل شوند.

این در مورد بخش های مهم نظریه انتقادی MMD است. یک سوال باقی می ماند: چگونه می توان از MMD-critic برای یادگیری ماشینی قابل تفسیر استفاده کرد؟

۴۳۷۴ MMD-critic می تواند به سه طریق قابلیت تفسیر را اضافه کند: با کمک به درک بهتر توزیع داده ها، با
۴۳۷۵ ساخت یک مدل قابل تفسیر؛ با ساختن یک مدل جعبه سیاه قابل تفسیر.

۴۳۷۶ اگر MMD-critic را روی داده های خود اعمال کنید تا نمونه های اولیه و انتقادات را بیابید، درک شما از داده ها
۴۳۷۷ را بهبود می بخشد، به خصوص اگر توزیع داده های پیچیده با موارد لبه داشته باشد. اما با MMD-critic می
۴۳۷۸ توانید به چیزهای بیشتری دست پیدا کنید!

۴۳۷۹ برای مثال، می توانید یک مدل پیش بینی قابل تفسیر ایجاد کنید: یک مدل به اصطلاح «نزدیک ترین نمونه
۴۳۸۰ اولیه».تابع پیش بینی به صورت زیر تعریف می شود:

$$\hat{f}(x) = \text{argmax}_{i \in S} k(x, x_i)$$

۴۳۸۱ به این معنی که نمونه اولیه A را از مجموعه نمونه های اولیه S انتخاب می کنیم که به نقطه داده جدید نزدیک
۴۳۸۲ است، به این معنا که بالاترین مقدار تابع هسته را به دست می دهد. خود نمونه اولیه به عنوان توضیحی برای
۴۳۸۳ پیش بینی بازگردانده می شود. این روش دارای سه پارامتر تنظیم است: نوع هسته، پارامتر مقیاس بندی هسته
۴۳۸۴ و تعداد نمونه های اولیه. تمام پارامترها را می توان در یک حلقه اعتبار سنجی متقاطع بهینه کرد. در این رویکرد
۴۳۸۵ از انتقادها استفاده نمی شود.

۴۳۸۷ به عنوان گزینه سوم، می توانیم از MMD-critic استفاده کنیم تا با بررسی نمونه های اولیه و انتقادات همراه با
۴۳۸۸ پیش بینی های مدل، هر مدل یادگیری ماشینی را در سطح جهانی قابل توضیح کنیم. روند کار به صورت زیر
۴۳۸۹ است:

۴۳۹۰ ۱- نمونه های اولیه و انتقادات را با MMD-critic بیابید.

۴۳۹۱ ۲- یک مدل یادگیری ماشینی را طبق معمول آموزش دهید.

۴۳۹۲ ۳- پیش بینی نتایج برای نمونه های اولیه و انتقادات با مدل یادگیری ماشین.

۴۳۹۳ ۴- تجزیه و تحلیل پیش بینی ها: در چه مواردی الگوریتم اشتباه بود؟ اکنون تعدادی مثال دارید که داده ها را به
۴۳۹۴ خوبی نشان می دهد و به شما کمک می کند تا نقاط ضعف مدل یادگیری ماشین را پیدا کنید.

۴۳۹۵ چگونه کمک می کند؟ زمانی را به خاطر دارید که طبقه بندی کننده تصویر گوگل، سیاه پستان را به عنوان
۴۳۹۶ گوریل شناسایی کرد؟ شاید آنها باید قبل از استقرار مدل تشخیص تصویر خود از روشی که در اینجا توضیح داده
۴۳۹۷ شده استفاده می کردند. فقط بررسی عملکرد مدل کافی نیست، زیرا اگر ۹۹٪ درست بود، این موضوع همچنان
۴۳۹۸ می تواند در ۱٪ باشد. و برچسب ها نیز ممکن است اشتباه باشند! بررسی همه داده های آموزشی و انجام یک

بررسی سلامت عقل در صورت مشکل‌ساز بودن پیش‌بینی ممکن است مشکل را آشکار کند، اما غیرممکن است.
اما انتخاب - مثلاً چند هزار - نمونه اولیه و انتقاد امکان پذیر است و می‌تواند مشکلی را در داده‌ها نشان دهد:
ممکن است نشان داده باشد که کمبود تصاویری از افراد با پوست تیره وجود دارد که نشان دهنده مشکل با تنوع
در مجموعه داده یا می‌توانست یک یا چند تصویر از یک فرد با پوست تیره را به عنوان نمونه اولیه یا (احتمالاً)
به عنوان انتقاد با طبقه بنده بدنام "گوریل" نشان دهد. من قول نمی‌دهم که منتقد MMD مطمئناً این نوع
اشتباهات را رهگیری کند، اما این یک بررسی عقلانی خوبی است.

۸,۷,۲ مثالها

مثال زیر از MMD-critic از یک مجموعه داده رقمی دست‌نویس استفاده می‌کند.
با نگاهی به نمونه‌های اولیه واقعی، ممکن است متوجه شوید که تعداد تصاویر در هر رقم متفاوت است. این به
این دلیل است که تعداد ثابتی از نمونه‌های اولیه در کل مجموعه داده جستجو شد و نه با تعداد ثابت در هر
کلاس. همانطور که انتظار می‌رفت، نمونه‌های اولیه روش‌های مختلفی را برای نوشتن ارقام نشان می‌دهند.



شکل ۸,۳۳: نمونه‌های اولیه برای مجموعه داده ارقام دست‌نویس.

۸,۷,۳ مزایا

در یک مطالعه کاربری، نویسنده‌گان MMD-critic تصاویری را به شرکت‌کنندگان دادند، که آنها باید به صورت بصری با یکی از دو مجموعه تصاویر مطابقت می‌دادند که هر کدام یکی از دو کلاس را نشان می‌داد (مثلاً دو نژاد سگ). شرکت کنندگان زمانی بهترین عملکرد را داشتند که مجموعه‌ها به جای تصاویر تصادفی یک کلاس، نمونه‌های اولیه و انتقادات را نشان دادند.

شما در انتخاب تعداد نمونه اولیه و انتقاد آزاد هستید.

MMD-critic با تخمین چگالی داده‌ها کار می‌کند. این با هر نوع داده و هر نوع مدل یادگیری ماشینی کار می‌کند.

پیاده سازی الگوریتم آسان است.

MMD-critic در روشهای افزایش تفسیرپذیری استفاده می‌شود بسیار انعطاف‌پذیر است. می‌توان از آن برای درک توزیع داده‌های پیچیده استفاده کرد. می‌توان از آن برای ساخت یک مدل یادگیری ماشینی قابل تفسیر استفاده کرد. یا می‌تواند تصمیم‌گیری در مورد مدل یادگیری ماشین جعبه سیاه را روشن کند.

یافتن انتقادات مستقل از فرآیند انتخاب نمونه‌های اولیه است. اما انتخاب نمونه‌های اولیه بر اساس MMD-critic منطقی است، زیرا در این صورت هم نمونه‌های اولیه و هم انتقادات با استفاده از روش مشابه مقایسه نمونه‌های اولیه و تراکم داده‌ها ایجاد می‌شوند.

۸,۷,۴ معایب

در حالی که، از نظر ریاضی، نمونه‌های اولیه و انتقادات به طور متفاوتی تعریف می‌شوند، تمایز آنها بر اساس یک مقدار برش (تعداد نمونه‌های اولیه) است. فرض کنید تعداد بسیار کمی از نمونه‌های اولیه را برای پوشش توزیع داده انتخاب کرده‌اید. انتقادات به حوزه‌هایی ختم می‌شود که به خوبی توضیح داده نشده‌اند. اما اگر بخواهید نمونه‌های اولیه بیشتری را اضافه کنید، آنها نیز در همان مناطق قرار می‌گیرند. هر تفسیری باید در نظر بگیرد که انتقادها به شدت به نمونه‌های اولیه موجود و مقدار قطعی (خودسرانه) تعداد نمونه‌های اولیه بستگی دارد.

شما باید تعداد نمونه اولیه و انتقادات را انتخاب کنید. به همان اندازه که این می‌تواند خوب باشد، یک نقطه ضعف نیز محسوب می‌شود. واقعاً به چند نمونه اولیه و انتقاد نیاز داریم؟ هرچی بیشتر بهتر؟ هر چه کمتر بهتر؟ یک راه حل این است که تعداد نمونه‌های اولیه و انتقادات را با اندازه گیری زمان برای کار نگاه کردن به تصاویر، که بستگی به کاربرد خاص دارد، انتخاب کنید. تنها زمانی که از MMD-critic برای ساختن یک طبقه بندی

کننده استفاده می کنیم، راهی برای بهینه سازی مستقیم آن داریم. یکی از راه حل ها می تواند نقشه های باشد که تعداد نمونه های اولیه را در محور X و اندازه گیری MMD_2 در محور Y نشان می دهد. ما تعداد نمونه های اولیه را انتخاب می کنیم که منحنی MMD_2 صاف می شود.

پارامترهای دیگر انتخاب کرنل و پارامتر مقیاس بندی هسته هستند. ما مشکل مشابهی با تعداد نمونه های اولیه و انتقادات داریم: چگونه یک هسته و پارامتر مقیاس پذیری آن را انتخاب کنیم؟ دوباره، وقتی از $MMD\text{-critic}$ به عنوان نزدیکترین طبقه بندی کننده نمونه اولیه استفاده می کنیم، می توانیم پارامترهای هسته را تنظیم کنیم. با این حال، برای موارد استفاده بدون نظرارت از $MMD\text{-critic}$ ، مشخص نیست. (شاید من در اینجا کمی خشن باشم، زیرا همه روش های بدون نظرارت این مشکل را دارند).

همه ویژگی ها را به عنوان ورودی می گیرد، بدون توجه به این واقعیت که برخی از ویژگی ها ممکن است برای پیش بینی نتیجه مورد علاقه مرتبط نباشند. یک راه حل این است که فقط از ویژگی های مرتبط استفاده کنید، برای مثال جاسازی تصویر به جای پیکسل های خام. این تا زمانی کار می کند که ما راهی برای نمایش نمونه اصلی بر روی نمایشی داشته باشیم که فقط حاوی اطلاعات مرتبط باشد.

تعدادی کد موجود است، اما هنوز به عنوان نرم افزار بسته بندی شده و مستند به خوبی پیاده سازی نشده است.

۸.۷.۵ کد و جایگزین
پیاده سازی $MMD\text{-critic}$ را می توان در مخزن GitHub نویسنده گان یافت.

اخيراً یک فرمت $MMD\text{-critic}$ توسعه داده شده است Protodash: نویسنده گان در انتشارات خود ادعای مزايايی نسبت به منتقدان MMD دارند. یک پیاده سازی Protodash در ابزار IBM AIX360 موجود است.

ساده ترین جایگزین برای يافتن نمونه های اولیه، k-medoids توسط کافمن و همکاران است. (۱۹۸۷). ۴۷.

فصل ۹ مدل محلی-روش های آگنوتیک

- ۴۵۷ روش های تفسیر محلی پیش بینی های فردی را توضیح می دهند. در این فصل، با روش های توضیح محلی زیر
۴۵۸ آشنا خواهید شد:
- ۴۵۹ منحنی های انتظار شرطی فردی، بلوک های سازنده نمودارهای وابستگی جزئی هستند و توضیح می دهند که
۴۶۰ چگونه تغییر یک ویژگی، پیش بینی را تغییر می دهد.
- ۴۶۱ مدل های جایگزین محلی (LIME) یک پیش بینی را با جایگزینی مدل پیچیده با یک مدل جایگزین قابل تفسیر
۴۶۲ محلی توضیح می دهند.
- ۴۶۳ قوانین محدوده (لنگرهای) قوانینی هستند که توصیف می کنند که کدام مقادیر ویژگی یک پیش بینی را ثابت
۴۶۴ می کنند، به این معنا که پیش بینی را در جای خود قفل می کنند.
- ۴۶۵ توضیحات خلاف واقع یک پیش بینی را با بررسی اینکه کدام ویژگی برای دستیابی به یک پیش بینی مطلوب
۴۶۶ نیاز به تغییر دارد، توضیح می دهد.
- ۴۶۷ مقادیر Shapley یک روش انتساب است که پیش بینی را نسبتاً به ویژگی های فردی اختصاص می دهد.
- ۴۶۸ SHAP یکی دیگر از روش های محاسباتی برای مقادیر Shapley است، اما همچنین روش های تفسیر کلی را
۴۶۹ بر اساس ترکیبی از مقادیر Shapley در میان داده ها پیشنهاد می کند.
- ۴۷۰ مقادیر LIME و Shapley روش های انتساب هستند، به طوری که پیش بینی یک نمونه واحد به عنوان
۴۷۱ مجموع اثرات ویژگی توصیف می شود. روش های دیگر، مانند توضیحات خلاف واقع، مبتنی بر مثال هستند
- ۴۷۲
- ۴۷۳

۱.انتظار شرطی فردی (ICE)

نماودارهای انتظار شرطی فردی (ICE) یک خط را برای هر نمونه نشان می دهد که نشان می دهد چگونه پیش بینی نمونه با تغییر یک ویژگی تغییر می کند.

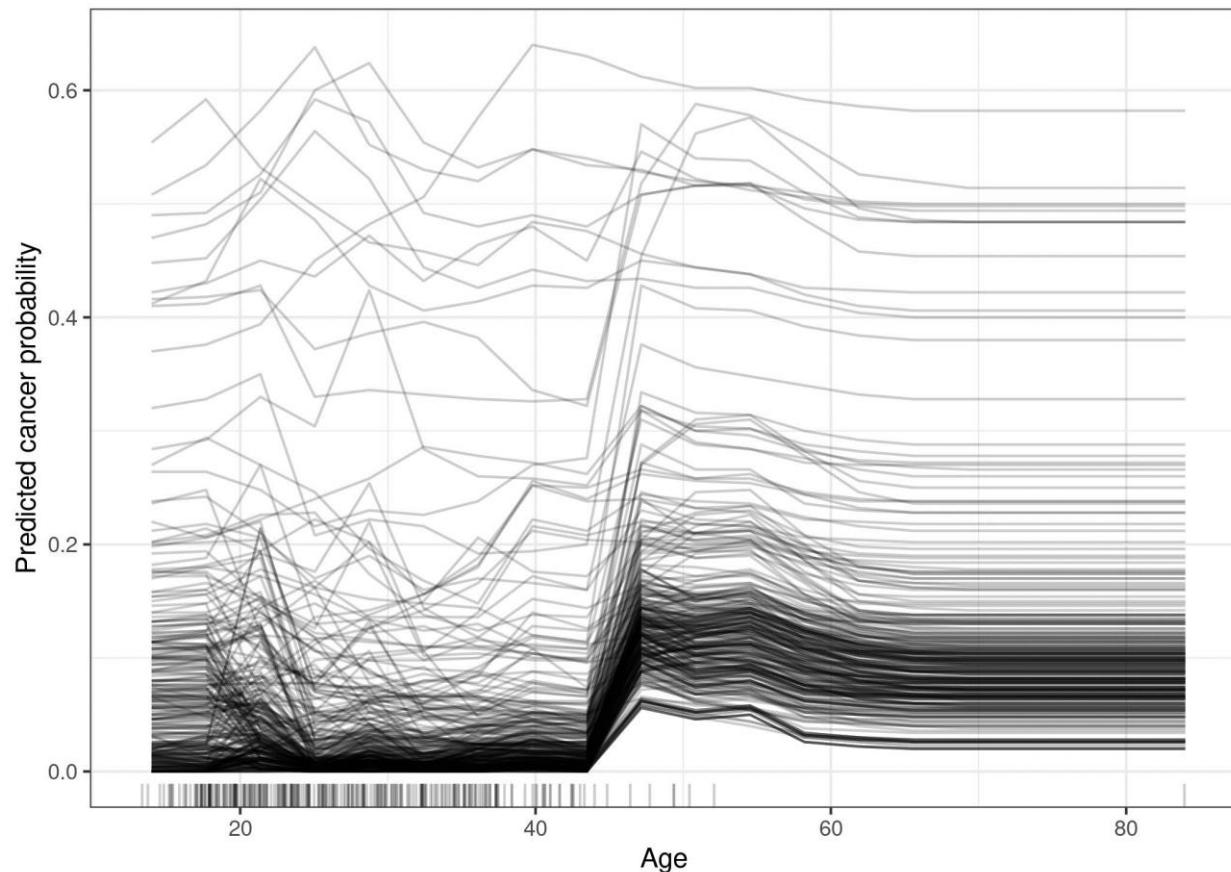
نمودار وابستگی جزئی برای اثر متوسط یک ویژگی یک روش جهانی است زیرا بر روی نمونه های خاص تمرکز نمی کند، بلکه بر میانگین کلی تمرکز می کند. معادل یک PDP برای نمونه های داده فردی، نماودار انتظار شرطی فردی (ICE) نامیده می شود. (Goldstein et al. 2017 48) نماودار ICE وابستگی پیش بینی به یک ویژگی را برای هر کدام به تصویر می کشد نمونه به طور جداگانه، منجر به یک خط در هر نمونه، در مقایسه با یک خط کلی در نماودارهای وابستگی جزئی PDP. میانگین خطوط یک نماودار ICE است. مقادیر یک خط (و یک نمونه) را می توان با ثابت نگه داشتن سایر ویژگی ها، ایجاد انواعی از این نمونه با جایگزینی مقدار ویژگی با مقادیر یک شبکه و پیش بینی با مدل جعبه سیاه برای این نمونه های جدید محاسبه کرد. نتیجه مجموعه ای از نقاط برای مثال با مقدار ویژگی از شبکه و پیش بینی های مربوطه است.

نگاه کردن به انتظارات فردی به جای وابستگی های جزئی چیست؟ نماودارهای وابستگی جزئی می توانند یک رابطه ناهمگن ایجاد شده توسط فعل و انفعالات را پنهان کنند PDP .ها می توانند به شما نشان دهنند که میانگین رابطه بین یک ویژگی و پیش بینی چگونه است. این تنها زمانی به خوبی کار می کند که تعامل بین ویژگی هایی که PDP برای آنها محاسبه می شود و سایر ویژگی ها ضعیف باشد. در صورت تعامل، طرح ICE بینش بسیار بیشتری را ارائه می دهد.

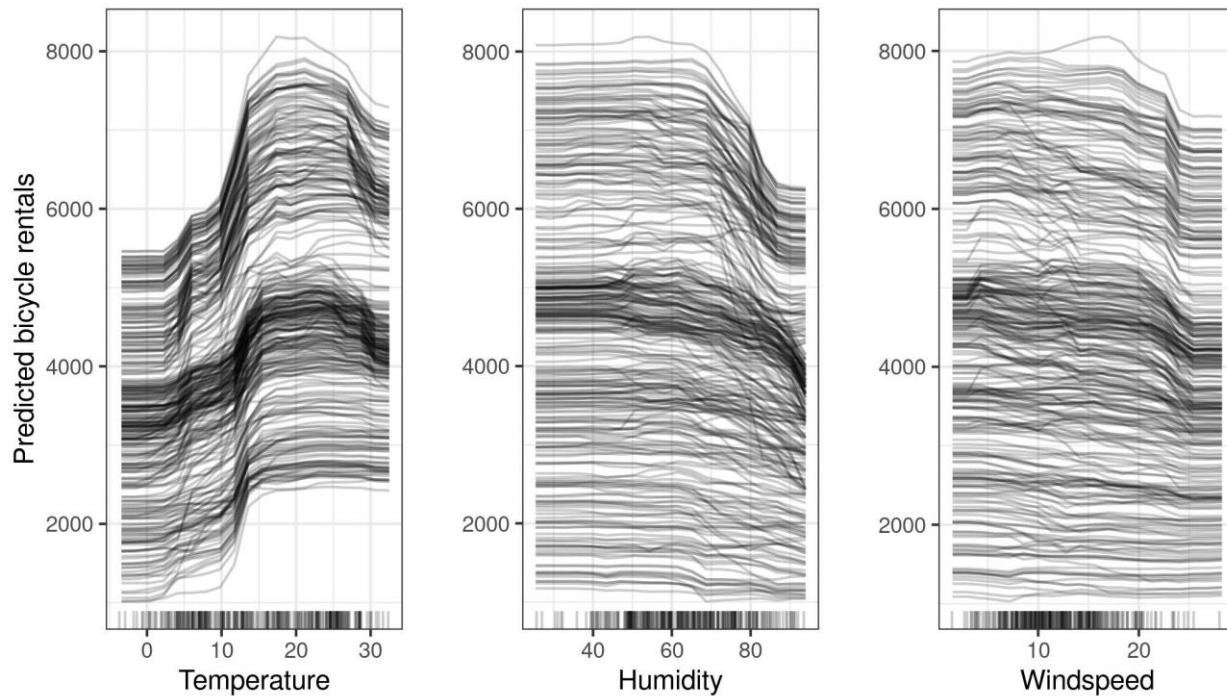
یک تعریف رسمی تر: در نماودارهای ICE ، برای هر نمونه در } ایکس(من (اس،ایکس(من (سی { (ن من 1= منحنی f من (اس علیه آن طرح ریزی شده است ایکس(من (اس ، در حالی که ایکس(من (سی ثابت باقی می ماند.

۹.۱.۱مثال ها

بیایید به مجموعه داده های سرطان دهانه رحم بازگردیم و ببینیم که چگونه پیش بینی هر نمونه با ویژگی "سن" مرتبط است. ما یک جنگل تصادفی را تجزیه و تحلیل خواهیم کرد که احتمال سرطان را برای یک زن با توجه به عوامل خطر پیش بینی می کند. در طرح وابستگی جزئی مشاهده کرده ایم که احتمال ابتلا به سرطان در حدود سن ۵۰ سالگی افزایش می باید، اما آیا این برای هر زن در مجموعه داده صادق است؟ نماودار ICE نشان می دهد که برای اکثر زنان اثر سن از الگوی متوسط افزایش در سن ۵۰ سالگی پیروی می کند، اما برخی استثناهای وجود دارد: برای محدود زنانی که احتمال پیش بینی شده بالایی در سنین جوانی دارند، احتمال سرطان پیش بینی شده تغییر نمی کند. بسیار با افزایش سن



- ۴۵۰۰ شکل ۹,۱: نمودار ICE احتمال سرطان دهانه رحم بر اساس سن. هر خط نشان دهنده یک زن است. برای اکثر زنان با افزایش سن احتمال سرطان پیش بینی شده افزایش می یابد. برای برخی از زنان با احتمال سرطان پیش بینی شده بالای ۰,۴، پیش بینی در سن بالاتر تغییر چندانی نمی کند.
- ۴۵۰۱
- ۴۵۰۲
- ۴۵۰۳
- ۴۵۰۴ شکل بعدی نمودارهای ICE را برای پیش بینی اجاره دوچرخه نشان می دهد . مدل پیش بینی زیربنایی یک
- ۴۵۰۵ جنگل تصادفی است.



شکل ۹.۲: نمودارهای ICE از اجاره دوچرخه پیش بینی شده بر اساس شرایط آب و هوایی. همان اثرات را می توان در نمودارهای وابستگی جزئی مشاهده کرد.

به نظر می رسد همه منحنی ها مسیر یکسانی را دنبال می کنند، بنابراین هیچ تعامل آشکاری وجود ندارد. این بدان معنی است که PDP در حال حاضر خلاصه خوبی از روابط بین ویژگی های نمایش داده شده و تعداد پیش بینی شده دوچرخه است.

۹.۱.۱.۱ طرح یخ مرکزی

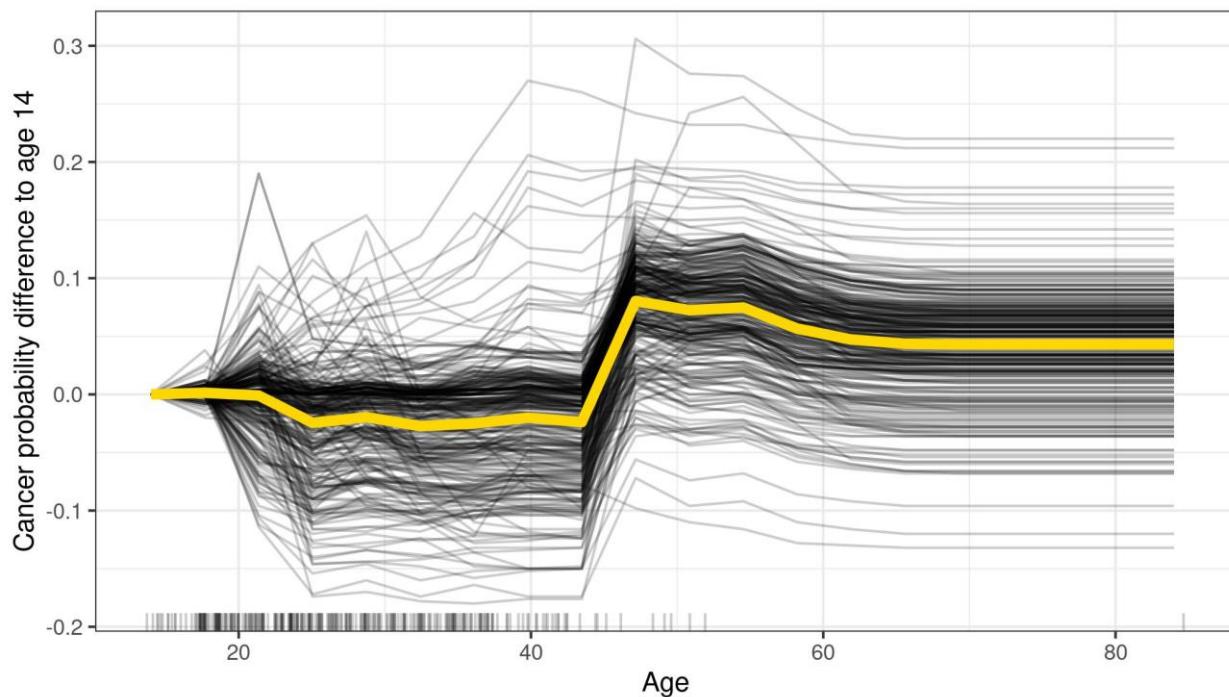
در نمودارهای ICE مشکلی وجود دارد: گاهی اوقات تشخیص اینکه آیا منحنی های ICE بین افراد متفاوت است یا خیر، دشوار است، زیرا آنها با پیش بینی های متفاوت شروع می شوند. یک راه حل ساده این است که منحنی ها را در نقطه خاصی از ویژگی متمرکز کنید و فقط تفاوت پیش بینی را تا این نقطه نشان دهید. نمودار حاصل را نمودار ICE مرکزی (c-ICE) می نامند. لنگر انداختن منحنی ها در انتهای پایین ویژگی انتخاب خوبی است. منحنی های جدید به صورت زیر تعریف می شوند:

$$\hat{f}_{cent}^{(i)} = \hat{f}^{(i)} - \mathbf{1}\hat{f}(x^a, x_C^{(i)})$$

بردار $\mathbf{1}$ با تعداد ابعاد مناسب (معمولًاً یک یا دو) است $\mathbf{1}^T$. مدل برآش شده و $a \times n$ نقطه لنگر است.

۹.۱.۱.۲ مثال

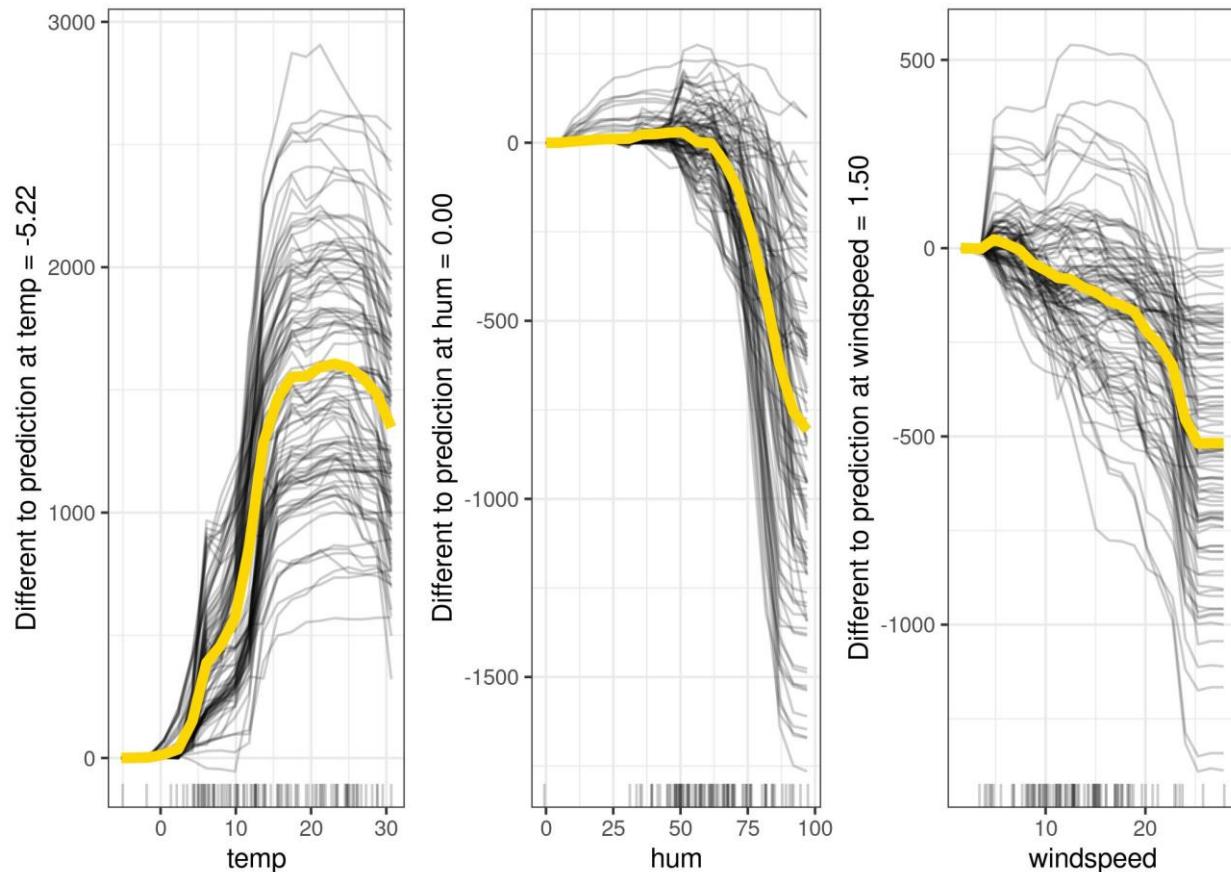
- ۴۵۲۰
- ۴۵۲۱ برای مثال، نمودار ICE سرطان دهانه رحم را برای سن در نظر بگیرید و خطوط روی جوانترین سن مشاهده شده متumer کز کنید:
- ۴۵۲۲



- ۴۵۲۳
- ۴۵۲۴ شکل ۹.۳: نمودار ICE مرکزی برای احتمال سرطان پیش بینی شده بر اساس سن. خطوط روی ۰ در سن ۱۴ سالگی ثابت می شوند. در مقایسه با سن ۱۴ سالگی، پیش بینی ها برای اکثر زنان تا سن ۴۵ سالگی بدون تغییر باقی می ماند، جایی که احتمال پیش بینی شده افزایش می یابد.
- ۴۵۲۵
- ۴۵۲۶

- ۴۵۲۷ نمودارهای ICE در مرکز، مقایسه منحنی های نمونه های جداگانه را آسان تر می کند. این می تواند مفید باشد
- ۴۵۲۸ اگر بخواهیم تغییر مطلق یک مقدار پیش بینی شده را مشاهده نکنیم، اما تفاوت در پیش بینی را در مقایسه با
- ۴۵۲۹ یک نقطه ثابت از محدوده ویژگی مشاهده کنیم.

- ۴۵۳۰ بیایید نگاهی به نمودارهای ICE متumer کز برای پیش بینی اجاره دوچرخه بیندازیم:



٤٥٣١

شکل ۹.۴: نمودارهای مرکز ICE تعداد پیش‌بینی شده دوچرخه‌ها بر اساس شرایط آب و هوایی. خطوط تفاوت در پیش‌بینی را در مقایسه با پیش‌بینی با مقدار ویژگی مربوطه در حداقل مشاهده شده نشان می‌دهد.

٤٥٣٤

۹.۱.۳ نمودار ICE مشتق

راه دیگر برای آسان‌تر کردن تشخیص ناهمگونی از نظر بصری، نگاه کردن به مشتق‌ات منفرد تابع پیش‌بینی با توجه به یک ویژگی است. نمودار حاصل، نمودار ICE مشتق (d-ICE) نامیده می‌شود. مشتق‌ات یک تابع (یا منحنی) به شما می‌گوید که آیا تغییرات رخ می‌دهند و در کدام جهت رخ می‌دهند. با نمودار ICE مشتق، به راحتی می‌توان محدوده‌هایی از مقادیر ویژگی را که در آن پیش‌بینی‌های جعبه سیاه برای (حداقل برخی) موارد تغییر می‌کند، تشخیص داد. اگر هیچ تعاملی بین ویژگی تحلیل شده وجود نداشته باشد ایکس اس و سایر ویژگی‌ها ایکس سی، سپس تابع پیش‌بینی را می‌توان به صورت زیر بیان کرد:

٤٥٤١

$$\hat{f}(x) = \hat{f}(x_S, x_C) = g(x_S) + h(x_C), \quad \text{with} \quad \frac{\delta \hat{f}(x)}{\delta x_S} = g'(x_S)$$

بدون فعل و انفعالات، مشتقات جزئی منفرد باید برای همه موارد یکسان باشند. اگر تفاوت دارند، به دلیل تعامل است و در نمودار d-ICE قابل مشاهده است. علاوه بر نمایش منحنی های منفرد برای مشتق تابع پیش بینی با توجه به ویژگی در S ، نشان دادن انحراف استاندارد مشتق به برجسته کردن مناطق در ویژگی در S با ناهمگنی در مشتق براورد شده کمک می کند. نمودار ICE مشتق برای محاسبه زمان زیادی طول می کشد و نسبتاً غیر عملی است.

۹.۱.۲ مزايا

منحنی های انتظار شرطی فردی حتی از نمودارهای وابستگی جزئی قابل درک تر هستند. اگر ویژگی مورد نظر را تغیير دهیم، یک خط پیش بینی ها را برای یک نمونه نشان می دهد.

برخلاف نمودارهای وابستگی جزئی، منحنی های ICE می توانند روابط ناهمگن را آشکار کنند.

۹.۱.۳ معایب

منحنی های ICE فقط می توانند یک ویژگی را به طور معنی دار نشان دهند، زیرا دو ویژگی به ترسیم چندین سطح همپوشانی نیاز دارند و شما چیزی در طرح نخواهید دید.

منحنی های ICE از همان مشکل PDP رنج می برند: اگر ویژگی مورد نظر با ویژگی های دیگر همبستگی داشته باشد، ممکن است برخی از نقاط در خطوط براساس توزیع ویژگی مشترک، نقاط داده نامعتبر باشند.

اگر منحنی های ICE زیادی رسم شود، طرح می تواند بیش از حد شلوغ شود و شما چیزی نخواهید دید. راه حل: یا مقداری شفافیت به خطوط اضافه کنید یا فقط نمونه ای از خطوط را بکشید.

در نمودارهای ICE ممکن است مشاهده میانگین آسان نباشد . این یک راه حل ساده دارد: منحنی های انتظار شرطی فردی را با نمودار وابستگی جزئی ترکیب کنید.

۹.۱.۴ نرم افزار و جایگزین

نمودارهای ICE در بسته های R (iml) و ICEbox49 (pdp) یکی دیگر از بسته های R که کاری بسیار شبیه به ICE انجام می دهد. condvis در پایتون، طرح های وابستگی جزئی در یادگیری scikit با نسخه ۰,۴۰ ساخته شده اند.

۹.۲ جایگزین محلی (LIME)

۴۵۶۵ مدل‌های جایگزین محلی مدل‌های قابل تفسیری هستند که برای توضیح پیش‌بینی‌های فردی مدل‌های
 ۴۵۶۶ یادگیری ماشین جعبه سیاه استفاده می‌شوند. توضیحات مدل قابل تفسیر محلی ۵۰ (LIME) مقاله‌ای است که
 ۴۵۶۷ در آن نویسنده‌گان اجرای ملموسی از مدل‌های جایگزین محلی را پیشنهاد می‌کنند. مدل‌های جایگزین برای
 ۴۵۶۸ تقریبی پیش‌بینی‌های مدل جعبه سیاه زیربنایی آموزش دیده‌اند. به جای آموزش یک مدل جانشین جهانی،
 ۴۵۶۹ LIME ۴۵۷۰ بر آموزش مدل‌های جایگزین محلی برای توضیح پیش‌بینی‌های فردی تمرکز می‌کند.

۴۵۷۱ ایده کاملاً شهودی است. ابتدا داده‌های آموزشی را فراموش کنید و تصور کنید که فقط مدل جعبه سیاه را
 ۴۵۷۲ دارید که می‌توانید نقاط داده را وارد کنید و پیش‌بینی‌های مدل را بدست آورید. می‌توانید هر چند وقت
 ۴۵۷۳ یکبار که می‌خواهید جعبه را بررسی کنید. هدف شما درک این موضوع است که چرا مدل یادگیری ماشین
 ۴۵۷۴ پیش‌بینی خاصی انجام داده است LIME. آزمایش می‌کند که وقتی تغییراتی از داده‌های خود را در مدل
 ۴۵۷۵ یادگیری ماشین می‌دهید، چه اتفاقی برای پیش‌بینی‌ها می‌افتد LIME. یک مجموعه داده جدید متشكل از
 ۴۵۷۶ نمونه‌های آشفته و پیش‌بینی‌های مربوط به مدل جعبه سیاه تولید می‌کند. در این مجموعه داده جدید
 ۴۵۷۷ LIME سپس یک مدل قابل تفسیر را آموزش می‌دهد که با نزدیکی نمونه‌های نمونه‌برداری شده به نمونه مورد
 ۴۵۷۸ نظر وزن می‌شود. مدل قابل تفسیر می‌تواند هر چیزی از فصل مدل‌های قابل تفسیر باشد، برای مثال Lasso یا
 ۴۵۷۹ درخت تصمیم. مدل آموخته شده باید تقریب خوبی از پیش‌بینی‌های مدل یادگیری ماشین به صورت محلی
 ۴۵۸۰ باشد، اما لزومی ندارد که یک تقریب جهانی خوب باشد. به این نوع دقت، وفاداری محلی نیز می‌گویند.

۴۵۸۱ از نظر ریاضی، مدل‌های جایگزین محلی با محدودیت تفسیرپذیری را می‌توان به صورت زیر بیان کرد:

$$\text{explanation}(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

۴۵۸۲ مدل توضیحی برای مثال X (مدل) g مثلاً مدل رگرسیون خطی است که تلفات L را به حداقل می‌رساند (مثلاً
 ۴۵۸۳ میانگین مربعات خط)، که میزان نزدیک بودن توضیح را به پیش‌بینی مدل اصلی (f) مثلاً یک مدل
 ۴۵۸۴ xgboost (مثلاً یک مدل $\Omega(g)$ پایین نگه داشته می‌شود (مثلاً ویژگی‌های کمتری را ترجیح می‌دهند
 ۴۵۸۵ G). (خانواده توضیحات ممکن است، برای مثال تمام مدل‌های رگرسیون خطی ممکن. اندازه گیری
 ۴۵۸۶ مجاورت π ایکس تعیین می‌کند که همسایگی اطراف مثال X چقدر است که برای توضیح در نظر می‌گیریم. در
 ۴۵۸۷ عمل، LIME فقط بخش تلفات را بهینه می‌کند. کاربر باید پیچیدگی را تعیین کند، به عنوان مثال با انتخاب
 ۴۵۸۸ حداقل تعداد ویژگی‌هایی که مدل رگرسیون خطی ممکن است استفاده کند.

۴۵۸۹ ۴۵۹۰ دستور العمل برای آموزش مدل‌های جایگزین محلی:

۴۵۹۱ نمونه مورد علاقه خود را که می خواهید توضیحی درباره پیش بینی جعبه سیاه آن داشته باشد را انتخاب
۴۵۹۲ کنید.

۴۵۹۳ داده های خود را مختل کنید و پیش بینی های جعبه سیاه را برای این نقاط جدید دریافت کنید.

۴۵۹۴ نمونه های جدید را با توجه به نزدیکی آنها به نمونه مورد نظر وزن کنید.

۴۵۹۵ یک مدل وزن دار و قابل تفسیر روی مجموعه داده با تغییرات آموزش دهید.

۴۵۹۶ پیش بینی را با تفسیر مدل محلی توضیح دهید.

۴۵۹۷ به عنوان مثال، در پیاده سازی های فعلی در R و Python ، رگرسیون خطی را می توان به عنوان مدل
۴۵۹۸ جایگزین قابل تفسیر انتخاب کرد. از قبل، باید K را انتخاب کنید، تعداد ویژگی هایی که می خواهید در مدل
۴۵۹۹ قابل تفسیر خود داشته باشید. هرچه K کمتر باشد، تفسیر مدل آسان تر است K . بالاتر به طور بالقوه مدل هایی
۴۶۰۰ با وفاداری بالاتر تولید می کند. روش های مختلفی برای آموزش مدل هایی با ویژگی های دقیقا K وجود دارد.
۴۶۰۱ یک انتخاب خوب کمند است . مدل Lasso با پارامتر تنظیم بالا λ مدلی بدون هیچ ویژگی به دست می دهد.
۴۶۰۲ با بازآموزی مدل های کمند با کاهش تدریجی λ ، یکی پس از دیگری، ویژگی ها تخمین وزنی را دریافت می
۴۶۰۳ کنند که با صفر متفاوت است. اگر K ویژگی در مدل وجود داشته باشد، به تعداد ویژگی های مورد نظر رسیده
۴۶۰۴ اید. راهبردهای دیگر انتخاب ویژگی ها به جلو یا عقب است. این بدان معناست که شما یا با مدل کامل (= شامل
۴۶۰۵ همه ویژگی ها) یا با مدلی که فقط وقفه دارد شروع کنید و سپس تست کنید که کدام ویژگی با اضافه یا حذف
۴۶۰۶ بیشترین پیشرفت را به همراه خواهد داشت تا زمانی که به مدلی با ویژگی های K برسید.

۴۶۰۷ چگونه تغییرات داده ها را بدست می آورید؟ این بستگی به نوع داده دارد که می تواند متن، تصویر یا داده های
۴۶۰۸ جدولی باشد. برای متن و تصاویر، راه حل این است که تک کلمات یا سوپرپیکسل ها را روشن یا خاموش کنید.
۴۶۰۹ در مورد داده های جدولی، LIME با ایجاد اختلال در هر ویژگی به صورت جداگانه، نمونه های جدیدی را ایجاد
۴۶۱۰ می کند و از یک توزیع نرمال با میانگین و انحراف استاندارد گرفته شده از ویژگی استخراج می شود.

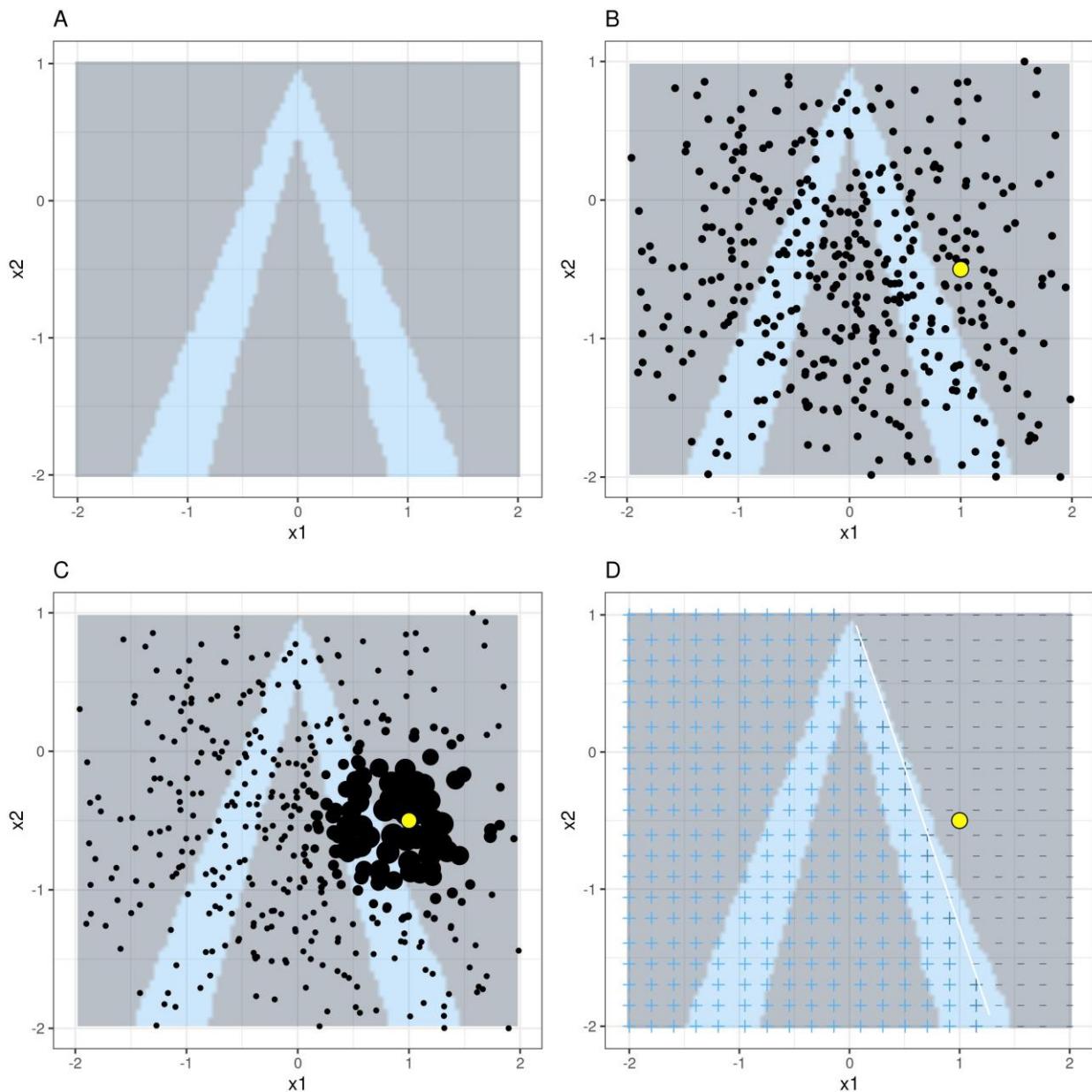
۴۶۱۱ ۹,۲,۱ برای داده های جدولی LIME

۴۶۱۲ داده های جدولی داده هایی هستند که در جداول قرار می گیرند و هر ردیف نشان دهنده یک نمونه و هر ستون
۴۶۱۳ یک ویژگی است. نمونه های LIME در اطراف نمونه مورد نظر گرفته نمی شوند، بلکه از مرکز انبوه داده های
۴۶۱۴ آموزشی گرفته می شوند، که مشکل ساز است. اما این احتمال را افزایش می دهد که نتیجه برخی از
۴۶۱۵ پیش بینی های نقاط نمونه با نقطه داده مورد علاقه متفاوت است و LIME می تواند حداقل توضیحی را بیاموزد.

۴۶۱۶

۴۶۱۷

بهتر است به صورت تصویری توضیح دهید که چگونه نمونه گیری و آموزش مدل محلی کار می کند:



۴۶۱۸

۴۶۱۹

شکل ۹.۵: الگوریتم LIME برای داده های جدولی. الف) پیش‌بینی‌های تصادفی جنگل با ویژگی‌های x_1 و x_2 .

۴۶۲۰

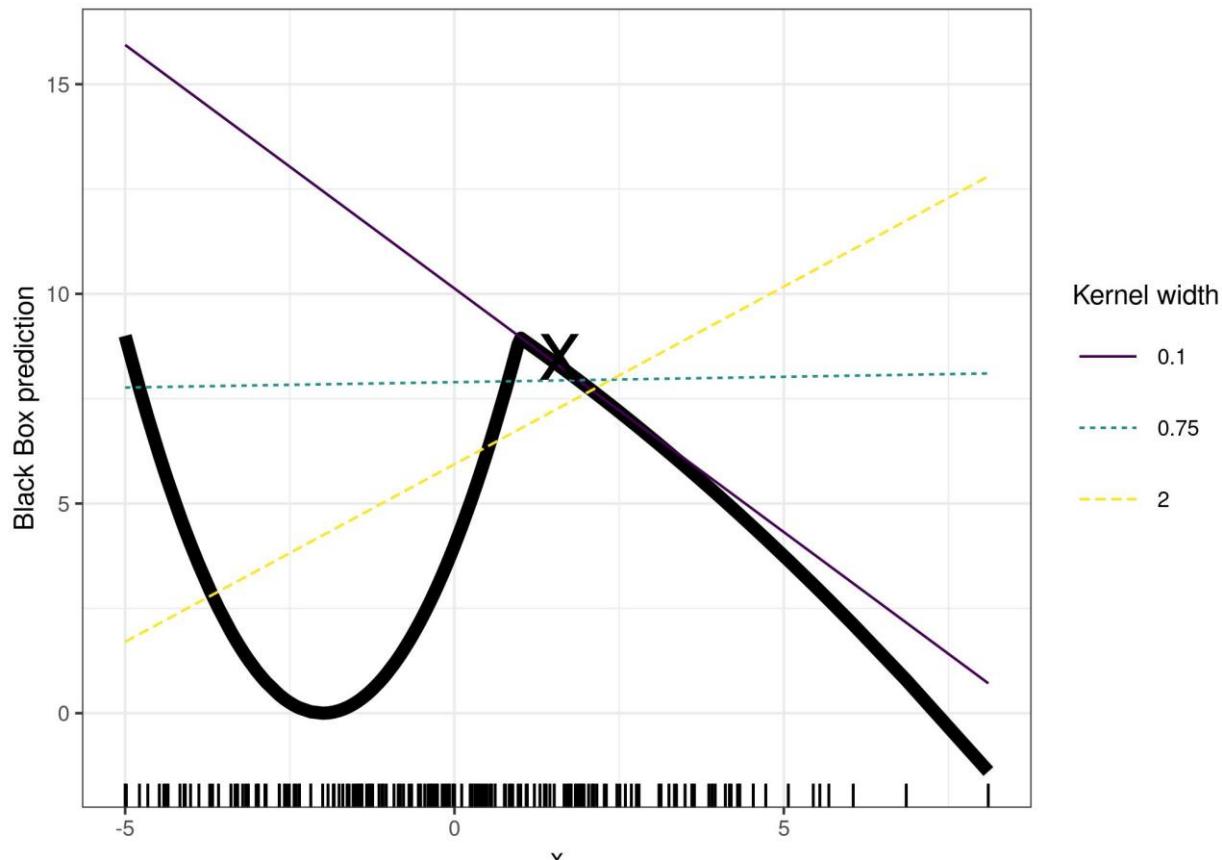
کلاس های پیش بینی شده: ۱ (تاریک) یا ۰ (روشن). ب) نمونه مورد علاقه (نقطه بزرگ) و داده های نمونه

۴۶۲۱

برداری شده از یک توزیع نرمال (نقاط کوچک). ج) وزن بیشتری را به نقاط نزدیک به نمونه مورد نظر اختصاص

۴۶۲۲ دهید. د) علائم شبکه طبقه بندی مدل محلی آموخته شده از نمونه های وزنی را نشان می دهد. خط سفید مرز
۴۶۲۳ تصمیم را مشخص می کند. ($P(class=1) = 0.5$)

۴۶۲۴ مثل همیشه، شیطان در جزئیات است. تعریف یک محله معنادار در اطراف یک نقطه دشوار است LIME. در
۴۶۲۵ حال حاضر از یک هسته هموارسازی نمایی برای تعریف همسایگی استفاده می کند. یک هسته هموارسازی
۴۶۲۶ تابعی است که دو نمونه داده را می گیرد و یک اندازه گیری مجاورت را برمی گرداند. عرض هسته تعیین می کند
۴۶۲۷ که همسایگی چقدر بزرگ است: عرض هسته کوچک به این معنی است که یک نمونه باید بسیار نزدیک باشد تا
۴۶۲۸ بر مدل محلی تأثیر بگذارد، عرض هسته بزرگتر به این معنی است که نمونه هایی که دورتر هستند نیز روی مدل
۴۶۲۹ تأثیر می گذارند. اگر به پیاده سازی پایتون LIME نگاه کنید (فایل lime/lime_tabular.py) خواهید دید که
۴۶۳۰ از یک هسته هموارسازی نمایی (بر روی داده های نرمال شده) استفاده می کند و عرض هسته ۰,۷۵ برابر ریشه
۴۶۳۱ دوم تعداد ستون های داده های آموزشی است. به نظر می رسد یک خط کد بی گناه است، اما مانند یک فیل
۴۶۳۲ است که در اتاق نشیمن شما در کنار ظروف چینی خوبی که از پدربزرگ و مادربزرگتان گرفته اید، نشسته است.
۴۶۳۳ مشکل بزرگ این است که ما راه خوبی برای یافتن بهترین هسته یا عرض نداریم. و ۰,۷۵ حتی از کجا می آید؟
۴۶۳۴ در سناریوهای خاصی، همانطور که در شکل زیر نشان داده شده است، می توانید به راحتی با تغییر عرض هسته،
۴۶۳۵ توضیح خود را تغییر دهید:



۴۶۳۶

۴۶۳۷

۴۶۳۸

۴۶۳۹

۴۶۴۰

شکل ۹.۶: توضیح پیش بینی نمونه $x = 1.6$. پیش بینی های مدل جعبه سیاه بسته به یک ویژگی منفرد به صورت یک خط ضخیم و توزیع داده ها با قالیچه ها نشان داده می شود. سه مدل جایگزین محلی با عرض هسته های مختلف محاسبه شده است. مدل رگرسیون خطی حاصل به عرض هسته بستگی دارد: آیا این ویژگی برای $x = 1.6$ تأثیر منفی، مثبت یا بدون تأثیر دارد؟

۴۶۴۱

۴۶۴۲

۴۶۴۳

۴۶۴۴

مثال فقط یک ویژگی را نشان می دهد. در فضاهای با ویژگی با ابعاد بالا بدتر می شود. همچنین بسیار نامشخص است که آیا اندازه‌گیری فاصله باید همه ویژگی‌ها را یکسان در نظر بگیرد. آیا یک واحد فاصله برای ویژگی x_1 با یک واحد برای ویژگی x_2 یکسان است؟ اندازه‌گیری‌های فاصله کاملاً دلخواه هستند و فاصله‌ها در ابعاد مختلف (معروف به ویژگی‌های) ممکن است اصلاً قابل مقایسه نباشند.

۴۶۴۵

۹.۲.۱.۱ مثال

۴۶۴۶

۴۶۴۷

۴۶۴۸

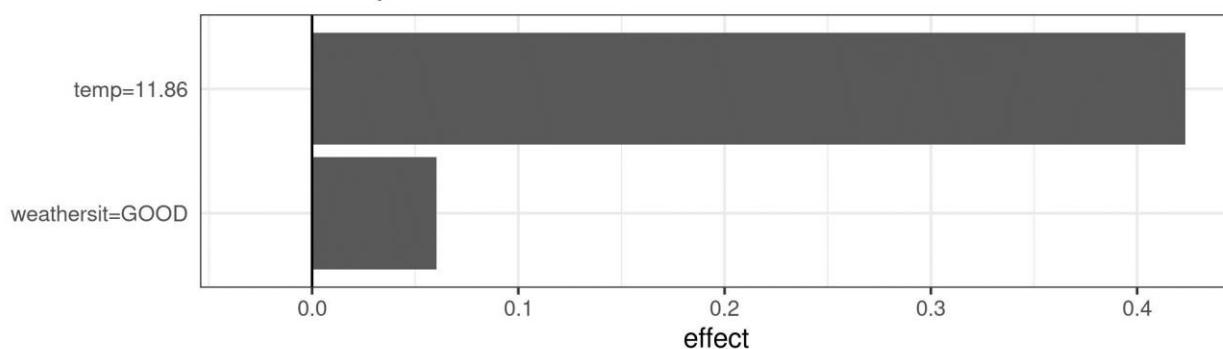
اجازه دهید به یک مثال عینی نگاه کنیم. به داده‌های اجاره دوچرخه برمی‌گردیم و مسئله پیش‌بینی را به یک طبقه‌بندی تبدیل می‌کنیم: پس از در نظر گرفتن روندی که کرایه دوچرخه در طول زمان محبوب‌تر شده است، می‌خواهیم در یک روز مشخص بدانیم که آیا تعداد دوچرخه‌های اجاره‌شده خواهد بود یا خیر. بالا یا زیر خط

روند. همچنین می‌توانید «بالا» را به عنوان بالاتر از میانگین تعداد دوچرخه‌ها تعبیر کنید، اما برای روند تنظیم شده است.

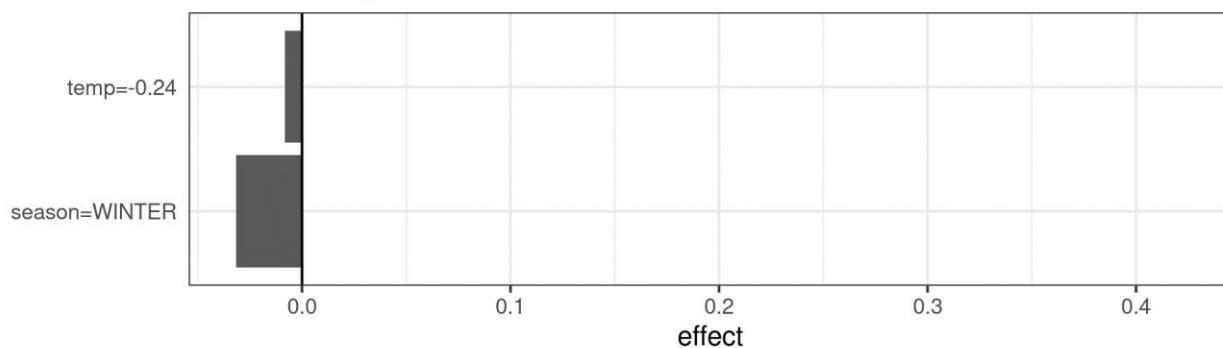
ابتدا یک جنگل تصادفی با ۱۰۰ درخت را در کار طبقه بندی آموزش می‌دهیم. بر اساس اطلاعات آب و هوا و تقویم، در چه روزی تعداد دوچرخه‌های اجاره‌ای بالاتر از میانگین بدون روند خواهد بود؟

توضیحات با ۲ ویژگی ایجاد شده است. نتایج مدل‌های خطی محلی پراکنده برای دو نمونه با کلاس‌های پیش‌بینی شده متفاوت آموزش داده شد:

Actual prediction: 0.89
LocalModel prediction: 0.44



Actual prediction: 0.01
LocalModel prediction: -0.03



شکل ۹.۷: توضیحات LIME برای دو نمونه از مجموعه داده‌های اجاره دوچرخه. دمای گرمتر و وضعیت آب و هوای خوب تأثیر مثبتی بر پیش‌بینی دارد. محور X اثر ویژگی را نشان می‌دهد: وزن ضربدر مقدار واقعی ویژگی.

از شکل مشخص می‌شود که تفسیر ویژگی‌های طبقه بندی آسان‌تر از ویژگی‌های عددی است. یک راه حل این است که ویژگی‌های عددی را به bin‌ها دسته بندی کنیم.

LIME برای متن ۹,۲,۲

برای متن با LIME برای داده های جدولی متفاوت است. تغییرات داده ها به طور متفاوتی تولید می شوند: با شروع از متن اصلی، متون جدید با حذف تصادفی کلمات از متن اصلی ایجاد می شوند. مجموعه داده با ویژگی های باینری برای هر کلمه نشان داده می شود. یک ویژگی اگر کلمه مربوطه گنجانده شود ۱ و اگر حذف شده باشد ۰ است.

۹,۲,۲,۱ مثال

در این مثال ما نظرات YouTube را به عنوان هرزنامه یا عادی طبقه بندی می کنیم.

مدل جعبه سیاه یک درخت تصمیم عمیق است که بر روی ماتریس کلمه سند آموزش داده شده است. هر نظر یک سند (= یک ردیف) و هر ستون تعداد تکرار یک کلمه داده شده است. درخت های تصمیم گیری کوتاه به راحتی قابل درک هستند، اما در این مورد درخت بسیار عمیق است. همچنین به جای این درخت می توانست یک شبکه عصبی مکرر یا یک ماشین بردار پشتیبان آموزش داده بر روی جاسازی کلمات (بردارهای انتزاعی) وجود داشته باشد. اجازه دهید به دو نظر این مجموعه داده و کلاس های مربوطه نگاه کنیم (۱ برای هرزنامه، ۰ برای نظر عادی):

	CONTENT	CLASS
267	PSY is a good guy	0
173	For Christmas Song visit my channell ;)	1

گام بعدی ایجاد برخی تغییرات از مجموعه داده های مورد استفاده در یک مدل محلی است. به عنوان مثال، برخی از تغییرات یکی از نظرات:

For	Christmas	Song	visit	my	channell	;)	prob	weight
1	0	1	1	0	0	1	0.17	0.57
0	1	1	1	1	0	1	0.17	0.71
1	0	0	1	1	1	1	0.99	0.71
1	0	1	1	1	1	1	0.99	0.86
0	1	1	1	0	0	1	0.17	0.57

هر ستون مربوط به یک کلمه در جمله است. هر ردیف یک تغییر است، ۱ به این معنی است که کلمه بخشی از این تغییر است و ۰ به معنای حذف کلمه است. جمله مربوط به یکی از تغییرات "Christmas Song visit my channell ;)" است. ستون "prob" احتمال پیش بینی شده هرزنامه را برای هر یک از تغییرات جمله نشان می دهد. ستون "وزن" نزدیکی تغییر به جمله اصلی را نشان می دهد که به صورت ۱ منهای نسبت کلمات حذف شده محاسبه می شود، برای مثال اگر ۱ کلمه از ۷ کلمه حذف شود، نزدیکی $1 - \frac{6}{7} = 0.14$ است.

در اینجا دو جمله (یک هرزنامه، یکی بدون هرزنامه) با وزن محلی تخمین زده شده توسط الگوریتم LIME آمده است:

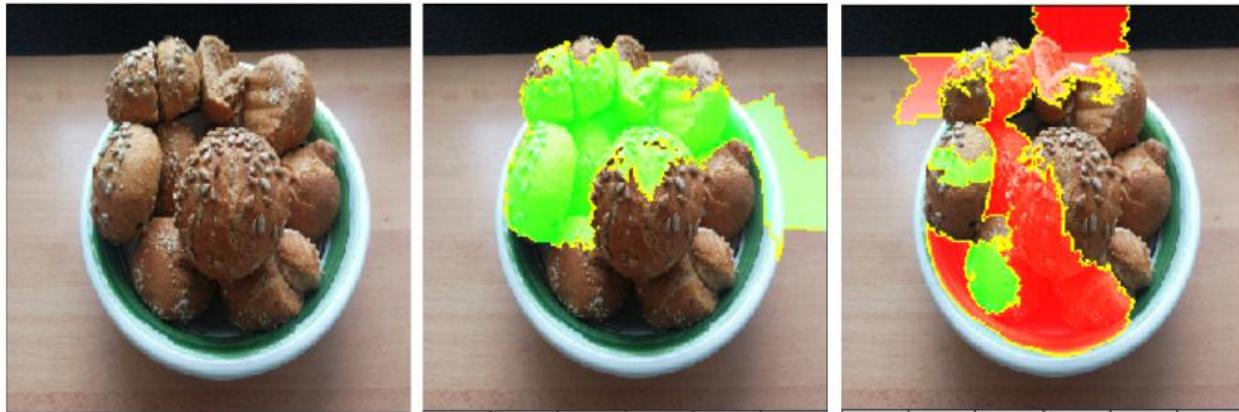
case	label_prob	feature	feature_weight
1	0.1701170	PSY	0.000000
1	0.1701170	guy	0.000000
1	0.1701170	good	0.000000
2	0.9939024	channell	6.180747
2	0.9939024	:)	0.000000
2	0.9939024	visit	0.000000

کلمه "کانال" نشان دهنده احتمال بالای اسپم است. برای نظر غیر هرزنامه، وزن غیر صفر تخمین زده نشد، زیرا مهم نیست کدام کلمه حذف شود، کلاس پیش‌بینی شده ثابت می‌ماند.

LIME برای تصاویر ۹,۲,۳ این بخش توسط Verena Haunschmid نوشته شده است.

LIME برای تصاویر متفاوت از LIME برای داده‌ها و متن جدولی عمل می‌کند. به طور شهودی، ایجاد مزاحمت برای پیکسل‌های مجزا چندان منطقی نخواهد بود، زیرا بیش از یک پیکسل در یک کلاس مشارکت دارند. تغییر تصادفی پیکسل‌های فردی احتمالاً پیش‌بینی‌ها را تغییر زیادی نمی‌دهد. بنابراین، تغییراتی در تصاویر با تقسیم بندی تصویر به "سوپرپیکسل" و خاموش یا روشن کردن سوپرپیکسل‌ها ایجاد می‌شود. سوپرپیکسل‌ها پیکسل‌های به هم پیوسته با رنگ‌های مشابه هستند و می‌توان با جایگزین کردن هر پیکسل با یک رنگ تعریف شده توسط کاربر مانند خاکستری، آن را خاموش کرد. کاربر همچنین می‌تواند یک احتمال برای خاموش کردن یک سوپرپیکسل در هر جایگشت مشخص کند.

۹,۲,۳,۱ در این مثال ما به طبقه بندی ساخته شده توسط شبکه عصبی Inception V3 نگاه می‌کنیم. تصویر استفاده شده مقداری نان را نشان می‌دهد که من پختم که در یک کاسه است. از آنجایی که می‌توانیم در هر تصویر چندین برچسب پیش‌بینی شده داشته باشیم (مرتب‌سازی شده بر اساس احتمال)، می‌توانیم برچسب‌های بالایی را توضیح دهیم. بالاترین پیش‌بینی "Bagel" با احتمال ۷۷٪ و پس از آن "توت فرنگی" با احتمال ۴٪ است. تصاویر زیر برای "Bagel" و "Strawberry" توضیحات LIME را نشان می‌دهد. توضیحات را می‌توان مستقیماً روی نمونه‌های تصویر نمایش داد. سبز به این معنی است که این قسمت از تصویر احتمال برچسب را افزایش می‌دهد و قرمز به معنای کاهش است.



۴۷۰۵

۴۷۰۶

۴۷۰۷

۴۷۰۸

۴۷۰۹

۴۷۱۰

۴۷۱۱

۴۷۱۲

۴۷۱۳

۴۷۱۴

۴۷۱۵

۴۷۱۶

۴۷۱۷

۴۷۱۸

۴۷۱۹

۴۷۲۰

۴۷۲۱

۴۷۲۲

۴۷۲۳

۴۷۲۴

شکل ۹,۸: سمت چپ: تصویر یک کاسه نان. وسط و راست: توضیحات LIME برای ۲ کلاس برتر (نان شیرینی، توت فرنگی) برای طبقه بندی تصاویر ساخته شده توسط شبکه عصبی Inception V3 Google.

پیش‌بینی و توضیح برای "Bagel" بسیار معقول است، حتی اگر پیش‌بینی اشتباه باشد - اینها به وضوح هیچ چیز شیرینی نیستند زیرا سوراخ در وسط آن وجود ندارد.

۹,۲,۴ مزایا

حتی اگر مدل اصلی یادگیری ماشین را جایگزین کنید، همچنان می‌توانید از همان مدل محلی و قابل تفسیر برای توضیح استفاده کنید. فرض کنید افرادی که به توضیحات نگاه می‌کنند درخت تصمیم را بهتر درک می‌کنند. از آنجایی که شما از مدل‌های جایگزین محلی استفاده می‌کنید، از درخت‌های تصمیم به عنوان توضیح استفاده می‌کنید بدون اینکه واقعاً مجبور باشید از درخت تصمیم برای پیش‌بینی‌ها استفاده کنید. برای مثال می‌توانید از SVM استفاده کنید. و اگر معلوم شد که یک مدل xgboost بهتر کار می‌کند، می‌توانید SVM را جایگزین کنید و همچنان از درخت تصمیم برای توضیح پیش‌بینی‌ها استفاده کنید.

مدل‌های جایگزین محلی از ادبیات و تجربه آموزش و تفسیر مدل‌های قابل تفسیر بهره می‌برند.

هنگام استفاده از کمند یا درختان کوتاه، توضیحات حاصل کوتاه (=انتخابی) و احتمالاً متضاد هستند. بنابراین، آنها توضیحات انسان دوستانه می‌دهند. به همین دلیل است که من LIME را بیشتر در برنامه‌هایی می‌بینم که دریافت‌کننده توضیح یک فرد عادی یا فردی با زمان بسیار کم است. برای تخصیص کامل کافی نیست، بنابراین من LIME را در سناریوهای انطباق که ممکن است از نظر قانونی ملزم به توضیح کامل یک پیش‌بینی باشد، نمی‌بینم. همچنین برای اشکال زدایی مدل‌های یادگیری ماشینی، داشتن همه دلایل به جای چند دلیل مفید است.

LIME یکی از معود روش‌هایی است که برای داده‌های جدولی، متن و تصاویر کار می‌کند.

۴۷۲۵ معیار وفاداری (مدل قابل تفسیر تا چه حد به پیش‌بینی‌های جعبه سیاه تقریب می‌کند) به ما ایده خوبی درباره
۴۷۲۶ اینکه مدل قابل تفسیر در توضیح پیش‌بینی‌های جعبه سیاه در همسایگی نمونه داده مورد نظر چقدر قابل
۴۷۲۷ اعتماد است، می‌دهد.

۴۷۲۸ در پایتون (کتابخانه lime و) R بسته آهک و بسته (iml پیاده سازی شده است و استفاده از آن
۴۷۲۹ بسیار آسان است.

۴۷۳۰ توضیحات ایجاد شده با مدل‌های جایگزین محلی می‌تواند از ویژگی‌های دیگری (قابل تفسیر) نسبت به مدل
۴۷۳۱ اصلی استفاده کند.. البته، این ویژگی‌های قابل تفسیر باید از نمونه‌های داده مشتق شوند. یک طبقه‌بندی‌کننده
۴۷۳۲ متن می‌تواند بر تعابیه‌های کلمات انتزاعی به عنوان ویژگی تکیه کند، اما توضیح می‌تواند بر اساس وجود یا عدم
۴۷۳۳ وجود کلمات در یک جمله باشد. یک مدل رگرسیون می‌تواند بر تبدیل غیر قابل تفسیر برخی از ویژگی‌ها تکیه
۴۷۳۴ کند، اما توضیحات را می‌توان با ویژگی‌های اصلی ایجاد کرد. به عنوان مثال، مدل رگرسیون را می‌توان بر روی
۴۷۳۵ اجزای یک تجزیه و تحلیل مؤلفه اصلی (PCA) پاسخ‌های یک نظرسنجی آموزش داد، اما LIME ممکن است در
۴۷۳۶ مورد سؤالات نظرسنجی اصلی آموزش داده شود. استفاده از ویژگی‌های قابل تفسیر برای LIME می‌تواند یک
۴۷۳۷ مزیت بزرگ نسبت به روش‌های دیگر باشد، به خصوص زمانی که مدل با ویژگی‌های غیر قابل تفسیر آموزش
۴۷۳۸ داده شده است.

۴۷۳۹ ۹,۲,۵ معایب
۴۷۴۰ هنگام استفاده از LIME با داده‌های جدولی، تعریف صحیح محله یک مشکل بسیار بزرگ و حل نشده است. به
۴۷۴۱ نظر من این بزرگترین مشکل LIME است و دلیل اینکه توصیه می‌کنم از LIME فقط با دقت زیاد استفاده
۴۷۴۲ کنید. برای هر برنامه باید تنظیمات هسته مختلف را امتحان کنید و خودتان ببینید که آیا توضیحات منطقی
۴۷۴۳ هستند یا خیر. متأسفانه، این بهترین توصیه ای است که می‌توانم برای یافتن پنهانی هسته خوب داشته باشم.

۴۷۴۴ نمونه برداری را می‌توان در اجرای فعلی LIME بهبود بخشید. نقاط داده از یک توزیع گاوی نمونه برداری می‌
۴۷۴۵ شوند و همبستگی بین ویژگی‌ها را نادیده می‌گیرند. این می‌تواند به نقاط داده غیرمحتمل منجر شود که
۴۷۴۶ سپس می‌توان از آنها برای یادگیری مدل‌های توضیح محلی استفاده کرد.

۴۷۴۷ پیچیدگی مدل توضیح باید از قبل تعریف شود. این فقط یک شکایت کوچک است، زیرا در نهایت کاربر همیشه
۴۷۴۸ باید مصالحه بین وفاداری و پراکندگی را تعریف کند.

۴۷۴۹ مشکل واقعاً بزرگ دیگر بی ثباتی توضیحات است. در مقاله ۵۱ نویسنده‌گان نشان دادند که توضیحات دو نقطه
۴۷۵۰ بسیار نزدیک در یک محیط شبیه سازی شده بسیار متفاوت است. همچنین، طبق تجربه من، اگر فرآیند نمونه

برداری را تکرار کنید، توضیحاتی که می آید می تواند متفاوت باشد. بی ثباتی به این معناست که اعتماد به توضیحات دشوار است و باید بسیار انتقادپذیر باشد.

توضیحات LIME می تواند توسط دانشمند داده دستکاری شود تا سوگیری ها را پنهان کند ۵۲ . امکان دستکاری اعتماد به توضیحات تولید شده با LIME را دشوارتر می کند.

نتیجه‌گیری: مدل‌های جایگزین محلی، با LIME به عنوان یک پیاده‌سازی بتن، بسیار امیدوارکننده هستند. اما این روش هنوز در مرحله توسعه است و بسیاری از مشکلات باید حل شود تا بتوان به طور ایمن از آن استفاده کرد.

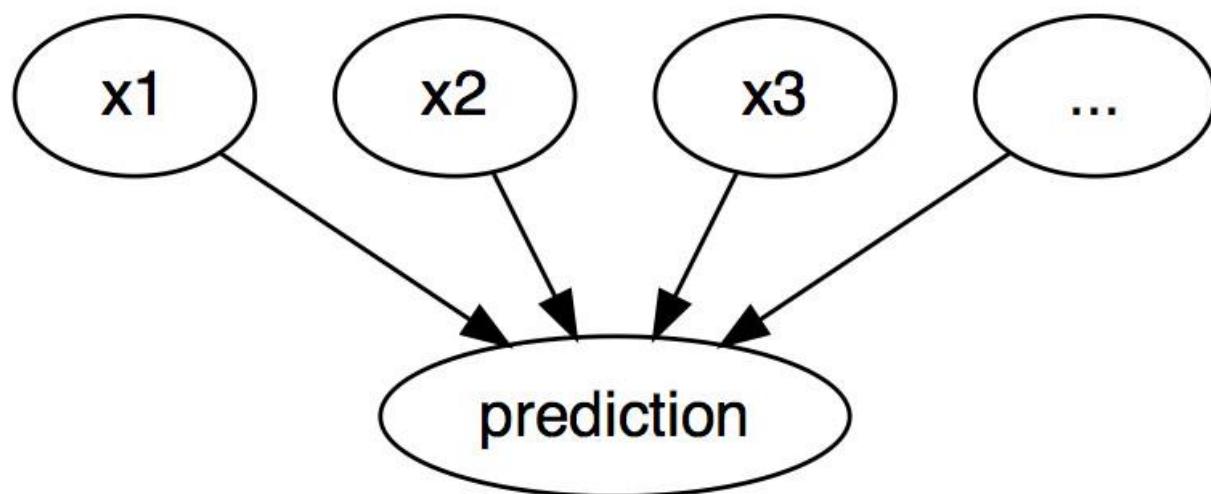
۴۷۵۸

۹.۳ توضیحات خلاف واقع

نویسنده‌گان: سوزان دنل و کریستوف مولنار

یک توضیح خلاف واقع یک وضعیت علی را به این شکل توصیف می‌کند: "اگر X رخ نمی‌داد، ۷ رخ نمی‌داد.".
به عنوان مثال: "اگر جرعه‌ای از این قهقهه داغ را ننوشیده بودم، زبانم نمی‌سوخت." رویداد ۷ این است که زبانم را سوزاندم. دلیل X این است که من یک قهقهه داغ خوردم. تفکر در خلاف واقع مستلزم تصور واقعیتی فرضی است که با واقعیت‌های مشاهده شده در تضاد است (مثلًاً دنیایی که در آن قهقهه داغ را ننوشیده‌ام)، از این رو نام آن را «ضد واقعیت» گذاشته‌اند. توانایی تفکر خلاف واقع، ما انسان‌ها را در مقایسه با سایر حیوانات بسیار باهوش می‌کند.

در یادگیری ماشینی قابل تفسیر، می‌توان از توضیحات خلاف واقع برای توضیح پیش‌بینی‌های نمونه‌های فردی استفاده کرد. «رویداد» نتیجه پیش‌بینی‌شده یک نمونه است، «علت‌ها» مقادیر ویژگی‌های خاص این نمونه هستند که به مدل وارد شده و یک پیش‌بینی مشخص را «سبب» می‌کنند. که به صورت نمودار نمایش داده می‌شود، رابطه بین ورودی‌ها و پیش‌بینی بسیار ساده است: مقادیر ویژگی باعث پیش‌بینی می‌شود.



شکل ۹.۹: روابط علی بین ورودی‌های یک مدل یادگیری ماشین و پیش‌بینی‌ها، زمانی که مدل صرفاً به عنوان یک جعبه سیاه دیده می‌شود. ورودی‌ها باعث پیش‌بینی می‌شوند (الزماءً منعکس کننده رابطه علی واقعی داده‌ها نیستند).

حتی اگر در واقعیت رابطه بین ورودی‌ها و نتیجه‌ای که باید پیش‌بینی شود ممکن است علی نباشد، می‌توانیم ورودی‌های یک مدل را به عنوان علت پیش‌بینی ببینیم.

با توجه به این نمودار ساده، به راحتی می‌توان فهمید که چگونه می‌توانیم خلاف واقع‌ها را برای پیش‌بینی‌های مدل‌های یادگیری ماشین شبیه‌سازی کنیم: ما به سادگی مقادیر ویژگی‌های یک نمونه را قبل از انجام پیش‌بینی‌ها تغییر می‌دهیم و چگونگی تغییر پیش‌بینی را تحلیل می‌کنیم. ما به سناریوهایی علاقه مند هستیم که در آن پیش‌بینی به شیوه‌ای مرتبط تغییر می‌کند، مانند تلنگر در کلاس پیش‌بینی شده (به عنوان مثال، درخواست اعتبار پذیرفته یا رد می‌شود)، یا در آن پیش‌بینی به آستانه خاصی می‌رسد (برای مثال، احتمال سرطان به آن می‌رسد. ۱۰ درصد). توضیح خلاف واقع یک پیش‌بینی، کوچک‌ترین تغییر در مقادیر ویژگی را توصیف می‌کند که پیش‌بینی را به یک خروجی از پیش‌تعریف‌شده تغییر می‌دهد.

هر دو روش توضیح خلاف واقع مدل-آگنوستیک و مدل خاص وجود دارد، اما در این فصل ما بر روی روش‌های مدل-آگنوستیک تمرکز می‌کنیم که فقط با ورودی‌ها و خروجی‌های مدل (و نه ساختار داخلی مدل‌های خاص) کار می‌کنند. این روش‌ها همچنین در فصل مدل-آگنوستیک احساس می‌کنند، زیرا تفسیر را می‌توان به صورت خلاصه‌ای از تفاوت‌ها در مقادیر ویژگی بیان کرد («تغییر ویژگی‌های A و B برای تغییر پیش‌بینی»). اما توضیح خلاف واقع خود یک نمونه جدید است، بنابراین در این فصل زندگی می‌کند («با شروع از مثال X، A و B را تغییر دهید تا یک نمونه خلاف واقع به دست آورید»). برخلاف نمونه‌های اولیه، خلاف واقع‌ها نباید نمونه‌های واقعی از داده‌های آموزشی باشند، بلکه می‌توانند ترکیب جدیدی از مقادیر ویژگی باشند.

قبل از بحث در مورد چگونگی ایجاد خلاف واقع، می‌خواهیم در مورد موارد استفاده از خلاف واقع و اینکه چگونه یک توضیح خلاف واقع خوب به نظر می‌رسد، صحبت کنم.

در این مثال اول، پیتر برای وام درخواست می‌کند و توسط نرم‌افزار بانکی (با قدرت یادگیری ماشینی) رد می‌شود. او تعجب می‌کند که چرا درخواستش رد شد و چگونه ممکن است شанс خود را برای دریافت وام افزایش دهد. سؤال «چرا» را می‌توان به عنوان خلاف واقع فرمول‌بندی کرد: کوچک‌ترین تغییر در ویژگی‌ها (درآمد، تعداد کارت‌های اعتباری، سن، ...) که پیش‌بینی را از رد به تأیید تغییر می‌دهد چیست؟ یک پاسخ ممکن می‌تواند این باشد: اگر پیتر ۱۰۰۰۰ بیشتر در سال درآمد داشت، وام را دریافت می‌کرد. یا اگر پیتر کارت‌های اعتباری کمتری داشت و پنج سال پیش وام را نکول نکرده بود، وام را دریافت می‌کرد. پیتر هرگز دلایل رد را نمی‌داند، زیرا بانک علاقه‌ای به شفافیت ندارد، اما این داستان دیگری است.

در مثال دوم خود، می‌خواهیم مدلی را توضیح دهیم که یک نتیجه پیوسته را با توضیحات خلاف واقع پیش‌بینی می‌کند. آنا می‌خواهد آپارتمانش را اجاره کند، اما مطمئن نیست که چقدر برای آن هزینه بگیرد، بنابراین تصمیم می‌گیرد یک مدل یادگیری ماشینی برای پیش‌بینی اجاره‌ها آموزش دهد. البته از آنجایی که آنا یک دانشمند داده است، مشکلات خود را از این طریق حل می‌کند. پس از وارد کردن تمام جزئیات در مورد اندازه،

مکان، مجاز بودن حیوانات خانگی و غیره، مدل به او می گوید که می تواند ۹۰۰ یورو شارژ کند. او انتظار ۱۰۰۰ یورو یا بیشتر را داشت، اما به مدل خود اعتماد می کند و تصمیم می گیرد با ارزش های ویژگی های آپارتمان بازی کند تا ببیند چگونه می تواند ارزش آپارتمان را بهبود بخشد. او متوجه می شود که آپارتمان را می توان بیش از ۱۰۰۰ یورو اجاره کرد، اگر ۱۵ متر مربع بود. بزرگتر دانش جالب، اما غیر قابل عمل، زیرا او نمی تواند آپارتمان خود را بزرگ کند. در نهایت، با تغییر دادن فقط مقادیر ویژگی های تحت کنترل خود (آشپزخانه داخلی بله/خیر، حیوانات خانگی مجاز هستند بله/خیر، نوع کف و غیره)، متوجه می شود که اگر اجازه حیوانات خانگی را بدهد و پنجره هایی با عایق بهتر نصب کند، او می تواند ۱۰۰۰ یورو شارژ کند. آنا به طور شهودی با عوامل خلاف واقع کار کرده است تا نتیجه را تغییر دهد.

ضدافتک ها توضیحاتی انسان پسند هستند، زیرا آنها با نمونه فعلی متضاد هستند و به دلیل انتخابی بودن آنها، به این معنی که آنها معمولاً روی تعداد کمی از تغییرات ویژگی تمرکز می کنند. اما خلاف واقع ها از «اثر راشومون» رنج می برند. راشومون یک فیلم ژاپنی است که در آن قتل یک سامورایی توسط افراد مختلف روایت می شود. هر یک از داستان ها نتیجه را به یک اندازه به خوبی توضیح می دهد، اما داستان ها با یکدیگر تناقض دارند. همین امر می تواند در مورد خلاف واقع نیز اتفاق بیفت، زیرا معمولاً چندین توضیح خلاف واقع مختلف وجود دارد. هر خلاف واقع «داستان» متفاوتی از چگونگی دستیابی به یک نتیجه معین می گوید. یک خلاف واقع ممکن است بگوید ویژگی A را تغییر دهید، خلاف واقع دیگر ممکن است بگوید A را همان طور باقی بگذارید اما ویژگی B را تغییر دهید که این یک تناقض است.

در مورد معیارها، چگونه توضیح خلاف واقع خوب را تعریف کنیم؟ ابتدا، کاربر یک توضیح خلاف واقع، یک تغییر مرتبط در پیش‌بینی یک نمونه (= واقعیت جایگزین) را تعریف می‌کند. اولین نیاز آشکار این است که یک نمونه خلاف واقع، پیش‌بینی از پیش تعریف شده را تا حد امکان به دقت تولید کند.. همیشه نمی توان با پیش‌بینی از پیش تعریف شده، یک خلاف واقع پیدا کرد. برای مثال، در یک تنظیم طبقه بندی با دو کلاس، یک کلاس نادر و یک کلاس مکرر، مدل ممکن است همیشه یک نمونه را به عنوان کلاس مکرر طبقه بندی کند. تغییر مقادیر ویژگی به طوری که برچسب پیش‌بینی شده از کلاس مکرر به کلاس نادر تغییر کند ممکن است غیرممکن باشد. بنابراین ما می خواهیم این شرط را که پیش‌بینی خلاف واقع باید دقیقاً با نتیجه از پیش تعریف شده مطابقت داشته باشد، کاهش دهیم. در مثال طبقه‌بندی، می‌توانیم به دنبال خلاف واقع باشیم که در آن احتمال پیش‌بینی شده کلاس نادر به جای ۲ درصد فعلی به ۱۰ درصد افزایش می‌یابد. پس سؤال این است که حداقل تغییرات در ویژگی ها چیست به طوری که احتمال پیش‌بینی شده از ۲٪ به ۱۰٪ (یا نزدیک به ۱۰٪) تغییر می کند؟

یکی دیگر از معیارهای کیفی این است که یک خلاف واقع باید تا حد امکان مشابه نمونه مربوط به مقادیر ویژگی باشد . فاصله بین دو نمونه را می توان به عنوان مثال با فاصله منهتن یا فاصله Gower اندازه گیری کرد اگر هم ویژگی های گستته و هم پیوسته داشته باشیم. خلاف واقع نه تنها باید به نمونه اصلی نزدیک باشد، بلکه باید تا حد امکان ویژگی های کمتری را تغییر دهد . برای سنجش میزان خوب بودن توضیح خلاف واقع در این متريک، می توانيم به سادگی تعداد ویژگی های تغيير يافته را بشماريم یا به تعبير رياضي فانتзи، آن را اندازه گيري کنيم^۱ . هنچار بین مثال خلاف واقع و واقعی

ثالثاً، اغلب مطلوب است که چندين توضیح خلاف واقع متنوع ايجاد شود ، به طوری که موضوع تصمیم گيرنده به چندين روش قابل دوام برای ايجاد یک نتيجه متفاوت دسترسی پیدا کند. به عنوان مثال، در ادامه مثال وام ما، یک توضیح خلاف واقع ممکن است فقط دوبرابر کردن درآمد برای دریافت وام را پیشنهاد کند، در حالی که خلاف واقع ممکن است انتقال به یک شهر مجاور را پیشنهاد کند و درآمد را با مقدار کمی افزایش دهد تا وام دریافت کند. می توان اشاره کرد که در حالی که اولین خلاف واقع ممکن است برای برخی امكان پذير باشد، دومی ممکن است برای دیگران قابل اجراتر باشد. بنابراین، علاوه بر ارائه یک موضوع تصمیم با روش های مختلف برای رسیدن به نتيجه مطلوب، تنوع همچنین افراد "متنوع" را قادر می سازد تا ویژگی هایی را که برای آنها مناسب است تغیير دهند.

آخرین شرط اين است که یک نمونه خلاف واقع باید مقادير مشخصه اي داشته باشد که محتمل است . ايجاد توضیح خلاف واقع برای مثال اجاره که در آن اندازه یک آپارتمان منفی است یا تعداد اتاق ها روی ۲۰۰ اتاق تنظیم شده است، منطقی نیست. حتی بهتر است زمانی که خلاف واقع بر اساس توزيع مشترک داده ها باشد. به عنوان مثال، یک آپارتمان با ۱۰ اتاق و ۲۰ متر مربع نباید به عنوان توضیح خلاف واقع در نظر گرفته شود. در حالت ايده آل، اگر تعداد متر مربع افزایش يابد، افزایش تعداد اتاق ها نيز باید پیشنهاد شود.

۹.۳.۱ ايجاد توضیحات خلاف واقع

یک رویکرد ساده و ساده لوحانه برای ايجاد توضیحات خلاف واقع، جستجو با آزمون و خطأ است. اين رویکرد شامل تغيير تصادفي مقادير ویژگی نمونه مورد علاقه و توقف زمانی است که خروجي مورد نظر پيش‌بياني می‌شود. مانند مثالی که آنا سعی کرد نسخه اي از آپارتمان خود را پیدا کند که بتواند برای آن اجاره بيشتری بگيرد. اما رویکردهای بهتری نسبت به آزمون و خطأ وجود دارد. ابتدا یکتابع ضرر را بر اساس معیارهای ذکر شده در بالا تعريف می کنيم. اين ضرر به عنوان ورودی نمونه مورد علاقه، خلاف واقع و نتيجه مطلوب (معمولا) را می گيرد. سپس، می توانيم توضیح خلاف واقع را پیدا کنيم که اين تلفات را با استفاده از یک الگوريتم

۴۸۵۷ بهینه‌سازی به حداقل می‌رساند. بسیاری از روش‌ها به این شکل پیش می‌روند، اما در تعریف تابع ضرر و روش
۴۸۵۸ بهینه‌سازی متفاوت هستند.

۴۸۵۹ در ادامه، به دو مورد از آنها می‌پردازیم: اول، یکی از واچتر و همکاران. (۲۰۱۷) ۵۳، که تبیین خلاف واقع را به
۴۸۶۰ عنوان یک روش تفسیری معرفی کردند و دوم، توضیح دندل و همکاران. (۲۰۲۰) ۵۴ که هر چهار معیار ذکر
۴۸۶۱ شده در بالا را در نظر می‌گیرد.

۱,۱,۳,۹ روش توسطه واچتر و همکاران.

۴۸۶۲ واچتر و همکاران پیشنهاد به حداقل رساندن ضرر زیر:

$$L(x, x', y', \lambda) = \lambda \cdot (\hat{f}(x') - y')^2 + d(x, x')$$

۴۸۶۴ عبارت اول فاصله درجه دوم بین پیش‌بینی مدل برای x' خلاف واقع و نتیجه مطلوب y' است که کاربر باید از
۴۸۶۵ قبل آن را تعریف کند. جمله دوم فاصله d بین مثال x که باید توضیح داده شود و x' خلاف واقع است. زیان
۴۸۶۶ اندازه‌گیری می‌کند که نتیجه پیش‌بینی شده خلاف واقع تا چه اندازه با نتیجه از پیش تعریف شده فاصله دارد و
۴۸۶۷ خلاف واقع چقدر با نمونه مورد علاقه فاصله دارد. تابع فاصله d به عنوان فاصله منهتن وزن شده با میانگین
۴۸۶۸ معکوس انحراف مطلق (MAD) هر ویژگی تعریف می‌شود.

$$d(x, x') = \sum_{j=1}^p \frac{|x_j - x'_j|}{MAD_j}$$

۴۸۶۹ فاصله کل مجموع تمام فواصل p از نظر ویژگی است، یعنی تفاوت مطلق مقادیر ویژگی بین مثال x و خلاف
۴۸۷۰ واقع x' فاصل از نظر ویژگی با معکوس انحراف مطلق میانه ویژگی \bar{z} بر روی مجموعه داده تعریف شده به
۴۸۷۱ صورت زیر مقیاس می‌شوند:

$$MAD_j = \text{median}_{i \in \{1, \dots, n\}} (|x_{i,j} - \text{median}_{l \in \{1, \dots, n\}} (x_{l,j})|)$$

۴۸۷۲ میانه یک بردار مقداری است که در آن نیمی از مقادیر بردار بزرگتر و نیمی دیگر کوچکتر باشند MAD. معادل
۴۸۷۳ واریانس یک ویژگی است، اما به جای استفاده از میانگین به عنوان مرکز و جمع بر روی فواصل مربع، از میانه به
۴۸۷۴ عنوان مرکز و جمع در فواصل مطلق استفاده می‌کنیم. تابع فاصله پیشنهادی این مزیت را نسبت به فاصله
۴۸۷۵ اقلیدسی دارد که نسبت به نقاط پرت قوی تر است. مقیاس بندی با MAD برای رساندن همه ویژگی‌ها به یک
۴۸۷۶ مقیاس ضروری است - مهم نیست که اندازه یک آپارتمان را در متر مربع یا فوت مربع اندازه گیری کنید.

پارامتر λ فاصله را در پیش بینی (ترم اول) با فاصله در مقادیر ویژگی (ترم دوم) متعادل می کند. ضرر برای یک معین حل می شود λ و یک X خلاف واقع را برمی گرداند. ارزش بالاتر از λ به این معنی است که ما خلاف واقع ها را با پیش بینی های نزدیک به نتیجه مطلوب λ ترجیح می دهیم، یک مقدار کمتر به این معنی است که ما خلاف واقع های X را ترجیح می دهیم که در مقادیر ویژگی بسیار شبیه به X هستند. اگر λ بسیار بزرگ است، نمونه ای با پیش بینی نزدیک به λ انتخاب خواهد شد، صرف نظر از اینکه چقدر از X فاصله دارد. در نهایت، کاربر باید تصمیم بگیرد که چگونه بین شرطی که پیش بینی خلاف واقع با نتیجه دلخواه مطابقت دارد، تعادل برقرار کند. نویسنده ای روش به جای انتخاب مقداری برای λ برای انتخاب یک تلورانس ϵ برای چقدر دور از λ پیش بینی نمونه خلاف واقع مجاز است. این محدودیت را می توان به صورت زیر نوشت:

$$|f(x') - y'| \leq \epsilon$$

برای به حداقل رساندن اینتابع از دست دادن، می توان از هر الگوریتم بهینه سازی مناسبی مانند-Nelder-Mead استفاده کرد. اگر به گرادیان های مدل یادگیری ماشین دسترسی دارید، می توانید از روش های مبتنی بر گرادیان مانند ADAM استفاده کنید. نمونه X توضیح داده شود، خروجی مورد نظر λ و پارامتر تلورانس ϵ باید از قبل تنظیم شود. تابع زیان برای X به حداقل می رسد و (محلی) بهینه λ^* در حالی که افزایش می یابد، بر می گردد λ^* تا زمانی که یک راه حل به اندازه کافی نزدیک پیدا شود (= در پارامتر تحمل):

$$\arg \min_{x'} \max_{\lambda} L(x, x', y', \lambda).$$

به طور کلی، دستور تهیه خلاف واقع ساده است:

۱- یک نمونه X را برای توضیح انتخاب کنید، نتیجه مورد نظر λ ، یک تلورانس ϵ و یک مقدار اولیه (کم) برای λ

۲- یک نمونه تصادفی را به عنوان خلاف واقع اولیه نمونه بگیرید

۳- زیان را با نمونه اولیه ضد عملی به عنوان نقطه شروع بهینه کنید.

۴- در حالی که $|f(X) - f(\text{ایکس})| < \epsilon$ ایکس:

- افزایش دادن. λ

- ضرر را با خلاف واقع فعلی به عنوان نقطه شروع بهینه کنید.

- خلاف واقع را که ضرر را به حداقل می رساند، برگردانید.

۴۹۰۳ ۵-مراحل ۲ را تکرار کنید و لیستی از موارد خلاف واقع یا مواردی که ضرر را به حداقل می‌رساند را
 ۴۹۰۴ برگردانید.

۴۹۰۵ روش پیشنهادی دارای معایبی است. این فقط معیارهای اول و دوم را در نظر می‌گیرد نه دو مورد آخر ("تولید
 ۴۹۰۶ موارد متضاد تنها با چند تغییر ویژگی و مقادیر احتمالی ویژگی d). راه حل‌های پراکنده را ترجیح نمی‌دهد زیرا
 ۴۹۰۷ افزایش ۱۰ ویژگی در ۱ همان فاصله را با افزایش یک ویژگی به ۱۰ به X می‌دهد. ترکیب‌های غیرواقعی ویژگی
 ۴۹۰۸ جریمه نمی‌شوند.

۴۹۰۹ این روش ویژگی‌های طبقه بندی شده با سطوح مختلف را به خوبی مدیریت نمی‌کند. نویسنده‌گان روش
 ۴۹۱۰ پیشنهاد کردند که روش را به طور جداگانه برای هر ترکیبی از مقادیر ویژگی ویژگی‌های دسته‌بندی اجرا کنید،
 ۴۹۱۱ اما اگر چندین ویژگی دسته‌بندی با مقادیر زیاد داشته باشد، این امر منجر به انفجار ترکیبی می‌شود. به عنوان
 ۴۹۱۲ مثال، شش ویژگی طبقه بندی شده با ده سطح منحصر به فرد به معنای یک میلیون اجرا است.

۴۹۱۳ اجازه دهد اکنون نگاهی به رویکرد دیگری برای غلبه بر این مسائل بیندازیم.

۴۹۱۴ ۹,۳,۱,۲ روش دندل و همکاران .
 ۴۹۱۵ دندل و همکاران پیشنهاد می‌شود به طور همزمان یک ضرر چهار هدف به حداقل برسد:

۴۹۱۶ $L(x, x', y', X^{obs}) = (o_1(\hat{f}(x'), y'), o_2(x, x'), o_3(x, x'), o_4(x', X^{obs}))$
 ۴۹۱۷ هر یک از چهار هدف ۱ به ۴ با یکی از چهار معیار ذکر شده در بالا مطابقت دارد. هدف اول ۱ نشان می‌
 ۴۹۱۸ دهد که پیش‌بینی X' خلاف واقع ما باید تا حد امکان به پیش‌بینی y' مورد نظر ما نزدیک باشد. بنابراین ما می‌
 ۴۹۱۹ خواهیم فاصله بین آنها را به حداقل برسانیم(f)^۸ ایکس (" x' و y' ، در اینجا با متریک منهتن محاسبه می‌شود) ۱
 ۴۹۲۰ هنجار):

۴۹۲۱
$$o_1(\hat{f}(x'), y') = \begin{cases} 0 & \text{if } \hat{f}(x') \in y' \\ \inf_{y' \in y'} |\hat{f}(x') - y'| & \text{else} \end{cases}$$

 ۴۹۲۲ هدف دوم ۲ نشان می‌دهد که خلاف واقع ما باید تا حد امکان مشابه نمونه ما باشد ایکس . فاصله بین X' و X
 ۴۹۲۳ را به عنوان فاصله Gower کمیت می‌کند:

۴۹۲۴
$$o_2(x, x') = \frac{1}{p} \sum_{j=1}^p \delta_G(x_j, x'_j)$$

 ۴۹۲۵ با p تعداد ویژگی‌ها است. ارزش δ جی بستگی به نوع ویژگی دارد ایکس: j

$$\delta_G(x_j, x'_j) = \begin{cases} \frac{1}{R_j} |x_j - x'_j| & \text{if } x_j \text{ numerical} \\ \mathbb{I}_{x_j \neq x'_j} & \text{if } x_j \text{ categorical} \end{cases}$$

۴۹۲۶
۴۹۲۷ تقسیم فاصله یک ویژگی عددی \bar{x} توسط آرزو، محدوده مقدار مشاهده شده، مقیاس ها δ جی برای همه ویژگی های بین ۰ و ۱.

۴۹۲۹ فاصله Gower می تواند هم ویژگی های عددی و هم ویژگی های طبقه ای را کنترل کند، اما تعداد ویژگی های ۴۹۳۰ تغییر یافته را محاسبه نمی کند. بنابراین، تعداد ویژگی ها را در هدف سوم می شماریم ۳۰ با استفاده از L_0 هنجار:

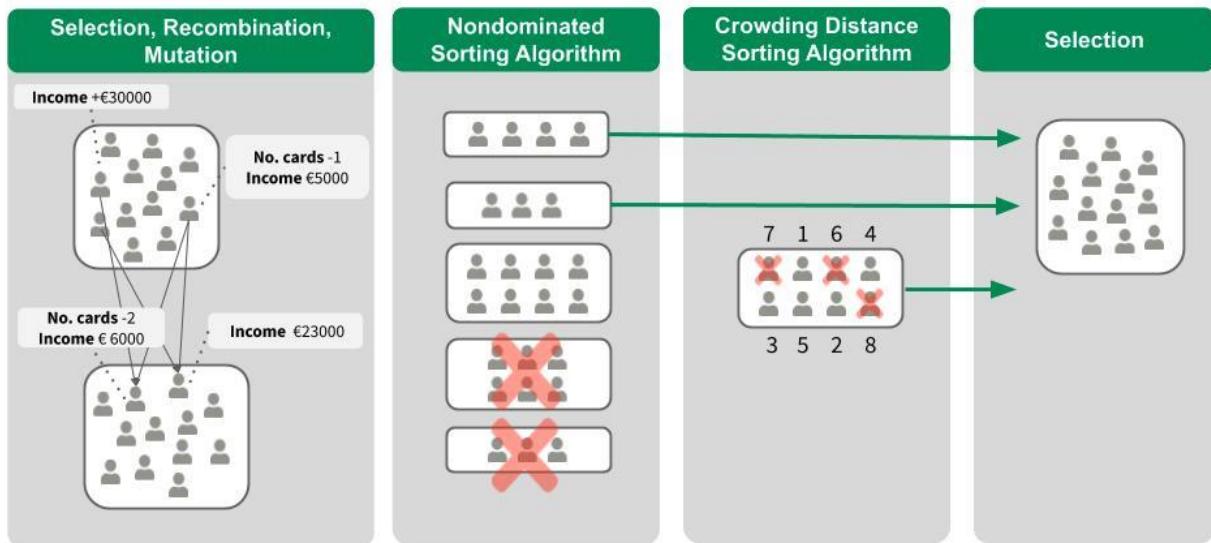
۴۹۳۱
۴۹۳۲ با به حداقل رساندن ۳۰ هدف ما سومین معیار است - تغییرات پراکنده ویژگی.

۴۹۳۳ هدف چهارم ۴۰ نشان می دهد که خلاف واقع های ما باید مقادیر/ترکیب های ویژگی احتمالی داشته باشند. ما ۴۹۳۴ می توانیم حدس بزنیم که یک نقطه داده چقدر "احتمال" از داده های آموزشی یا مجموعه داده دیگری استفاده ۴۹۳۵ می کند. ما این مجموعه داده را به عنوان نشان می دهیم ایکس ۰ ب س. به عنوان تقریبی برای احتمال، ۴۰ ۴۹۳۶ میانگین فاصله Gower بین x^i و نزدیکترین نقطه داده مشاهده شده را اندازه می گیرد ایکس $[1] \in [0, 1]$. ایکس ۰ ب ۴۹۳۷ س:

۴۹۳۸
۴۹۳۹ در مقایسه با واچتر و همکاران، (ایکس، ایکس $"^i$) ایکس ۰ ب س (هیچ عبارات تعادلی/وزنی مانند λ مانمی ۴۹۴۰ خواهیم چهار هدف را از بین ببریم ۱۰، ۲۰، ۳۰ و ۴۰ با جمع بندی و وزن دهی به یک هدف واحد، اما ما می ۴۹۴۱ خواهیم هر چهار عبارت را به طور همزمان بهینه کنیم.

۴۹۴۲ چطور می توانیم انجامش دهیم؟ ما از الگوریتم ژنتیک مرتب سازی بدون مالکیت ۵۵ یا کوتاه NSGA-II استفاده ۴۹۴۳ می کنیم NSGA-II. یک الگوریتم الهام گرفته از طبیعت است که قانون داروین را در مورد "بقاء قویترین" ۴۹۴۴ اعمال می کند. ما تناسب یک خلاف واقع را با بردار مقادیر اهداف آن نشان می دهیم $(0, 1, 20, 30, 40)$. هرچه ۴۹۴۵ مقادیر اهداف برای یک خلاف واقع کمتر باشد، "مناسب تر" است.

۴۹۴۶ این الگوریتم از چهار مرحله تشکیل شده است که تا زمانی که یک معیار توقف برآورده شود، تکرار می شود، به ۴۹۴۷ عنوان مثال، حداکثر تعداد تکرار / نسل. شکل زیر چهار مرحله یک نسل را به تصویر می کشد.



۴۹۴۸

۴۹۴۹

شکل ۹،۱۰: تجسم یک نسل از الگوریتم NSGA-II.

در نسل اول، گروهی از نامزدهای خلاف واقع با تغییر تصادفی برخی از ویژگی‌ها در مقایسه با مثال \times ما، مقداردهی اولیه می‌شوند. با رعایت مثال اعتباری بالا، یک مخالف می‌تواند افزایش درآمد را ۳۰۰۰۰ یورو پیشنهاد کند، در حالی که یکی دیگر پیشنهاد می‌کند که در پنج سال گذشته نکول نداشته باشد و سن را تا ۱۰ کاهش دهد. همه مقادیر دیگر ویژگی برابر با مقادیر \times هستند. سپس هر نامزد با استفاده از چهار تابع هدف فوق ارزیابی می‌شود. از بین آنها، ما به صورت تصادفی چند کاندیدا را انتخاب می‌کنیم، که در آن کاندیداهای متناسب با احتمال بیشتری انتخاب می‌شوند. نامزدها به صورت جفتی دوباره ترکیب می‌شوند تا با میانگین‌گیری مقادیر مشخصه عددی یا با عبور از ویژگی‌های طبقه‌بندی، فرزندانی شبیه به آنها تولید کنند. علاوه بر این،

از دو گروه به دست آمده، یکی با والدین و دیگری با فرزندان، ما فقط بهترین نیمه را با استفاده از دو الگوریتم مرتب سازی می‌خواهیم. الگوریتم مرتب سازی غیر غالب، نامزدها را بر اساس مقادیر هدف آنها مرتب می‌کند. اگر نامزدها به همان اندازه خوب باشند، الگوریتم مرتب سازی فاصله ازدحام کاندیداهای را بر اساس تنوع آنها مرتب می‌کند.

با توجه به رتبه‌بندی دو الگوریتم مرتب سازی، امیدوار کننده‌ترین و/یا متنوع‌ترین نیمی از نامزدها را انتخاب می‌کنیم. ما از این مجموعه برای نسل بعدی استفاده می‌کنیم و دوباره با فرآیند انتخاب، نوترکیب و جهش شروع می‌کنیم. با تکرار مکرر مراحل، امیدواریم به مجموعه متنوعی از نامزدهای امیدوار کننده با مقادیر هدف پایین نزدیک شویم. از این مجموعه می‌توانیم مواردی را انتخاب کنیم که از آنها رضایت بیشتری داریم، یا

۴۹۶۶ می توانیم خلاصه ای از همه موارد خلاف واقع را با برجسته کردن این که کدام و چند بار ویژگی ها تغییر کرده اند
۴۹۶۷ ارائه دهیم.

۴۹۶۸ **۹.۳.۲ مثال**
۴۹۶۹ مثال زیر بر اساس نمونه داده اعتباری در Dandl و همکاران است. (۲۰۲۰). مجموعه داده Rيسك اعتباری
۴۹۷۰ آلمان را می توان در پلتفرم چالش های یادگیری ماشینی kaggle.com پیدا کرد.

۴۹۷۱ نویسنده گان یک ماشین بردار پشتیبان (با هسته پایه شعاعی) را آموزش دادند تا احتمال اینکه یک مشتری
۴۹۷۲ Rيسك اعتباری خوبی دارد را پیش بینی کند. مجموعه داده مربوطه دارای ۵۲۲ مشاهدات کامل و ۹ ویژگی
۴۹۷۳ حاوی اطلاعات اعتباری و مشتری است.

۴۹۷۴ هدف یافتن توضیحات خلاف واقع برای مشتری با مقادیر ویژگی زیر است:

age	sex	job	housing	savings	amount	duration	purpose
58	f	unskilled	free	little	6143	48	car

۴۹۷۵ پیش بینی می کند که زن دارای Rيسك اعتباری خوبی با احتمال ۲۴,۲ درصد است. خلاف واقع ها باید پاسخ
۴۹۷۶ دهنده که چگونه ویژگی های ورودی باید تغییر کنند تا احتمال پیش بینی شده بزرگتر از ۵۰٪ بدست آید؟

۴۹۷۷ **جدول زیر ده بهترین خلاف واقع را نشان می دهد:**

age	sex	job	amount	duration	o_2	o_3	o_4	$\hat{f}(x')$
		skilled		-20	0.108	2	0.036	0.501
		skilled		-24	0.114	2	0.029	0.525
		skilled		-22	0.111	2	0.033	0.513
-6		skilled		-24	0.126	3	0.018	0.505
-3		skilled		-24	0.120	3	0.024	0.515
-1		skilled		-24	0.116	3	0.027	0.522
-3	m			-24	0.195	3	0.012	0.501
-6	m			-25	0.202	3	0.011	0.501
-30	m	skilled		-24	0.285	4	0.005	0.590
-4	m		-1254	-24	0.204	4	0.002	0.506

۴۹۷۹ پنج ستون اول شامل تغییرات ویژگی پیشنهادی است (فقط ویژگی های تغییر یافته نمایش داده می شود)، سه
۴۹۸۰ ستون بعدی مقادیر هدف را نشان می دهد) ۱ در تمام موارد برابر با ۰ است) و آخرین ستون احتمال پیش
۴۹۸۱ بینی شده را نشان می دهد.
۴۹۸۲

۴۹۸۳ همه خلاف واقع ها احتمالات بیشتر از ۵۰٪ را پیش بینی کرده اند و بر یکدیگر تسلط ندارند. بدون تسلط به این
۴۹۸۴ معنی است که هیچ یک از خلاف واقع ها در همه اهداف دارای مقادیر کوچکتری نسبت به سایر موارد خلاف
۴۹۸۵ واقع نیستند. ما می توانیم خلاف واقع های خود را به عنوان مجموعه ای از راه حل های مبادله ای در نظر
۴۹۸۶ بگیریم.

۴۹۸۷ همه آنها کاهش مدت زمان را از ۴۸ ماه به حداقل ۲۳ ماه پیشنهاد می کنند، برخی از آنها پیشنهاد می کنند که
۴۹۸۸ زن باید به جای غیر ماهر شدن، ماهر شود. برخی خلاف واقع ها حتی پیشنهاد می کنند که جنسیت را از زن به
۴۹۸۹ مرد تغییر دهید که نشان دهنده تعصب جنسیتی مدل است. این تغییر همیشه با کاهش سن بین یک تا ۳۰ سال همراه است.
۴۹۹۰ همچنین می توانیم بینیم که، اگرچه برخی خلاف واقع ها تغییراتی را در چهار ویژگی پیشنهاد
۴۹۹۱ می کنند، اما این خلاف واقع ها آنهایی هستند که به داده های آموزشی نزدیک تر هستند.

۴۹۹۲ ۹.۳.۳ مزایا

۴۹۹۳ تفسیر توضیحات خلاف واقع بسیار روشن است . اگر مقادیر ویژگی یک نمونه بر اساس خلاف واقع تغییر کند،
۴۹۹۴ پیش بینی از پیش بینی از تعریف شده تغییر می کند. هیچ فرض اضافی و هیچ جادوی در پس زمینه وجود
۴۹۹۵ ندارد. این همچنین به این معنی است که به اندازه روش هایی مانند LIME خطرناک نیست ، جایی که مشخص
۴۹۹۶ نیست تا چه حد می توانیم مدل محلی را برای تفسیر تعبیر کنیم.

۴۹۹۷ روش خلاف واقع یک نمونه جدید ایجاد می کند، اما می توانیم یک خلاف واقع را با گزارش دادن اینکه کدام
۴۹۹۸ مقادیر ویژگی تغییر کرده است، خلاصه کنیم. این دو گزینه برای گزارش نتایج به ما می دهد . می توانید نمونه
۴۹۹۹ خلاف واقع را گزارش دهید یا مشخص کنید که کدام ویژگی بین نمونه مورد علاقه و نمونه خلاف واقع تغییر
۵۰۰۰ کرده است.

۵۰۰۱ روش خلاف واقع نیازی به داده یا مدل ندارد . این فقط نیاز به دسترسی بهتابع پیش بینی مدل
۵۰۰۲ دارد، که برای مثال از طریق یک وب API نیز کار می کند. این برای شرکت هایی که توسط اشخاص ثالث
۵۰۰۳ حسابرسی می شوند یا بدون افشای مدل یا داده ها توضیحاتی را برای کاربران ارائه می دهند جذاب است. یک
۵۰۰۴ شرکت به دلیل اسرار تجاری یا دلایل حفاظت از داده ها، علاقه مند به محافظت از مدل و داده است. توضیحات
۵۰۰۵ خلاف واقع تعادلی بین توضیح پیش بینی های مدل و حفاظت از منافع مالک مدل ارائه می دهد.

۵۰۰۶ این روش همچنین با سیستم هایی کار می کند که از یادگیری ماشینی استفاده نمی کنند . ما می توانیم برای
۵۰۰۷ هر سیستمی که ورودی ها را دریافت می کند و خروجی ها را بر می گرداند، خلاف واقع ایجاد کنیم. سیستمی

۵۰۰۸ که اجاره آپارتمان را پیش‌بینی می‌کند همچنین می‌تواند شامل قوانین دست‌نویس باشد و توضیحات خلاف واقع
۵۰۰۹ همچنان کارساز است.

۵۰۱۰ پیاده‌سازی روش توضیح خلاف واقع نسبتاً آسان است، زیرا اساساً یک تابع ضرر (با یک یا چند هدف) است که
۵۰۱۱ می‌تواند با کتابخانه‌های بهینه‌ساز استاندارد بهینه شود. برخی جزئیات اضافی باید در نظر گرفته شود، مانند
۵۰۱۲ محدود کردن مقادیر ویژگی به محدوده‌های معنی دار (مثلاً فقط اندازه‌های آپارتمان مثبت).

۹,۳,۴ معايب

۵۰۱۴ برای هر نمونه معمولاً چندین توضیح خلاف واقع (اثر راشومون) پیدا خواهد کرد. این ناخوشایند است - بیشتر
۵۰۱۵ مردم توضیحات ساده را به پیچیدگی دنیای واقعی ترجیح می‌دهند. همچنین یک چالش عملی است. فرض
۵۰۱۶ کنید برای یک نمونه ۲۳ توضیح خلاف واقع ایجاد کردیم. آیا ما همه آنها را گزارش می‌کنیم؟ تنها بهترین؟ اگر
۵۰۱۷ همه آنها نسبتاً "خوب" اما بسیار متفاوت باشند چه؟ برای هر پروژه باید دوباره به این سوالات پاسخ داد.
۵۰۱۸ همچنین داشتن چندین توضیح خلاف واقع می‌تواند سودمند باشد، زیرا انسان‌ها می‌توانند آن‌ها را انتخاب
۵۰۱۹ کنند که با دانش قبلی‌شان مطابقت دارند.

۹,۳,۵ نرم افزار و جايگزين

۵۰۲۱ روش تبیین ضدواقعی چنددهده توسط دندل و همکاران. در یک مخزن GitHub پیاده سازی شده است.

۵۰۲۲ در بسته پایتون، نویسنده‌گان Alibi یک روش خلاف واقع ساده و همچنین یک روش توسعه یافته را پیاده سازی
۵۰۲۳ کرده‌اند که از نمونه‌های اولیه کلاس برای بهبود تفسیرپذیری و همگرایی خروجی‌های الگوریتم استفاده می‌کند
۵۰۲۴ .

۵۰۲۵ کریمی و همکاران (۲۰۲۰) ۵۷ همچنین پیاده سازی پایتون از الگوریتم MACE خود را در یک مخزن
۵۰۲۶ GitHub ارائه کرد. آنها معیارهای لازم برای خلاف واقع‌های مناسب را به فرمول‌های منطقی ترجمه کردند و از
۵۰۲۷ حل‌کننده‌های رضایت‌پذیری برای یافتن خلاف واقع‌هایی که آن‌ها را برآورده می‌کنند، استفاده کردند.

۵۰۲۸ متیالال و همکاران (۲۰۲۰) ۵۸ DICE تبیین متضاد متنوع) را برای تولید مجموعه متنوعی از توضیحات خلاف
۵۰۲۹ واقع بر اساس فرآیندهای نقطه تعیین کننده توسعه داد DICE . هم یک روش مدل-آنالوگیک و هم یک روش
۵۰۳۰ مبتنی بر گرادیان را پیاده سازی می‌کند.

۵۰۳۱ روش دیگر برای جستجوی خلاف واقع، الگوریتم Growing Spheres توسط Laugel و همکاران است.
۵۰۳۲ (۲۰۱۷) ۵۹ . آنها از کلمه خلاف واقع در مقاله خود استفاده نمی‌کنند، اما روش کاملاً مشابه است. آنها
۵۰۳۳ همچنین یک تابع ضرر را تعریف می‌کنند که به نفع خلاف واقع‌ها با کمترین تغییرات ممکن در مقادیر ویژگی

۵۰۳۴ است. به جای بهینه سازی مستقیم تابع، آنها پیشنهاد می کنند ابتدا یک کره در اطراف نقطه مورد نظر ترسیم
کنید، از نقاط آن کره نمونه برداری کنید و بررسی کنید که آیا یکی از نقاط نمونه برداری شده پیش بینی مورد
نظر را به دست می دهد یا خیر. سپس کره را بر این اساس منقبض یا منبسط می کنند تا زمانی که یک خلاف
واقع (پراکنده) پیدا شود و در نهایت برگردانده شود.

۵۰۳۵ ۵۰۳۶ ۵۰۳۷ ۵۰۳۸ ۵۰۳۹ مجریان توسط ریسیرو و همکاران. (۲۰۱۸) ۶۰ برعکس خلاف واقع هستند، به فصل مربوط به قوانین محدوده
(لنگرهای) مراجعه کنید.

۵۰۴۰

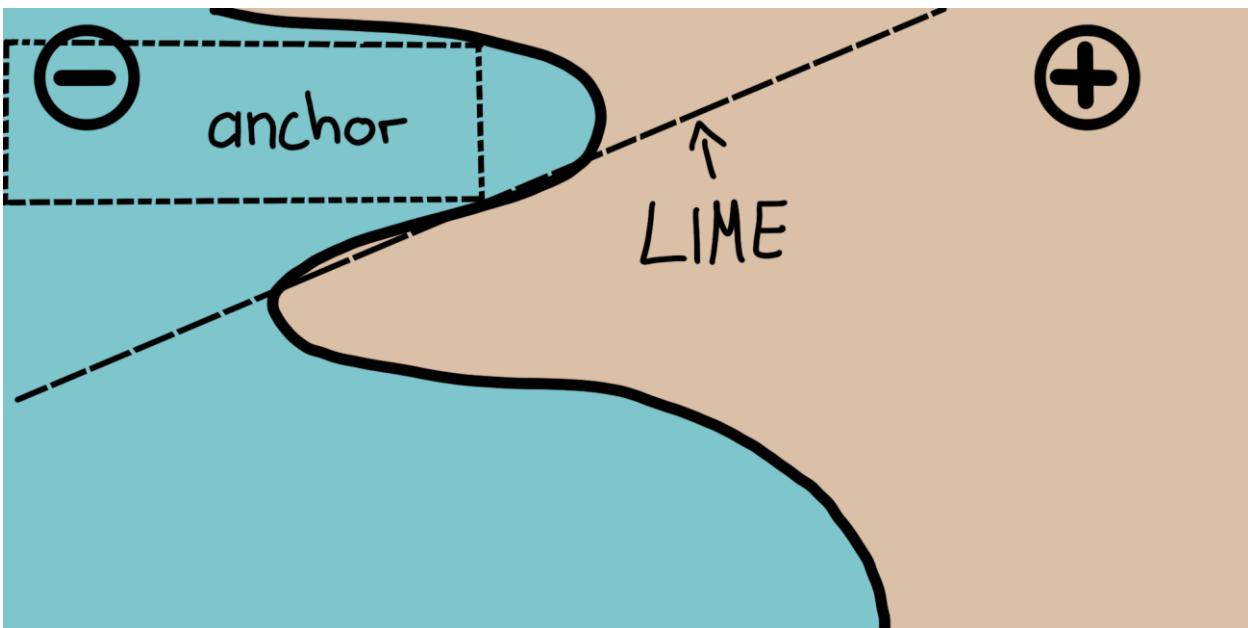
۹.۴ قوانین محدوده (لنگرها)

نویسنده‌گان : Tobias Goerke & Magdalena Lang

روش لنگر، پیش‌بینی‌های فردی هر مدل طبقه‌بندی جعبه سیاه را با یافتن یک قانون تصمیم‌گیری که پیش‌بینی را به اندازه کافی «لنگر» می‌کند، توضیح می‌دهد. در صورتی که تغییرات در سایر مقادیر ویژگی بر پیش‌بینی تأثیری نداشته باشد، یک قانون یک پیش‌بینی را ثابت می‌کند Anchors. از تکنیک‌های یادگیری تقویتی در ترکیب با الگوریتم جستجوی نمودار استفاده می‌کند تا تعداد تماس‌های مدل (و در نتیجه زمان اجرا مورد نیاز) را به حداقل کاهش دهد و در عین حال قادر به بازیابی از بهینه محلی باشد. ریپیرو، سینگ و گسترین این الگوریتم را در سال ۲۰۱۸ پیشنهاد کردند - همان محققانی که الگوریتم LIME را معرفی کردند.

مانند سلف خود، رویکرد لنگرها یک استراتژی مبتنی بر اغتشاش را برای ایجاد توضیحات محلی برای پیش‌بینی مدل‌های یادگیری ماشین جعبه سیاه به کار می‌گیرد. با این حال، به جای مدل‌های جایگزینی که توسط LIME استفاده می‌شود، توضیحات به دست آمده به عنوان قواعد IF-THEN قابل درک بیان می‌شوند که لنگر نامیده می‌شوند . این قوانین قابل استفاده مجدد هستند، زیرا دارای محدوده هستند: لنگرها شامل مفهوم پوشش است که دقیقاً در مورد کدام نمونه‌های دیگر، احتمالاً دیده نشده، اعمال می‌شود. یافتن لنگرها شامل یک مشکل اکتشافی یا راهزن چند مسلح است که منشا آن در رشته یادگیری تقویتی است. برای این منظور، همسایگان، یا اغتشاشات، برای هر نمونه‌ای که توضیح داده می‌شود، ایجاد و ارزیابی می‌شود. انجام این کار به رویکرد اجازه می‌دهد تا ساختار جعبه سیاه و پارامترهای داخلی آن را نادیده بگیرد تا این پارامترها هم مشاهده نشده و هم بدون تغییر باقی بمانند. بنابراین، الگوریتم مدل-آگنوستیک است ، به این معنی که می‌توان آن را برای هر کلاس از مدل اعمال کرد.

در مقاله خود، نویسنده‌گان هر دو الگوریتم خود را مقایسه می‌کنند و تجسم می‌کنند که چگونه این الگوریتم‌ها با همسایگی یک نمونه متفاوت مشورت می‌کنند تا نتایج را به دست آورند. برای این کار، شکل زیر هم LIME و هم لنگرها را به صورت محلی نشان می‌دهد که یک طبقه‌بندی‌کننده با اینتری پیچیده را توضیح می‌دهد (پیش‌بینی می‌کند - یا +) با استفاده از دو نمونه نمونه. نتایج LIME نشان نمی‌دهد که چقدر وفادار هستند، زیرا LIME تنها یک مرز تصمیم‌گیری خطی را می‌آموزد که بهترین مدل را با توجه به فضای اغتشاش تقریب می‌کند . D. با توجه به فضای اغتشاش یکسان، رویکرد لنگرها توضیحاتی را می‌سازد که پوشش آن با رفتار مدل تطبیق داده می‌شود و رویکرد به وضوح مرزهای آنها را بیان می‌کند. بنابراین، آنها از نظر طراحی وفادار هستند و دقیقاً برای کدام موارد معتبر هستند. این ویژگی لنگرها را بصری و آسان برای درک می‌کند.



شکل ۱۱: آهک در مقابل لنگرها - تجسم اسباب بازی. شکلی از ریبیرو، سینگ و گسترن (۲۰۱۸).

همانطور که قبلاً ذکر شد، نتایج یا توضیحات الگوریتم در قالب قوانینی به نام لنگر آمده است. مثال ساده زیر چنین لنگری را نشان می‌دهد. به عنوان مثال، فرض کنید یک مدل جعبه سیاه دو متغیره به ما داده می‌شود که پیش‌بینی می‌کند مسافری از فاجعه تایتانیک جان سالم به در برده است یا خیر. اکنون می‌خواهیم بدانیم چرا مدل برای یک فرد خاص پیش‌بینی می‌کند که زنده بماند. الگوریتم لنگرها یک توضیح نتیجه را مانند آنچه در زیر نشان داده شده است ارائه می‌دهد.

Feature	Value
Age	20
Sex	female
Class	first
Ticket price	300\$
More attributes	...
Survived	true

و توضیح لنگر مربوطه این است:

اگر $\text{SEX} = \text{female}$ و $\text{Class} = \text{first}$ و $\text{Survived} = \text{true}$ سپس $\text{Age} = 20$ و $\text{Ticket price} = 300\$$ پیش‌بینی کنید

این مثال نشان می‌دهد که چگونه لنگرها می‌توانند بینش‌های اساسی را در مورد پیش‌بینی یک مدل و استدلال زیربنایی آن ارائه دهند. نتیجه نشان می‌دهد که کدام ویژگی‌ها توسط مدل در نظر گرفته شده است که در این مورد، جنس زن و درجه یک است. انسان‌ها که برای صحت اهمیت دارند، می‌توانند از این قانون برای

اعتبارسنجی رفتار مدل استفاده کنند. لنگر علاوه بر این به ما می گوید که در ۱۵ درصد موارد فضای اغتشاش اعمال می شود. در این موارد، توضیح ۹۷٪ دقیق است، به این معنی که محمول های نمایش داده شده تقریباً به طور انحصاری مسئول نتیجه پیش بینی شده هستند.
 یک لنگر آ به طور رسمی به شرح زیر تعریف می شود:

$$\mathbb{E}_{\mathcal{D}_x(z|A)}[1_{f(x)=\hat{f}(z)}] \geq \tau, A(x) = 1$$
 که در آن:
 -ایکس نمونه ای را نشان می دهد که توضیح داده می شود (به عنوان مثال یک ردیف در مجموعه داده های جدولی).
 -آ مجموعه ای از محمولات است، یعنی قاعده یا لنگر حاصل، به گونه ای که آیکس = ۱ هنگامی که تمام محمولات ویژگی تعریف شده توسط آ مرتبط به ایکس مقادیر ویژگی
 نشان دهنده مدل طبقه بندی است که باید توضیح داده شود (به عنوان مثال یک مدل شبکه عصبی مصنوعی). می توان برای پیش بینی یک برچسب برای آن پرس و جو کرد
 ایکس و آشفتگی های آن
 |ایکس | آ (نشان دهنده توزیع همسایگان از ایکس ، تطابق آ . ک τ یک آستانه دقت را مشخص می کند. فقط قوانینی که حداقل به وفاداری محلی دست می یابند τ نتیجه معتبر محسوب می شوند.
 توصیف رسمی ممکن است ترسناک باشد و می تواند در کلمات بیان شود:
 با توجه به یک نمونه ایکس توضیح داده شود، یک قانون یا یک لنگر آ یافت می شود، به گونه ای که به آن اعمال می شود ایکس ، در حالی که همان کلاس برای ایکس حداقل برای کسری پیش بینی می شود τ از ایکس همسایگان که در آن همان آ قابل اجرا است. دقت یک قانون از ارزیابی همسایگان یا اغتشاشات ناشی می شود (به شرح زیر $A(x)$) (با استفاده از مدل یادگیری ماشین ارائه شده (که با تابع نشانگر مشخص می شود $f(x)$) $=f(x)$).

۹,۴,۱ یافتن لنگرها

اگرچه ممکن است توصیف ریاضی لنگرها واضح و ساده به نظر برسد، ساختن قوانین خاص غیرممکن است. نیاز به ارزیابی دارد $f(z) = \max_{z \in D} f(z)$ (برای همه $z \in D$) که در فضاهای ورودی پیوسته یا بزرگ امکان پذیر نیست. بنابراین، نویسندهای پیشنهاد می‌کنند که پارامتر را معرفی کنند $\delta \leq 0$ برای ایجاد یک تعریف احتمالی به این ترتیب، نمونه‌ها تا زمانی که اطمینان آماری در مورد دقت آنها وجود داشته باشد، ترسیم می‌شوند. تعریف احتمالی به شرح زیر است:

$$P(\text{prec}(A) \geq \tau) \geq 1 - \delta \quad \text{with} \quad \text{prec}(A) = \mathbb{E}_{D_s(z|A)}[1_{f(x)=f(z)}]$$

دو تعریف قبلی با مفهوم پوشش ترکیب شده و گسترش یافته است. منطق آن شامل یافتن قوانینی است که ترجیحاً برای بخش بزرگی از فضای ورودی مدل اعمال می‌شود. پوشش به طور رسمی به عنوان احتمال یک لنگر برای اعمال به همسایگانش، یعنی فضای اغتشاش آن تعریف می‌شود:

$$\text{cov}(A) = \mathbb{E}_{D_{(i)}}[A(z)]$$

گنجاندن این عنصر به تعریف نهایی انکر با در نظر گرفتن حداکثر کردن پوشش منجر می‌شود:

$$\max_{A \text{ s.t. } P(\text{prec}(A) \geq \tau) \geq 1 - \delta} \text{cov}(A)$$

بنابراین، روند رسیدگی برای قانونی تلاش می‌کند که بالاترین پوشش را در بین همه قوانین واجد شرایط داشته باشد (همه قوانینی که آستانه دقت را با توجه به تعریف احتمالی برآورده می‌کنند). تصور می‌شود که این قوانین مهم‌تر هستند، زیرا بخش بزرگ‌تری از مدل را توصیف می‌کنند. توجه داشته باشید که قوانین با محمول‌های بیشتر نسبت به قوانین با محمول‌های کمتر دقت بیشتری دارند. به طور خاص، قانونی که هر ویژگی را برطرف می‌کند ایکس همسایگی ارزیابی شده را به نمونه‌های یکسان کاهش می‌دهد. بنابراین، مدل همه همسایگان را به طور مساوی طبقه‌بندی می‌کند و دقت قاعده آن است ۱. در عین حال، قاعده‌ای که بسیاری از ویژگی‌ها را رفع می‌کند، بیش از حد خاص است و فقط برای چند نمونه قابل اجرا است. از این‌رو، بین دقت و پوشش تعادل وجود دارد.

رویکرد لنگرها از چهار جزء اصلی برای یافتن توضیحات استفاده می‌کند.

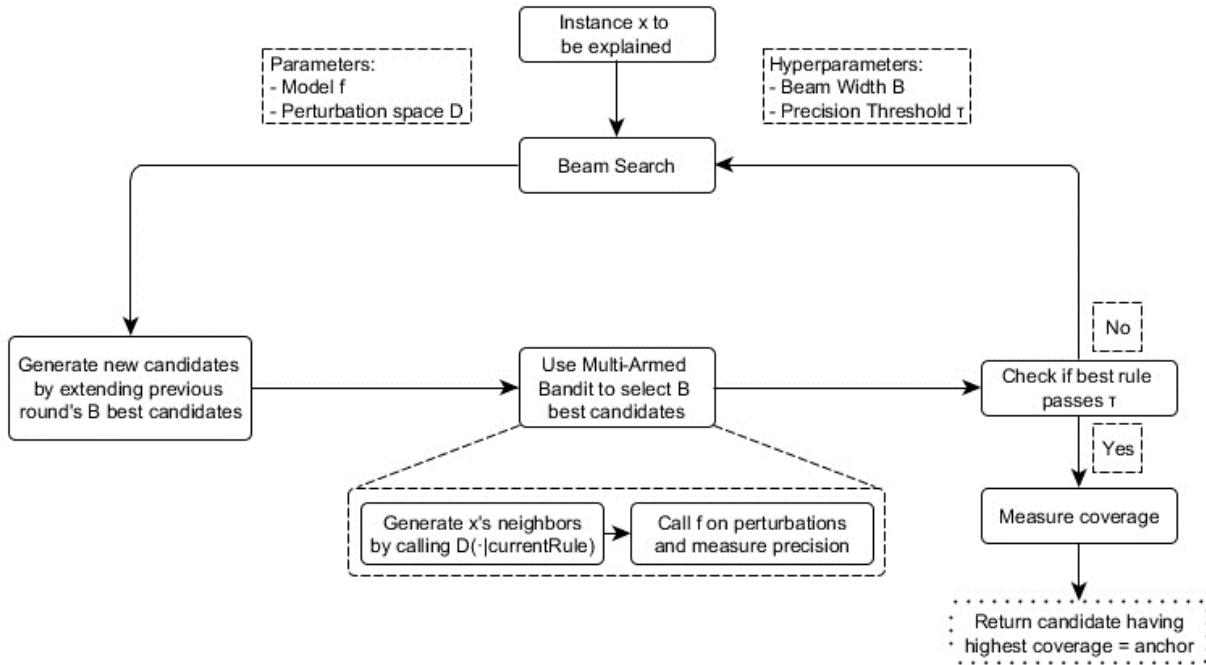
Candidate Generation : نامزدهای توضیح جدیدی را ایجاد می‌کند. در دور اول، یک نامزد در هر ویژگی از ایکس ایجاد می‌شود و مقدار مربوطه اغتشاشات احتمالی را برطرف می‌کند. در هر دور دیگر، بهترین نامزدهای دور قبلی با یک گزاره مشخصه که هنوز در آن وجود ندارد، گسترش می‌یابد.

-بهترین شناسایی کاندیدا : قوانین نامزد باید با توجه به اینکه کدام قانون توضیح می دهد با هم مقایسه شوند
ایکس بهترین. برای این منظور، آشتفتگی هایی که با قانون مشاهده شده در حال حاضر مطابقت دارند ایجاد و با
فراخوانی مدل ارزیابی می شوند. با این حال، این تماسها باید به حداقل برسد تا سربار محاسباتی محدود شود.
به همین دلیل است که در هسته این مؤلفه، یک راهزن چند مسلح با اکتشاف خالص (MAB)؛ KL-LUCB)
وجود دارد. به طور دقیق MAB (ها برای کاوش و بهرهبرداری مؤثر از استراتژی های مختلف (که در قیاس با
ماشین های اسلات بازو نامیده می شوند) با استفاده از انتخاب متوالی استفاده می شوند. در تنظیمات داده شده،
هر قانون نامزد باید به عنوان بازویی دیده شود که می توان آن را کشید. هر بار که کشیده می شود، همسایگان
مربوطه مورد ارزیابی قرار می گیرند، و از این طریق اطلاعات بیشتری در مورد بازده قانون نامزد به دست می
آوریم (دقت در مورد لنگر). بنابراین دقت بیان می کند که قاعده تا چه حد نمونه ای را که باید توضیح داده شود،
توصیف می کند.

-اعتبار سنجی دقیق کاندیدا : در صورتی که هنوز اطمینان آماری وجود نداشته باشد که نامزد بیش از حد مجاز
باشد، نمونه های بیشتری می گیرد ۲. آستانه.

-جستجوی پرتو اصلاح شده : همه اجزای فوق در یک جستجوی پرتو جمع آوری می شوند، که یک الگوریتم
جستجوی نمودار و گونه ای از الگوریتم عرض اول است. حمل می کند ب بهترین نامزدهای هر دور به دور بعدی
(جایی که ب عرض پرتو نامیده می شود . اینها ب سپس بهترین قوانین برای ایجاد قوانین جدید استفاده می
شود. جستجوی پرتو حداکثر انجام می شود $f_{\text{آتیتو}}(t)$ ها تا t (دور، زیرا هر ویژگی حداکثر یک بار
می تواند در یک قانون گنجانده شود. بنابراین، در هر دور من ، دقیقاً نامزدها را ایجاد می کند من محمول می ند
و بهترین B را انتخاب می کند. بنابراین با تنظیم ب بالا، الگوریتم به احتمال زیاد از بهینه محلی اجتناب می
کند. به نوبه خود، این به تعداد بالایی از فراخوانی های مدل نیاز دارد و در نتیجه بار محاسباتی را افزایش می
دهد.

این چهار جزء در شکل زیر نشان داده شده است.



شکل ۹،۱۲: اجزای الگوریتم لنگرها و روابط متقابل آنها (ساده شده)

۵۱۴۸

۵۱۴۹

۵۱۵۰ این رویکرد دستور العمل ظاهراً کاملی برای استخراج کارآمد اطلاعات آماری صحیح در مورد اینکه چرا هر
 ۵۱۵۱ سیستمی یک نمونه را به روشی که انجام می‌داد طبقه‌بندی می‌کند است. به طور سیستماتیک با ورودی مدل
 ۵۱۵۲ آزمایش می‌کند و با مشاهده خروجی‌های مربوطه نتیجه می‌گیرد. برای کاهش تعداد تماس‌های انجام‌شده با
 ۵۱۵۳ مدل، بر روش‌های یادگیری ماشینی (MABs) به خوبی تثبیت شده و تحقیق شده متکی است. این به نوبه
 ۵۱۵۴ خود، زمان اجرای الگوریتم را به شدت کاهش می‌دهد.

۵۱۵۵

۹،۴،۲ پیچیدگی و زمان اجرا

۵۱۵۶ دانستن رفتار زمان اجرا مجانبی رویکرد لنگرها به ارزیابی اینکه چقدر انتظار می‌رود در مسائل خاص عملکرد
 ۵۱۵۷ خوبی داشته باشد، کمک می‌کند. اجازه دهید ب نشان دهنده عرض تیر و p تعداد تمام ویژگی‌ها سپس
 ۵۱۵۸ الگوریتم لنگرها تابع موارد زیر است:

۵۱۵۹

$$\mathcal{O}(B \cdot p^3 + p^2 \cdot \mathcal{O}_{\text{MAB}[B,p,B]})$$

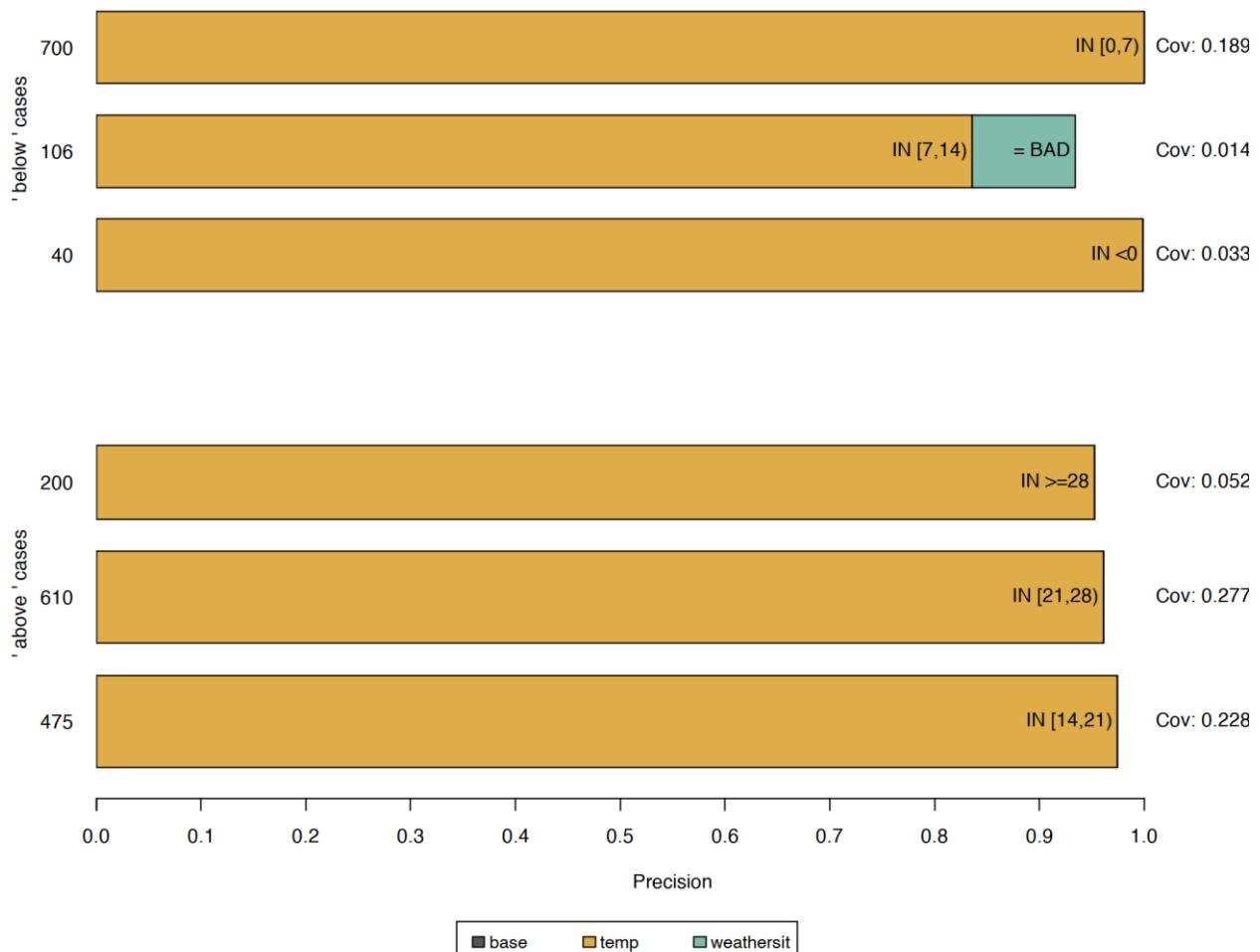
۵۱۶۰ این مرز از فرآپارامترهای مستقل از مسئله، مانند اطمینان آماری انتزاعی می‌شود. δ نادیده گرفتن
 ۵۱۶۱ فرآپارامترها به کاهش پیچیدگی مرز کمک می‌کند (برای اطلاعات بیشتر به مقاله اصلی مراجعه کنید). از
 ۵۱۶۲ آنجایی که MAB استخراج می‌کند ب بهترین از $B \cdot p$ کاندیداها در هر دور، اکثر MAB‌ها و زمان اجرا آنها را

۵۱۶۳ چند برابر می کنند پ ۲ فاکتور بیشتر از هر پارامتر دیگری بنابراین آشکار می شود: کارایی الگوریتم زمانی که
۵۱۶۴ ویژگی های زیادی وجود دارد کاهش می یابد.

۹.۴.۳ مثال داده های جدولی

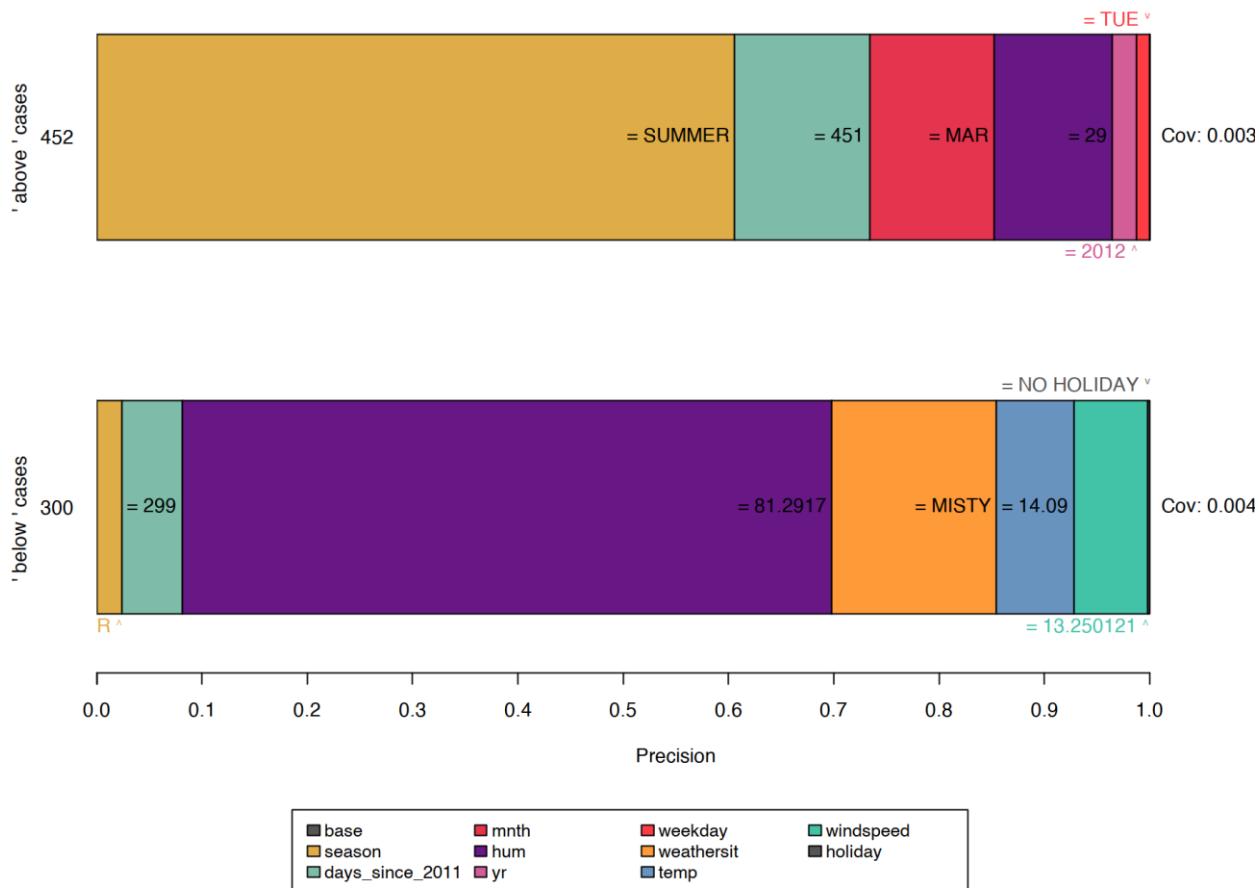
۵۱۶۵ داده های جدولی، داده های ساختاری هستند که با جداول نشان داده می شوند، که در آن ستون ها ویژگی ها و
۵۱۶۶ نمونه های ردیف ها را نشان می دهند. برای مثال، ما از داده های اجاره دوچرخه برای نشان دادن پتانسیل رویکرد
۵۱۶۷ لنگرها برای توضیح پیش بینی های ML برای نمونه های انتخاب شده استفاده می کنیم. برای این، ما رگرسیون را
۵۱۶۸ به یک مسئله طبقه بندی تبدیل می کنیم و یک جنگل تصادفی را به عنوان مدل جعبه سیاه خود آموزش می
۵۱۶۹ دهیم. این برای طبقه بندی است که آیا تعداد دوچرخه های کرایه شده بالاتر یا پایین تر از خط روند است.
۵۱۷۰

۵۱۷۱ قبل از ایجاد توضیحات لنگر، باید یک تابع اغتشاش تعریف شود. یک راه آسان برای انجام این کار استفاده از یک
۵۱۷۲ فضای آشفتگی پیش فرض بصری برای موارد توضیح جدولی است که می تواند با نمونه برداری از داده های
۵۱۷۳ آموزشی ساخته شود. هنگام ایجاد اختلال در یک نمونه، این رویکرد پیش فرض مقادیر ویژگی را که تابع
۵۱۷۴ محمولات لنگر هستند حفظ می کند، در حالی که ویژگی های غیر ثابت را با مقادیری که از نمونه های نمونه گیری
۵۱۷۵ تصادفی دیگری با احتمال مشخص گرفته شده جایگزین می کند. این فرآیند نمونه های جدیدی را به دست
۵۱۷۶ می دهد که مشابه موارد توضیح داده شده هستند، اما مقادیری را از نمونه های تصادفی دیگر اتخاذ کرده اند.
۵۱۷۷ بنابراین، آنها شبیه همسایگان نمونه توضیح داده شده هستند.



۵۱۷۸ شکل ۹,۱۳: لنگرهایی که شش نمونه از مجموعه داده های اجاره دوچرخه را توضیح می دهند. هر ردیف یک
 ۵۱۷۹ توضیح یا لنگر را نشان می دهد و هر نوار محمولات ویژگی موجود در آن را نشان می دهد. محور X دقت یک
 ۵۱۸۰ قانون را نشان می دهد و ضخامت یک میله با پوشش آن مطابقت دارد. قاعده «پایه» هیچ محمولی ندارد. این
 ۵۱۸۱ لنگرها نشان می دهند که مدل عمدتاً دما برای پیش بینی در نظر می گیرد.
 ۵۱۸۲
 ۵۱۸۳ نتایج به طور غریزی قابل تفسیر هستند و برای هر نمونه توضیح داده شده نشان می دهند که کدام ویژگی برای
 ۵۱۸۴ پیش بینی مدل مهمتر است. از آنجایی که لنگرها فقط دارای چند محمول هستند، علاوه بر این، پوشش بالایی
 ۵۱۸۵ دارند و از این رو در موارد دیگر کاربرد دارند. قوانین نشان داده شده در بالا با ایجاد شده است $\tau = 0.9$.
 ۵۱۸۶ بنابراین، ما از لنگرهایی درخواست می کنیم که اغتشاشات ارزیابی شده آنها به طور صادقانه برچسب را با دقت
 ۵۱۸۷ حداقل پشتیبانی کند. ۹۰٪ همچنین از گسسته سازی برای افزایش بیان و کاربرد ویژگی های عددی استفاده
 ۵۱۸۸ شد.

۵۱۸۹ همه قوانین قبلی برای نمونه هایی ایجاد شده اند که مدل با اطمینان بر اساس چند ویژگی تصمیم می گیرد. با
 ۵۱۹۰ این حال، نمونه های دیگر به طور مشخص توسط مدل طبقه بندی نمی شوند زیرا ویژگی های بیشتر اهمیت
 ۵۱۹۱ دارند. در چنین مواردی، لنگرهای خاص تر می شوند، ویژگی های بیشتری را شامل می شوند و برای نمونه های
 ۵۱۹۲ کمتری اعمال می شوند.



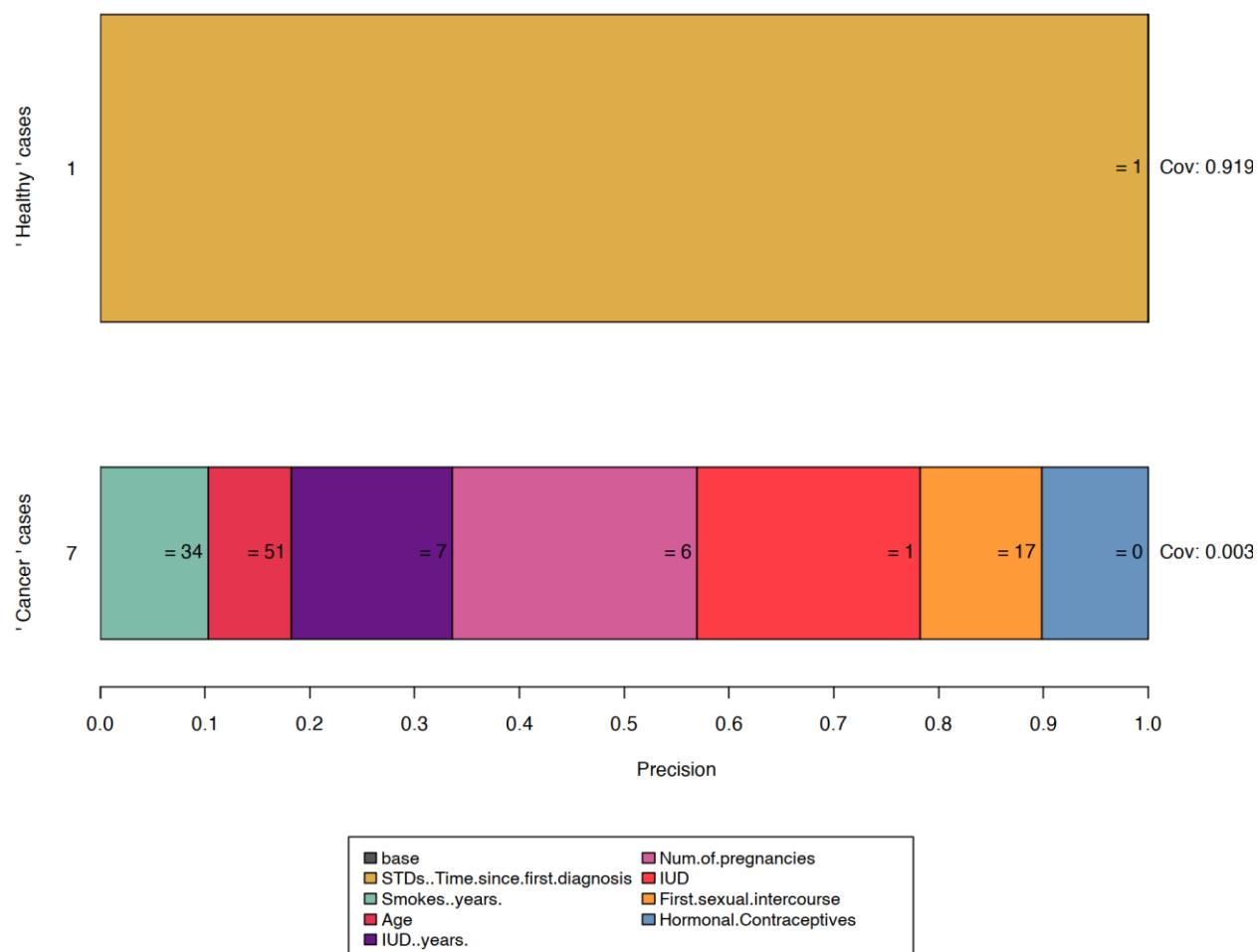
۵۱۹۳ شکل ۹,۱۴: توضیح موارد نزدیک به مرزهای تصمیم منجر به قوانین خاصی می شود که شامل تعداد بیشتری از
 ۵۱۹۴ محمولات ویژگی و پوشش کمتر است. همچنین، قانون خالی، یعنی ویژگی پایه، اهمیت کمتری پیدا می کند.
 ۵۱۹۵ این می تواند به عنوان یک سیگنال برای یک مرز تصمیم تفسیر شود، زیرا نمونه در یک محله فرار قرار دارد.
 ۵۱۹۶

۵۱۹۷ در حالی که انتخاب فضای اغتشاش پیش فرض یک انتخاب راحت است، ممکن است تأثیر زیادی بر الگوریتم
 ۵۱۹۸ داشته باشد و در نتیجه منجر به نتایج مغرضانه شود. به عنوان مثال، اگر مجموعه قطار نامتعادل باشد (تعداد
 ۵۱۹۹ نمونه های نامساوی از هر کلاس وجود دارد)، فضای اغتشاش نیز وجود دارد. این شرایط بیشتر بر قواعد یابی و
 ۵۲۰۰ دقیق نتیجه تأثیر می گذارد.

۵۲۰۱ مجموعه داده های سرطان دهانه رحم یک مثال عالی از این وضعیت است. اعمال الگوریتم لنگرها منجر به یکی
 ۵۲۰۲ از شرایط زیر می شود:

۵۲۰۳ - توضیح نمونه هایی که برچسب سالم دارند، قوانین خالی را به دست می دهد زیرا همه همسایگان تولید شده
 ۵۲۰۴ به سالم ارزیابی می کنند.

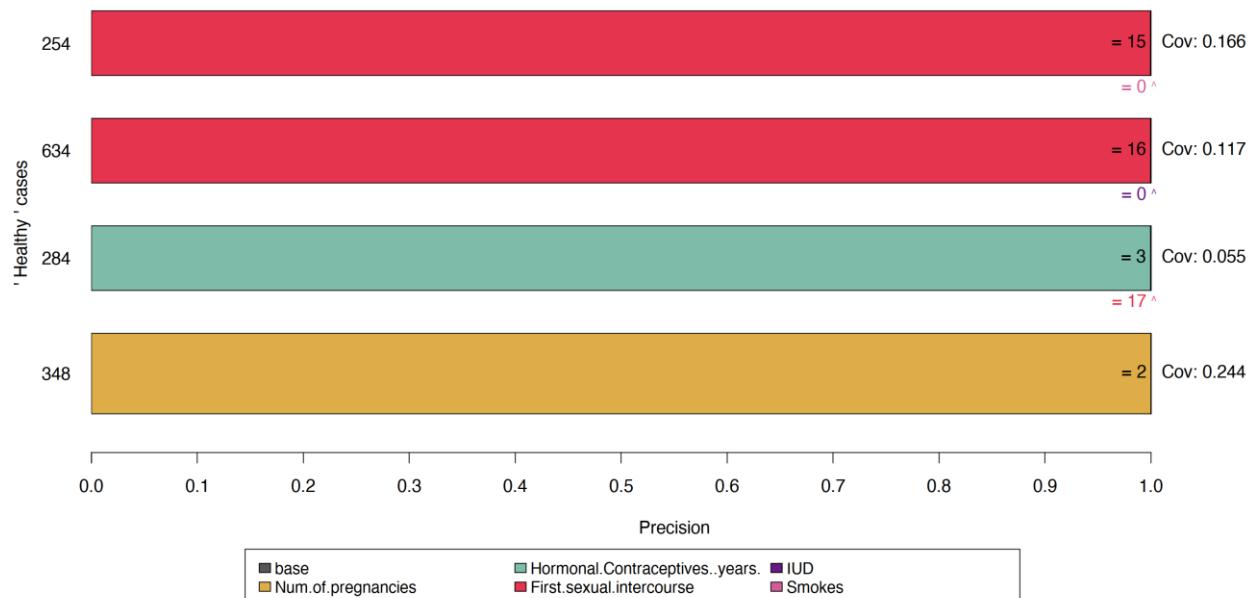
۵۲۰۵ - توضیحات برای نمونه هایی که سرطان برچسب گذاری شده اند، بیش از حد خاص هستند، به عنوان مثال،
 ۵۲۰۶ محموله های ویژگی بسیاری را شامل می شوند، زیرا فضای اختشاش عمدتاً مقدار نمونه های سالم را پوشش
 ۵۲۰۷ می دهد.



۵۲۰۸
 ۵۲۰۹ شکل ۹.۱۵: ساخت لنگرها در فضاهای اختشاش نامتعادل منجر به نتایج غیر قابل بیان می شود.

۵۲۱۰ این نتیجه ممکن است ناخواسته باشد و می توان به روش های مختلفی به آن نزدیک شد. به عنوان مثال، یک
 ۵۲۱۱ فضای اغتشاش سفارشی را می توان تعریف کرد. این اغتشاش سفارشی می تواند نمونه های متفاوتی داشته
 ۵۲۱۲ باشد، به عنوان مثال از یک مجموعه داده نامتعادل یا یک توزیع نرمال. با این حال، این یک عارضه جانبی دارد:
 ۵۲۱۳ همسایگان نمونه گیری شده نماینده نیستند و دامنه پوشش را تغییر می دهند. از طرف دیگر، می توانیم اطمینان
 ۵۲۱۴ را تغییر دهیم δ و مقادیر پارامتر خطای ϵ این امر باعث می شود MAB نمونه های بیشتری بکشد و در
 ۵۲۱۵ نهایت منجر به نمونه برداری بیشتر از اقلیت به صورت مطلق می شود.

۵۲۱۶ برای این مثال، ما از زیرمجموعه ای از مجموعه سلطان دهانه رحم استفاده می کنیم که در آن اکثر موارد سلطان
 ۵۲۱۷ بر چسب گذاری شده اند. سپس چارچوبی برای ایجاد یک فضای اغتشاش مربوطه از آن داریم. اغتشاشات در حال
 ۵۲۱۸ حاضر بیشتر به پیش بینی های مختلف منجر می شوند و الگوریتم لنگرها می تواند ویژگی های مهم را شناسایی
 ۵۲۱۹ کند. با این حال، باید تعریف پوشش را در نظر گرفت: این پوشش فقط در فضای اغتشاش تعریف می شود. در
 ۵۲۲۰ مثال های قبلی از مجموعه قطار به عنوان پایه فضای اغتشاش استفاده کردیم. از آنجایی که ما در اینجا فقط از
 ۵۲۲۱ یک زیرمجموعه استفاده می کنیم، پوشش بالا لزوماً نشان دهنده اهمیت بالای قانون در سطح جهانی نیست.



۵۲۲۲ شکل ۹,۱۶: متعادل کردن مجموعه داده ها قبل از ساختن لنگرها، استدلال مدل را برای تصمیم گیری در موارد
 ۵۲۲۳ اقلیت نشان می دهد.
 ۵۲۲۴

۹,۴,۴ مزایا

۵۲۲۵ رویکرد لنگرها مزایای متعددی را نسبت به LIME ارائه می دهد. اولاً، درک خروجی الگوریتم آسان تر است، زیرا
۵۲۲۶ قوانین به راحتی قابل تفسیر هستند (حتی برای افراد عادی).
۵۲۲۷

۵۲۲۸ علاوه بر این، لنگرها قابل تنظیم هستند و حتی با گنجاندن مفهوم پوشش، اندازه ای از اهمیت را بیان می کنند.
۵۲۲۹ دوم، رویکرد لنگرها زمانی کار می کند که پیش‌بینی‌های مدل غیرخطی یا پیچیده در همسایگی یک نمونه
۵۲۳۰ باشند. از آنجایی که این رویکرد به جای برازش مدل‌های جایگزین، تکنیک‌های یادگیری تقویتی را به کار
۵۲۳۱ می‌گیرد، احتمال کمتری دارد که مدل را نادیده بگیرد.

۵۲۳۲ جدای از آن، الگوریتم مدل-آگنوستیک است و بنابراین برای هر مدلی قابل استفاده است.

۵۲۳۳ علاوه بر این، بسیار کارآمد است زیرا می توان با استفاده از MAB هایی که از نمونه برداری دسته ای پشتیبانی
۵۲۳۴ می کنند (مثلًا BatchSAR موازی سازی کرد).

۹,۴,۵ معایب

۵۲۳۵ این الگوریتم از یک تنظیم بسیار قابل تنظیم و تاثیرگذار رنج می برد، درست مانند اکثر توضیح دهنده‌گان مبتنی
۵۲۳۶ بر اغتشاش. نه تنها فرآپارامترهایی مانند عرض پرتو یا آستانه دقت باید تنظیم شوند تا نتایج معنی داری به
۵۲۳۷ دست آورند، بلکه تابع اغتشاش نیز باید به صراحت برای یک دامنه/مورد استفاده طراحی شود. به این فکر کنید
۵۲۳۸ که چگونه داده های جدولی آشفته می شوند و به این فکر کنید که چگونه مفاهیم مشابه را در داده های
۵۲۳۹ تصویری اعمال کنید (نکته: اینها قابل اعمال نیستند). خوبشخтанه، رویکردهای پیش‌فرض ممکن است در برخی
۵۲۴۰ از حوزه‌ها (مثلًا جدولی) مورد استفاده قرار گیرند، که تنظیم توضیحات اولیه را تسهیل می کند.
۵۲۴۱

۵۲۴۲ همچنین، بسیاری از سناریوها نیاز به گستره سازی دارند، زیرا در غیر این صورت نتایج بسیار خاص هستند،
۵۲۴۳ پوشش کمی دارند و به درک مدل کمک نمی کنند. در حالی که گستره سازی می تواند کمک کند، اما اگر
۵۲۴۴ بی دقت استفاده شود، ممکن است مزهای تصمیم را محو کند و در نتیجه دقیقاً اثر معکوس داشته باشد. از
۵۲۴۵ آنجایی که بهترین تکنیک گستره سازی وجود ندارد، کاربران باید قبل از تصمیم گیری در مورد نحوه گستره
۵۲۴۶ سازی داده ها برای به دست آوردن نتایج ضعیف از داده ها آگاه باشند.

۵۲۴۷ ساختن لنگرها به فراخوانی های زیادی به مدل ML نیاز دارد ، درست مانند همه توضیح دهنده‌گان مبتنی بر
۵۲۴۸ اغتشاش. در حالی که الگوریتم MAB ها را برای به حداقل رساندن تعداد تماس ها مستقر می کند، زمان اجرای
۵۲۴۹ آن هنوز هم بسیار به عملکرد مدل بستگی دارد و بنابراین بسیار متغیر است.

در نهایت، مفهوم پوشش در برخی حوزه ها تعریف نشده است . به عنوان مثال، هیچ تعریف واضح یا جهانی در مورد اینکه چگونه سوپرپیکسل در یک تصویر با آن در سایر تصاویر مقایسه می شود، وجود ندارد.

۹.۴.۶ نرم افزار و جایگزین

در حال حاضر، دو پیادهسازی در دسترس است `anchor` :، یک بسته پایتون (همچنین توسط `Alibi` یکپارچه شده است) و یک پیادهسازی جاوا . اولی مرجع نویسندهان الگوریتم لنگرها و دومی یک پیاده سازی با کارایی بالا است که با یک رابط R به نام `anchors` ارائه می شود که برای مثال های این فصل استفاده شد. در حال حاضر، پیاده سازی `anchors` فقط از داده های جدولی پشتیبانی می کند. با این حال، لنگرها ممکن است از نظر تئوری برای هر دامنه یا نوع داده ای ساخته شوند.

۹,۵ ارزش های شپلی

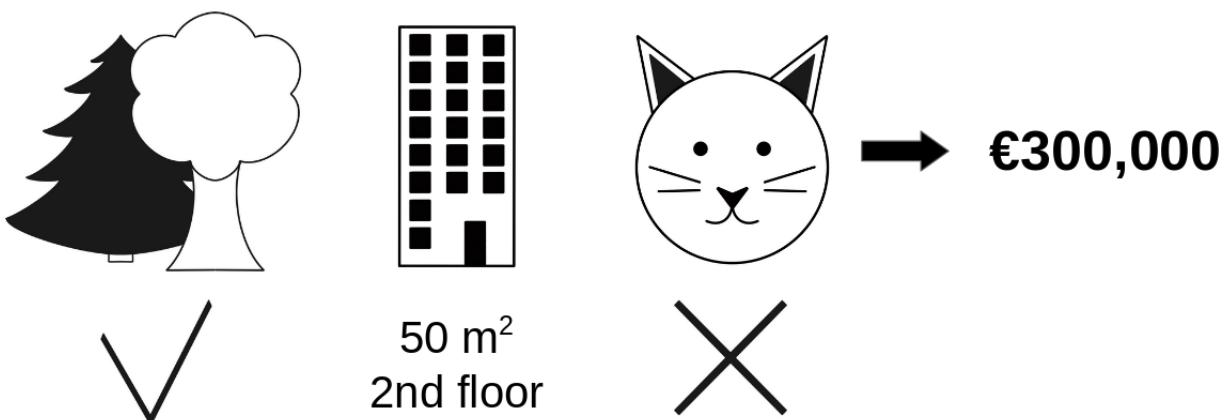
یک پیش‌بینی را می‌توان با این فرض توضیح داد که هر مقدار ویژگی نمونه یک «بازیکن» در بازی است که پیش‌بینی آن پرداخت است. مقادیر - Shapley روشی از تئوری بازی های ائتلافی - به ما می‌گوید که چگونه "پرداخت" را به طور عادلانه بین ویژگی ها توزیع کنیم.

به دنبال یک کتاب عمیق و کاربردی در مورد ارزش‌های SHAP و Shapley هستید؟ من شما را تحت پوشش قرار دادم.

۹,۶,۱ ایده کلی

سناریوی زیر را فرض کنید:

شما یک مدل یادگیری ماشینی برای پیش‌بینی قیمت آپارتمان آموزش داده اید. برای یک آپارتمان خاص ۳۰۰۰۰۰ یورو پیش‌بینی می‌شود و شما باید این پیش‌بینی را توضیح دهید. آپارتمان دارای مساحت ۵۰ متر مربع است ، در طبقه ۲ واقع شده است، دارای پارک در نزدیکی است و گربه ممنوع است:



شکل ۹,۱۷: قیمت پیش‌بینی شده برای ۵۰ متر ۲ آپارتمان طبقه دوم با پارک نزدیک و ممنوعیت گربه ۳۰۰۰۰۰ یورو است. هدف ما توضیح این است که چگونه هر یک از این مقادیر ویژگی به پیش‌بینی کمک کرده است.

میانگین پیش‌بینی برای همه آپارتمان ها ۳۱۰۰۰ یورو است. هر مقدار ویژگی در مقایسه با میانگین پیش‌بینی چقدر در پیش‌بینی نقش داشته است؟

پاسخ برای مدل های رگرسیون خطی ساده است. تأثیر هر ویژگی وزن ویژگی ضربدر مقدار ویژگی است. این فقط به دلیل خطی بودن مدل کار می کند. برای مدل های پیچیده تر، ما به راه حل متفاوتی نیاز داریم. به عنوان مثال، LIME مدل های محلی را برای تخمین اثرات پیشنهاد می کند. راه حل دیگر از نظریه بازی های مشارکتی ناشی می شود: ارزش Shapley ، که توسط (Shapley 1953) ابداع شد ، روشی برای تخصیص پرداخت ها به بازیکنان بسته به سهم آنها در کل پرداخت است. بازیکنان به صورت انتلافی همکاری می کنند و از این همکاری سود مشخصی دریافت می کنند.

بازیکنان؟ بازی؟ پرداخت؟ ارتباط با پیش بینی های یادگیری ماشین و قابلیت تفسیر چیست؟ "بازی" وظیفه پیش بینی برای یک نمونه واحد از مجموعه داده است. "سود" پیش بینی واقعی برای این مثال منهای پیش بینی میانگین برای همه موارد است. «بازیکن ها» مقادیر ویژگی نمونه ای هستند که برای دریافت سود (= مقدار معینی) را پیش بینی می کنند. در مثال آپارتمان ما، مقادیر ویژگی park-nearby ، و برای دستیابی به پیش بینی ۳۰۰۰۰۰ یورو با هم کار کردند. cat-banned هدف ما توضیح تفاوت بین پیش بینی واقعی (۳۰۰۰۰۰ یورو) و میانگین پیش بینی (۳۱۰۰۰۰ یورو) است: اختلاف ۱۰۰۰۰ یورو - area-50floor-2nd

2nd

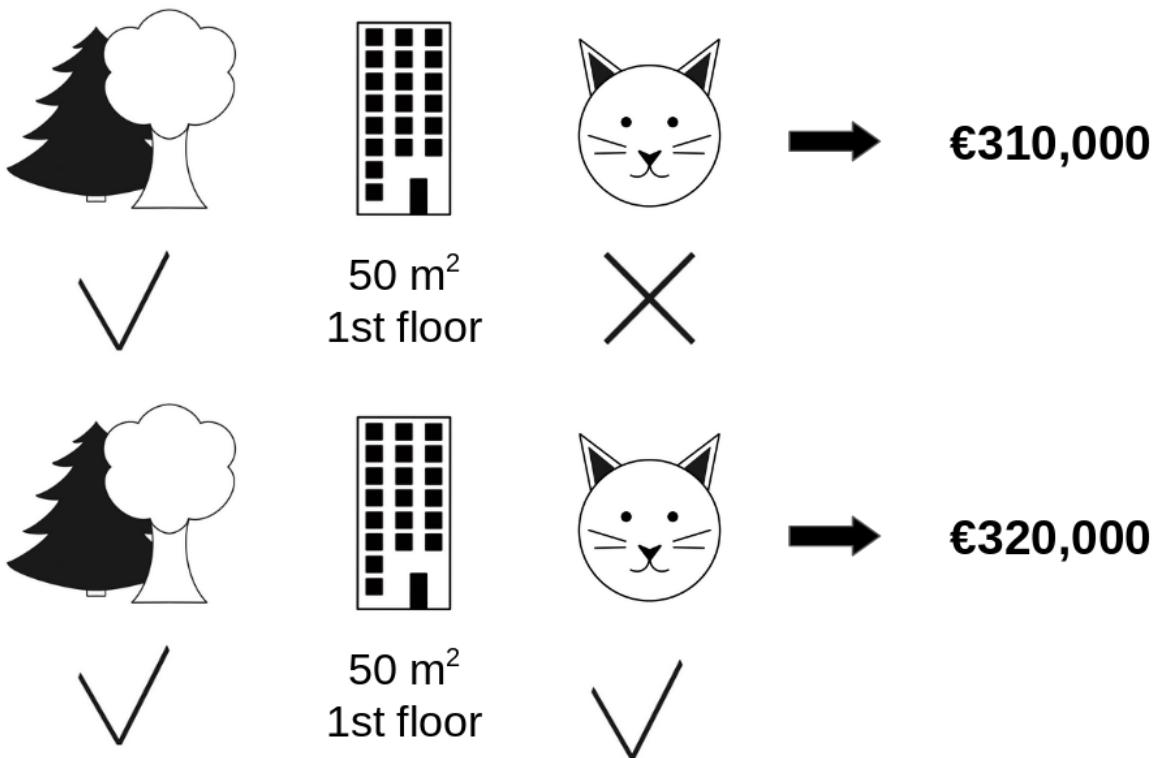
پاسخ می تواند این باشد park-nearby30000 : یورو کمک شده area-5010000 یورو کمک کرد . یورو کمک کرد floor-2nd0 یورو کمک کرد cat-banned کمک - ۵۰۰۰۰ یورو. مجموع کمک ها به - ۱۰۰۰۰ یورو می رسد، که پیش بینی نهایی منهای میانگین قیمت آپارتمان پیش بینی شده است.

چگونه مقدار Shapley را برای یک ویژگی محاسبه کنیم؟

مقدار Shapley میانگین سهم حاشیه ای یک مقدار ویژگی در تمام ائتلاف های ممکن است. الان همه چی واضحه؟

در شکل زیر سهم cat-banned مقدار ویژگی را هنگامی که به ائتلافی از park-nearby و اضافه می شود، ارزیابی می کنیم area-50. ما فقط آن را شبیه سازی می کنیم park-nearby و با رسم تصادفی آپارتمان دیگری از داده ها و استفاده از مقدار آن برای ویژگی طبقه، در یک ائتلاف هستیم cat-banned. area-50 مقدار floor-2nd به طور تصادفی ترسیم شده جایگزین شد. سپس قیمت آپارتمان را با این ترکیب (۳۱۰۰۰ یورو) پیش بینی می کنیم. در مرحله دوم، با جایگزین کردن آن با مقدار تصادفی ویژگی مجاز / ممنوع گربه از آپارتمان ترسیم شده به طور تصادفی، از ائتلاف حذف می کنیم. در مثال بود cat-allowed، اما می توانست cat-banned دوباره باشد. قیمت آپارتمان را برای ائتلاف park-nearby و پیش بینی می کنیم area-50(320000 یورو). سهم cat-banned310000 یورو - ۳۲۰۰۰ یورو = -

۵۳۰۳ یورو بود. این تخمین به مقادیر آپارتمانی که به طور تصادفی کشیده شده است، بستگی دارد که به عنوان
 ۵۳۰۴ «اهداکننده» برای مقادیر ویژگی‌های گریه و کف عمل می‌کرد. اگر این مرحله نمونه‌گیری را تکرار کنیم و
 ۵۳۰۵ مشارکت‌ها را میانگین‌گیری کنیم، تخمین‌های بهتری به دست خواهیم آورد.

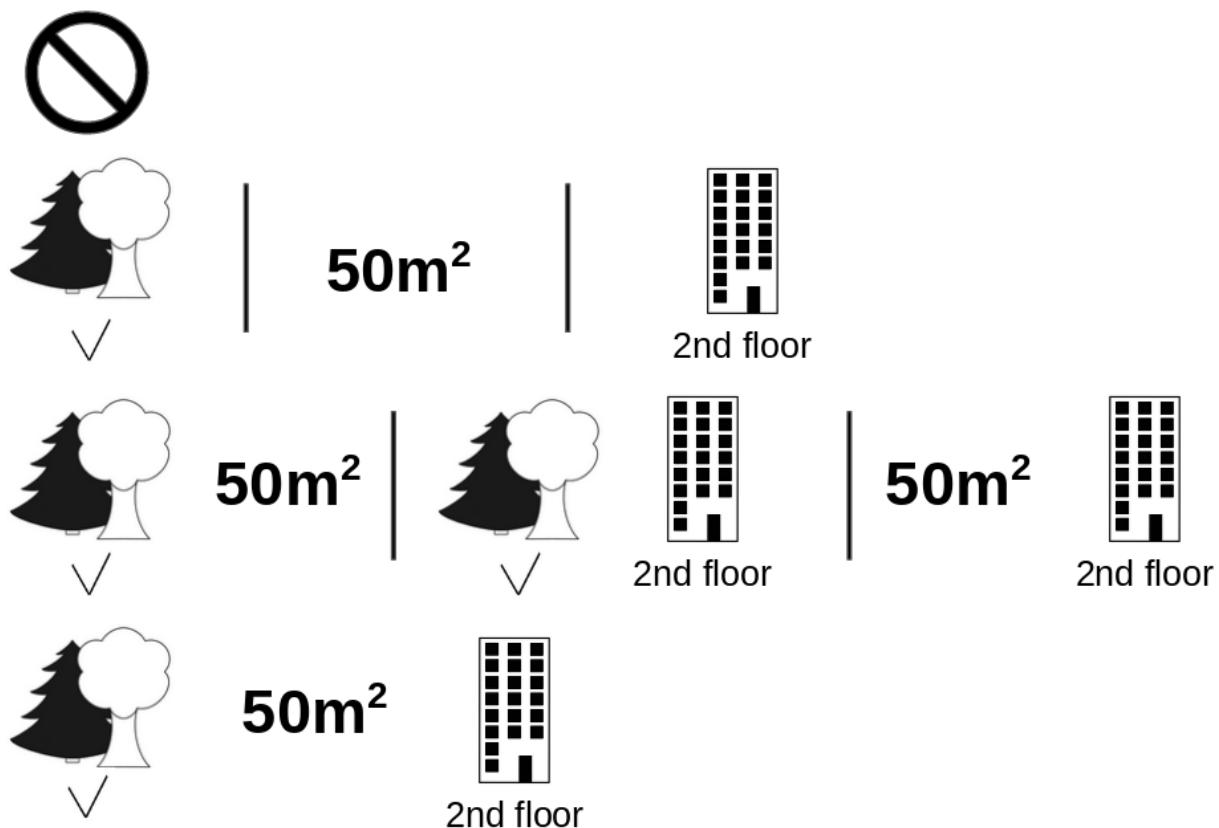


۵۳۰۶
 ۵۳۰۷ شکل ۹,۱۸: یک تکرار نمونه برای تخمین سهم cat-banned در پیش‌بینی هنگامی که به ائتلاف-
 ۵۳۰۸ area-50 و nearby اضافه می‌شود.
 ۵۳۰۹

ما این محاسبه را برای همه ائتلاف‌های ممکن تکرار می‌کنیم. مقدار Shapley میانگین تمام مشارکت‌های
 ۵۳۱۰ حاشیه‌ای به همه ائتلاف‌های ممکن است. زمان محاسبه با تعداد ویژگی‌ها به طور تصاعدی افزایش می‌یابد.
 ۵۳۱۱ یک راه حل برای مدیریت زمان محاسبات، محاسبه مشارکت تنها برای چند نمونه از ائتلاف‌های ممکن است.

۵۳۱۲ شکل زیر تمام ائتلاف‌های مقادیر ویژگی را نشان می‌دهد که برای تعیین مقدار Shapley برای-
 ۵۳۱۳ cat-banned. ردیف اول ائتلاف را بدون هیچ مقدار مشخصه نشان می‌دهد. ردیف‌های دوم، سوم و چهارم
 ۵۳۱۴ ائتلاف‌های متفاوتی را با افزایش اندازه ائتلاف نشان می‌دهند که با «» از هم جدا شده‌اند. در مجموع، ائتلاف
 ۵۳۱۵ های زیر ممکن است:

۵۳۱۶	
۵۳۱۷	No feature values
۵۳۱۸	park-nearby
۵۳۱۹	area-50
۵۳۲۰	floor-2nd
۵۳۲۱	park-nearby+area-50
۵۳۲۲	park-nearby+floor-2nd
۵۳۲۳	area-50+floor-2nd
۵۳۲۴	park-nearby+ area-50+ floor-2nd.
۵۳۲۵	برای هر یک از این ائتلاف‌ها، قیمت آپارتمان پیش‌بینی شده را با و بدون ارزش ویژگی محاسبه می‌کنیم- <i>cat</i> -
۵۳۲۶	و مابه التفاوت را می‌گیریم تا سهم نهایی را بدست آوریم. مقدار <i>Shapley</i> میانگین (وزنی) مشارکت
۵۳۲۷	های حاشیه‌ای است. ما مقادیر ویژگی‌هایی را که در ائتلاف نیستند با مقادیر ویژگی تصادفی از مجموعه
۵۳۲۸	داده آپارتمان جایگزین می‌کنیم تا از مدل یادگیری ماشینی پیش‌بینی کنیم.



شکل ۹.۱۹: همه ۸ ائتلاف مورد نیاز برای محاسبه مقدار دقیق Shapley مقدار cat-banned ویژگی.

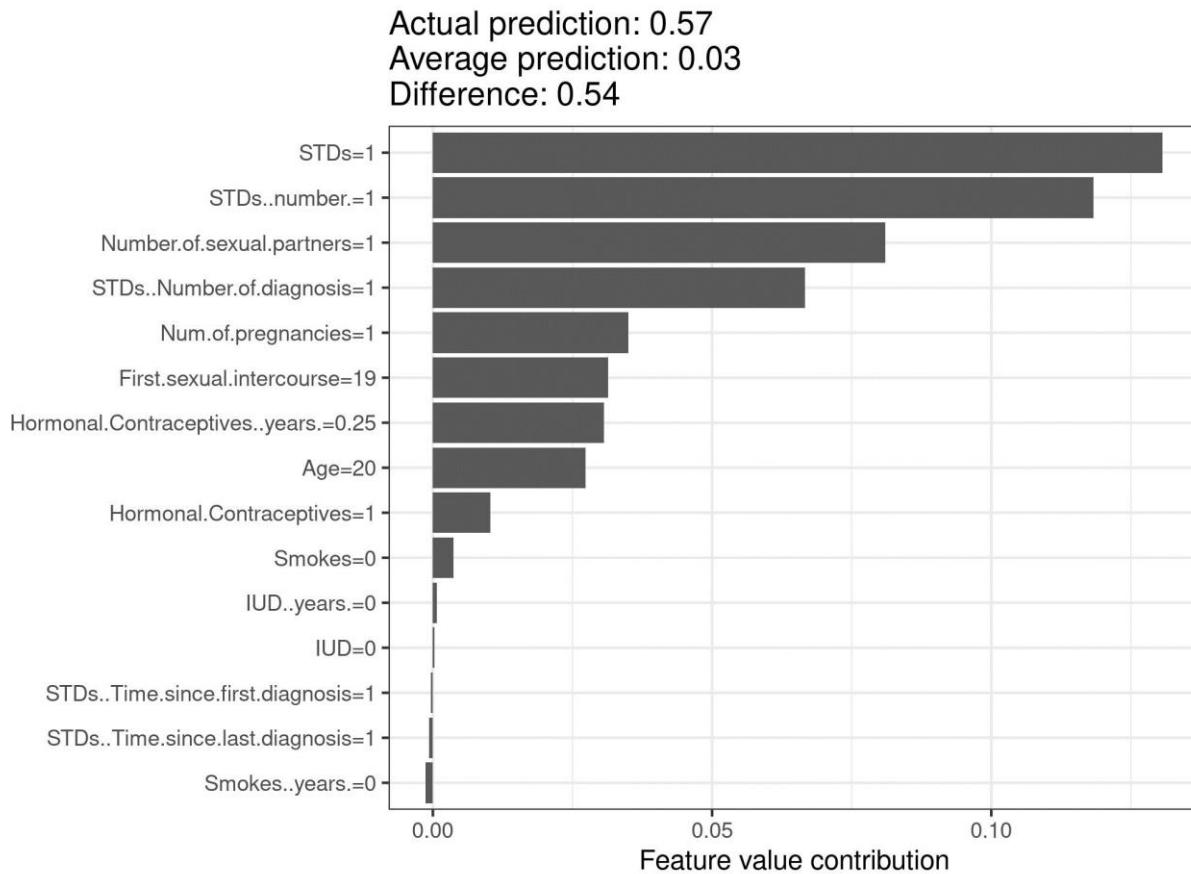
اگر مقادیر Shapley را برای همه مقادیر ویژگی تخمین بزنیم، توزیع کامل پیش بینی (منهای میانگین) را در بین مقادیر ویژگی بدست می آوریم.

۹.۵.۲ مثال ها و تفسیر

تفسیر مقدار Shapley برای مقدار ویژگی \hat{z} این است: مقدار- \hat{z} -امین ویژگی کمک شده است \hat{z} برای پیش بینی این نمونه خاص در مقایسه با میانگین پیش بینی مجموعه داده.

مقدار Shapley هم برای طبقه بندی (اگر با احتمالات سر و کار داریم) و هم برای رگرسیون کار می کند.

ما از مقدار Shapley برای تجزیه و تحلیل پیش بینی های یک مدل جنگل تصادفی که سرطان دهانه رحم را پیش بینی می کند، استفاده می کنیم:

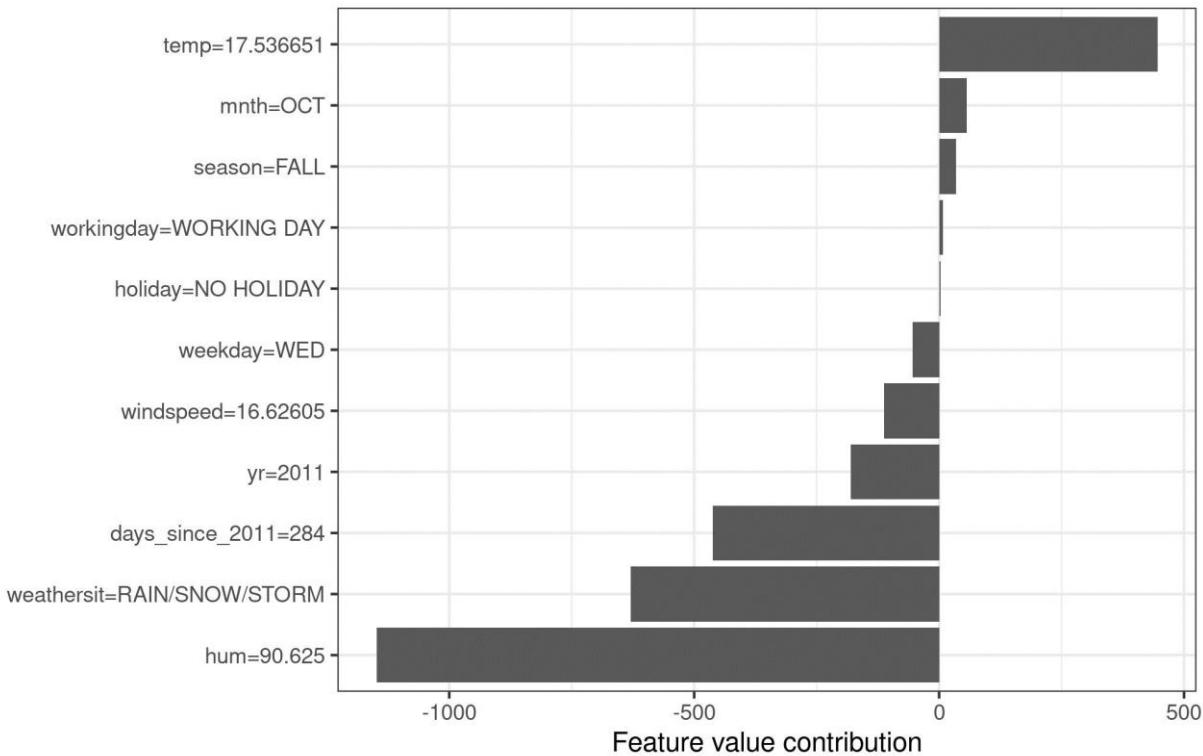


۵۳۳۹

شکل ۹,۲۰: مقادیر Shapley برای یک زن در مجموعه داده سرطان دهانه رحم. با پیش بینی ۰,۵۷، احتمال سرطان این زن ۰,۵۴ بالاتر از میانگین پیش بینی ۰,۰۳ است. تعداد STD های تشخیص داده شده احتمال را بیشتر افزایش می دهد. مجموع مشارکت ها تفاوت بین پیش بینی واقعی و متوسط (۰,۵۴) را نشان می دهد.

برای مجموعه داده های اجاره دوچرخه، ما همچنین یک جنگل تصادفی را آموزش می دهیم تا با توجه به اطلاعات آب و هوای و تقویم، تعداد دوچرخه های اجاره ای را برای یک روز پیش بینی کند. توضیحات ایجاد شده برای پیش بینی تصادفی جنگل یک روز خاص:

Actual prediction: 2409
 Average prediction: 4518
 Difference: -2108



۵۳۴۶

۵۳۴۷

شکل ۹,۲۱: مقدادیر Shapley برای روز ۲۸۵. با پیش بینی ۲۴۰۹ دوچرخه اجاره ای، این روز ۲۱۰۸- کمتر از میانگین پیش بینی ۴۵۱۸ است. وضعیت آب و هوا و رطوبت بیشترین سهم منفی را داشتند. دمای هوا در این روز سهم مثبتی داشت. مجموع مقدادیر شپلی تفاوت پیش بینی واقعی و میانگین (-۲۱۰۸) را به دست می دهد.

۵۳۵۰

۵۳۵۱

مراقب باشید که مقدار Shapley را به درستی تفسیر کنید: مقدار Shapley سهم متوسط یک مقدار ویژگی در پیش بینی در ائتلاف های مختلف است. مقدار Shapley تفاوتی در پیش بینی زمانی نیست که ویژگی را از مدل حذف کنیم.

۵۳۵۲

۵۳۵۳

۹,۵,۳ ارزش Shapley در جزئیات

۵۳۵۴

این بخش بیشتر به تعریف و محاسبه مقدار Shapley برای خواننده کنگکاو می رود. اگر به جزئیات فنی علاقه ندارید، از این بخش رد شوید و مستقیماً به "مزایا و معایب" بروید.

۵۳۵۶

۵۳۵۷

ما علاقه مندیم که ببینیم هر ویژگی چگونه بر پیش بینی یک نقطه داده تأثیر می گذارد. در یک مدل خطی محاسبه اثرات فردی آسان است. در اینجا یک پیش بینی مدل خطی برای یک نمونه داده به نظر می رسد:

$$\hat{f}(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

۵۳۵۸
۵۳۵۹ که در آن X نمونه ای است که می خواهیم مشارکت ها را برای آن محاسبه کنیم. هر یک ایکس یک مقدار
۵۳۶۰ ویژگی است، با $\beta_j = \sum_{i=1}^p \beta_i x_{ij}$ وزن مربوط به ویژگی j است.

۵۳۶۱ سهم j از ویژگی j ام در پیش بینی \hat{f} ایکس (است:

۵۳۶۲
$$\phi_j(\hat{f}) = \beta_j x_j - E(\beta_j X_j) = \beta_j x_j - \beta_j E(X_j)$$

۵۳۶۳ جایی که β_j ایکس (برآورد اثر میانگین برای ویژگی j است. سهم تفاوت بین اثر ویژگی منهای اثر متوسط
۵۳۶۴ است. خوب! اکنون می دانیم که هر ویژگی چقدر در پیش بینی نقش داشته است. اگر تمام مشارکت های ویژگی
۵۳۶۵ را برای یک نمونه جمع کنیم، نتیجه به شرح زیر است:

$$\begin{aligned} \sum_{j=1}^p \phi_j(\hat{f}) &= \sum_{j=1}^p (\beta_j x_j - E(\beta_j X_j)) \\ &= (\beta_0 + \sum_{j=1}^p \beta_j x_j) - (\beta_0 + \sum_{j=1}^p E(\beta_j X_j)) \\ &= \hat{f}(x) - E(\hat{f}(X)) \end{aligned}$$

۵۳۶۶ این مقدار پیش بینی شده برای نقطه داده x منهای میانگین مقدار پیش بینی شده است. مشارکت ویژگی می
۵۳۶۷ تواند منفی باشد.

۵۳۶۹ آیا می توانیم برای هر مدلی همین کار را انجام دهیم؟ بسیار عالی خواهد بود که این را به عنوان یک ابزار مدل-
۵۳۷۰ آگنوستیک داشته باشیم. از آنجایی که معمولاً در مدل های دیگر وزن مشابه نداریم، به راه حل متفاوتی نیاز
۵۳۷۱ داریم.

۵۳۷۲ کمک از مکان های غیرمنتظره می آید: نظریه بازی های مشارکتی. مقدار Shapley راه حلی برای محاسبه
۵۳۷۳ مشارکت ویژگی ها برای پیش بینی های منفرد برای هر مدل یادگیری ماشینی است.

۵۳۷۴ [Shapley ارزش](#) ۹,۵,۳,۱
۵۳۷۵ مقدار Shapley از طریق یکتابع مقدار تعریف می شود آل از بازیکنان در S .

۵۳۷۶ ارزش Shapley یک مقدار ویژگی، سهم آن در پرداخت است، وزن دهی شده و جمعبندی شده بر روی تمام
۵۳۷۷ ترکیب های ارزش ویژگی ممکن:

$$\phi_j(val) = \sum_{S \subseteq \{1, \dots, p\} \setminus \{j\}} \frac{|S|! (p - |S| - 1)!}{p!} (val(S \cup \{j\}) - val(S))$$

که در آن S زیرمجموعه‌ای از ویژگی‌های استفاده شده در مدل است، بردار مقادیر ویژگی نمونه مورد توضیح و p تعداد ویژگی‌ها است⁷. آلایکس(اس) پیش‌بینی مقادیر ویژگی در مجموعه S است که نسبت به ویژگی‌هایی که در مجموعه S نشده‌اند به حاشیه رفته‌اند:

$$val_x(S) = \int \hat{f}(x_1, \dots, x_p) d\mathbb{P}_{x \notin S} - E_X(\hat{f}(X))$$

شما در واقع چندین ادغام را برای هر ویژگی که حاوی S نیست انجام می‌دهید. یک مثال عینی: مدل یادگیری ماشین با 4×1 , $x_2 \times 3$, $x_3 \times 4$ کار می‌کند و ما پیش‌بینی ائتلاف S متشکل از مقادیر ویژگی 1×3 و 3×3 را ارزیابی می‌کنیم:

$$val_x(S) = val_x(\{1, 3\}) = \int_{\mathbb{R}} \int_{\mathbb{R}} \hat{f}(x_1, X_2, x_3, X_4) d\mathbb{P}_{X_2, X_4} - E_X(\hat{f}(X))$$

این به نظر می‌رسد شبیه به کمک‌های ویژگی در مدل خطی!

با کاربردهای زیاد کلمه "ارزش" گیج نشوید: مقدار ویژگی مقدار عددی یا مقوله‌ای یک ویژگی و نمونه است. مقدار Shapley سهم ویژگی در پیش‌بینی است.تابع ارزش تابع پرداخت برای ائتلاف بازیکنان (مقادیر ویژگی) است.

مقدار Shapley تنها روش انتساب است که ویژگی‌های Dummy, Symmetry, Efficiency و Additivity را برآورده می‌کند که در مجموع می‌توان آن‌ها را تعریفی از پرداخت منصفانه در نظر گرفت. کارایی سهم ویژگی‌ها باید با اختلاف پیش‌بینی برای x و میانگین جمع شود.

$$\sum_{j=1}^p \phi_j = \hat{f}(x) - E_X(\hat{f}(X))$$

تقارن سهم دو مقدار ویژگی j و k باید یکسان باشد اگر به طور مساوی در همه ائتلاف‌های ممکن مشارکت داشته باشند. اگر

$$val(S \cup \{j\}) = val(S \cup \{k\})$$

برای همه

$$S \subseteq \{1, \dots, p\} \setminus \{j, k\}$$

سپس

$$\phi_j = \phi_k$$

ساختگی ویژگی \hat{J} که مقدار پیش‌بینی شده را تغییر نمی‌دهد - صرف نظر از اینکه به کدام اختلاف مقادیر ویژگی اضافه می‌شود - باید مقدار Shapley 0 داشته باشد.

$$val(S \cup \{j\}) = val(S)$$

برای همه

$$S \subseteq \{1, \dots, p\}$$

سپس

$$\phi_j = 0$$

افزودنی برای یک بازی با پرداخت‌های ترکیبی $Shapley$ مربوطه به شرح زیر است:

$$\phi_j + \phi_j^+$$

فرض کنید شما یک جنگل تصادفی را آموزش داده اید، به این معنی که پیش‌بینی میانگین بسیاری از درختان تصمیم‌گیری است. ویژگی Additivity تضمین می‌کند که برای یک مقدار ویژگی، می‌توانید مقدار $Shapley$ را برای هر درخت به صورت جداگانه محاسبه کنید، آنها را میانگین بگیرید و مقدار $Shapley$ را برای مقدار ویژگی برای جنگل تصادفی دریافت کنید.

۹,۵,۳,۲ شهود

یک روش شهودی برای درک مقدار $Shapley$ ، تصویر زیر است: مقادیر ویژگی به ترتیب تصادفی وارد اتاق می‌شوند. همه مقادیر ویژگی در اتاق در بازی شرکت می‌کنند (= به پیش‌بینی کمک می‌کنند). مقدار $Shapley$ یک مقدار مشخصه میانگین تغییر در پیش‌بینی است که اختلاف موجود در اتاق زمانی که مقدار ویژگی به آنها می‌پیوندد دریافت می‌کند.

۹,۵,۳,۳ براورد ارزش $Shapley$

همه اختلاف‌ها (مجموعه‌های) ممکن از مقادیر ویژگی باید با و بدون ویژگی \hat{J} ام ارزیابی شوند تا مقدار دقیق $Shapley$ محاسبه شود. برای بیش از چند ویژگی، راه حل دقیق این مشکل مشکل ساز می‌شود زیرا تعداد اختلاف‌های احتمالی به طور تصاعدی افزایش می‌یابد و ویژگی‌های بیشتری اضافه می‌شود. استرامبلج و همکاران (۲۰۱۴) تقریبی را با نمونه گیری مونت کارلو پیشنهاد می‌کنند:

$$\hat{\phi}_j = \frac{1}{M} \sum_{m=1}^M \left(\hat{f}(x_{+j}^m) - \hat{f}(x_{-j}^m) \right)$$

جایی که $\wedge f$ ایکس‌متر ($j + \text{پیش‌بینی } X$) است، اما با تعدادی تصادفی از مقادیر ویژگی که با مقادیر ویژگی از نقطه داده تصادفی Z جایگزین شده‌اند، به جز مقدار مربوط به ویژگی j بردار X ایکس متر Z -تقریباً مشابه است ایکس مترا Z ، اما ارزش ایکس متر Z نمونه برداری شده است. هر یک از این نمونه‌های جدید نوعی «هیولای فرانکشتاین» است که از دو نمونه مونتاژ شده است. توجه داشته باشید که در الگوریتم زیر، ترتیب ویژگی‌ها عملاً تغییر نمی‌کند - هر ویژگی وقتی به تابع پیش‌بینی منتقل می‌شود در همان موقعیت برداری باقی می‌ماند. در اینجا از ترتیب فقط به عنوان یک «ترفند» استفاده می‌شود: با دادن یک نظم جدید به ویژگی‌ها، یک مکانیسم تصادفی دریافت می‌کنیم که به ما کمک می‌کند «هیولای فرانکشتاین» را کنار هم قرار دهیم. برای ویژگی‌هایی که در سمت چپ ویژگی ظاهر می‌شوند ایکس Z ، مقادیر را از مشاهدات اصلی می‌گیریم و برای ویژگی‌های سمت راست، مقادیر را از یک نمونه تصادفی می‌گیریم.

5435 تخمین تقریبی Shapley برای مقدار تک ویژگی:

5436 - خروجی: مقدار Shapley برای مقدار ویژگی Z

5437 - مورد نیاز: تعداد تکرار M ، نمونه مورد علاقه X ، شاخص ویژگی j ، ماتریس داده X و مدل یادگیری ماشین f

5438 -- برای همه $m = 1, \dots, M$

5439 --- نمونه تصادفی Z را از ماتریس داده X رسم کنید

5440 --- یک جایگشت تصادفی O از مقادیر ویژگی را انتخاب کنید

5441 --- نمونه سفارش: $X = (j_1, \dots, j_m)$ ، $i = (i_1, \dots, i_m)$

5442 --- سفارش نمونه: $Z = (z_1, \dots, z_m)$

5443 --- دو نمونه جدید بسازید

5444 --- با: $Z = (z_{i_1}, \dots, z_{i_m})$ و $Z' = (z_{i_1'}, \dots, z_{i_m'})$

5445 --- بدون: $Z = (z_{i_1}, \dots, z_{i_m})$ و $Z' = (z_{i_1'}, \dots, z_{i_m'})$

5446 --- محاسبه سهم حاشیه ای φ : $\varphi = f(Z) - f(Z')$

5447 -- مقدار Shapley را به عنوان میانگین محاسبه کنید: $\varphi_j = \frac{1}{m} \sum \varphi_i$

ابتدا یک نمونه مورد علاقه X ، یک ویژگی Z و تعداد تکرار M را انتخاب کنید. برای هر تکرار، یک نمونه تصادفی Z از داده ها انتخاب شده و ترتیب تصادفی ویژگی ها تولید می شود. دو نمونه جدید با ترکیب مقادیر از نمونه مورد علاقه X و نمونه Z ایجاد می شود. نمونه ایکس Z +نمونه مورد علاقه است، اما تمام مقادیر به ترتیب بعد از ویژگی Z با مقادیر ویژگی از نمونه Z جایگزین می شوند. نمونه ایکس Z -مث این هست که ایکس Z +، اما علاوه بر این ویژگی Z با مقدار ویژگی Z از نمونه Z جایگزین شده است. تفاوت پیش بینی از جعبه سیاه محاسبه می شود:

$$\phi_j^m = \hat{f}(x_{+j}^m) - \hat{f}(x_{-j}^m)$$

همه این تفاوت ها به طور میانگین محاسبه می شوند و به این نتیجه می رسد:

$$\phi_j(x) = \frac{1}{M} \sum_{m=1}^M \phi_j^m$$

میانگین گیری به طور ضمنی نمونه ها را با توزیع احتمال X وزن می کند.

این روش باید برای هر یک از ویژگی ها تکرار شود تا تمام مقادیر Shapley به دست آید.

۹,۵,۴ مزايا

تفاوت بین پیش بینی و پیش بینی میانگین به طور عادلانه بین مقادیر ویژگی نمونه توزیع می شود - ویژگی کارایی مقادیر Shapley. این ویژگی مقدار Shapley را از سایر روش ها مانند LIME متمایز می کند. تضمین نمی کند که پیش بینی به طور عادلانه بین ویژگی ها توزیع شده است. مقدار Shapley ممکن است تنها روش برای ارائه توضیح کامل باشد. در شرایطی که قانون مستلزم تبیین پذیری است - مانند "حق توضیح" اتحادیه اروپا - ارزش Shapley ممکن است تنها روش سازگار قانونی باشد، زیرا مبتنی بر یک نظریه محکم است و تأثیرات را به طور عادلانه توزیع می کند. من وکیل نیستم، بنابراین این فقط شهود من را در مورد الزامات منعکس می کند.

مقدار Shapley به توضیح متضاد اجازه می دهد . به جای مقایسه یک پیش بینی با میانگین پیش بینی کل مجموعه داده، می توانید آن را با یک زیر مجموعه یا حتی با یک نقطه داده مقایسه کنید. این تضاد نیز چیزی است که مدل های محلی مانند LIME ندارند.

مقدار Shapley تنها روش توضیحی با یک نظریه استوار است . بدیهیات - کارایی، تقارن، ساختگی، افزایشی - به توضیح یک پایه معقول می دهد. روش هایی مانند LIME رفتار خطی مدل یادگیری ماشین را به صورت محلی فرض می کنند، اما هیچ نظریه ای مبنی بر اینکه چرا این کار باید کار کند وجود ندارد.

توضیح دادن یک پیش بینی به عنوان یک بازی که توسط مقادیر ویژگی انجام می شود، شگفت آور است.

۵۴۷۳

۵۴۷۴ مقدار Shapley به زمان محاسباتی زیادی نیاز دارد . در ۹۹,۹ درصد از مشکلات دنیای واقعی، تنها راه حل
 ۵۴۷۵ تقریبی امکان پذیر است. محاسبه دقیق مقدار Shapley از نظر محاسباتی گران است زیرا $k \geq 2$ وجود
 ۵۴۷۶ دارد ائتلاف‌های احتمالی مقادیر ویژگی و «عدم وجود» یک ویژگی باید با ترسیم نمونه‌های تصادفی شبیه‌سازی
 ۵۴۷۷ شود، که واریانس تخمین مقادیر Shapley را افزایش می‌دهد. تعداد تصادفی ائتلاف‌ها با نمونه برداری از
 ۵۴۷۸ ائتلاف‌ها و محدود کردن تعداد تکرارها انجام می‌شود. کاهش M زمان محاسبه را کاهش می‌دهد، اما واریانس
 ۵۴۷۹ مقدار Shapley را افزایش می‌دهد. هیچ قانون سرانگشتی خوبی برای تعداد تکرارها وجود ندارد M . باید به
 ۵۴۸۰ اندازه کافی بزرگ باشد تا مقادیر Shapley را دقیقاً تخمین بزنند، اما به اندازه کافی کوچک باشد تا محاسبات را
 ۵۴۸۱ در یک زمان معقول کامل کند. انتخاب M بر اساس کرانف Chernoff باید امکان پذیر باشد، اما من هیچ مقاله
 ۵۴۸۲ ای در مورد انجام این کار برای مقادیر Shapley برای پیش‌بینی‌های یادگیری ماشین ندیده‌ام.

۵۴۸۳ مقدار Shapley را می‌توان به اشتباه تفسیر کرد . مقدار Shapley یک مقدار ویژگی، تفاوت مقدار
 ۵۴۸۴ پیش‌بینی‌شده پس از حذف ویژگی از آموزش مدل نیست. تفسیر مقدار Shapley این است: با توجه به مجموعه
 ۵۴۸۵ فعلی مقادیر ویژگی، سهم یک مقدار ویژگی در تفاوت بین پیش‌بینی واقعی و میانگین پیش‌بینی، مقدار تخمینی
 ۵۴۸۶ است Shapley.

۵۴۸۷ اگر به دنبال توضیحات پراکنده هستید (توضیحاتی که ویژگی‌های کمی دارند)، مقدار Shapley روش توضیح
 ۵۴۸۸ اشتباهی است. توضیحات ایجاد شده با روش ارزش Shapley همیشه از تمام ویژگی‌ها استفاده می‌کنند .
 ۵۴۸۹ انسان‌ها توضیحات انتخابی را ترجیح می‌دهند، مانند آنچه توسط LIME ارائه شده است . ممکن است
 ۵۴۹۰ انتخاب بهتری برای توضیحاتی باشد که افراد غیرمتخصص باید با آن سروکار داشته باشند. راه حل دیگر
 ۵۴۹۱ است که توسط Lundberg و Lee (2016) معرفی شده است که بر اساس مقدار Shapley است، اما می‌
 ۵۴۹۲ تواند توضیحاتی را با ویژگی‌های کمی ارائه دهد.

۵۴۹۳ مقدار Shapley یک مقدار ساده برای هر ویژگی برمی‌گرداند، اما هیچ مدل پیش‌بینی مانند LIME ندارد. این
 ۵۴۹۴ بدان معناست که نمی‌توان از آن برای بیان تغییرات در پیش‌بینی تغییرات در ورودی استفاده کرد، مانند: "اگر
 ۵۴۹۵ سالانه ۳۰۰ یورو بیشتر درآمد داشته باشم، امتیاز اعتبری من ۵ امتیاز افزایش می‌یابد".

۵۴۹۶ یکی دیگر از معایب این است که اگر می‌خواهید مقدار Shapley را برای یک نمونه داده جدید محاسبه کنید،
 ۵۴۹۷ نیاز به دسترسی به داده‌ها دارید . دسترسی به تابع پیش‌بینی کافی نیست، زیرا به داده‌ها نیاز دارید تا
 ۵۴۹۸ بخش‌هایی از نمونه مورد علاقه را با مقادیر نمونه‌های تصادفی داده‌ها جایگزین کنید. تنها در صورتی می‌توان از

این امر جلوگیری کرد که بتوانید نمونه های داده ای ایجاد کنید که شبیه نمونه های داده واقعی هستند اما نمونه های واقعی از داده های آموزشی نیستند.

مانند بسیاری از روش های دیگر تفسیر مبتنی بر جایگشت، روش ارزش Shapley از گنجاندن نمونه های داده غیر واقعی رنج می برد. وقتی ویژگی ها با هم مرتبط هستند برای شبیه سازی اینکه یک مقدار مشخصه در یک ائتلاف وجود ندارد، ویژگی را به حاشیه می برمی. این با نمونه برداری از مقادیر توزیع حاشیه ای ویژگی به دست می آید. این تا زمانی که ویژگی ها مستقل باشند خوب است. وقتی ویژگی ها وابسته هستند، ممکن است مقادیر مشخصه ای را که برای این نمونه معنی ندارند نمونه برداری کنیم. اما ما از آنها برای محاسبه مقدار Shapley ویژگی استفاده می کنیم. یک راه حل ممکن است این باشد که ویژگی های همبسته را با هم جابجا کنید و یک مقدار Shapley متقابل برای آنها بدست آورید. انطباق دیگر، نمونه گیری مشروط است: ویژگی ها مشروط به ویژگی هایی که قبلاً در تیم هستند، نمونه برداری می شوند. در حالی که نمونه گیری شرطی مشکل نقاط داده غیر واقعی را برطرف می کند، یک مسئله جدید معرفی می شود: مقادیر حاصل دیگر مقادیر Shapley برای بازی ما نیستند، زیرا آنها اصل تقارن را نقض می کنند. همانطور که Sundararajan و همکاران دریافتند. (۲۰۱۹) ۶۶ و بیشتر توسط Janzing و همکاران مورد بحث قرار گرفته است. (۲۰۲۰) ۶۷

۹.۵ نرم افزار و جایگزین

مقادیر Shapley در هر دو بسته iml و بسته fastshap برای R پیاده سازی می شوند. در Julia، می توانید از استفاده کنید Shapley.jl.

یک روش تخمین جایگزین برای مقادیر Shapley، در فصل بعدی ارائه شده است.

رویکرد دیگری شکسته نام دارد که در breakDown بسته R 68 پیاده سازی شده است . BreakDown همچنین سهم هر ویژگی را در پیش بینی نشان می دهد، اما آنها را گام به گام محاسبه می کند. اجازه دهید از قیاس بازی دوباره استفاده کنیم: ما با یک تیم خالی شروع می کنیم، مقدار ویژگی را که بیشترین سهم را در پیش بینی دارد اضافه می کنیم و تا زمانی که همه مقادیر ویژگی اضافه شوند، تکرار می کنیم. این که هر مقدار ویژگی چقدر کمک می کند به مقادیر ویژگی مربوطه بستگی دارد که قبلاً در "تیم" وجود دارد، که اشکال بزرگ روش شکست است. این سریعتر از روش ارزش Shapley است و برای مدل های بدون تعامل، نتایج یکسان است.

۹.۶ SHapley توضیحات افزودنی

SHapley توضیحات افزودنی (SHAP) توسط لوندبرگ و لی (۲۰۱۷) ۶۹ روشی برای توضیح پیش‌بینی‌های فردی است SHAP. بر اساس بازی از لحاظ نظری بهینه مقادیر Shapley است.

به دنبال یک کتاب عمیق و کاربردی در مورد ارزش‌های Shapley و SHAP هستید؟ من شما را تحت پوشش قرار دادم.

دو دلیل وجود دارد که SHAP فصل خودش را دارد و زیرفصل مقادیر Shapley نیست. ابتدا، نویسنده‌گان SHAP KernelSHAP را پیشنهاد کردند، یک رویکرد تخمینی جایگزین و مبتنی بر هسته برای مقادیر Shapley با الهام از مدل‌های جایگزین محلی. و آنها TreeSHAP را پیشنهاد کردند، یک رویکرد تخمین کارآمد برای مدل‌های مبتنی بر درخت. دوم، SHAP با بسیاری از روش‌های تفسیر جهانی مبتنی بر تجمعیع مقادیر Shapley ارائه می‌شود. این فصل هم رویکردهای برآوردهای جدید و هم روش‌های تفسیر جهانی را توضیح می‌دهد.

توصیه می‌کنم ابتدا فصل‌های مربوط به مقادیر Shapley و مدل‌های محلی (LIME) را بخوانید.

۹.۶.۱ تعریف

هدف SHAP توضیح پیش‌بینی یک نمونه \mathbf{x} با محاسبه سهم هر ویژگی در پیش‌بینی است. روش توضیح SHAP مقادیر Shapley را از تئوری بازی‌های ائتلافی محاسبه می‌کند. مقادیر ویژگی یک نمونه داده به عنوان بازیکن در یک ائتلاف عمل می‌کند. مقادیر Shapley به ما می‌گویند که چگونه «پرداخت» (= پیش‌بینی) را بین ویژگی‌ها به طور عادلانه توزیع کنیم. یک پخش کننده می‌تواند یک مقدار ویژگی فردی باشد، به عنوان مثال برای داده‌های جدولی. یک بازیکن همچنین می‌تواند گروهی از مقادیر ویژگی باشد. به عنوان مثال برای توضیح یک تصویر، پیکسل‌ها را می‌توان به سوپرپیکسل‌ها گروه بندی کرد و پیش‌بینی را بین آنها توزیع کرد. یکی از نوآوری‌هایی که SHAP در جدول آورده است این است که توضیح مقدار Shapley به عنوان یک روش انتساب ویژگی افزودنی، یک مدل خطی نشان داده می‌شود. این نما مقادیر LIME و Shapley را به هم متصل می‌کند SHAP. توضیح را به صورت زیر مشخص می‌کند:

$$g(\mathbf{z}') = \phi_0 + \sum_{j=1}^M \phi_j z'_j$$

جایی که g مدل توضیحی است، $\{\mathbf{z}'\}_{j=0}^M$ بردار ائتلاف است، M حداکثر اندازه ائتلاف و ϕ آر انتساب ویژگی برای یک ویژگی j ، مقادیر Shapley است. آنچه من "بردار ائتلاف" می‌نامم در مقاله "SHAP" ویژگی‌های ساده شده" نامیده می‌شود. فکر می‌کنم این نام انتخاب شده است، زیرا برای مثال داده‌های تصویر، تصاویر

در سطح پیکسل نمایش داده نمی‌شوند، بلکه در سوپرپیکسل‌ها جمع می‌شوند. فکر می‌کنم در مورد \mathbb{Z} به عنوان توصیف کننده ائتلاف‌ها مفید است: در بردار ائتلاف، ورودی ۱ به این معنی است که مقدار ویژگی مربوطه "حالا" و ۰ "غایب" است. اگر در مورد مقادیر Shapley بدانید، این باید برای شما آشنا به نظر برسد. برای محاسبه مقادیر Shapley، شبیه سازی می‌کنیم که فقط برخی از مقادیر ویژگی در حال پخش هستند ("حال") و برخی دیگر نیستند ("غایب"). نمایش به عنوان یک مدل خطی از ائتلاف ترفندی برای محاسبه است $S' \Phi$ برای x ، نمونه مورد علاقه، بردار ائتلاف x' بردار همه ۱‌ها است، یعنی همه مقادیر ویژگی "حال" هستند. فرمول به این صورت ساده می‌شود:

$$g(x') = \phi_0 + \sum_{j=1}^M \phi_j$$

شما می‌توانید این فرمول را با نماد مشابه در فصل ارزش Shapley پیدا کنید. اطلاعات بیشتر در مورد برآورده واقعی بعداً ارائه می‌شود. اجازه دهید ابتدا در مورد خواص آن صحبت کنیم Φ قبل از اینکه به جزئیات تخمین آنها بپردازیم.

مقادیر Shapley تنها راه حلی است که ویژگی‌های کارایی، تقارن، ساختگی و افزایش را برآورده می‌کند. همچنین این موارد را برآورده می‌کند، زیرا مقادیر Shapley را محاسبه می‌کند. در مقاله SHAP تفاوت‌هایی بین ویژگی‌های Shapley و ویژگی‌های SHAP خواهد دید. سه ویژگی مطلوب زیر را شرح می‌دهد:

۱ (دقت محلی)

$$\hat{f}(x) = g(x') = \phi_0 + \sum_{j=1}^M \phi_j x'_j$$

اگر تعریف کنید $\Phi = E[f(X)]$ (و همه را تنظیم کنید ایکس j به ۱)، این ویژگی کارایی Shapley است. فقط با نامی دیگر و با استفاده از بردار ائتلاف.

$$\hat{f}(x) = \phi_0 + \sum_{j=1}^M \phi_j x'_j = E_X(\hat{f}(X)) + \sum_{j=1}^M \phi_j$$

۲ (غیبت)

$$x'_j = 0 \Rightarrow \phi_j = 0$$

می‌گوید که یک ویژگی از دست رفته یک نسبت صفر دریافت می‌کند. توجه داشته باشید که ایکس j به ائتلاف‌هایی اشاره دارد که در آن مقدار ۰ نشان دهنده عدم وجود یک مقدار ویژگی است. در نماد

۵۵۷۴ ائتلاف، همه مقادیر ویژگی‌ها ایکس z "نمونه‌ای که باید توضیح داده شود باید '۱' باشد. وجود ۰ به این معنی
 ۵۵۷۵ است که مقدار ویژگی برای نمونه مورد علاقه وجود ندارد. این ویژگی در میان ویژگی‌های مقادیر «عادی» «
 ۵۵۷۶ Shapley نیست. پس چرا برای SHAP به آن نیاز داریم؟ لوندبرگ آن را "مالیات جزئی دفترداری" می‌نامد.
 ۵۵۷۷ یک ویژگی از دست رفته می‌تواند - در تئوری - دارای یک مقدار Shapley دلخواه بدون آسیب رساندن به
 ۵۵۷۸ ویژگی دقت محلی باشد، زیرا با ضرب می‌شود ایکس $= j^0 \dots$. ویژگی Missingness باعث می‌شود که
 ۵۵۷۹ ویژگی‌های گمشده یک مقدار Shapley برابر با ۰ دریافت کنند. در عمل، این فقط برای ویژگی‌هایی که ثابت
 ۵۵۸۰ هستند مرتبط است.

۵۵۸۱ (سازگاری

۵۵۸۲ اجازه دهد f^0 ایکس $(z) = f^0(z)$ ساعت ایکس $((z))$ و $-z^j$ نشان می‌دهد که $z^j = 0$. برای هر دو مدل f و f' که
 ۵۵۸۳ ارضا شوند:

$$f'_x(z') - f'_x(z'_{-j}) \geq \hat{f}_x(z') - \hat{f}_x(z'_{-j})$$

۵۵۸۴ برای همه ورودی‌ها $z^0, z^1 \in \{0, 1\}^m$ ، سپس:

۵۵۸۶ $\phi_j(\hat{f}', x) \geq \phi_j(\hat{f}, x)$
 ۵۵۸۷ ویژگی سازگاری می‌گوید که اگر یک مدل تغییر کند به طوری که سهم حاشیه‌ای یک مقدار مشخصه افزایش
 ۵۵۸۸ یابد یا ثابت بماند (صرف نظر از سایر ویژگی‌ها)، مقدار Shapley نیز افزایش می‌یابد یا ثابت می‌ماند. از
 ۵۵۸۹ Symmetry و Dummy، Shapley Linearity، Consistency ویژگی‌های دنبال می‌شوند، همانطور که
 ۵۵۹۰ در ضمیمه Lee و Lundberg توضیح داده شده است.

۵۵۹۱ KernelSHAP^{۹,۶,۲}

۵۵۹۲ برای یک مثال x سهم هر یک از ویژگی‌ها در پیش‌بینی را تخمین می‌زند KernelSHAP
 ۵۵۹۳ شامل پنج مرحله است:

۵۵۹۴ - ائتلاف‌های نمونه z^k که $k = 1, \dots, m$ ، $z^k = 1$ (ویژگی موجود در ائتلاف، ۰ = ویژگی وجود ندارد).

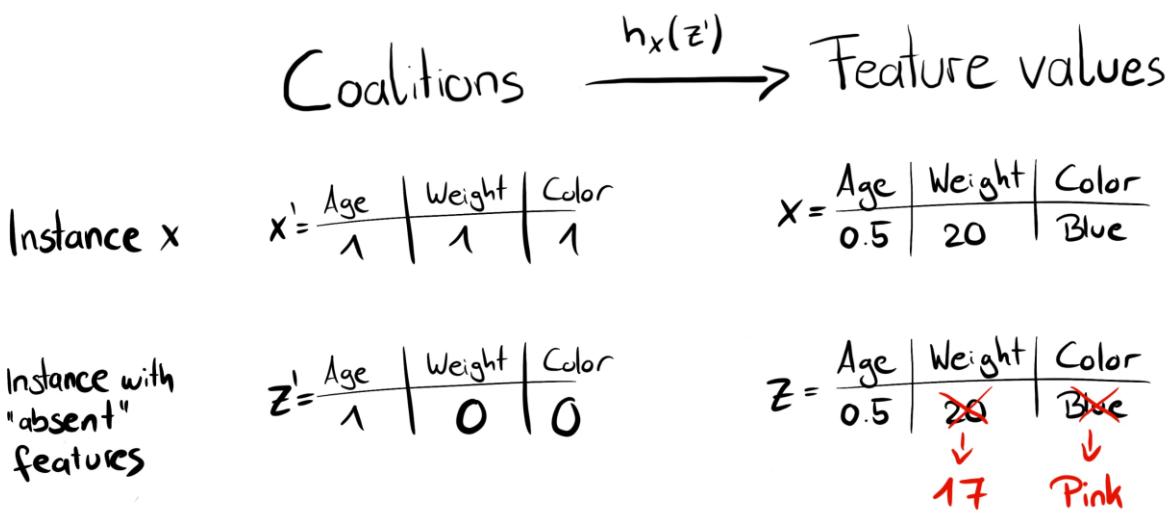
۵۵۹۵ - برای هر کدام پیش‌بینی دریافت کنید z که با اولین تبدیل z' که فضای ویژگی اصلی و سپس اعمال مدل
 ۵۵۹۶ $(z^k)^{f^0, f^1}$ ساعت ایکس (z) که

۵۵۹۷ وزن هر کدام را محاسبه کنید z که با هسته SHAP.

۵۵۹۸ مدل خطی وزنی متناسب.

۵۵۹۹ مقادیر Shapley را برگردانید که ضرایب از مدل خطی.

۵۶۰۰ می‌توانیم با چرخش‌های مکرر سکه یک ائتلاف تصادفی ایجاد کنیم تا زمانی که زنجیره‌ای از ۰ و ۱ داشته باشیم.
۵۶۰۱ به عنوان مثال، بردار $(1, 0, 0)$ به این معنی است که ما یک ائتلاف از ویژگی‌های اول و سوم داریم. ائتلاف‌های
۵۶۰۲ نمونه K به مجموعه داده‌ای برای مدل رگرسیون تبدیل می‌شوند. هدف مدل رگرسیون، پیش‌بینی یک ائتلاف
۵۶۰۳ است. (شما می‌گویید «صبر کنید!» «مدل روی این داده‌های ائتلاف باین‌ری آموزش ندیده است و نمی‌تواند
۵۶۰۴ برای آنها پیش‌بینی کند.») برای رسیدن از ائتلاف مقادیر ویژگی به نمونه‌های داده معتبر، به یکتابع نیاز
۵۶۰۵ داریم. ساعت ایکس $= z^{(z)}$ جایی که ساعت ایکس $= z^{(1)}$ کارکرد ساعت‌ایکس ۱ ها را به مقدار متناظر
۵۶۰۶ از نمونه X که می‌خواهیم توضیح دهیم نگاشت می‌کند. برای داده‌های جدولی، ۰ ها را به مقادیر نمونه دیگری
۵۶۰۷ که از داده‌ها نمونه برداری می‌کنیم، نگاشت می‌کند. این بدان معنی است که ما "ارزش ویژگی وجود ندارد" را
۵۶۰۸ با "مقدار ویژگی با مقدار ویژگی تصادفی از داده‌ها جایگزین شده است" برابر می‌کنیم. برای داده‌های جدولی،
۵۶۰۹ شکل زیر نگاشت از ائتلاف‌ها به مقادیر ویژگی را به تصویر می‌کشد:



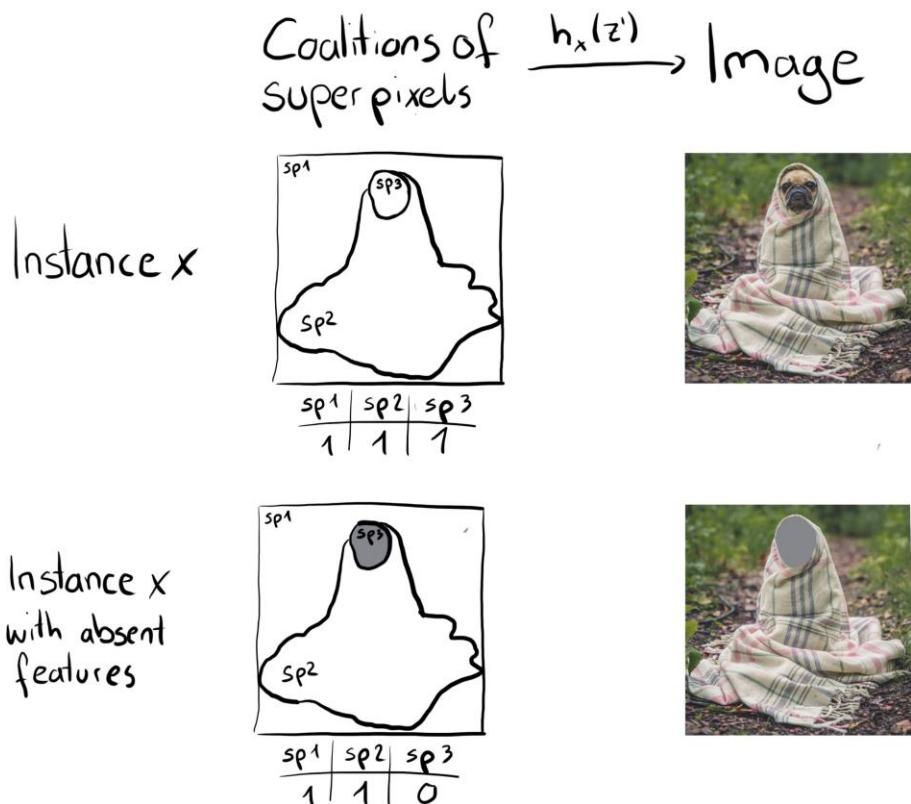
۵۶۱۰ ۵۶۱۱ شکل ۹,۲۲: تابع h_x یک ائتلاف را به یک نمونه معتبر نگاشت می‌کند. برای ویژگی‌های فعلی (۱)،
۵۶۱۲ h_x به مقادیر ویژگی X نگاشت می‌شود. برای ویژگی‌های غایب (۰)، h_x به مقادیر یک نمونه داده
۵۶۱۳ نمونه‌گیری تصادفی نگاشت می‌شود.

۵۶۱۴ ساعت ایکس برای داده‌های جدولی ویژگی رفتار می‌کند ایکس (۰ ایکس) (۱-سایر ویژگی‌ها) به عنوان مستقل و
۵۶۱۵ ادغام بر توزیع حاشیه‌ای:

$$\hat{f}(h_z(z')) = E_{X \sim j}[\hat{f}(x)]$$

نمونه برداری از توزیع حاشیه‌ای به معنای نادیده گرفتن ساختار وابستگی بین ویژگی‌های موجود و غایب است.
بنابراین KernelSHAP از مشکلی مشابه با همه روش‌های تفسیر مبتنی بر جایگشت رنج می‌برد. این تخمین به
موارد غیر محتمل اهمیت زیادی می‌دهد. نتایج می‌توانند غیر قابل اعتماد شوند. اما نمونه برداری از توزیع
حاشیه‌ای ضروری است. راه حل نمونه برداری از توزیع شرطی است که تابع مقدار و در نتیجه بازی را که مقادیر
Shapley راه حل آن است تغییر می‌دهد. در نتیجه، مقادیر Shapley تفسیر متفاوتی دارند: برای مثال، یک
ویژگی که ممکن است اصلاً توسط مدل استفاده نشده باشد، می‌تواند یک مقدار Shapley غیر صفر داشته باشد
که از نمونه‌گیری شرطی استفاده می‌شود. برای بازی حاشیه‌ای، این مقدار ویژگی همیشه یک مقدار Shapley
برابر با ۰ دریافت می‌کند.

برای تصاویر، شکل زیر یک تابع نقشه برداری ممکن را توضیح می‌دهد:



۵۶۲۶

شکل ۹.۲۳:تابع $\$h_x$ ائتلاف های سوپرپیکسل ها (sp) را به تصاویر ترسیم می کند. سوپرپیکسل ها گروهی از پیکسل ها هستند. برای ویژگی های فعلی (۱)، $\$h_x$ قسمت مربوطه از تصویر اصلی را برمی گرداند. برای ویژگی های غایب (۰)، $\$h_x$ ناحیه مربوطه را خاکستری می کند. تعیین میانگین رنگ پیکسل های اطراف یا موارد مشابه نیز یک گزینه خواهد بود.

تفاوت بزرگ با LIME وزن نمونه ها در مدل رگرسیون است LIME. نمونه ها را با توجه به نزدیک بودن آنها به نمونه اصلی وزن می کند. هر چه ۰ در بردار ائتلاف بیشتر باشد، وزن در LIME کوچکتر است SHAP. نمونه های نمونه برداری شده را با توجه به وزنی که ائتلاف در تخمین ارزش Shapley بدست می آورد وزن می کند. ائتلاف های کوچک (چند ۱) و ائتلاف های بزرگ (یعنی بسیاری از ۱) بیشترین وزن را دارند. شهود پشت آن این است: ما بیشتر در مورد ویژگی های فردی می آموزیم اگر بتوانیم اثرات آنها را به صورت جدأگانه مطالعه کنیم. اگر یک ائتلاف از یک ویژگی واحد تشکیل شده باشد، می توانیم در مورد تأثیر اصلی جدا شده این ویژگی بر پیش‌بینی بیاموزیم. اگر یک ائتلاف از همه ویژگی ها به جز یک ویژگی تشکیل شده باشد، می توانیم در مورد تأثیر کلی این ویژگی (اثر اصلی به اضافه تعاملات ویژگی) اطلاعات کسب کنیم. اگر یک ائتلاف از نیمی از ویژگی ها تشکیل شده باشد، ما اطلاعات کمی در مورد سهم یک ویژگی فردی می دانیم، زیرا ائتلاف های احتمالی زیادی با نیمی از ویژگی ها وجود دارد. برای دستیابی به وزن بندی مطابق با شیلی، لوندبرگ و همکاران. پیشنهاد هسته: SHAP

$$\pi_x(z') = \frac{(M-1)}{\binom{M}{|z'|} |z'| (M-|z'|)}$$

در اینجا، M حداکثر اندازه ائتلاف و $|z'|$ تعداد ویژگی های موجود در مثال z' . لوندبرگ و لی نشان می دهند که رگرسیون خطی با این وزن هسته، مقادیر شیلی را به دست می دهد. اگر از هسته LIME با SHAP در داده های ائتلاف استفاده کنید، LIME مقادیر Shapley را نیز تخمین می زند!

ما می توانیم در مورد نمونه گیری از ائتلاف ها کمی هوشمندتر باشیم: کوچک ترین و بزرگترین ائتلاف ها بیشترین وزن را به خود اختصاص می دهند. ما با استفاده از برحی از بودجه نمونه برداری K برای گنجاندن این ائتلاف های پر وزن به جای نمونه برداری کورکرانه، تخمین های ارزش شیلی بهتری دریافت می کنیم. ما با همه ائتلاف های ممکن با ویژگی های ۱ و $M-1$ شروع می کنیم که در مجموع ۲ برابر M ائتلاف می شود. وقتی بودجه کافی باقی مانده است (بودجه فعلی $2M - K$ است)، می توانیم ائتلاف هایی با ۲ ویژگی و با ویژگی های $M-2$ وغیره را شامل کنیم. از اندازه های ائتلاف باقی مانده، با وزن های تنظیم شده مجدد نمونه برداری می کنیم.

ما داده ها، هدف و وزن ها را داریم. همه چیزهایی که برای ساختن مدل رگرسیون خطی وزنی خود نیاز داریم:

$$g(z') = \phi_0 + \sum_{j=1}^M \phi_j z'_j$$

۵۶۵۳

۵۶۵۴

ما مدل خطی g را با بهینه سازی تابع ضرر L زیر آموزش می دهیم:

۵۶۵۵

۵۶۵۶

که در آن Z داده های آموزشی است. این مجموع خسته کننده قدیمی خطاهای مربعی است که ما معمولاً برای مدل های خطی بهینه می کنیم. ضرایب تخمینی مدل، s_j 's، مقادیر Shapley هستند.

۵۶۵۸

۵۶۵۹

۵۶۶۰

۵۶۶۱

از آنجایی که ما در یک تنظیم رگرسیون خطی هستیم، می توانیم از ابزارهای استاندارد برای رگرسیون نیز استفاده کنیم. برای مثال، می توانیم اصطلاحات منظم سازی را اضافه کنیم تا مدل پراکنده شود. اگر یک پنالتی L1 به ضرر L اضافه کنیم، می توانیم توضیحات پراکنده ایجاد کنیم. (من خیلی مطمئن نیستم که آیا ضرایب حاصل هنوز هم مقادیر Shapley معتبر هستند یا خیر.).

۵۶۶۲

TreeSHAP ۹,۶,۳

۵۶۶۳

۵۶۶۴

۵۶۶۵

۵۶۶۶

لوندبرگ و همکاران (۲۰۱۸) TreeSHAP پیشنهادی، گونه‌ای از SHAP برای مدل‌های یادگیری ماشینی مبتنی بر درخت مانند درخت‌های تصمیم‌گیری، جنگل‌های تصادفی و درخت‌های تقویت‌شده گرادیان . TreeSHAP به عنوان یک جایگزین سریع و مخصوص مدل برای KernelSHAP معرفی شد، اما مشخص شد که می‌تواند ویژگی‌های نامشهودی را تولید کند.

۵۶۶۷

۵۶۶۸

۵۶۶۹

۵۶۷۰

۵۶۷۱

تابع مقدار را با استفاده از انتظار شرطی تعریف می کنیم $E[\text{ایکس} | \text{ایکس}(f)] - E[\text{ایکس} | \text{ایکس}(f)]$ که در جای انتظار حاشیه ای مشکل انتظار شرطی این است که ویژگی‌هایی که هیچ تأثیری بر تابع پیش‌بینی f ندارند، می‌توانند تخمین TreeSHAP متفاوت از صفر دریافت کنند که توسط Sundararajan و همکاران نشان داده شده است. (۲۰۱۹) Janzing و همکاران. (۲۰۱۹) و شرطی TreeSHAP با ویژگی دیگری که در واقع بر پیش‌بینی تأثیر دارد، مرتبط باشد.

۵۶۷۲

۵۶۷۳

۵۶۷۴

چقدر سریعتر است؟ در مقایسه با KernelSHAP دقیق، پیچیدگی محاسباتی را کاهش می دهد $O(TLD)$ (به $O(DT)$)، که در آن T تعداد درختان، L حداکثر تعداد برگ در هر درخت و D حداکثر عمق هر درخت است.

۵۶۷۵

از انتظار شرطی استفاده می کنیم $E[\text{ایکس} | \text{ایکس}(f)] - E[\text{ایکس} | \text{ایکس}(f)]$

برای تخمین اثرات من به شما شهودی در مورد اینکه چگونه می‌توانیم پیش‌بینی مورد انتظار را برای یک درخت واحد، یک نمونه X و زیرمجموعه ویژگی S محاسبه کنیم. که نمونه X می‌افتد، پیش‌بینی مورد انتظار خواهد بود. اگر پیش‌بینی را با هیچ ویژگی شرطی نکنیم - اگر S خالی بود - از میانگین وزنی پیش‌بینی‌های تمام گره‌های پایانه استفاده می‌کنیم. اگر S شامل برخی ویژگی‌ها، اما نه همه، باشد، پیش‌بینی گره‌های غیرقابل دسترس را نادیده می‌گیریم. دست نیافتنی به این معنی است که مسیر تصمیم گیری که به این گره منتهی می‌شود با مقادیر موجود در تضاد است ایکس اس . از گره‌های پایانی باقی‌مانده، پیش‌بینی‌های وزن شده بر اساس اندازه گره (یعنی تعداد نمونه‌های آموزشی در آن گره) را میانگین می‌گیریم. میانگین گره‌های پایانی باقی‌مانده، وزن‌دهی شده با تعداد نمونه‌های هر گره، پیش‌بینی مورد انتظار برای X داده شده S است. مشکل این است که ما باید این رویه را برای هر زیرمجموعه S ممکن از مقادیر ویژگی اعمال کنیم TreeSHAP . در زمان چند جمله ای به جای نمایی محاسبه می‌کند. ایده اصلی این است که همه زیرمجموعه‌های ممکن S را به طور همزمان به پایین درخت فشار دهید. برای هر گره تصمیم باید تعداد زیرمجموعه‌ها را پیگیری کنیم. این بستگی به زیر مجموعه‌های گره والد و ویژگی تقسیم دارد. به عنوان مثال، هنگامی که اولین تقسیم در یک درخت روی ویژگی X_3 باشد، آنگاه تمام زیرمجموعه‌هایی که دارای ویژگی X_3 هستند به یک گره (گرهی که X می‌رود) می‌روند. زیرمجموعه‌هایی که دارای ویژگی X_3 نیستند با کاهش وزن به هر دو گره می‌روند. متأسفانه زیر مجموعه‌هایی با اندازه‌های مختلف وزن متفاوتی دارند. الگوریتم باید وزن کلی زیرمجموعه‌ها را در هر گره پیگیری کند. این الگوریتم را پیچیده می‌کند. برای جزئیات TreeSHAP به مقاله اصلی مراجعه می‌کنم. محاسبات را می‌توان به درختان بیشتری گسترش داد: به لطف ویژگی Additivity مقادیر Shapley ،

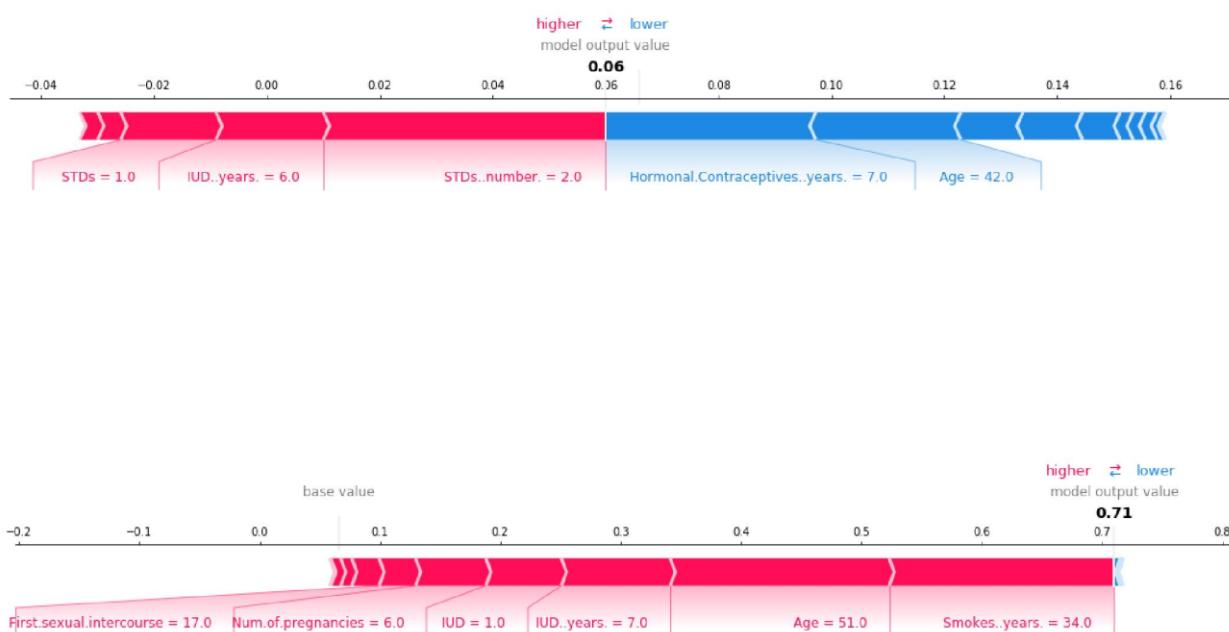
در مرحله بعد، به توضیحات SHAP در عمل نگاه خواهیم کرد.

۹,۶,۴ مثالها

من یک طبقه‌بندی تصادفی جنگل را با ۱۰۰ درخت آموزش دادم تا خطر ابتلا به سلطان دهانه رحم را پیش‌بینی کند . ما از SHAP برای توضیح پیش‌بینی‌های فردی استفاده خواهیم کرد. ما می‌توانیم از روش تخمین سریع TreeSHAP به جای روش کندر KernelSHAP استفاده کنیم، زیرا یک جنگل تصادفی مجموعه‌ای از درختان است. اما این مثال به جای تکیه بر توزیع شرطی، از توزیع حاشیه‌ای استفاده می‌کند. این در بسته بندی توضیح داده شده است، اما در مقاله اصلی نیست.تابع Python TreeSHAP با توزیع حاشیه‌ای کندر است، اما همچنان سریعتر از KernelSHAP است، زیرا به صورت خطی با ردیف‌های داده مقیاس می‌شود.

از آنجا که ما در اینجا از توزیع حاشیه‌ای استفاده می‌کنیم، تفسیر همان است که در فصل مقدار Shapley آمده است. اما با بسته شکل Python تجسم متفاوتی ارائه می‌شود: می‌توانید ویژگی‌های ویژگی‌هایی مانند مقادیر Shapley را به عنوان «نیروها» تجسم کنید. هر مقدار ویژگی نیرویی است که پیش بینی را افزایش یا کاهش می‌دهد. پیش بینی از پایه شروع می‌شود. خط پایه برای مقادیر Shapley میانگین همه پیش بینی‌ها است. در نمودار، هر مقدار Shapley یک فلش است که برای افزایش (مقدار مثبت) یا کاهش (مقدار منفی) پیش بینی فشار می‌آورد. این نیروها در پیش‌بینی واقعی نمونه داده، یکدیگر را متعادل می‌کنند.

شکل زیر نمودار نیروی توضیحی SHAP را برای دو زن از مجموعه داده سرطان دهانه رحم نشان می‌دهد:



شکل ۹.۲۴: مقادیر SHAP برای توضیح احتمالات پیش‌بینی شده سرطان دو فرد. خط پایه - میانگین احتمال پیش‌بینی شده - ۰,۰۶۶ است. اولین زن دارای خطر کم پیش‌بینی ۰,۰۶ است. اثرات افزایش خطر مانند بیماری‌های مقاربته با کاهش اثراتی مانند سن جبران می‌شود. زن دوم دارای خطر بالای ۰,۷۱ پیش‌بینی شده است. سن ۵۱ سالگی و سیگار کشیدن ۳۴ سال خطر ابتلا به سرطان را افزایش می‌دهد.

اینها توضیحاتی برای پیش‌بینی‌های فردی بود.

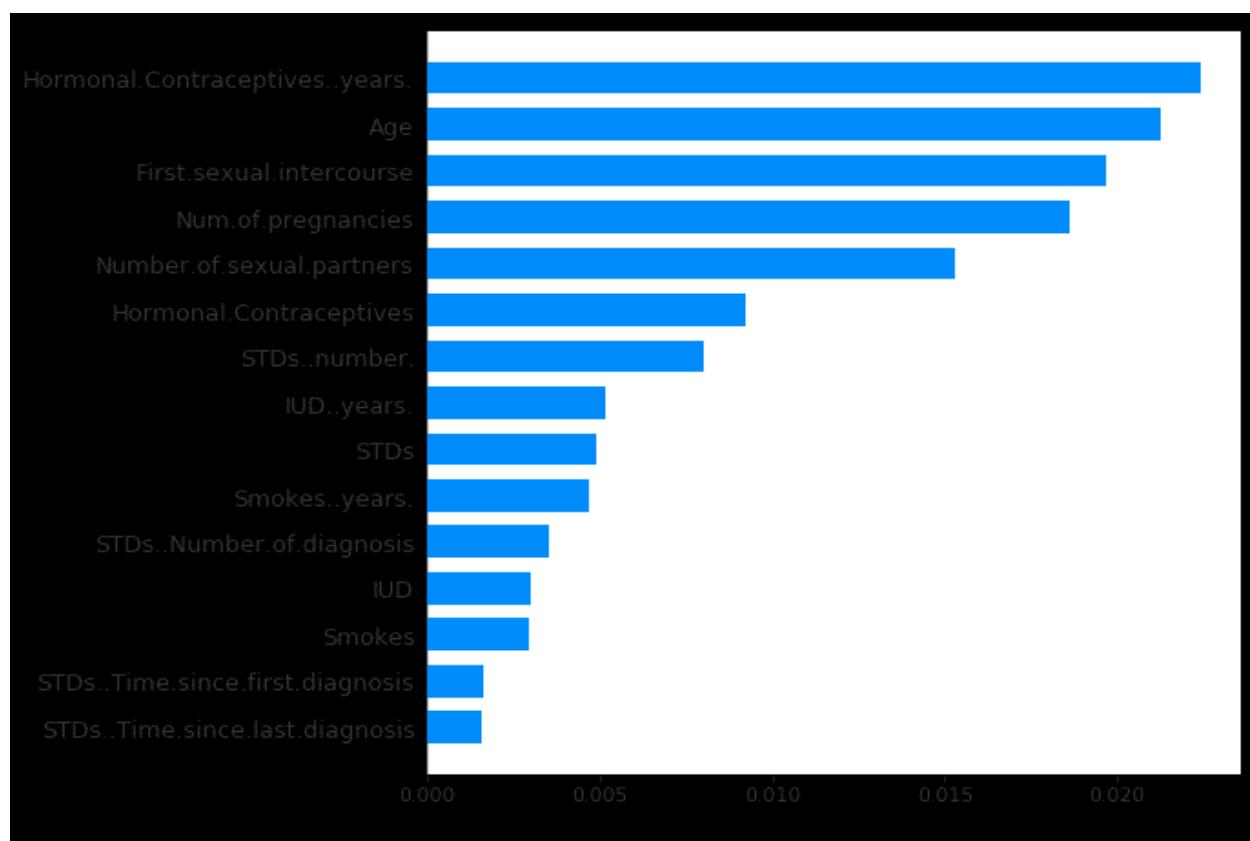
۵۷۱۵ مقادیر Shapley را می توان در توضیحات کلی ترکیب کرد. اگر SHAP را برای هر نمونه اجرا کنیم، ماتریسی
 ۵۷۱۶ از مقادیر Shapley را دریافت می کنیم. این ماتریس دارای یک ردیف برای هر نمونه داده و یک ستون در هر
 ۵۷۱۷ ویژگی است. ما می توانیم کل مدل را با تجزیه و تحلیل مقادیر Shapley در این ماتریس تفسیر کنیم.
 ۵۷۱۸ ما با اهمیت ویژگی SHAP شروع می کنیم.

۹,۶,۵ اهمیت ویژگی SHAP

۵۷۲۰ ایده اهمیت ویژگی SHAP ساده است: ویژگی هایی با مقادیر بزرگ Shapley مطلق مهم هستند. از آنجایی که
 ۵۷۲۱ ما اهمیت جهانی را می خواهیم، مقادیر مطلق Shapley را برای هر ویژگی در میان داده ها میانگین می کنیم:

۵۷۲۲
$$I_j = \frac{1}{n} \sum_{i=1}^n |\phi_j^{(i)}|$$

 ۵۷۲۳ در مرحله بعد، ویژگی ها را با کاهش اهمیت مرتب می کنیم و آنها را رسم می کنیم. شکل زیر اهمیت ویژگی
 ۵۷۲۴ SHAP را برای جنگل تصادفی که قبلاً برای پیش‌بینی سرطان دهانه رحم آموزش داده شده را نشان می‌دهد.



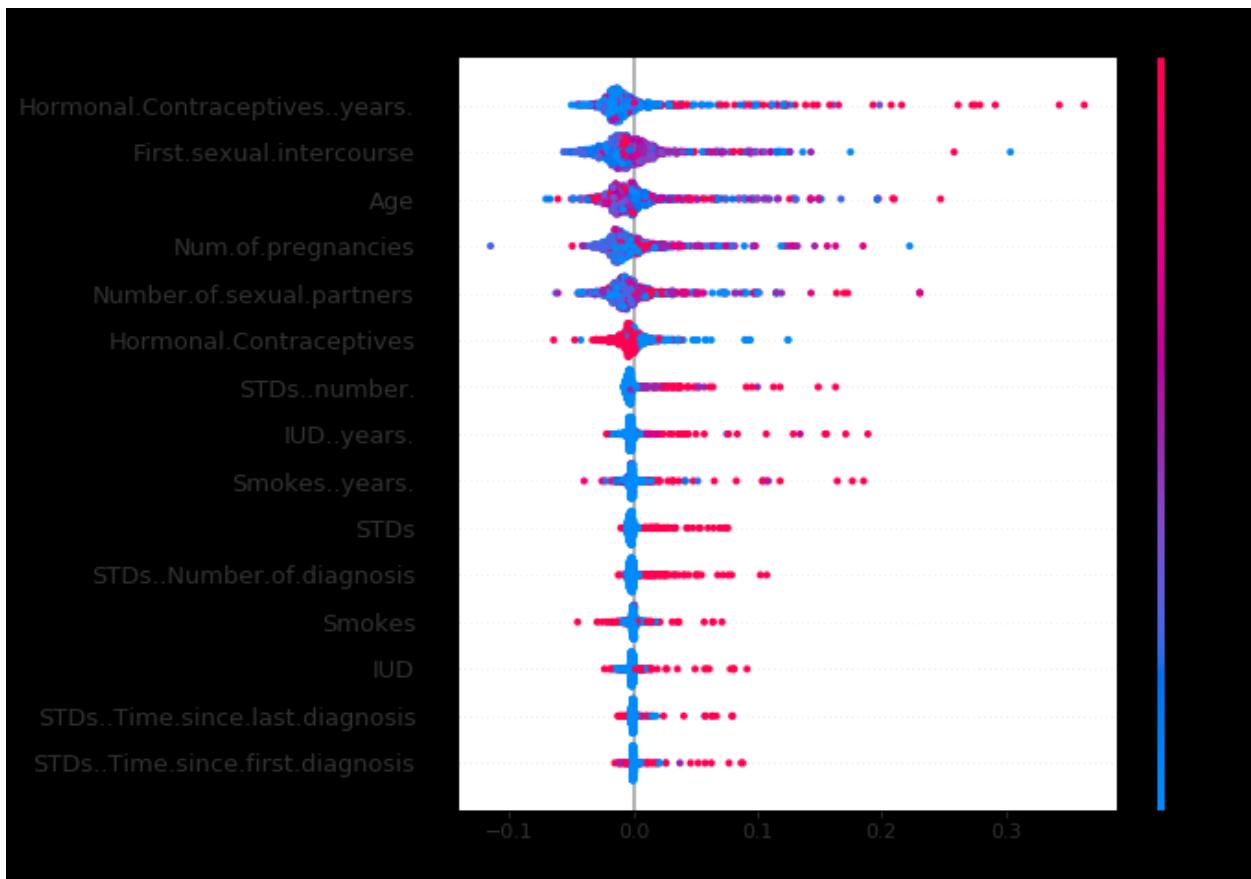
۵۷۲۶ شکل ۹,۲۵: اهمیت ویژگی SHAP به عنوان میانگین مقادیر مطلق Shapley اندازه گیری می شود. تعداد
۵۷۲۷ سالهای استفاده از داروهای ضد بارداری هورمونی مهم ترین ویژگی بود که احتمال سرطان مطلق پیش بینی
۵۷۲۸ شده را به طور متوسط ۲,۴ درصد (۰,۰ ۲۴) در محور X تغییر داد.

۵۷۲۹ اهمیت ویژگی SHAP جایگزینی برای اهمیت ویژگی جایگشتی است . تفاوت زیادی بین هر دو معیار اهمیت
۵۷۳۰ وجود دارد: اهمیت ویژگی جایگشت بر اساس کاهش عملکرد مدل است SHAP . بر اساس مقدار اسناد ویژگی
۵۷۳۱ است.

۵۷۳۲ نمودار اهمیت ویژگی مفید است، اما حاوی اطلاعاتی فراتر از اهمیت نیست. برای یک طرح آموزنده تر، در ادامه
۵۷۳۳ به طرح خلاصه نگاه خواهیم کرد.

۹,۶ طرح خلاصه SHAP

۵۷۳۵ طرح خلاصه اهمیت ویژگی را با جلوه های ویژگی ترکیب می کند. هر نقطه در نمودار خلاصه یک مقدار
۵۷۳۶ برای یک ویژگی و یک نمونه است. موقعیت روی محور Y توسط ویژگی و در محور X با مقدار
۵۷۳۷ تعیین می شود. رنگ نشان دهنده ارزش ویژگی از کم به بالا است. نقاط همپوشانی در جهت محور Y
۵۷۳۸ تکان می خورند، بنابراین ما یک حس از توزیع مقادیر Shapley در هر ویژگی دریافت می کنیم. ویژگی ها با
۵۷۳۹ توجه به اهمیت آنها مرتب شده اند.



۵۷۴۰

طرح خلاصه SHAP. سال‌های کم استفاده از داروهای ضد بارداری هورمونی خطر سرطان پیش‌بینی شده را کاهش می‌دهد، تعداد زیاد سال‌ها این خطر را افزایش می‌دهد. یادآوری همیشگی شما: همه اثرات رفتار مدل را توصیف می‌کنند و لزوماً در دنیای واقعی علت نیستند.

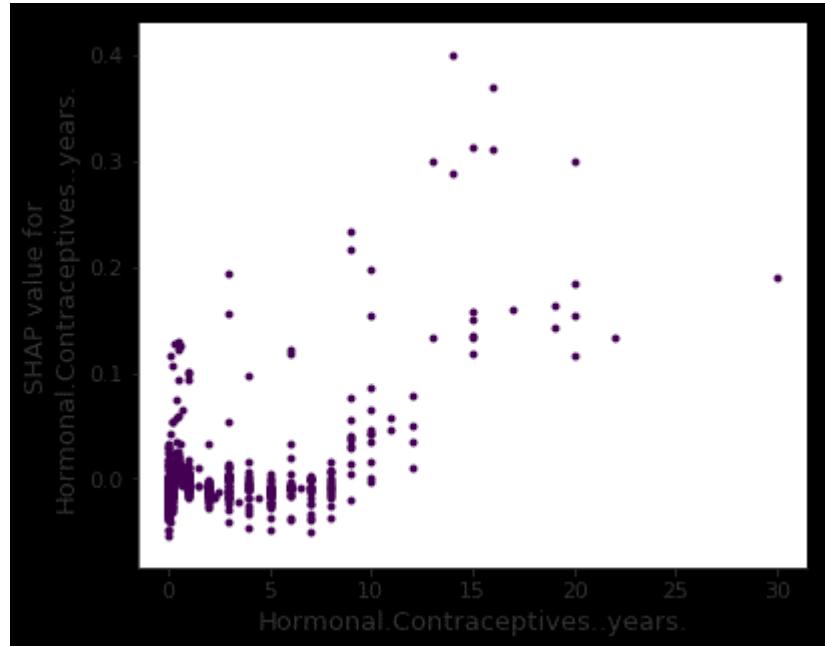
شکل ۹,۲۶: نمودار خلاصه SHAP. سال‌های کم استفاده از داروهای ضد بارداری هورمونی خطر سرطان پیش‌بینی شده را کاهش می‌دهد، تعداد زیاد سال‌ها این خطر را افزایش می‌دهد. یادآوری همیشگی شما: همه اثرات رفتار مدل را توصیف می‌کنند و لزوماً در دنیای واقعی علت نیستند.

در نمودار خلاصه، اولین نشانه‌های رابطه بین ارزش یک ویژگی و تأثیر آن بر پیش‌بینی را می‌بینیم. اما برای دیدن شکل دقیق رابطه، باید به نمودارهای وابستگی SHAP نگاه کیم.

۹,۶,۷ طرح وابستگی SHAP
وابستگی ویژگی SHAP ممکن است ساده‌ترین طرح تفسیر کلی باشد: ۱) یک ویژگی را انتخاب کنید. ۲) برای هر نمونه داده، یک نقطه با مقدار ویژگی در محور X و مقدار Shapley مربوطه در محور Y رسم کنید. ۳) انجام شد.

از نظر ریاضی، طرح حاوی نکات زیر است) } :ایکس(من j)، من n} من = 1

شکل زیر وابستگی ویژگی SHAP را برای سال ها به داروهای ضد بارداری هورمونی نشان می دهد:



شکل ۹,۲۷: نمودار وابستگی SHAP برای سالها به داروهای ضد بارداری هورمونی. در مقایسه با ۰ سال، چند سال احتمال پیش‌بینی شده کمتر و تعداد سال‌های زیاد احتمال سرطان پیش‌بینی شده را افزایش می‌دهد.

نمودارهای وابستگی SHAP جایگزینی برای نمودارهای وابستگی جزئی و اثرات محلی انباشته شده هستند . در حالی که نمودار PDP و ALE اثرات متوسط را نشان می دهن، وابستگی SHAP نیز واریانس را در محور ۷ نشان می دهد. به خصوص در صورت برهمکنش ها، نمودار وابستگی SHAP در محور ۷ پراکنده تر خواهد بود. نمودار وابستگی را می توان با برحسبه کردن این تعاملات ویژگی بهبود بخشد.

۹,۶,۸ ارزش های تعامل SHAP

اثر متقابل اثر ویژگی ترکیبی اضافی پس از محاسبه اثرات ویژگی های فردی است. شاخص تعامل Shapley از نظریه بازی به صورت زیر تعریف می شود:

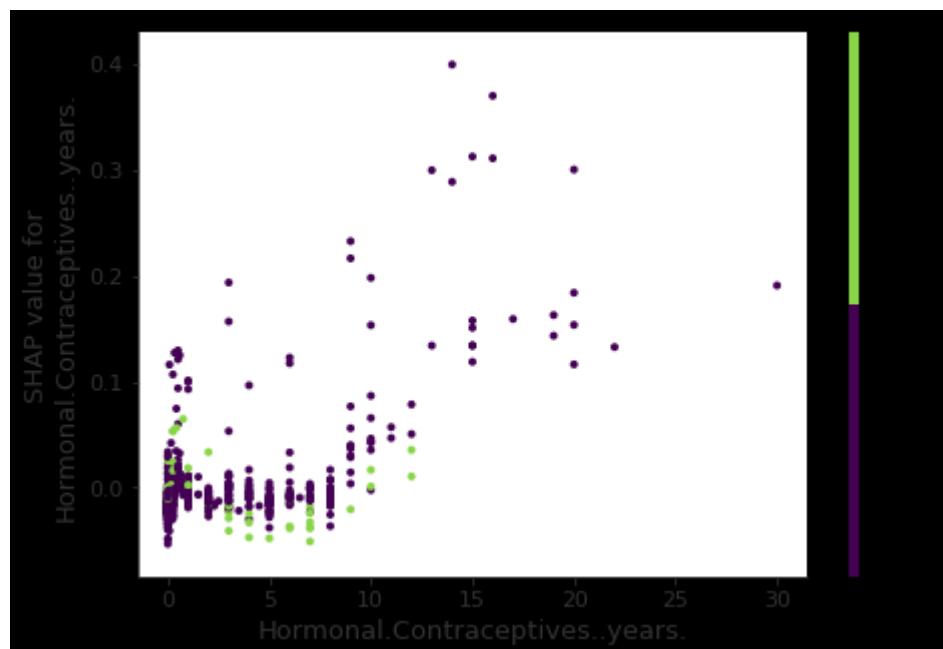
$$\phi_{i,j} = \sum_{S \subseteq \{i,j\}} \frac{|S|!(M - |S| - 2)!}{2(M - 1)!} \delta_{ij}(S)$$

چه زمانی من ≠ و :

$$\delta_{ij}(S) = \hat{f}_x(S \cup \{i, j\}) - \hat{f}_x(S \cup \{i\}) - \hat{f}_x(S \cup \{j\}) + \hat{f}_x(S)$$

۵۷۶۸ این فرمول اثر اصلی ویژگی ها را کم می کند تا پس از محاسبه اثرات فردی، اثر متقابل خالص را دریافت کنیم. ما
۵۷۶۹ مقادیر را روی همه ائتلاف های مشخصه S میانگین می کنیم، مانند محاسبه مقدار Shapley وقتی مقادیر
۵۷۷۰ تعامل SHAP را برای همه ویژگی ها محاسبه می کنیم، برای هر نمونه یک ماتریس با ابعاد $M \times M$ بدست می
۵۷۷۱ آوریم که M تعداد ویژگی ها است.

۵۷۷۲ چگونه می توانیم از شاخص تعامل استفاده کنیم؟ به عنوان مثال، برای رنگ آمیزی خودکار نمودار وابستگی
۵۷۷۳ SHAP با قوی ترین تعامل: ویژگی



۵۷۷۴ شکل ۹,۲۸: نمودار وابستگی ویژگی SHAP با تجسم تعامل. سال ها استفاده از داروهای ضد بارداری هورمونی با
۵۷۷۵ بیماری های مقاربتهای تداخل دارد. در موارد نزدیک به صفر سال، وقوع STD خطر سرطان پیش بینی شده را
۵۷۷۶ افزایش می دهد. برای سال های بیشتر در مورد داروهای ضد بارداری، وقوع STD خطر پیش بینی شده را کاهش
۵۷۷۷ می دهد. باز هم، این یک مدل علی نیست. اثرات ممکن است به دلیل گیج کننده باشد (مثالاً بیماری های
۵۷۷۸ مقاربته و خطر کمتر سرطان می تواند با مراجعه بیشتر به پزشک مرتبط باشد).

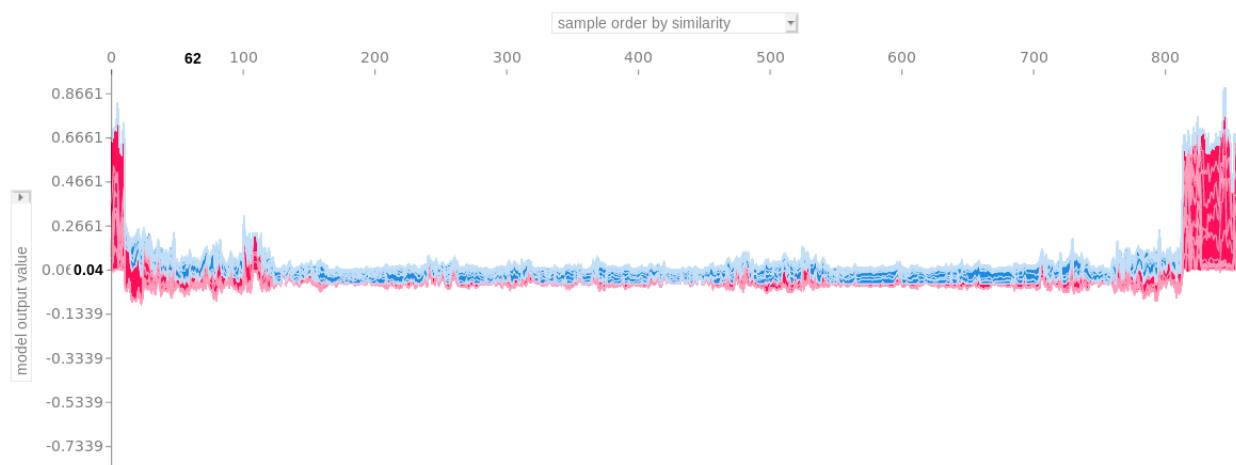
۹,۶,۹ خوشه بندی مقادیر Shapley

۵۷۸۰ شما می توانید داده های خود را با کمک مقادیر Shapley خوشه بندی کنید. هدف از خوشه بندی یافتن گروه
۵۷۸۱ هایی از نمونه های مشابه است. به طور معمول، خوشه بندی بر اساس ویژگی ها است. ویژگی ها اغلب در
۵۷۸۲ مقیاس های مختلف هستند. برای مثال، ارتفاع ممکن است بر حسب متر، شدت رنگ از ۰ تا ۱۰۰ و مقداری
۵۷۸۳

خروجی سنسور بین ۱- و ۱ اندازه گیری شود. مشکل در محاسبه فاصله بین نمونه های با چنین ویژگی های متفاوت و غیر قابل مقایسه است.

خوشه بندی SHAP با خوشه بندی مقادیر Shapley هر نمونه کار می کند. این بدان معنی است که شما نمونه ها را با تشابه توضیح خوشه بندی می کنید. همه مقادیر SHAP واحد یکسانی دارند - واحد فضای پیش بینی. شما می توانید از هر روش خوشه بندی استفاده کنید. مثال زیر از خوشه بندی تجمعی سلسله مرتبی برای مرتب سازی نمونه ها استفاده می کند.

طرح شامل نمودارهای نیرو زیادی است که هر کدام پیش بینی یک نمونه را توضیح می دهد. نمودارهای نیرو را به صورت عمودی می چرخانیم و با توجه به شباهت خوشه ای آنها در کنار هم قرار می دهیم.



شکل ۹,۲۹: توضیحات SHAP انباشته که بر اساس تشابه توضیح خوشه بندی شده اند. هر موقعیت در محور X نمونه ای از داده ها است. مقادیر قرمز SHAP پیش بینی را افزایش می دهد، مقادیر آبی آن را کاهش می دهد. یک خوشه برجسته است: در سمت راست گروهی با خطر سرطان پیش بینی شده بالا قرار دارد.

۹,۶,۱۰ مزایا

از آنجایی که SHAP مقادیر Shapley را محاسبه می کند، تمام مزایای مقادیر Shapley اعمال می شود : SHAP یک پایه نظری محکم در تئوری بازی ها دارد. پیش بینی به طور عادلانه بین مقادیر ویژگی توزیع شده است. ما توضیحات متضادی را دریافت می کنیم که پیش بینی را با پیش بینی میانگین مقایسه می کند.

SHAP مقادیر LIME و Shapley را به هم متصل می کند . این برای درک بهتر هر دو روش بسیار مفید است. همچنین به متحد کردن زمینه یادگیری ماشینی قابل تفسیر کمک می کند.

5802 یک پیاده سازی سریع برای مدل های مبتنی بر درخت دارد . من معتقدم که این کلید محبوبیت
5803 SHAP بود، زیرا بزرگترین مانع برای پذیرش مقادیر Shapley محاسبه کند است.

5804 محاسبات سریع امکان محاسبه بسیاری از مقادیر Shapley مورد نیاز برای تفاسیر مدل جهانی را فراهم می
5805 کند . روش های تفسیر جهانی شامل اهمیت ویژگی، وابستگی ویژگی، تعاملات، خوشبندی و نمودارهای خلاصه
5806 است. با SHAP ، تفاسیر جهانی با توضیحات محلی سازگار است، زیرا مقادیر " واحد اتمی " تفاسیر
5807 جهانی هستند. اگر از LIME برای توضیحات محلی و نمودارهای وابستگی جزئی به اضافه اهمیت ویژگی
5808 جایگشت برای توضیحات کلی استفاده می کنید، شما قادر به ایجاد یک پایه مشترک هستید.

5809 ۹,۶,۱۱ معایب

5810 KernelSHAP کند است . این باعث می شود وقتی می خواهید مقادیر Shapley را برای بسیاری از نمونه ها
5811 محاسبه کنید، استفاده از KernelSHAP غیر عملی است. همچنین تمام روش های جهانی SHAP مانند اهمیت
5812 ویژگی SHAP نیاز به محاسبه مقادیر Shapley برای بسیاری از نمونه ها دارند.

5813 KernelSHAP وابستگی ویژگی را نادیده می گیرد . اکثر روش های دیگر تفسیر مبتنی بر جایگشت این مشکل
5814 را دارند. با جایگزینی مقادیر ویژگی با مقادیر نمونه های تصادفی، معمولاً نمونه برداری تصادفی از توزیع حاشیه
5815 ای آسان تر است. با این حال، اگر ویژگی ها وابسته باشند، به عنوان مثال همبسته، این منجر به اعمال وزن بیش
5816 از حد بر روی نقاط داده غیر محتمل می شود TreeSHAP . این مشکل را با مدل سازی صریح پیش بینی مورد
5817 انتظار شرطی حل می کند.

5818 TreeSHAP می تواند ویژگی های نامشهود را تولید کند . در حالی که TreeSHAP مشکل برونویابی به نقاط
5819 داده غیر محتمل را حل می کند، این کار را با تغییرتابع مقدار انجام می دهد و بنابراین کمی بازی را تغییر می
5820 دهد TreeSHAP . تابع مقدار را با تکیه بر پیش بینی مورد انتظار مشروط تغییر می دهد. با تغییر در تابع
5821 مقدار، ویژگی هایی که هیچ تاثیری بر پیش بینی ندارند می توانند یک مقدار TreeSHAP متفاوت از صفر
5822 دریافت کنند.

5823 معایب مقادیر Shapley در مورد SHAP نیز صدق می کند: مقادیر Shapley می توانند اشتباه تفسیر شوند و
5824 برای محاسبه آنها برای داده های جدید (به جز TreeSHAP) به داده ها نیاز است.

5825 امکان ایجاد تعابیر عمدى گمراه کننده با SHAP وجود دارد که می تواند سوگیری ها را پنهان کند ۷۳ . اگر
5826 شما دانشمند داده ای هستید که توضیحات را ایجاد می کند، این یک مشکل واقعی نیست (حتی اگر شما

دانشمند داده شیطانی باشد که می خواهد توضیحات گمراه کننده ایجاد کند، یک مزیت خواهد بود). برای گیرندگان توضیح SHAP، این یک نقطه ضعف است: آنها نمی توانند در مورد صحت توضیح مطمئن باشند.

۹,۶,۱۲ نرم افزار

نویسندهای SHAP را در بسته پایتون shap پیاده سازی کردند.

کتاب تفسیر مدل های یادگیری ماشین با SHAP کاربرد SHAP با استفاده از shap بسته را به طور عمیق پوشش می دهد.

این پیاده سازی برای مدل های مبتنی بر درخت در کتابخانه یادگیری ماشینی Scikit-Learn برای پایتون کار می کند. همچنین برای مثال های این فصل از بسته بندی Shap استفاده شده است SHAP. در چارچوب های تقویت درخت LightGBM و xgboost ادغام شده است . در R ، بسته های fastshap و shapper وجود دارد SHAP . نیز در بسته R xgboost گنجانده شده است.

فصل ۱۱ نگاهی به توب کریستالی

۵۸۳۸ آینده یادگیری ماشینی قابل تفسیر چیست؟ این فصل یک تمرین فکری ذهنی و یک حدس ذهنی است که
۵۸۳۹ یادگیری ماشینی قابل تفسیر چگونه توسعه خواهد یافت. من کتاب را با داستان های کوتاه نسبتاً بدینه باز
۵۸۴۰ کردم و می خواهم با نگاهی خوش بینانه تر به پایان برسم.
۵۸۴۱

۵۸۴۲ من «پیش‌بینی‌های» خود را بر سه فرض استوار کرده‌ام:

۵۸۴۳ دیجیتال سازی: هر گونه اطلاعات (جالب) دیجیتالی می شود. به پول نقد الکترونیکی و معاملات آنلاین فکر
۵۸۴۴ کنید. به کتاب های الکترونیکی، موسیقی و ویدئو فکر کنید. به تمام داده های حسی در مورد محیط، رفتار
۵۸۴۵ انسانی، فرآیندهای تولید صنعتی و غیره فکر کنید. حرکت‌های دیجیتالی کردن همه چیز عبارتند از:
۵۸۴۶ رایانه‌ها/حسگرها/ذخیره‌سازی‌های ارزان، جلوه‌های مقیاس‌پذیر (برنده همه چیز را می‌گیرد)، مدل‌های تجاری
۵۸۴۷ جدید، زنجیره‌های ارزش مدولار، فشار هزینه و بسیاری موارد دیگر.

۵۸۴۸ اتوماسیون: هنگامی که یک کار می تواند خودکار باشد و هزینه اتوماسیون کمتر از هزینه انجام کار در طول
۵۸۴۹ زمان باشد، کار خودکار می شود. حتی قبل از معرفی کامپیوتر ما درجه خاصی از اتوماسیون را داشتیم. به عنوان
۵۸۵۰ مثال، ماشین بافندگی خودکار یا ماشین بخار با اسب بخار خودکار، اما کامپیوترها و دیجیتالی شدن،
۵۸۵۱ اتوماسیون را به سطحی بالاتر می برد. صرفاً این واقعیت که می‌توانید برای حلقه‌ها برنامه‌نویسی کنید، ماکروهای
۵۸۵۲ اکسل بنویسید، پاسخ‌های ایمیل را خودکار کنید، و غیره، نشان می‌دهد که یک فرد چقدر می‌تواند خودکار کند.
۵۸۵۳ ماشین‌های بليت خريد بليط قطار را خودکار می‌کنند (ديگر نيازی به صندوق دار نیست)، ماشین‌های لباسشویی
۵۸۵۴ شستشوی لباس‌ها را خودکار می‌کنند، سفارش‌های دائمی معاملات پرداخت را خودکار می‌کنند و غیره. خودکار
۵۸۵۵ کردن وظایف زمان و پول را آزاد می‌کند، بنابراین انگیزه اقتصادی و شخصی زیادی برای خودکار کردن کارها
۵۸۵۶ وجود دارد. ما در حال حاضر در حال مشاهده اتوماسیون ترجمه زبان، رانندگی و تا حدی حتی کشف علمی
۵۸۵۷ هستیم.

۵۸۵۸ تعریف اشتباه: ما نمی‌توانیم یک هدف را با تمام محدودیت‌هایش کاملاً مشخص کنیم. به جن در بطری فکر
۵۸۵۹ کنید که همیشه خواسته های شما را به معنای واقعی کلمه می‌پذیرد: "من می خواهم ثروتمندترین فرد جهان
۵۸۶۰ باشم!" -> شما تبدیل به ثروتمندترین فرد می‌شوید، اما به عنوان یک عارضه جانبی، ارزی که در اختیار دارید
۵۸۶۱ به دلیل تورم سقوط می‌کند.

۵۸۶۲ "من می خواهم تا آخر عمرم خوشحال باشم!" -> ۵ دقیقه بعد خیلی احساس خوشبختی می‌کنی، بعد جن تو
۵۸۶۳ را می‌کشد.

۵۸۶۴ "آرزوی صلح جهانی دارم!" -> جن همه انسان ها را می کشد.

۵۸۶۵ ما اهداف را به اشتباه مشخص می کنیم، یا به این دلیل که همه محدودیت ها را نمی دانیم یا به این دلیل که
۵۸۶۶ نمی توانیم آنها را اندازه گیری کنیم. باید به شرکت ها به عنوان نمونه ای از مشخصات هدف ناقص نگاه کنیم.
۵۸۶۷ یک شرکت هدف ساده کسب درآمد برای سهامداران خود دارد. اما این مشخصات هدف واقعی را با تمام
۵۸۶۸ محدودیت هایش که ما واقعاً برای آن تلاش می کنیم، نشان نمی دهد: به عنوان مثال، ما از شرکتی که مردم را
۵۸۶۹ برای کسب درآمد می کشد، رودخانه ها را مسموم می کند یا صرفاً پول خود را چاپ می کند قدردانی نمی کنیم. ما
۵۸۷۰ قوانین، مقررات، تحریم ها، رویه های انطباق، اتحادیه های کارگری و موارد دیگر را برای اصلاح مشخصات هدف
۵۸۷۱ ناقص ابداع کرده ایم. نمونه دیگری که می توانید خودتان تجربه کنید، گیره کاغذ است، بازی ای که در آن با یک
۵۸۷۲ ماشین با هدف تولید هرچه بیشتر گیره کاغذ بازی می کنید. هشدار: اعتیاد آور است. من نمی خواهم آن را
۵۸۷۳ خیلی خراب کنم، اما باید بگوییم که همه چیز خیلی سریع از کنترل خارج می شود. در یادگیری ماشین،
۵۸۷۴ نقص در مشخصات هدف ناشی از انتزاع داده های ناقص (جمعیت های مغرضانه، خطاهای اندازه گیری، ...)
۵۸۷۵ توابع از دست دادن نامحدود، عدم آگاهی از محدودیت ها، تغییر توزیع بین داده های آموزشی و برنامه کاربردی
۵۸۷۶ و بسیاری موارد دیگر است.

۵۸۷۷ دیجیتالی شدن اتوماسیون رانندگی است. مشخصات هدف ناقص با اتوماسیون در تعارض است. من ادعا می کنم
۵۸۷۸ که این تعارض تا حدی با روش های تفسیری میانجی گری می شود.

۵۸۷۹ صحنه برای پیش بینی های ما آماده شده است، توب کریستالی آماده است، اکنون ما نگاه می کنیم که زمین به
۵۸۸۰ کجا می تواند برسد

۵۸۸۱

فصل ۱۳ با استناد به این کتاب

۵۸۸۲ اگر این کتاب را برای پست و بلاگ، مقاله تحقیقاتی یا محصول خود مفید دیدید، ممنون می شوم اگر به این
۵۸۸۳ کتاب استناد کنید. شما می توانید کتاب را به این صورت استناد کنید:
۵۸۸۴

۵۸۸۵ Molnar, C. (2022). Interpretable Machine Learning:
۵۸۸۶ A Guide for Making Black Box Models Explainable (2nd ed.).
۵۸۸۷ christophm.github.io/interpretable-ml-book/

۵۸۸۸ یا از ورودی **bibtex** زیر استفاده کنید:

۵۸۸۹ @book{molnar2022,
۵۸۹۰ title = {Interpretable Machine Learning},
۵۸۹۱ author = {Christoph Molnar},
۵۸۹۲ year = {2022},
۵۸۹۳ subtitle = {A Guide for Making Black Box Models Explainable},
۵۸۹۴ edition = {2},
۵۸۹۵ url = {https://christophm.github.io/interpretable-ml-book}

۵۸۹۶ }

۵۸۹۷ من همیشه کنجکاو هستم که کجا و چگونه از روش های تفسیر در صنعت و تحقیق استفاده می شود. اگر از
۵۸۹۸ کتاب به عنوان مرجع استفاده می کنید، خیلی خوب می شود اگر یک خط برای من بنویسید و بگویید برای چه.
۵۸۹۹ این البته اختیاری است و فقط برای ارضای کنجکاوی خودم و تحریک مبادلات جالب است. ایمیل من
۵۹۰۰ . christoph.molnar.ai@gmail.com است.

۵۹۰۱

فصل ۱۴ ترجمه ها

به ترجمه کتاب علاقه دارید؟

۵۹۰۴ این کتاب تحت مجوز Creative Commons Attribution-NonCommercial-ShareAlike 4.0 بین
۵۹۰۵ المللی مجوز دارد . یعنی شما مجاز به ترجمه و قرار دادن آن در اینترنت هستید. شما باید من را به عنوان
۵۹۰۶ نویسنده اصلی ذکر کنید و اجازه فروش کتاب را ندارید.

۵۹۰۷ اگر علاقه مند به ترجمه کتاب هستید می توانید پیام بنویسید و ترجمه شما را اینجا لینک کنم. آدرس من
۵۹۰۸ christoph.molnar.ai@gmail.com است.

۵۹۰۹ فهرست ترجمه ها

۵۹۱۰ باهسا اندونزی

۵۹۱۱ ترجمه کامل توسط Smart City & Cybersecurity Laboratory, و Hatma Suryotrisongko
۵۹۱۲ Information Technology, ITS .

۵۹۱۳ چینی:

۵۹۱۴ ترجمه کامل نسخه دوم توسط CSDN از Jiazen ، یک جامعه آنلاین برنامه نویسان.

۵۹۱۵ ترجمه های کامل توسط Mingchao Zhu . نسخه الکترونیکی و چاپی این ترجمه موجود است.

۵۹۱۶ ترجمه اکثر فصول توسط CSDN.

۵۹۱۷ ترجمه چند فصل . این وب سایت همچنین شامل سوالات و پاسخ های کاربران مختلف است.

۵۹۱۸ زبانی

۵۹۱۹ ترجمه کامل توسط HACARUS. ریوجی ماسوئی و تیم.

۵۹۲۰ کره ای:

۵۹۲۱ ترجمه کامل کره ای توسط TooTouch

۵۹۲۲ ترجمه جزئی کره ای توسط An Subin

۵۹۲۳ اسپانیایی

- ۵۹۲۴ ترجمه کامل اسپانیایی توسط فدریکو فلیگر
- ۵۹۲۵ ویتنامی
- ۵۹۲۶ ترجمه کامل Hoang Tri Le ،Hung-Quang Nguyen .Duy-Tung Nguyen ، Giang Nguyen.
- ۵۹۲۷ Nguyen.
- ۵۹۲۸ اگر ترجمه دیگری از کتاب یا هر فصل دیگری می شناسید، ممنون می شوم که درباره آن بشنوم و آن را در
- ۵۹۲۹ اینجا فهرست کنید. می توانید از طریق ایمیل با من تماس بگیرید: christoph.molnar.ai@gmail.com
- ۵۹۳۰ .
- ۵۹۳۱

فصل ۱۵ سپاسگزاریها

۵۹۳۲ نوشتن این کتاب بسیار سرگرم کننده بود (و هنوز هم هست). اما این کار هم زیاد است و از حمایتی که دریافت
۵۹۳۳ کردم بسیار خوشحالم.
۵۹۳۴
۵۹۳۵ بزرگترین تشکر من از کاترین است که سخت ترین کار را از نظر ساعت و تلاش داشت: او کتاب را از ابتدا تا انتهای
۵۹۳۶ تصحیح کرد و بسیاری از اشتباهات املایی و تناقضات را کشف کرد که من هرگز آنها را پیدا نمی کردم. من از
۵۹۳۷ حمایت او بسیار سپاسگزارم.
۵۹۳۸ از همه نویسندهای مهمان تشکر می کنم. من واقعاً متعجب شدم وقتی فهمیدم مردم علاقه مند به مشارکت در
۵۹۳۹ این کتاب هستند. و به لطف تلاش آنها، می توان مطالب کتاب را بهبود بخشید! توبیاس گورک و ماجdalana لانگ
۵۹۴۰ فصلی را در مورد قوانین محدوده (لنگرهای) نوشته‌اند. فانگزو لی در بخش تشخیص مفاهیم مشارکت داشت. و
۵۹۴۱ سوزان دنل فصل مربوط به مثال‌های خلاف واقع را بسیار بهبود بخشید. آخرین اما نه کم اهمیت، Verena
۵۹۴۲ Haunschmid بازخورد و اصلاحات را مستقیماً در GitHub ارائه کردند تشکر کنم!
۵۹۴۳
۵۹۴۴ علاوه بر این، من می خواهم از همه کسانی که تصاویر را ایجاد کردند تشکر کنم: جلد توسط دوست من
۵۹۴۵ @YvonneDoinel طراحی شده است . گرافیک‌های فصل Shapley Value توسط هایدی سیبولد و
۵۹۴۶ همچنین نمونه لاک پشت در فصل نمونه‌های متخصص ایجاد شده است Verena Haunschmid . این
۵۹۴۷ گرافیک را در فصل RuleFit ایجاد کرد.
۵۹۴۸ همچنین از همسر و خانواده ام که همیشه از من حمایت کردند تشکر می کنم. به خصوص همسرم مجبور بود به
۵۹۴۹ صحبت‌های من درباره کتاب گوش دهد. او به من کمک کرد تا تصمیمات زیادی در مورد نوشتن کتاب بگیرم.
۵۹۵۰ نحوه انتشار این کتاب کمی غیر متعارف است. اولاً، این نه تنها به عنوان جلد شومیز و کتاب الکترونیکی، بلکه به
۵۹۵۱ عنوان یک وب سایت، کاملاً رایگان در دسترس است. نرم افزاری که من برای ایجاد این کتاب استفاده کردم، نام
۵۹۵۲ bookdown دارد، نوشته شده توسط Yihui Xie ، که بسیاری از بسته‌های R را ایجاد کرد که ترکیب کد R
۵۹۵۳ و متن را آسان می کند. خیلی ممنون! من کتاب را به عنوان کتاب در دست انتشار منتشر کردم که به من کمک
۵۹۵۴ زیادی کرد تا بازخورد دریافت کنم و در طول مسیر از آن درآمدزایی کنم. همچنین می خواهم از شما خواننده
۵۹۵۵ عزیز تشکر کنم که این کتاب را بدون نام ناشر بزرگی خواندید.

من از بودجه تحقیقاتم در مورد یادگیری ماشینی قابل تفسیر توسط وزارت علوم و هنرهای ایالت باواریا در
چارچوب مرکز دیجیتال سازی باواریا (ZD.B) و موسسه تحقیقاتی باواریا برای تحول دیجیتال (bidt) سپاسگزارم.

۵۹۵۹