چرا مدل‌های مبتنی بر درخت در داده‌های جدولی بهتر از یادگیری عمیق عمل می‌کنند؟

پویا اردِه خانی
دانشجو مهندسی کامپیوتر
دانشکدگان فارابی دانشگاه تهران
pouya.ardehkhani@ut.ac.ir

متخصصان یادگیری ماشین در سراسر جهان، در حوزه‌های مختلف، به طور مداوم پدیده‌ای را مشاهده کرده‌اند: مدل‌های مبتنی بر درخت، مانند Random Forest، در تحلیل داده‌های جدولی از یادگیری عمیق/ شبکه‌های عصبی بهتر عمل می‌کنند. حال بهتر است در ابتدا داده‌های جدولی را توضیح دهیم.

داده جدولی به اطلاعات سازمان‌دهی شده در یک جدول یا ساختار صفحه گسترده، که در آن داده‌ها در ردیف‌ها و ستون‌ها مرتب شده‌اند، اشاره دارد. هر ردیف معمولاً یک مشاهده یا نمونه خاص را نشان می‌دهد، در حالی که ستون‌ها ویژگی‌های مختلف مرتبط با آن مشاهدات را نشان می‌دهند. داده‌های جدولی معمولاً در پایگاه‌های داده، صفحات گسترده یا جداشده با کاما یافت می‌شوند. نمونه‌ها شامل مجموعه داده‌هایی با اطلاعات در مورد مشتریان، تراکنش‌های مالی، آزمایش‌های علمی یا هر سناریویی است که در آن داده‌ها را می‌توان در قالب جدول ساختاریافته سازمان‌دهی کرد. برای مثال به شکل یک دقت کنید.

Play golf dataset

Independent variables				Dep. var
OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	FALSE	Don't Play
sunny	80	90	TRUE	Don't Play
overcast	83	78	FALSE	Play
rain	70	96	FALSE	Play
rain	68	80	FALSE	Play
rain	65	70	TRUE	Don't Play
overcast	64	65	TRUE	Play
sunny	72	95	FALSE	Don't Play
sunny	69	70	FALSE	Play
rain	75	80	FALSE	Play
sunny	75	70	TRUE	Play
overcast	72	90	TRUE	Play
overcast	81	75	FALSE	Play
rain	71	80	TRUE	Don't Play

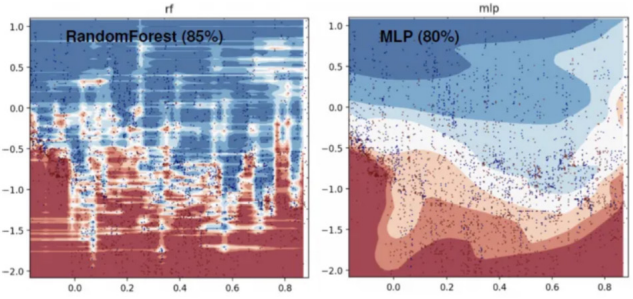
شکل ۱: نمونه یک دیتای جدولی

حال، بیایید به سوال کلیدی که شما را به اینجا رسانده بپردازیم: چرا روش‌های مبتنی بر درخت بهتر از یادگیری عمیق عمل می‌کنند؟

دلیل اول) شبکه‌های عصبی به سمت راه‌حل‌های بیش از حد صاف سوگیری می کنند.

نتایج حاکی از آن است که توابع هدف در مجموعه داده‌ها ویژگی‌های غیر هموار را نشان می‌دهند که در مقایسه با مدل‌های مبتنی بر درخت، چالشی را برای شبکه‌های عصبی در برازش این توابع نامنظم ایجاد می‌کنند. این‌ها با یافته‌هایی همسو هستند که نشان می‌دهند شبکه‌های عصبی تمایل دارند به سمت توابع فرکانس پایین سوگیری داشته باشند. در مقابل، مدل‌های مبتنی بر درخت تصمیم، که توابع ثابت تکه‌ای را یاد می‌گیرند، چنین تعصبی را نشان نمی‌دهند. منظم‌سازی کافی و بهینه‌سازی دقیق ممکن است شبکه‌های عصبی را قادر سازد تا الگوهای نامنظم را به‌طور مؤثر ثبت کنند.

به زبان ساده، زمانی که به توابع/مرزهای تصمیم غیر-صاف می‌پردازیم، شبکه‌های عصبی دچار مشکل می‌شوند تا توابع مناسب بهترین تطابق را ایجاد کنند. Random Forest با الگوهای عجیب، خشن، یا غیر منظم بهتر عمل می‌کنند. اگر بخواهم حدس بزنم که چرا، یک دلیل ممکن است استفاده از گرادیان در شبکه‌های عصبی باشد. گرادیان‌ها بر اساس فضاهای جستجوی مشتقی استوار هستند که به تعریف صاف هستند. توابع نوک‌نما، شکسته و تصادفی نمی‌توانند مشتق گرفته شوند. برای مثال ملموس‌تر از تفاوت در مرزهای تصمیم بین روش های مبتنی بر درخت (Random Forests) و یادگیری عمیق به تصویر زیر نگاهی بیندازید.



شکل ۲: عملکرد بهتر RF ها را می توان به مرزهای تصمیم گیری دقیق تری که تولید می کنند نسبت داد.

همانطور که دقت می کنیم، می توانیم ببینیم که RandomForest می تواند الگوهای نامنظم را در محور x (که با ویژگی تاریخ مطابقت دارد) یاد بگیرد که MLP آنها را نمی‌آموزد. ما این تفاوت را برای پارامترهای پیش‌فرض نشان می‌دهیم، اما به نظر می‌رسد که این یک رفتار معمولی از شبکه‌های عصبی است و در واقع یافتن فراپارامترها برای یادگیری موفقیت‌آمیز این الگوها دشوار است، هرچند غیرممکن نیست

زمانی که متوجه می‌شوید که روش‌های مبتنی بر درخت هزینه‌های تنظیم بسیار پایین‌تری دارند، این امر حتی قابل توجه‌تر می‌شود و وقتی صحبت از راه‌حل‌های سریع و ارزان قیمت به میان می‌آید، آنها را بسیار بهتر می‌کند.

دلیل دوم) ویژگی های غیر اطلاعاتی بر NNهای MLP مانند، بیشترین تأثیر را می‌گذارند.

اول به این می‌پردازیم که ویژگی‌های غیر اطلاعاتی چه چیزی هستند. ویژگی‌های غیر اطلاعاتی که به عنوان ویژگی‌های نامربوط یا زائد نیز شناخته می‌شوند، متغیرهایی در یک مجموعه داده هستند که اطلاعات معنی‌داری را برای کار مورد نظر ارائه نمی‌کنند. این ویژگی‌ها بینش ارزشمندی را ارائه نمی‌دهند یا بین نمونه‌های مختلف تبعیض قائل نمی‌شوند. گنجاندن ویژگی‌های غیر اطلاعاتی در یک مدل به طور بالقوه می‌تواند نویز ایجاد کند، پیچیدگی محاسباتی را افزایش دهد و منجر به overfitting شود.

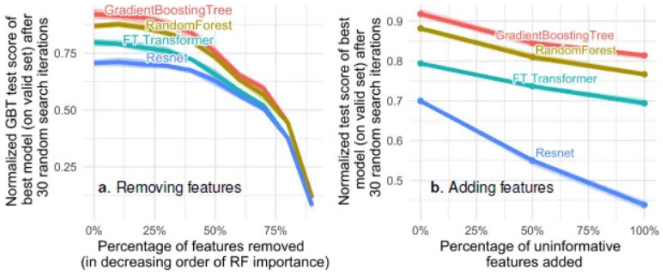
یک عامل بسیار مهم دیگر، به ویژه برای کسانی که با مجموعه‌داده‌های بزرگ کار می‌کنند، این است که همزمان چندین ارتباط را کد می‌کنند. اگر ویژگی‌های نامرتبط را به شبکه عصبی خود وارد کنید، نتایج بسیار بد خواهند بود (و شما ممکن است بسیاری از منابع خود را در آموزش مدل‌های خود هدر بدهید). به همین دلیل صرف زمان کافی بر روی تحلیل و اکتشاف دامنه (EDA/Domain Exploration) بسیار مهم است. این کمک می کند تا ویژگی‌ها را درک کنید و اطمینان حاصل کنید که همه چیز به طور صحیح اجرا می‌شود.

هنگام افزودن و حذف ویژگی‌های بی‌اهمیت (به‌طور دقیق‌تر، کم‌اهمیت) با

استفاده از روش تصادفی، بر اساس نتایج بدست آمده، دو نکته جالب ظاهر شد:

حذف تعداد زیادی از ویژگی‌ها باعث کاهش چشمگیری در تفاوت عملکرد بین مدل‌ها شد. این به وضوح نشان می‌دهد که یکی از مزایای بزرگ درخت‌ها، توانایی آن‌ها در مقاومت در برابر تأثیرات ویژگی‌های بدتر است.

افزودن ویژگی‌های تصادفی به مجموعه داده، نشان‌دهنده کاهش شدیدتری در شبکه‌ها نسبت به روش‌های مبتنی بر درخت است. به ویژه، ResNet به شدت تحت تأثیر این ویژگی‌های بی‌اهمیت قرار می‌گیرد. فرض می‌شود که مکانیزم توجه در ترنسفورمر تا حدی از این تأثیرات حفاظت می‌کند.



شکل ۳: دقت آزمون هنگام حذف (a) یا اضافه (b) ویژگی‌های بی‌اطلاع‌کننده تغییر می‌کند. ویژگی‌ها به ترتیب صعودی اهمیت حذف می‌شوند (محاسبه شده با یک جنگل تصادفی). ویژگی‌های اضافه شده از گوسی‌های استاندارد نمونه‌برداری می‌شوند که با هدف و دیگر ویژگی‌ها همبسته نیستند. امتیازها در دیتاست‌ها میانگین گرفته می‌شوند و نوارها متناظر با حداقل و حداکثر امتیاز در میان ۳۰ ترتیب تصادفی مختلف (شروع با مدل‌های پیش‌فرض) هستند.

یک توضیح ممکن برای این پدیده می‌تواند در روش طراحی درخت تصمیم، به ویژه در مفاهیم Information Gain و آنتروپی در درخت تصمیم، باشد.

آنتروپی، در زمینه درخت‌های تصمیم و نظریه اطلاعات، یک معیار از عدم قطعیت یا ناهماهنگی در یک مجموعه داده است. در زمینه الگوریتم‌های درخت تصمیم، مانند آن‌های استفاده شده در یادگیری ماشین، آنتروپی به عنوان یک معیار رایج برای تصمیم در مورد چگونگی تقسیم یک مجموعه داده به زیرمجموعه‌ها استفاده می‌شود. آنتروپی زمانی حداکثر است که کلاس‌ها در مجموعه داده به صورت متساوی توزیع شده‌اند و نشان‌دهنده عدم قطعیت بیشتری است. از سوی دیگر، آنتروپی زمانی حداقل (صفر) است که مجموعه داده خالص است، به این معنا که همه نمونه‌ها به یک کلاس تعلق دارند و عدم قطعیتی وجود ندارد. در زمینه درخت‌های تصمیم، هدف تقسیم داده به نحوی است که آنتروپی کاهش یابد و به درخت کمک کند تا تصمیماتی اتخاذ کند که به تدریج داده را بهتر سازماندهی و طبقه‌بندی کند.

Information Gain یک معیار کارایی برای یک ویژگی خاص در دسته‌بندی داده‌هاست. در زمینه درخت تصمیم، هدف این است که مجموعه داده را به زیرمجموعه‌هایی تقسیم کنیم که از نظر متغیر هدف، به حد امکان یکنواخت باشند. افزایش اطلاعات بالا نشان می‌دهد که ویژگی انتخاب شده در کاهش عدم اطمینان در مورد دسته‌بندی موثر است. الگوریتم‌های درخت تصمیم از افزایش اطلاعات به عنوان یک معیار استفاده می‌کنند تا ویژگی بهترین جهت تقسیم را در هر گره تعیین کنند، با هدف ایجاد زیرمجموعه‌هایی که از نظر متغیر هدف یکنواخت‌تر هستند.

این مفاهیم به درخت‌های تصمیم این امکان را می‌دهند که بهترین مسیرها را با مقایسه ویژگی‌های باقی‌مانده انتخاب کنند و ویژگی‌ای را انتخاب کنند که بهترین انتخاب‌ها را فراهم می‌کند.

دلیل سوم) شبکه‌های عصبی نسبت به چرخش بی‌تغییر هستند. داده واقعی این ویژگی را ندارد.

شبکه‌های عصبی نسبت به چرخش بی‌تغییر هستند. این بدان معناست که اگر دیتاست را چرخانیده باشید، عملکرد آن‌ها تغییر نخواهد کرد. پس از چرخاندن دیتاست‌ها، رتبه‌بندی عملکرد یادگیرنده‌های مختلف معکوس می‌شود، به گونه‌ای که ResNet ها (که بدترین بودند) به عنوان برترین عناصر ظاهر می‌شوند. آن‌ها عملکرد اصلی خود را حفظ می‌کنند، در حالی که تمام یادگیرنده‌های دیگر میزان قابل توجهی از عملکرد خود را از دست می‌دهند. به شکل زیر دقت کنید.

شکل ۴: عملکرد مدل‌ها بعد از اعمال چرخش

حال این سوال مطرح میشود که در واقع، چه معنایی دارد که مجموعه داده‌ها را چرخانده‌ایم؟ چرخش داده معمولاً به فرآیند تغییر یا تغییر جهت یا ترتیب نقاط داده در یک مجموعه داده اشاره دارد. این تکنیکی است که اغلب در یادگیری ماشین و تجزیه و تحلیل داده‌ها برای معرفی تنوع، کاهش تعصب یا افزایش قابلیت‌های تعمیم مدل استفاده می‌شود. چرخش می‌تواند شامل بهم‌زدن ترتیب نقاط داده، تغییر توزیع مقادیر، یا تبدیل مجموعه داده به گونه‌ای باشد که ویژگی‌های اساسی آن را حفظ کند و در عین حال دیدگاه متفاوتی را برای اهداف آموزشی و ارزیابی ارائه دهد.

در همین حین، بیایید بررسی کنیم که چرا تغییرات واریانس چرخشی اهمیت دارند. استفاده از ترکیب‌های خطی از ویژگی‌ها (که این امر باعث تغییرناپذیری در ResNets می‌شود) ممکن است واقعیت و ارتباطات ویژگی‌ها را به نادرستی بازنمایی کند. یک مبنای طبیعی وجود دارد (در اینجا، مبنای اصلی) که بهترین ویژگی‌های مربوط به داده را کدگذاری می‌کند و توسط مدل‌هایی که از چرخش بی‌تغییر هستند و احتمالاً ویژگی‌ها را با ویژگی‌های آماری بسیار متفاوت مخلوط می‌کنند، قابل بازیابی نیست.

نتیجه گیری

در مجموع، این متن نشان می‌دهد که مدل‌های مبتنی بر درخت، به ویژه در تحلیل داده‌های جدولی، نسبت به شبکه‌های عصبی/یادگیری عمیق برتری دارند. دلایل این برتری شامل مقاومت به سوگیری بیش از حد شبکه‌های عصبی، اثر بی‌تغییری در برابر چرخش داده، و توانایی بهتر در مدیریت ویژگی‌های غیراطلاعاتی در مدل‌های مبتنی بر درخت است. همچنین، مدل‌های مبتنی بر درخت به راحتی می‌توانند الگوهای نامنظم و غیرهموار در داده‌های جدولی را یاد بگیرند، در حالی که شبکه‌های عصبی با چالش‌های بیشتری در برازش این توابع نامنظم مواجه می‌شوند. این نتایج مطرح می‌کنند که در موارد خاصی، انتخاب مدل‌های مبتنی بر درخت ممکن است به دلیل عملکرد بهتر در مواجه با خصوصیات خاص داده‌های جدولی مورد ترجیح باشد.

منابع:

Grinsztajn, Léo, Edouard Oyallon, and Gaël Varoquaux. "Why do tree-based models still outperform deep learning on typical tabular data?." Advances in Neural Information Processing Systems ۵۲۰-۵۰۷ :(۲۰۲۲) ۳۵.