مقدمه ای‌بر یادگیری ماشین

سولویگ‌بدیلو[۱،*،†]، بالاژ بنفای[۱]، فابیان‌بیرزله[۱]، یاکوف آی. داویدوف[۱]، لوسی هاچینسون[۱]، تونی‌کام-تونگ[۱]، جولیان سیبور-پولستر[۱]، برنهارد استیرت[۱]، وجیتائو دیوید ژانگ[۱]

درچند سال گذشته، یادگیری ماشین (ML) و هوش مصنوعی موج جدیدی از تبلیغات را به خود دیده اند که ناشی از حجم عظیم و روزافزون داده هاو قدرت محاسباتی و همچنین کشف الگوریتم های یادگیری بهبود یافته است. با این حال، ایده یادگیری مفاهیم انتزاعی از داده ها توسطکامپیوتر و اعمال آنها در موقعیت های ناشناخته، چیز جدیدی نیست و حداقل از دهه 1950 وجود داشته است. بسیاری از این اصول اساسی‌برای جامعه فارماکومتریک و فارماکولوژی بالینی بسیار آشنا هستند. در این مقاله، می خواهیم ایده های اساسی یادگیری ماشین را به این‌جامعه معرفی کنیم تا خوانندگان ابزارهای ضروری مورد نیاز برای درک انتشارات مربوط به این موضوع را به دست آورند. اگرچه ما به جزئیاتو پیشینه نظری نخواهیم پرداخت، اما هدف ما این است که خوانندگان را به ادبیات مربوطه ارجاع دهیم و کاربردهای یادگیری ماشین‌را در زیست شناسی مولکولی و همچنین زمینه های فارماکومتریک و فارماکولوژی بالینی در چشم انداز قرار دهیم.

ظهوردسترسی به داده ها و رشد قدرت محاسباتی، همراه با ورود روش های‌نوین یادگیری، منجر به پیشرفت های چشمگیری در بسیاری‌از حوزه های علمی شده است. این شامل تحقیقات بیولوژیکی‌و بالینی می شود که کاربردهای آن از زیست شناسی مولکولی‌گرفته تا ...[۱]برای تجزیه و تحلیل داده های تصویری[۲]و عمل بالینی.[۳]با این حال، ایده یادگیری مفاهیم انتزاعی توسط کامپیوتر - مانندکاری که انسان ها دائماً انجام می دهند - حداقل از دهه ۱۹۵۰، زمانی‌که اولین شبکه های عصبی ساخته شدند، وجود داشته است.[۴] توسعه‌داده شدند. حتی قبل از آن، روش های دیگری مانند آمار بیزی وزنجیره های مارکوف با ایده ای مشابه مورد استفاده قرار می گرفتند.بسیاری از این روش ها با قراردادهای نامگذاری مختلف برای‌جامعه فارماکومتریک و فارماکولوژی بالینی شناخته شده اند. در سمت‌چپ، اصطلاحات یادگیری ماشین و در سمت راست، نامگذاری‌معمول آمار (بر اساس https://statweb.stanford.edu/~tibs/stat315a/glossary.pdf Tibshirani) را نشان می دهیم:

- شبکه، نمودارها⇔مدل
- وزنه ها⇔پارامترها
- یادگیری⇔مناسب سازی
- تعمیم⇔عملکرد مجموعه آزمون
- یادگیری تحت نظارت⇔رگرسیون یا طبقه بندی
- یادگیری بدون نظارت⇔تخمین چگالی، خوشه بندی
- ویژگی ها⇔متغیرهای کمکی یا توضیحی

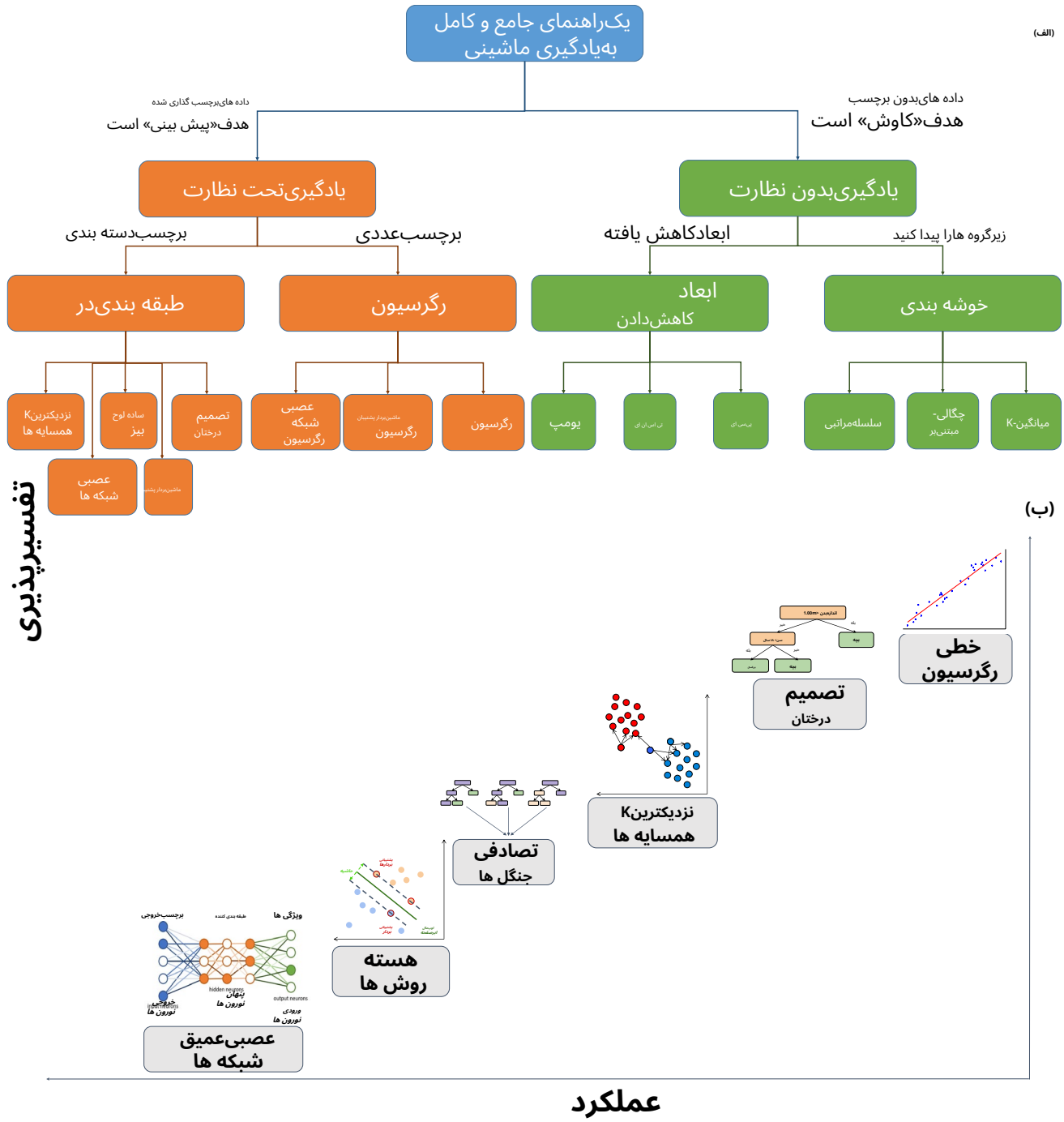
تفاوت‌اصلی با رویکردهای سنتی تر، تا حد زیادی در دو فرهنگ متمایز مدل سازی‌آماری نهفته است. بریمن تقریباً دو دهه پیش از این موضوع طفره‌رفته بود.[۵]در اینجا، ما تعریف او را با گنجاندن مدل های فیزیولوژیکی‌در یکی از فرهنگ ها گسترش می دهیم. به طور خاص، فرهنگ۱ شامل تعیین مدلی برای توصیف داده های مشاهده شده است وفرهنگ ۲ با اتخاذ یک رویکرد مدل سازی الگوریتمی، قصد دارد مسئله راحل کند، بنابراین ذاتاً منجر به مدل هایی با تعداد بیشتری از پارامترهای آزادو پیچیدگی های بیشتر می شود.

تعاملات.این پیچیدگی می تواند چالش هایی را برای تفسیر مدل ایجاد کند(به اصطلاح مشکل "جعبه سیاه"). رویکردهایی که معمولاً در کاربردهای‌فارماکومتری استفاده می شوند، در فرهنگ ۱ قرار می گیرند، که‌در آن یک مدل اساسی بر اساس اصول دارویی و درک خواص دارو فرض‌می شود. چنین مدل هایی معمولاً از نظر فیزیولوژیکی قابل تفسیر هستند.اکثر رویکردهای یادگیری ماشین (ML) در فرهنگ ۲ قرار می گیرند،که در آن هیچ مدل صریحی مشخص نشده است و یک کامپیوترمسئول شناسایی ارتباطات در داده های مشاهده شده است. تفسیراین مدل ها از نظر فیزیولوژیکی دشوار است، با این حال، در طول سال هاپیشرفت قابل توجهی در تفسیرپذیری مدل های ML حاصل شده‌است.[۶،۷]امروزه، بسیاری از جنبه های یک مدل جعبه سیاه را می توان‌با استفاده از ابزارهای مناسب تفسیر کرد.[۸]

دراین مقاله، هدف ما پشتیبانی از خوانندگان برای توسعه شهود مورد نیازبرای درک چگونگی یادگیری کامپیوترها یا کمک به انسان ها در شناسایی‌الگوها در داده ها است. ایده های بنیادی یادگیری ماشین برجسته‌شده اند، اما جزئیات و پیشینه نظری روش های یادگیری ماشین‌موجود را شرح نمی دهیم. خوانندگان علاقه مند را به مقالات یا کتاب های‌دیگر، مانند «عناصر یادگیری آماری» ارجاع می دهیم.[۹]که به عنوانESL شناخته می شود) و ما به نمونه هایی از کاربرد آنها در زیست شناسی‌مولکولی، کشف دارو، توسعه دارو و داروشناسی بالینی اشاره‌می کنیم.

ماابتدا مفاهیم نقاط داده، ویژگی ها، فضاهای ویژگی و معیارهای شباهت رامعرفی می کنیم و سپس عمیق تر به دو حوزه اصلی یادگیری ماشین، یعنی یادگیری‌بدون نظارت و یادگیری تحت نظارت، می پردازیم و جنبه های کلیدی ومثال ها را بررسی می کنیم. در مورد یادگیری بدون نظارت، کامپیوترها وظیفه‌دارند الگوهای ناشناخته در داده ها را بدون دانش از پیش موجود مانند گروه هایا کلاس ها شناسایی کنند، در حالی که در مورد یادگیری تحت نظارت،کامپیوترها وظیفه دارند یاد بگیرند که چگونه کلاس یا مقدار نقاط داده‌هنوز مشاهده نشده را بر اساس یک مفهوم (که اغلب "مدل" نیز نامیده می شود) که از یک مجموعه داده آموزشی استخراج شده است، پیش بینی کنند.**شکل ۱**طبقه بندی روش های مختلف شرح داده شده در این مقاله را نشان‌می دهد و می تواند

[۱]علوم دارویی، تحقیقات و توسعه اولیه داروسازی روشه (pRED)، مرکز نوآوری روشه بازل، بازل، سوئیس.
* مکاتبه: Solveig Badillo (solveig.badillo@roche.com) نویسندگان به
ترتیب‌حروف الفبا. همه نویسندگان به طور مساوی مشارکت داشتند.
[اصلاحیه در تاریخ ۶ مارس ۲۰۲۰، پس از اولین انتشار آنلاین اضافه شد: متن مشارکت نویسنده اضافه شد].
†اس.بی. در زمان نگارش این دست نوشته، در استخدام گروه سولادیس بود.
دریافت شده‌در ۸ اکتبر ۲۰۱۹؛ پذیرفته شده در ۱۵ ژانویه ۲۰۲۰. doi:10.1002/cpt.1796



شکل۱طبقه بندی و مرور کلی الگوریتم های اصلی یادگیری ماشین (ML).الف)طبقه بندی روش های مختلف ارائه شده. (ب)مروری بر روش های یادگیری ماشین. طیف روش های‌موجود از ساده تر و تفسیرپذیرتر تا الگوریتم های پیشرفته تر با عملکرد بالقوه بالاتر به قیمت تفسیرپذیری کمتر، متغیر است. موقعیت روش ها روی شکل کیفی است ودر عمل به تعداد پارامترهای آزاد، پیچیدگی مدل، نوع داده و تعریف دقیق تفسیرپذیری مورد استفاده بستگی دارد.۸.تقریب و تصویر منیفولد یکنواخت ؛UMAP، t؛ جاسازی همسایه‌تصادفی توزیع شده با tSNE، ماشین بردار پشتیبان؛ SVM، تحلیل مؤلفه های اصلی؛ PCA،

به‌عنوان مرجعی، هرچند غیرجامع، در مورد اینکه چه سناریویی برای استفاده از کدام ابزار یادگیری‌ماشین مناسب است، استفاده می شود. لطفاً توجه داشته باشید که تمام روش های بدون‌نظارت در مواردی که برچسب ها در دسترس باشند نیز قابل اجرا هستند.

داده هاو ویژگی ها

دریادگیری ماشین، ما با داده ها و مجموعه داده ها سروکار داریم. یک مجموعه داده از چندین نقطه‌داده (که گاهی نمونه نیز نامیده می شوند) تشکیل شده است، که در آن هر نقطه‌داده، موجودیتی را نشان می دهد که می خواهیم آن را تجزیه و تحلیل کنیم. بنابراین،یک نقطه داده می تواند هر چیزی مانند یک بیمار یا نمونه ای که از بافت سرطانی‌گرفته شده است را نشان دهد. بسیاری از مسائل مربوط به داده ها، جهانی‌هستند و نه تنها بر رویکردهای یادگیری ماشین، بلکه بر هر رشته کمی، از جمله‌فارماکومتری، تأثیر می گذارند.

برای‌گردآوری مجموعه داده ها، تعدادی از ویژگی ها (یعنی داده هایی که ویژگی های‌داده ها را توصیف می کنند) اندازه گیری و جمع آوری شده اند.

امتیاز). این ویژگی ها می توانند دسته بندی شده (مقادیر از پیش تعریف شده‌بدون ترتیب خاص مانند مرد و زن)، ترتیبی (مقادیر از پیش‌تعریف شده که دارای ترتیب ذاتی مانند مرحله بیماری هستند) یاعددی (مثلاً مقادیر واقعی) باشند. برای یک بیمار در یک محیط بالینی،این ویژگی ها می توانند (ترکیبی از) اطلاعات جمعیت شناختی‌بیمار، سابقه بیماری، نتایج آزمایش خون یا معیارهای‌پیچیده تر و با ابعاد بالا، مانند پروفایل های بیان ژن در یک بافت‌خاص یا تمام پلی مورفیسم های تک نوکلئوتیدی که نشان دهنده‌ژنوم منحصر به فرد بیمار هستند، باشند.

هرویژگی، یک بعُد از فضای ویژگی را نشان می دهد و مقدار مشخص‌یک ویژگی برای یک نقطه داده خاص، آن نقطه را در مکانی تعریف شده‌در این بعُد از فضا قرار می دهد. در مجموع، تمام مقادیر تمام‌ویژگی های یک نقطه داده، بردار ویژگی نامیده می شود. هرچه ویژگی های‌بیشتری برای مجموعه داده ها جمع آوری کرده باشیم، برداررویژگی حاصل و فضای ویژگی، ابعاد بالاتری خواهند داشت. بدیهی‌است که با افزایش ابعاد، تجسم تمام ابعاد فضای ویژگی، دشوارمی شود و ما باید برای شناسایی الگوهای مربوطه به کامپیوتر تکیه‌کنیم یا از روش های کاهش ابعاد استفاده کنیم، همانطور که بعداً در بخش «کاهش ابعاد» توضیح داده خواهد شد.

داروشناسان‌بالینی معمولاً با داده های طولی، مانند پروفایل های فارماکوکینتیک(PK) و فارماکودینامیک (PD)، که در آن ها وابستگی به زمان‌نقش محوری دارد، آشنا هستند. در واقع، مدل های مورد استفاده درفارماکومتری مبتنی بر معادلاتی هستند که می توانند بر اساس فیزیولوژی‌و فارماکولوژی توجیه شوند و بینش هایی در مورد تکامل زمانی‌سیستم ارائه دهند. این امر، به عنوان مثال، مشابه مسائل فیزیکی‌مانند پیش بینی آب و هوا است که در آن جریان هوا و دما منجر به‌رفتار زمانی خاصی از سیستم می شوند. در یادگیری ماشین، گنجاندن زمان‌به عنوان یک متغیر پیوسته متمایز در الگوریتم های مربوطه، همچنان‌چالش برانگیز است و حوزه ای از تحقیقات فعال است. در حال حاضر،چندین گزینه برای گنجاندن داده های وابسته به زمان در مجموعه داده های‌یادگیری ماشین وجود دارد: یا به طور مستقیم که در آن هر نقطه‌زمانی نشان دهنده یک ویژگی است، یا از طریق تبدیلاتی مانند تبدیل‌فوریه یا B-splines که منجر به ضرایب توابع اساسی می شود که می توانندبه عنوان ویژگی در نظر گرفته شوند. از طرف دیگر، شبکه های عصبی‌بازگشتی (RNN) می توانند برای مدیریت داده های طولی، همانطورکه در بخش "شبکه عصبی بازگشتی" ذکر شده است، استفاده شوند.با این حال، همه این رویکردها محدودیت گسسته سازی بعُد زمانی- چه مستقیم و چه غیرمستقیم - را دارند.

اکثرالگوریتم های یادگیری ماشین برای مدیریت مجموعه داده های با ابعاد بالاطراحی شده اند. از این رو، ویژگی های مشتق شده از داده های موجود اغلب‌شامل می شوند، مانند داده های تبدیل شده با لگاریتم، حاصلضرب ها ونسبت های ویژگی ها یا ترکیبات پیشرفته تر. چنین تبدیل داده ای یک مرحله‌پیش پردازش مهم است که می تواند تأثیر عمیقی بر عملکرد مدل داشته‌باشد. بنابراین، همیشه ایده خوبی است که از دانش و تخصص موجود درحوزه برای دستیابی به ویژگی های مرتبط استفاده شود، فرآیندی که گاهی اوقات‌به عنوان مهندسی ویژگی شناخته می شود.

کیفیت‌داده ها نقش حیاتی در یادگیری ماشینی ایفا می کند. روش های یادگیری‌ماشینی که با دقت انتخاب شده اند و بازرسی بصری، از مقادیر شدیدیا داده های پرت جلوگیری می کنند. با این حال، داده های گمشده می توانندچالش برانگیز باشند. همه روش ها از داده های گمشده پشتیبانی نمی کنندو در چنین مواردی، تبدیل داده ها می تواند به عنوان یک مرحله پیش پردازش‌مورد نیاز باشد. روش های مختلفی برای جایگذاری داده های گمشده‌وجود دارد که عملکرد آنها به مجموعه داده ها و روش مورد استفاده بستگی‌دارد.[10]بدیهی ترین رویکرد برای جانهی، جایگزینی یک مقدار گمشده با میانگین‌ویژگی در تمام نمونه هایی که در آنها تعریف شده است. با این حال، این‌گاهی اوقات می تواند باعث بیش برازش شود[11](همچنین به بخش «معیارهای‌عملکرد و مسئله بیش برازش» مراجعه کنید).

همچنین‌بررسی دقیق هرگونه سوگیری در داده ها (مثلاً سوگیری انتخاب) ضروری است. ترجیحاً، نمونه های ML باید یک زیرمجموعه تصادفی‌بدون سوگیری از جمعیت باشند. در عمل، این امر به ندرت اتفاق‌می افتد و برخی سوگیری ها در داده ها وجود دارد. این سوگیری هامی توانند بر توانایی مدل برای تعمیم فراتر از مجموعه داده های‌آموزشی (و حتی مجموعه داده های آزمایشی اگر هر دو سوگیری‌مشابهی داشته باشند) تأثیر بگذارند. نمونه ای از چنین مسئله‌تعمیمی، مدلی است که قرار است یاد بگیرد چگونه گرگ را ازهاسکی بر اساس ویژگی های حیوان تشخیص دهد، اما در نهایت معلوم‌می شود که به سادگی تکه های برف را روی عکس تشخیص می دهد.[6]رویکردهای مختلفی برای کاهش سوگیری وجود دارد (مثلاً می توان‌نمونه ها یا ویژگی های سوگیرانه را کم وزن کرد یا به طور کامل‌حذف کرد).[12]به طور خاص، نمرات گرایش هنگام تخمین اثر یک مداخله‌درمانی مفید هستند.[13]بررسی اهمیت ویژگی، اطلاعات ارزشمندی‌در مورد بزرگی و تأثیر بایاس ارائه می دهد.[6,7]که توصیه می شودبرای بررسی قابلیت اعتماد مدل های یادگیری ماشینی استفاده‌شود.

بسیاری‌از مجموعه داده های طبقه بندی بالینی نامتوازن هستند، به این معنی‌که یک یا چند کلاس کمتر از حد واقعی نمایش داده شده اند. این می تواندبرای بسیاری از الگوریتم های یادگیری ماشین، از جمله شبکه های عصبی‌مصنوعی و روش های تقویت گرادیان، مشکلاتی ایجاد کند. یک راه برای‌کاهش این مشکل، نمونه گیری کمتر/بیش از حد به ترتیب از کلاس اکثریت/اقلیت، یا تنظیم هزینه طبقه بندی نادرست در تابع هدف است.[14]

درنهایت، برای بسیاری از کاربردها، تعریف یک معیار شباهت یا فاصله‌بین دو نقطه داده در فضای ویژگی مهم است. ساده ترین معیارفاصله، فاصله اقلیدسی است:

$$\text{(الف، ب)}=\sqrt{\sum_{\text{من}=1}^{\text{ن}}(\text{یک‌من‌ب‌من})^{2}} \quad (1)$$

بین‌بردارهای ویژگی عددی دو نقطه داده *الف* و *ب*، برای ویژگی ها=1... ن، اما بسته به نوع داده ای که با آن سروکار داریم، می تواند معیارهای‌فاصله یا شباهت بسیار پیچیده تری مانند شباهت کسینوسی‌وجود داشته باشد.[15]
یاامتیازهای شباهت دو توالی بیولوژیکی.[16]

غذاهای‌اصلی

- تبدیل داده های ورودی و مهندسی ویژگی ها ممکن است مدل را بهبودبخشد.
- داده های گمشده نیاز به جانهی دارند.
- سوگیری های موجود در داده ها باید به دقت بررسی شوند.
- مجموعه داده های نامتوازن نیاز به اصلاح مدل دارند.
- معیارهای معنادار شباهت بین نمونه ها باید تعریف شوند.

یادگیری‌بدون نظارت

درتحلیل اکتشافی داده ها، ما اغلب «برچسب های» واقعی را نمی دانیم،یا ممکن است بخواهیم الگوهای طبیعی در حال ظهور در داده هارا بررسی کنیم. برای این منظور، می توانیم از روش های یادگیری بدون‌نظارت، مانند خوشه بندی، تشخیص الگوهای مکرر و کاهش ابعاد استفاده‌کنیم. در اینجا، ما به طور خاص تمرکز خواهیم کرد.

درمورد خوشه بندی و کاهش ابعاد، زیرا کاربردهای زیادی در زیست شناسی‌مولکولی و عمل بالینی دارند.

خوشه بندی

هدف‌از به کارگیری روش های خوشه بندی، شناسایی زیرگروه های مرتبط در یک مجموعه‌داده معین بدون داشتن فرضیه ای از پیش تعریف شده در مورد ویژگی هایی‌است که زیرگروه ها ممکن است داشته باشند. به عنوان مثال، در گروهی‌از بیماران مبتلا به یک بیماری خاص، ممکن است بخواهیم زیرگروه هایی را شناسایی‌کنیم که نشان دهنده مکانیسم های بیولوژیکی متمایزی هستند که بیماری‌را بر اساس اقدامات مولکولی انجام شده هدایت می کنند.[۱۷]

یک‌خوشه زیرمجموعه ای از داده ها است که به یکدیگر «شبیه» هستند،در حالی که نقاط متعلق به خوشه های مختلف «متفاوت تر» هستند.رویکردهای متعددی برای خوشه بندی وجود دارد که از الگوریتم های‌اساسی مختلفی برای گروه بندی نقاط داده بر اساس «شباهت»آنها استفاده می کنند. همه آنها مزایا و معایبی دارند و بسته‌به کاربرد و ویژگی های داده ها، باید با دقت انتخاب شوند.

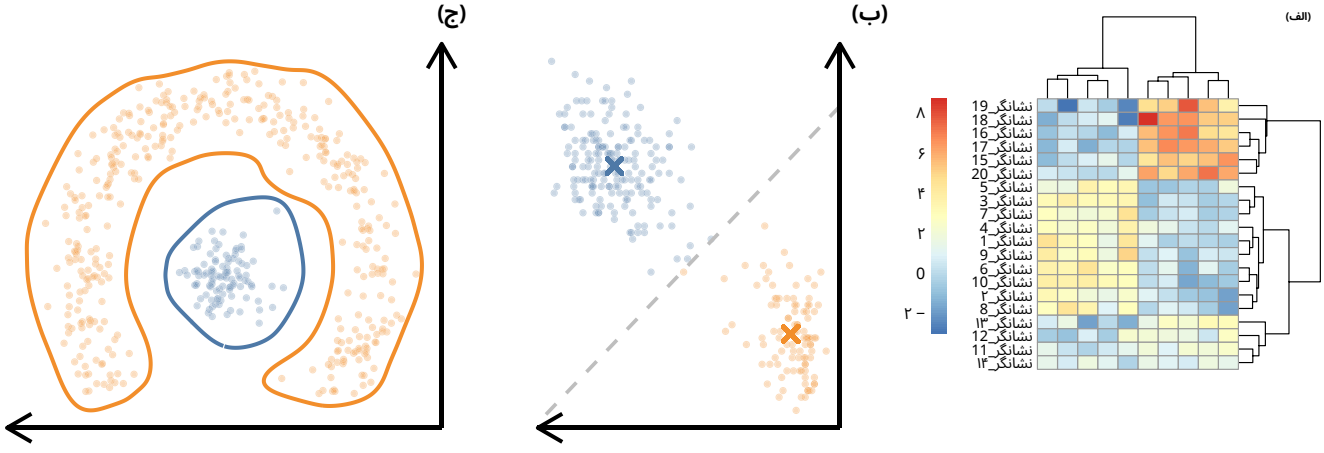
یک‌رویکرد ساده برای خوشه بندی‌ک-به معنی خوشه بندی است.[۱۸]در اینجا،تعداد خوشه هایی که باید شناسایی شوند، توسط یک پارامتر انتخابی کاربراز پیش تعریف شده است.ک‌هر خوشه توسط یک مرکز خوشه نمایش داده‌می شود، که یک نقطه داده مصنوعی است که نشان دهنده میانگین (یا میانه) مقدار تمام نقاط اختصاص داده شده به این خوشه است. در ابتدا،ک مراکزخوشه، که به عنوان "دانه ها" شناخته می شوند، به طور تصادفی در فضای‌ویژگی قرار می گیرند. سپس الگوریتم دو مرحله را تکرار می کند. در مرحله‌اول ("انتساب")، نقاط داده به خوشه ای که توسط نزدیکترین مرکز نشان‌داده شده است، اختصاص داده می شوند. در مرحله دوم ("تغییر مرکز")، موقعیت هر مرکز خوشه بر اساس ترکیب خوشه ها پس از مرحله اول به روز می شود.پس از تعدادی تکرار، این معمولاً به یک بهینه محلی همگرا می شودکه در آن انتساب خوشه ها تغییر نمی کند یا فقط به میزان کمی تغییرمی کند. نتیجه چنین فرآیندی در شکل نشان داده شده است. **شکل ۲ب**اگرچه این روش شهودی است، اما اشکال اصلی آن این است که معمولاً خوشه بندی‌به شدت تحت تأثیر مقدار ... قرار می گیرد.ک‌و اغلب تعداد واقعی‌خوشه ها در داده ها ناشناخته است*به طور پیشینی*از آنجا که در خوشه بندی‌به ندرت یک پاسخ درست یا غلط قطعی وجود دارد، بررسی بیشترخوشه بندی برای شناسایی‌خوشه های معنادار، که می تواند به ویژه با توجه به فضای ویژگی‌با ابعاد بالا چالش برانگیز باشد.

گروه‌دیگری از روش های خوشه بندی، خوشه بندی مبتنی بر چگالی است.[۱۹]در روش های‌مبتنی بر چگالی، یک خوشه بخشی از فضای ویژگی را نشان می دهد که درآن نقاط داده متراکم هستند. نقاط داده متعلق به مناطقی‌ازفضای ویژگی با چگالی‌کم، نویز در نظر گرفته می شوند. یکی از الگوریتم های خوشه بندی مبتنی بر چگالی‌شناخته شده، خوشه بندی مکانی مبتنی بر چگالی برنامه های دارای نویز است.[۲۰]خوشه بندی مبتنی بر چگالی نیازی به مقدار از پیش تعریف شده ای برای تعیین‌تعداد خوشه ها ندارد و نتیجه ای تکرارپذیر ارائه می دهد. علاوه بر این، قادر به‌شناسایی اشکال پیچیده خوشه ها، مانند آنچه در شکل نشان داده شده است، نیزمی باشد.**شکل ۲ج**.

درتحلیل خوشه بندی سلسله مراتبی، هدف ساخت سلسله مراتبی از خوشه هااست (ESL، فصل ۱۴).[۹]یک رویکرد ساده برای خوشه بندی سلسله مراتبی،اتصال همسایه است. ابتدا، تمام فواصل جفتی بین تمام نقاط داده درمجموعه داده ها محاسبه می شوند. سپس، در هر مرحله از یک فرآیند تکراری،دو نقطه داده با کمترین فاصله با هم گروه بندی می شوند. این منجر به‌یک ساختار خوشه بندی درخت مانند می شود، همانطور که در شکل نشان داده‌شده است.**شکل ۲ الف**در سمت چپ و بالای نقشه حرارتی که طول شاخه های‌درخت نشان دهنده فواصل نمونه ها است. برای رسیدن به مجموعه ای‌گسسته از خوشه ها مانندک‌یعنی باید یک آستانه فاصله انتخاب شودکه در آن درخت به صورت افقی بریده شود. باز هم، هیچ راه بهینه ای برای‌انتخاب چنین آستانه ای وجود ندارد و ممکن است راه حل های معقول زیادی‌وجود داشته باشد. خوشه بندی سلسله مراتبی می تواند به تنهایی یا در ترکیب‌با نقشه های حرارتی (مثلاً**شکل ۲ الف**) برای تجسم ویژگی های انتخاب‌شده یا همه آنها، به عنوان مثال، داده های بیان ژن.

کاهش‌ابعاد

تعدادویژگی ها و بنابراین، ابعاد فضای ویژگی می تواند بسیار زیاد باشدو ده ها هزار معیار در هر نمونه داشته باشد. این امر نه تنها تجسم‌داده ها را چالش برانگیز می کند، بلکه تجزیه و تحلیل آنها را نیز چالش برانگیزمی کند. به طور خاص، تجزیه و تحلیل مجموعه داده های‌با ابعاد بالا می تواند با پدیده ای به نام «نفرین ابعاد» مرتبط باشد.[۲۱]که به پراکندگی داده ها و ویژگی های هندسی غیرشهودی در داده های‌با ابعاد بالا اشاره دارد



شکل۲مروری بر نتایج رویکردهای مختلف خوشه بندی. (الف)نتایج یک خوشه بندی سلسله مراتبی دوبعدی را نشان می دهد. دو دندروگرام، شباهت بین نمونه ها و همچنین بین نشانگرهای‌اندازه گیری شده را تجسم می کنند. چنین تجسمی اغلب در زیست شناسی برای بیان ژن یا سایر خوانش های فناوری -omics استفاده می شود. (ب)نتیجه یک خوشه بندی‌کلاسیک را با استفاده از نشان می دهدک-یعنی با مقدار انتخابی‌ک=۲. خوشه های حاصل معمولاً محدب هستند و هر نقطه به یک خوشه اختصاص داده می شود، یعنی‌خوشه ای که با نزدیکترین نقطه مرکزی (با X مشخص شده است) نشان داده شده است. (ج)نتیجه خوشه بندی مبتنی بر چگالی را نشان می دهد. لطفاً توجه داشته باشید که این‌رویکرد می تواند اشکال خوشه های غیرمحدب، مانند خوشه نارنجی، را شناسایی کند.

فضاها.«نفرین ابعاد» چالش هایی را در اکثر رویکردهای تحلیل داده ها، از جمله یادگیری‌ماشین (ML) ایجاد می کند، اما محدود به آنها نیست.

برای‌کاهش چنین مشکلاتی، می توان از روش های کاهش ابعاد استفاده‌کرد. کاهش ابعاد می تواند با تبدیل هر نقطه داده با ابعاد بالا به‌دو یا چند بعد، ضمن حفظ اکثر تغییرات و فواصل نسبی، به تجسم‌داده ها کمک کند. علاوه بر این، حذف ویژگی های غیرمفید می تواندعملکرد مدل و زمان همگرایی را بهبود بخشد. اگرچه برخی از این‌روش ها، مانند تحلیل مؤلفه های اصلی، مدت ها قبل از ابداع اصطلاح‌یادگیری ماشین توسعه یافته اند،[۲۲]موارد دیگر، مانند جاسازی‌همسایه تصادفی توزیع شده با t[۲۳]یا تقریب و تصویر منیفولد یکنواخت،[۲۴]اخیراً توسعه یافته اند و به چالش های پیچیده ای که در تجزیه‌و تحلیل داده ها ایجاد می شوند، می پردازند. همچنین یک رویکردقدرتمند کاهش ابعاد مبتنی بر شبکه عصبی به نام خودرمزگذار وجوددارد. برای جزئیات بیشتر در مورد نحوه اعمال کاهش ابعاد در داده های‌زیست پزشکی، خواننده را به یک بررسی اخیر ارجاع می دهیم.[۲۵]

نمونه هایی‌از کاربردهای یادگیری ماشینی بدون نظارت

خوشه بندی‌به طور گسترده در تجزیه و تحلیل داده های با ابعاد بالا، مانندآزمایش های رونویسی، متابولومیک و پروتئومیک، مورد استفاده‌قرار می گیرد. معمولاً از خوشه بندی سلسله مراتبی برای شناسایی‌عوامل اصلی مؤثر بر خوانش ها و همچنین برای شناسایی ماژول هایی‌با درجه بالایی از تنظیم همسو استفاده می شود. در توالی یابی‌تک سلولی، از خوشه بندی غیرسلسله ای برای درک انواع سلول های‌موجود در نمونه استفاده می شود. خوشه بندی همچنین برای‌شناسایی روابط بین بیماران، بافت ها، بیماری ها یا حتی علائم بیماری‌استفاده می شود.[۲۶-۲۹]خود ترکیبات دارویی نیز ممکن است بر اساس‌بیان ژن، حساسیت و خواص پروتئین هدف دسته بندی شوند.[۳۰-۳۲]با هدف هدایت کشف دارو.

کاهش‌ابعاد به طور معمول در آزمایش های ترانسکریپتومیک و سایرآزمایش های -omics، معمولاً برای شناسایی داده های پرت و اثرات‌دسته ای بالقوه، استفاده می شود. در توالی یابی تک سلولی، تقریب‌منیفولد یکنواخت و تصویرسازی یا جاسازی همسایه تصادفی توزیع شدهt هم برای تجسم داده ها و هم برای خوشه بندی بعدی استفاده‌می شوند.[۲۴]کاهش ابعاد همچنین برای تجسم فضای شیمیایی‌با ابعاد بالا استفاده می شود.[۳۳]یا به عنوان یک مرحله پیش پردازش‌برای بهبود عملکرد یک مدل یادگیری ماشین.[۳۴]

غذاهای‌اصلی

- خوشه بندی می تواند برای درک ساختار داده ها با گروه بندی مشاهدات‌مشابه با هم مورد استفاده قرار گیرد.
- کخوشه بندی به معنای ابزاری ساده اما قدرتمند است، با این حال،تعداد خوشه ها باید از قبل مشخص شود.
- روش های مبتنی بر چگالی به تعداد از پیش تعیین شده ای از خوشه ها نیاز ندارندو امکان شناسایی الگوهای پیچیده در داده ها را فراهم می کنند.
- خوشه بندی سلسله مراتبی، یک نمای کلی از رابطه در سطوح مختلف ارائه‌می دهد.
- کاهش ابعاد نه تنها برای تجسم داده ها، بلکه برای حذف ویژگی های‌غیرمفید نیز استفاده می شود.

یادگیری‌تحت نظارت

دریک مسئله یادگیری نظارت شده، داده های آموزشی به همراه مشاهدات و مقادیرخروجی شناخته شده ی متناظر به کامپیوتر داده می شود. هدف، یادگیری قوانین‌کلی (که اغلب «مدل» نیز نامیده می شود) است که ورودی ها را به خروجی هانگاشت می کنند، به طوری که پیش بینی ... امکان پذیر باشد.

خروجی‌برای داده های جدید دیده نشده، که در آن مقادیر ورودی را مشاهده کرده ایم اما خروجی‌مرتبط با آنها را ندیده ایم.

دودسته اصلی از یادگیری نظارت شده وجود دارد: (۱) طبقه بندی که در آن مقادیرخروجی به صورت دسته بندی هستند، و (۲) رگرسیون که در آن مقادیر خروجی‌به صورت عددی هستند.

دربخش های بعدی، زمینه برازش مدل در یادگیری نظارت شده و مسئله‌رایج بیش برازش معرفی می شوند. سپس، توضیح می دهیم‌که چگونه عملکرد برای ابزارهای طبقه بندی و رگرسیون ارزیابی‌می شود (یعنی چگونه کیفیت نگاشت از ورودی ها به خروجی هاتوسط الگوریتم ارزیابی شود). این جنبه ضروری است، زیرا مزیت‌اتخاذ روش های یادگیری ماشین اغلب حول محور دستیابی به عملکردبالاتر با بده بستان تفسیرپذیری متمرکز است. درک معیارهای مختلف‌عملکرد، ارزیابی بهتر مزایای یک مدل پیشنهادی را امکان پذیرمی کند، برخلاف این فرض که یک راه حل یادگیری ماشین همیشه‌می تواند از یک رویکرد سنتی بهتر عمل کند.

سپس‌به بررسی برخی از روش های طبقه بندی و رگرسیون موجود می پردازیم،از سطح پایین شروع می کنیم، جایی که تفسیر مدل ها هنوزسرراست است، و به سمت رویکردهای یادگیری ماشین محورترپیش می رویم که در آن ها عملکرد، اغلب به قیمت از دست رفتن‌تفسیرپذیری، پیروز می شود.**شکل ۱**طیف روش های موجود را ازنظر عملکرد و قابلیت تفسیر خلاصه می کند. این بخش با مروری نه‌چندان جامع بر کاربردهای روش های یادگیری نظارت شده در زیست شناسی‌و به ویژه داروشناسی بالینی به پایان می رسد.

معیارهای‌عملکرد و مسئله ی بیش برازش

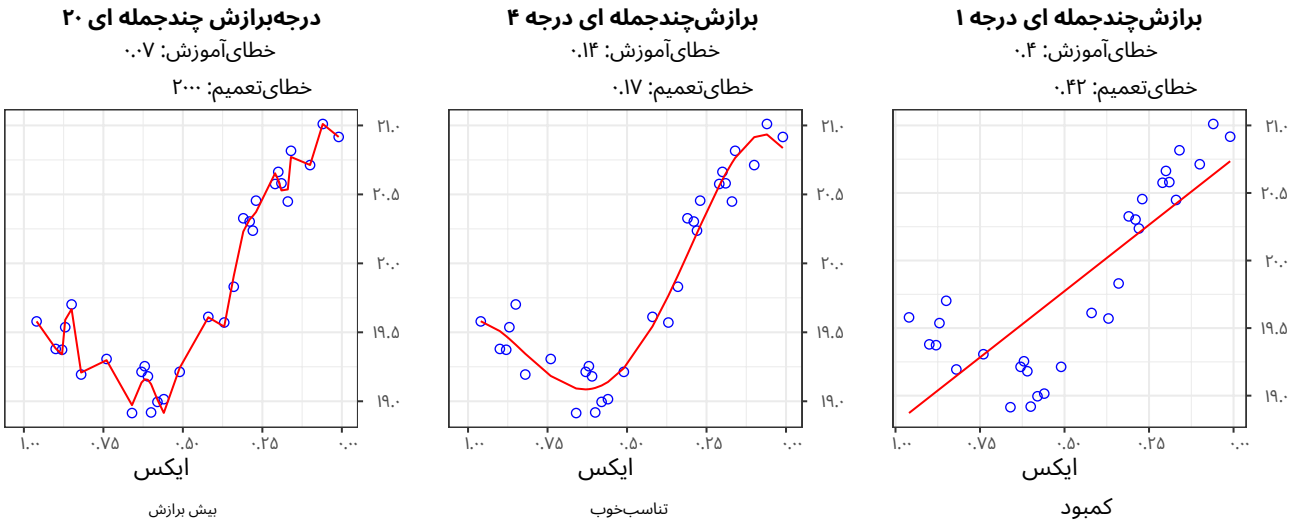
هدف‌یک الگوریتم یادگیری، یادگیری یک مفهوم یا تابع (= یک مدل) است که داده های آموزشی‌مشاهده شده را توصیف می کند و قادر است با اجتناب از کم برازش (underfitting) و بیش برازش (overfitting)، آن را روی داده های مستقل جدید تعمیم‌دهد.

عملکردیک مدل با روش هایی ارزیابی می شود که امکان ارزیابی مدل‌را فراهم می کنند (یعنی تخمین میزان عملکرد کلی یک مدل مشخص‌و انتخاب مدل؛ و تخمین عملکرد مدل های مختلف برای انتخاب‌مناسب ترین مدل). برخی از این روش ها در بخش های بعدی‌برجسته شده اند.

برازش‌مدل.پارامترهای مدل بر اساس داده های مشاهده شده در مجموعه‌آموزشی تخمین زده می شوند. برای استخراج مقادیر بهینه پارامترها(مثلاً برای ضرایب و وزن ها)، یک معیار فاصله بین مدل و داده هاتعریف و به صورت عددی کمینه می شود. مستقل از معیار انتخاب شده،هدف از برازش مدل همیشه تخمین پارامترها با کمینه کردن‌فاصله است که به آن فاصله نیز گفته می شود.*تابع زیان*یا*تابع هزینه*، با دو الزام:

- مدل باید مقادیر پیش بینی شده ای را ارائه دهد که به مقادیر مشاهده شده در مجموعه‌آموزش نزدیک باشند، در غیر این صورت می گوییم که*زیربرازش*دارد و از درجه بالایی‌برخوردار است*تعصب*.
- مدل باید فراتر از مجموعه آموزشی تعمیم یابد. مدلی که*بیش برازش*روی مجموعه‌آموزش خوب پیش بینی می کند اما روی یک مجموعه آزمون مستقل‌ضعیف است، اغلب به این دلیل که برای داده ها خیلی پیچیده است.در این مورد، ما همچنین در مورد*واریانس بالا*.

درادامه، تماس خواهیم گرفت*تابع هدف*هر تابعی که برای تخمین پارامترهای‌مدل بهینه شده باشد.



شکل۳نمایش مسئله ی کم برازش/بیش برازش در یک مورد رگرسیون ساده. نقاط داده با نقاط آبی و برازش مدل با خطوط قرمز نشان داده شده اند. کم برازش با یک مدل خطی(پانل سمت چپ)، برازش خوب با یک چندجمله ای درجه ۴ (پانل مرکزی) و بیش برازش با چندجمله ای درجه ۲۰ (پانل سمت راست) رخ می دهد. جذر میانگین مربعات خطابه عنوان تابع هدف برای ارزیابی خطای آموزش و خطای تعمیم انتخاب شده است که با استفاده از اعتبارسنجی متقابل ۱۰-fold ارزیابی می شود.

درحالت رگرسیون،**شکل ۳**مسئله ی کم برازش و بیش برازش را در زمینه ی رگرسیون‌نشان می دهد. کم برازش می تواند زمانی رخ دهد که مدل خیلی ساده باشدیا وقتی ویژگی های استخراج شده از داده ها به اندازه ی کافی آموزنده نباشند (**شکل ۳**، پنل سمت چپ). بیش برازش اغلب زمانی رخ می دهد که مدل بیش ازحد پیچیده باشد یا ویژگی های زیادی در یک مجموعه کوچک از مثال های آموزشی‌وجود داشته باشد (**شکل ۳**، پنل سمت راست).

این‌مشکل کم برازش/بیش برازش اغلب به عنوان بده بستان بایاس/واریانس‌نیز شناخته می شود، که از بیان خطای پیش بینی مورد انتظار، شامل‌هر دو عبارت بایاس و واریانس، ناشی می شود. بایاس نشانه ای ازمیانگین خطای مدل برای مجموعه های آموزشی مختلف است: این اختلاف‌بین میانگین مقادیر پیش بینی شده و میانگین واقعی است که ماسعی در پیش بینی آن داریم. واریانس نشان دهنده حساسیت مدل به مجموعه‌آموزشی است: برای یک نقطه معین، با پراکندگی مقادیر پیش بینی شده‌در اطراف میانگین آنها مطابقت دارد.

برای‌به حداقل رساندن خطای پیش بینی شده، بین به حداقل رساندن بایاس‌و واریانس، بده بستانی وجود دارد: افزایش پیچیدگی مدل، بایاس راکاهش می دهد اما واریانس را افزایش می دهد. برای ساخت مدل های‌ساده تر، تکنیک های مختلفی وجود دارد که تحت عنوان منظم سازی‌خلاصه شده اند. اصل این روش شامل اصلاح تابع هدف با اضافه‌کردن عبارات جریمه ای است که بر تخمین پارامتر تأثیر می گذارند.منظم سازی L1 و L2 رایج ترین آنها هستند (ESL، بخش های 3.4.1و 3.4.2).[۹]

دسته‌بندی های مختلف توابع زیان.هدف متفاوت توابعی‌را می توان برای اندازه گیری فاصله بین داده های مشاهده شده و مقادیرپیش بینی شده توسط مدل انتخاب کرد. برخی از معیارهای فاصله ای‌که در عمل استفاده می شوند را می توان به ... مرتبط کرد. *احتمال*درستنمایی نشان می دهد که مشاهده داده های ما طبق مدل انتخاب‌شده چقدر محتمل است. رایج ترین کاربرد درستنمایی، یافتن پارامترهایی‌است که مدل را به طور بهینه با داده ها برازش می دهند (یعنی‌تخمین های پارامتر درستنمایی حداکثری). معمولاً، لگاریتم منفی درستنمایی‌به حداقل می رسد و به عنوان تابع هدف در نظر گرفته می شودزیرا خواص عددی مطلوبی دارد. به طور مشابه، در معیارهای یادگیری‌ماشین، مانند میانگین‌مربعات خطا، تابع هدف لجستیک یا آنتروپی متقاطع، برای یافتن‌پارامترهای بهینه یا ارزیابی برازش مدل استفاده می شوند.

درعمل، محاسبه تحلیلی حداکثر درستنمایی یا حداقل زیان ممکن است‌امکان پذیر نباشد و اغلب لازم است از یک الگوریتم بهینه سازی‌عددی برای حل مسئله و یافتن بهترین مقادیر پارامتر استفاده‌شود.*گرادیان نزولی*چنین الگوریتمی است که در آن ابتدا یک تابع‌هدف تعریف می کنیم که می خواهیم آن را به حداقل برسانیم و سپس *به طور تکراری*مقادیر پارامترها را در جهتی که شدیدترین کاهش(مشتق مرتبه اول) تابع هدف رخ می دهد، به روزرسانی کنید تازمانی که همگرایی به حداقل فاصله حاصل شود. در سناریوی تابع هدف‌غیرمحدب، موفقیت یافتن یک حداقل سراسری، برخلاف رسیدن‌به برخی حداقل های محلی، به انتخاب مجموعه اولیه مقادیر پارامتر،نرخ یادگیری (یعنی اندازه گام هر تکرار) و معیار همگرایی بستگی‌دارد. خواننده می تواند به مرجع ... مراجعه کند.[۳۵]برای جزئیات‌بیشتر در مورد فرآیندهای بهینه سازی محدب و غیرمحدب. نزول‌گرادیان تصادفی یک ترفند اضافی است که می تواند با نمونه برداری‌تصادفی از یک مجموعه داده آموزشی و جمع کردن فواصل‌در این زیرمجموعه از نقاط داده آموزشی برای تقریب تابع هدف،بهینه سازی را سرعت بیشتری بخشد.

اصول‌کلی انتخاب و ارزیابی مدل.مشکل بیش برازش‌نشان می دهد که عملکرد مدل روی مجموعه داده آموزشی،شاخص خوبی برای عملکرد آن روی مجموعه داده جدید نیست.در ادامه اصول ارزیابی عملکرد مدل در یک محیط یادگیری نظارت شده‌را شرح خواهیم داد.

اصل‌کلی انتخاب مدل به شرح زیر است: وقتی داده های کافی وجود داشته باشد،آنها را به سه زیرمجموعه تقسیم می کنیم - مجموعه های آموزشی، اعتبارسنجی‌و آزمون. مجموعه آموزشی برای ساخت مدل های مختلف استفاده می شود،در حالی که مجموعه اعتبارسنجی متعاقباً برای انتخاب الگوریتم و در صورت‌نیاز، انتخاب ابرپارامترها استفاده می شود. سپس، مدلی که بهترین عملکردرا در مجموعه اعتبارسنجی دارد انتخاب می شود. در نهایت، مجموعه آزمون‌امکان ارزیابی خطای تعمیم را فراهم می کند که به آن خطای تعمیم نیز گفته می شود.*خطای آزمون*که خطای پیش بینی روی یک مجموعه داده آزمایشی است

که‌در طول آموزش استفاده نشده است.۹لازم به ذکر است که وقتی مجموعه‌داده های اصلی دارای سوگیری باشند، خطای تعمیم می تواندبیشتر از حد انتظار باشد (به بخش «داده ها و ویژگی ها» مراجعه‌کنید). اعتبارسنجی مدل در برابر یک مجموعه داده آزمایشی کاملاً مستقل، روش استاندارد طلایی برای ارزیابی تعمیم پذیری مدل است.

وقتی‌مجموعه داده ها برای استخراج یک مجموعه اعتبارسنجی مناسب بسیار کوچک‌است، به عنوان مثال، می توان از تکنیک های اعتبارسنجی متقابل برای انتخاب‌ابرپارامترهای مدل استفاده کرد. پس از کنار گذاشتن زیرمجموعه ای از داده هابرای آزمایش،ک‌اعتبارسنجی fold شامل تقسیم مجموعه آموزشی به ... است.ک‌زیرمجموعه ها،ک–ازیرمجموعه هایی که برای آموزش استفاده می شوند وزیرمجموعه آخری که برای ارزیابی عملکرد استفاده می شود. این فرآیند تکرار می شودک‌بار، هر کدام‌ک‌زیرمجموعه یک بار برای اعتبارسنجی استفاده می شود و سپس‌نمرات عملکرد هر زیرمجموعه برای هر مجموعه از ابرپارامترها برای آزمایش،میانگین گیری می شوند.ک‌روش اعتبارسنجی متقابل برابر در خلاصه شده‌است**شکل ۴**برای انتخاب بین الگوریتم های یادگیری مختلف۳۶می توان از اعتبارسنجی‌متقابل تو در تو استفاده کرد.

شاخص های‌پیچیدگی مدل در مقابل خوبی برازش.در

درفارماکومتریک، انتخاب مدل معمولاً بر اساس معیارهای کمی است‌که خلاصه می کنند مدل چقدر خوب با داده ها برازش دارد، که اغلب‌با جریمه هایی برای بیش برازش همراه است. رایج ترین آنها معیاراطلاعات آکائیک و معیار اطلاعات بیزی هستند. آنها تعداد پارامترهای‌مدل را جریمه می کنند و به خوبی برازش، که از طریق درستنمایی‌اندازه گیری می شود، پاداش می دهند. معیار اطلاعات آکائیک‌به صورت زیر فرموله می شود:

آی‌سی=2م–۲ لیتر (ℒ)، (2)

باتعداد پارامترهاموحداکثر احتمال .

درمقابل، معیار اطلاعات بیزی:

آی سی=ln (ن)م–۲ لیتر (ℒ)، (3)

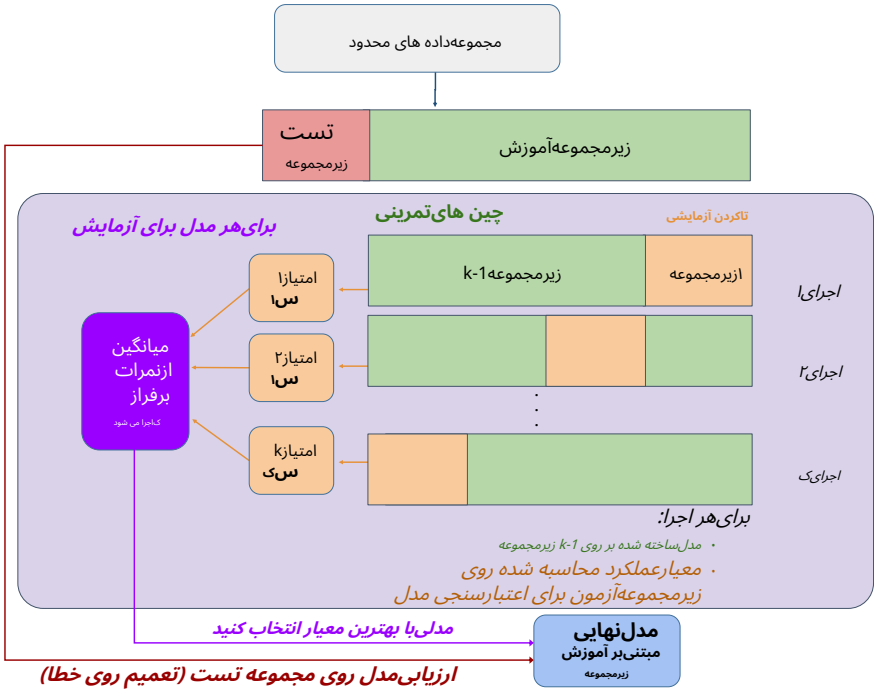
تعدادنقاط داده را در نظر می گیرند.

این‌رویکردهای انتخاب مدل به ندرت در یادگیری ماشینی استفاده می شوند،که بخشی از آن به دلیل پیچیدگی مجموعه داده ها و نقض فرضیات‌توزیعی مرتبط است. در عوض، رویکردهایی مانند اعتبارسنجی متقابل(خوشه بندی) بیشتر مورد استفاده قرار می گیرند.

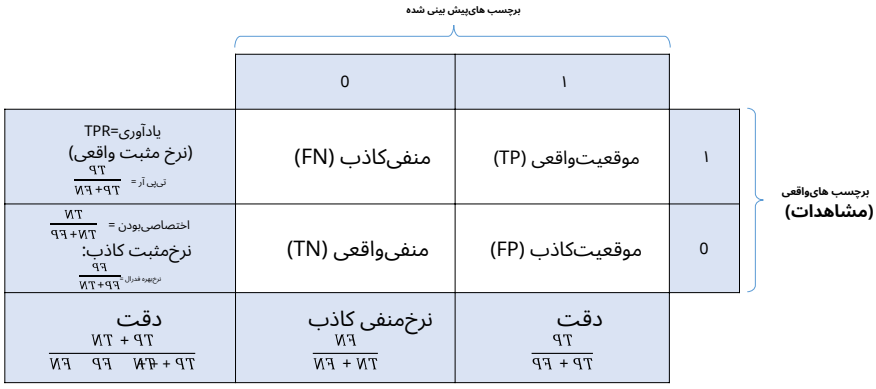
معیارهای‌عملکرد برای ارزیابی مدل.برای رگرسیون

درمدل ها، ما معمولاً از میانگین مربعات خطا یا انواع دیگر توابع هدف‌میانگین برای مقایسه عملکرد مدل در مجموعه آموزش و آزمون‌استفاده می کنیم. برای مسائل طبقه بندی دو کلاسه، معیارهای‌عملکرد رایج اغلب از «ماتریس سردرگمی» نشان داده شده در**شکل ۵**و به طور خلاصه در زیر توضیح داده شده است.

- دقت، مربوط به نسبت مقادیر مثبت درست پیش بینی شده به تعدادکل مقادیر مثبت پیش بینی شده.
- یادآوری، که نرخ مثبت واقعی (TPR) نیز نامیده می شود و مربوط به نسبت‌مقادیر مثبت پیش بینی شده ی صحیح به تعداد کل مقادیر مثبت‌در مجموعه داده ها است.
- نرخ مثبت کاذب (FPR)، مربوط به نسبت مقادیر منفی پیش بینی شده‌نادرست است.
- دقت، مربوط به تعداد مقادیر درست پیش بینی شده تقسیم بر تعدادکل مقادیر پیش بینی شده.
- مساحت زیر منحنی)AUC(ROC: منحنی های مشخصه عملیاتی گیرنده(ROC) وابستگی TPR (فراخوان) و FPR را نشان می دهند. چگالی.در طبقه بندی دودویی، هر نقطه روی منحنی ROC با انتخاب آستانه های‌مختلف برای طبقه بندی ... مکان یابی می شود.ی‌من‌در کلاس‌مثبت یا منفی. گوشه بالا سمت چپ منحنی ROC



شکل۴نمایش اصول کلی یادگیری نظارت شده در مورد یک مجموعه داده محدود. برای ارزیابی توانایی تعمیم یک الگوریتم یادگیری نظارت شده، داده ها به یک زیرمجموعه آموزشی‌که برای ساخت مدل استفاده می شود و یک زیرمجموعه آزمایشی که برای ارزیابی خطای تعمیم استفاده می شود، تفکیک می شوند.



شکل۵ماتریس درهم ریختگی برای مسائل دو کلاسه. ماتریس درهم ریختگی نشان می دهد که الگوریتم در پیش بینی برچسب ها در یک مسئله طبقه بندی دودویی که برچسب هامقادیر 0 (به نام «منفی») یا 1 (به نام «مثبت») را می گیرند، با ارزیابی برچسب های پیش بینی شده در مقابل برچسب های واقعی، چقدر موفق بوده است. هر نقطه‌داده در مجموعه آزمون متعلق به یکی از چهار دسته است و می توان معیارهای مختلفی را از این اعداد استخراج کرد.

حالت‌ایده آل این است که ۱۰۰٪ مقادیر مثبت به درستی طبقه بندی شوند (۱= TPR) و ۰٪ مقادیر مثبت به اشتباه در ۰ پیش بینی شوند (۰ = FPR). از آنجایی‌که ایده آل این است که TPR را به حداکثر و FPR را به حداقل برسانیم،مساحت زیر منحنی)AUC(ROC بزرگتر بهتر است.

برخی‌از این معیارها را می توان برای مسائل چندکلاسه، که در آن هابیش از دو برچسب مختلف در مجموعه داده ها وجود دارد، تعمیم‌داد. با این حال، معیارهای ذکر شده در بالا نسبت به پارامترهای‌مدل غیرپیوسته هستند، از این رو، بهینه سازی پارامتر ممکن‌است هنگام استفاده از آن ها به عنوان تابع هدف چالش برانگیزباشد. یک معیار جایگزین پیوسته و پرکاربرد که قبلاً در بخش «برازش‌مدل» ذکر شد، آنتروپی متقاطع است (ESL، فصل 9).۹که نه تنهامحتمل ترین پیش بینی، بلکه امتیاز پیش بینی (اعتماد به پیش بینی) را نیز در نظر می گیرد.

ک-نزدیکترین همسایه ها

مابررسی اجمالی خود را بر روی روش های یادگیری موجود با روشی آغاز می کنیم‌که مرحله یادگیری را به طور کامل حذف می کند و بنابراین، منجر به یک‌مدل صریح که از داده های آموزشی آموخته می شود، نمی شود. همانطور که‌بعداً بحث خواهیم کرد، این نیز یکی از بزرگترین کاستی های آن است. این نوع‌یادگیری اغلب به عنوان "یادگیری مبتنی بر نمونه" نیز شناخته می شود و درمثال خاص ما، "ک‌یادگیری نزدیکترین همسایه (kNN).۳۷

دراین رویکردها، یادگیری صرفاً شامل ذخیره تمام نقاط داده موجود و برچسب گذاری‌شده (یعنی داده های آموزشی) در یک پایگاه داده است. هنگامی‌که یک مثال جدید، اما طبقه بندی نشده، مشاهده می شود، الگوریتم آن‌را در ... قرار می دهد.ن‌فضای ویژگی چندبعدی بر اساس مقادیر ویژگی آن. برای‌هر نقطه داده در پایگاه داده، اکنون فاصله (مثلاً فاصله اقلیدسی یا فاصله های‌پیچیده تر دیگر) را تا این نقطه داده جدید محاسبه می کنیم تا آن راشناسایی کنیم.ک‌نزدیکترین همسایگان. در مرحله دوم، برچسب های شناخته‌شده این موارد را بررسی می کنیم‌ک‌شبکه های عصبی در پایگاه داده ما.فرض کنید ما انتخاب کرده ایم‌ک‌اگر نه باشد، مشاهده می کنیم که هفت تااز نزدیکترین همسایه ها به عنوان کلاس X برچسب گذاری شده اند در حالی که‌دو تا از آنها به عنوان کلاس Y برچسب گذاری شده اند. در این حالت، نقطه داده‌جدید خود را به کلاس X اختصاص می دهیم زیرا اکثر همسایه های آن از این‌کلاس هستند. بسط این رویکرد ساده، وزن دهی به اهمیت همسایه ها درطبقه بندی بر اساس فاصله آنها تا نقطه داده جدید است. با وجود اینکه بسیارسرراست و ساده است، اما ثابت می شود که ...

یک‌روش طبقه بندی بسیار مؤثر در عمل. این روش در آموزش (یعنی ذخیره داده هادر پایگاه داده) و پیاده سازی های کارآمد برای محاسبه ی ... بسیار کارآمداست.ک‌شبکه های عصبی (NN) وجود دارند.

خب،چالش های این رویکرد چیست؟ واضح ترین چالش این است که چون‌هیچ «مرحله یادگیری» وجود ندارد،ک‌رویکرد شبکه های عصبی، ویژگی هایی‌را که واقعاً برای پیش بینی کلاس یک مورد جدید مرتبط هستند،شناسایی نمی کند. بنابراین، حتی اگر در یک فضای ویژگی ۲۰ بعدی،که در آن فقط ۲ مورد ممکن است واقعاً برای طبقه بندی مرتبط باشند،فاصله با در نظر گرفتن هر ۲۰ بعد محاسبه خواهد شد. بنابراین،ک نزدیکترین‌نقاط داده ای که توسط پرس وجو برگردانده می شوند، به شدت‌تحت تأثیر ویژگی های نامربوط یا نویز قرار خواهند گرفت (همچنین‌به «کاهش ابعاد» در مورد نحوه حذف برخی از این ویژگی ها مراجعه‌کنید). در نتیجه، طبقه بندی حاصل به جای الگوی واقعی زیربنایی درداده ها، توسط نویز هدایت می شود. از این نظر، این رویکرد از همان چالشی‌رنج می برد که رویکردهای خوشه بندی (به بخش «خوشه بندی» مراجعه‌کنید) با آن مواجه هستند، که اغلب به عنوان «نفرین ابعاد» خلاصه‌می شوند.۲۱

بیزساده

دومین‌رویکرد یادگیری بسیار شهودی که می خواهیم معرفی کنیم، بیزساده است. این رویکرد مبتنی بر محاسبه آمار ساده از یک مجموعه‌داده آموزشی معین به عنوان مرحله یادگیری است که پس ازیک کاربرد ساده (اما ساده) از فرمول بیزی برای احتمال شرطی به منظوردستیابی به یک طبقه بندی انجام می شود. به دلیل سادگی آن،اغلب برای دستیابی به یک عملکرد طبقه بندی پایه نیز استفاده می شودکه سایر روش های پیچیده تر باید آن را بهبود بخشند. این رویکردرا می توان با یک مثال ساده به بهترین شکل توضیح داد.

فرض‌کنید مجموعه داده های آموزشی شامل بیمارانی داریم که از سرماخوردگی‌بی ضرر یا عفونت آنفولانزا رنج می برند. ما برای هر بیمار دو ویژگی، یعنی‌تب (بالا، پایین یا بدون تب) و درد (قوی، پایین یا بدون تب) را اندازه گیری کرده ایم.برای هر بیمار، از طریق آزمایش آزمایشگاهی می دانیم که آیا بیمار به عفونت‌آنفولانزا مبتلا بوده است یا خیر. اکنون می خواهیم از این داده ها یاد بگیریم‌و آن را برای تشخیص یک بیمار جدید (در جایی که هیچ آزمایش آزمایشگاهی‌در دسترس نداریم) با استفاده از رویکرد بیز ساده به کار ببریم.

به‌عنوان یک مرحله یادگیری، برای هر مقدار ویژگی، تعداد دفعات وقوع آن رادر گروه بیماران مبتلا به آنفولانزا و سرماخوردگی محاسبه می کنیم (مثلاً برای‌به دست آوردن احتمال تب بالا در شرایط ابتلای بیمار به آنفولانزا و غیره) .نتیجه این مرحله یادگیری را می توان در ... مشاهده کرد.**جدول ۱**.

جدول1 تصویرسازی بیز ساده: نمونه ای از نتایج مرحله یادگیری روی مجموعه داده آنفولانزا، که احتمال مقادیر ویژگی ها را با توجه به دسته بیمار نشان می دهد.

ویژگی ها	تب			درد		
کلاس ها	بالا	کم	خیر	قوی	کم	خیر
آنفولانزا(آنفولانزا)						
پ(آنفولانزا) = 0.1	پ(تب = بالا | آنفولانزا) = ۰.۹۵	پ(تب = پایین | آنفولانزا) = ۰.۰۵	پ(تب = خیر | آنفولانزا) = 0	پ(درد = شدید | آنفولانزا) = ۰.۷۵	پ(درد = کم | آنفولانزا) = ۰.۲۰	پ(درد = بدون | آنفولانزا) = ۰.۰۵
سرد						
پ(سرد) = ۰.۹	پ(تب = بالا | سرما) = ۰.۱	پ(تب = پایین | سرماخوردگی) = ۰.۴	پ(تب = خیر | سرماخوردگی) = ۰.۵	پ(درد = شدید | سرد) = ۰.۳	پ(درد = کم | سرد) = ۰.۳	پ(درد = بدون | سرما) = ۰.۴

جدول۱احتمال هر ویژگی را با توجه به دسته بیمار خلاصه می کند ونشان می دهد که در کل جمعیت بیمار، احتمال ابتلای یک بیمار به عفونتآنفولانزا ۰.۱ است، در حالی که احتمال ابتلا به سرماخوردگی معمولی۰.۹ است.

وقتیاین مقادیر را تولید کردیم و بنابراین «مرحله یادگیری» را با تجزیه وتحلیل مجموعه داده های خود تکمیل کردیم، بیز ساده لوحانه یک فرض ساده لوحانهمطرح می کند، و آن این است که همه این ویژگی ها از نظر شرطیمستقل از یکدیگر هستند. در واقعیت، این به ندرت صادق است وروش های یادگیری بیزی پیشرفته تری وجود دارند که این فرض را در نظرنمی گیرند. با این حال، این فرض امکان کاربرد ساده ای از قضیه بیزیرا فراهم می کند. برای جزئیات (یعنی فرمول ها) در مورد نحوه استخراجاین طبقه بندی کننده، خواننده را به مطالب بیشتر برای مطالعه (ESL، فصل 6) ارجاع می دهیم.۹به طور خلاصه، احتمال یک برچسب خاص(آنفولانزا یا سرماخوردگی) برای یک آیتم آزمایشی جدید را می توان به‌صورت حاصلضرب احتمالات ویژگی شرطی واحد (تب و درد) که برای نقطه‌داده مشاهده می شوند، در احتمال کلاس (آنفولانزا یا سرماخوردگی) محاسبه‌کرد. کلاسی که بیشترین احتمال پسین را دارد، به عنوان کلاس پیش بینی شده‌برای آیتم آزمایشی انتخاب می شود. با فرض اینکه یک فردآزمایشی با تشخیص ناشناخته آنفولانزا یا سرماخوردگی داریم و می دانیم‌که این فرد با تب بالا و تب بالا مراجعه می کند.

اگرسطح درد بالا باشد، احتمال ابتلا به آنفولانزا را به صورت زیر محاسبه می کنیم:

$$\text{پ(تب = بالا | آنفولانزا)}\cdot\text{پ(درد=شدید | آنفولانزا)}\cdot \text{پ(آنفولانزا)}=0.95\cdot 0.75\cdot 0.1=0.07125 \quad (4)$$

به‌همین ترتیب، احتمال سرماخوردگی را به صورت زیر محاسبه می کنیم:

$$\text{پ(تب = بالا | سرماخوردگی)}\cdot\text{پ(درد= شدید | سرد)}\cdot \text{پ(سرد)}=0.1\cdot 0.3\cdot 0.9=0.027 \quad (5)$$

برایبیماری که با تب بالا و درد عضلانی یا سردرد شدید به پزشک مراجعه‌می کند، این امر منجر به احتمال خلفی (غیر نرمال) برای عفونتآنفولانزا ۷.۱۲۵٪ و احتمال ۲.۷٪ برای سرماخوردگی معمولی می شود.بنابراین، بیمار بیشتر از سرماخوردگی از آنفولانزا رنج می برد.

بنابراین،بیز ساده لوحانه، از بسیاری جهات، چگونگی یادگیری انسان هااز تجربه را فرموله می کند.

درخت های‌تصمیم، جنگل های تصادفی و تقویت گرادیان

درخت های‌تصمیم، بلوک های سازنده ی ضروری برای بسیاری از الگوریتم های یادگیری‌ماشین هستند و حداقل ۵۰ سال است که مورد استفاده قرار می گیرند.۳۸،۳۹ایده پشت درخت های تصمیم گیری بسیار شهودی است و به بهترین شکل به‌صورت بصری نمایش داده می شود (مثلاً،**شکل ۱**سته به مسئله، گره های برگ‌درخت تصمیم دارای کلاس، احتمال یا پیوسته هستند.

مقادیردر صورت رگرسیون. در روزهای اولیه یادگیری ماشین، از درخت های تصمیم‌برای حل مسائل دارویی مانند دوزبندی، سم شناسی و تشخیص استفاده‌شده است.۴۰-۴۲اگرچه استفاده از درخت های تصمیم گیری شهودی است،اما سوال این است که چگونه می توان چنین درخت هایی را از داده های‌موجود ساخت. چند رویکرد معروف که ارزش ذکر کردن دارند عبارتنداز: CART۴۳و ID3.۴۴

درحال حاضر، درخت های تصمیم تقریباً هرگز در یادگیری ماشین به شکل اصلیخود استفاده نمی شوند. یکی از دلایل این امر این است که درخت های تصمیم‌مستعد بیش برازش هستند. با این وجود، درخت های تصمیم به سنگ‌بنای دو رویکرد پرکاربرد تبدیل شده اند: جنگل های تصمیم تصادفی و چارچوب های‌تقویت گرادیان.

هم‌جنگل های تصمیم گیری تصادفی و هم تقویت گرادیان مبتنی بر درخت از مجموعه ای(گروهی) از درخت های تصمیم گیری آموزش دیده برای پیش بینی متغیرنتیجه استفاده می کنند. تفاوت اساسی بین تقویت گرادیان مبتنی بر درخت‌و جنگل های تصمیم گیری تصادفی در نحوه ایجاد درختان است.

درمورد جنگل های تصادفی، الگوریتم صدها یا هزاران درخت تصمیم گیری‌عمیق ("پیش بینی کننده های قوی") می سازد. هر یک از این درخت هااحتمالاً بیش برازش شده اند، با این حال، با ترکیب خروجی های چندین‌درخت می توانیم مشکل آموزش بیش از حد را حل کنیم. برعکس، در یک‌الگوریتم افزایش گرادیان، مانند XGBoost یا CatBoost، هر یک از درخت هایک درخت تصمیم گیری کم عمق ("پیش بینی کننده ضعیف") هستندو الگوریتم به طور تکراری با اضافه کردن درختان بیشتر و بیشتر، خطای‌طبقه بندی را در طول زمان کاهش می دهد.

امروزه،روش های تقویت گرادیان، عملکرد بسیار خوبی را هم در نشریات‌و هم در مسابقات یادگیری ماشین نشان می دهند. حتی بدون تنظیم‌هایپرپارامتر، آنها معمولاً عملکرد عالی را با هزینه محاسباتی نسبتاً کم‌ارائه می دهند.۱۱از طرف دیگر، جنگل های تصادفی معمولاً کمتر مستعدبیش برازش هستند.۴۵و نیاز به تنظیم پارامترهای کمتری دارند.۴۶ این‌امر، جنگل های تصمیم گیری تصادفی را برای مجموعه داده های کوچک ترایا به عنوان یک روش پایه برای الگوبرداری جذاب می کند.

روش های‌گروه درختی می توانند برای وظایف طبقه بندی و همچنین برای رگرسیون‌استفاده شوند. در هر دو مورد، خروجی های درخت میانگین گیری می شوندکه می تواند یک تابع خروجی هموار ایجاد کند.

روش های‌هسته: ماشین های بردار پشتیبان و رگرسیون

روش های‌کرنل و به طور خاص تر، ماشین بردار پشتیبان (SVM) برای طبقه بندی‌و رگرسیون بردار پشتیبان (SVR) برای خروجی پیوسته، به دلیل‌توانایی شان در مقاوم بودن در برابر نویز و کار با مجموعه داده های باابعاد بالا که در ژنتیک، ترانسکریپتومیکس و پروتئومیکس یافت می شوند،کاربردهایی در زیست شناسی محاسباتی پیدا کرده اند.۴۷به طورمشخص در یک مثال جدیدتر، از SVR برای مشخص کردن ترکیبات سلولی‌از داده های حجیم رونویسی استفاده شد.۴۸

این‌بخش ابتدا مروری مختصر بر مفاهیم کلیدی ارائه می دهد که مفاهیم‌تبدیلات هسته، یک تابع هدف با ناحیه بدون اتلاف و یک عبارت‌منظم سازی را برجسته می کند.[۴۹،۵۰]تأکید بر ارائه استدلال پشت‌این موضوع خواهد بود که چرا این روش در برخورد با ورودی های‌چندگانه که اثرات آنها بر خروجی ناشناخته است و می توان‌فرض کرد که به توابع غیرخطی تبدیل می شوند، روشی تطبیق پذیرتراست.

پیشینه.مشابه تمام روش های رگرسیون، هدف SVR فرض تابعی روی ورودی(ها) است که می تواند به تخمین خروجی مشاهده شده کمک کند.به همین ترتیب، برای SVM، هدف یافتن مرز تصمیم گیری بهینه‌است که کلاس ها را از هم جدا می کند. همانطور که از نامش پیداست،مفهوم اصلی پشت SVM/رگرسیون، توانایی انتخاب عینی زیرمجموعه ای‌از داده های آموزشی به نام ... است. *بردارهای پشتیبانی*این بردارهای پشتیبان، مدل را تعریف می کنند که معمولاً یک‌ابرصفحه در فضای ویژگی است. برای دستیابی به این هدف، باید چندین‌مفهوم معرفی شوند.

- یک-*از دست دادن غیر حساس*تابع اجازه می دهد تا باقیمانده های کمتر از ، بدون‌اتلاف در نظر گرفته شوند و بنابراین، بخشی از بردارهای پشتیبان در نظر گرفته‌شده برای تخمین تابع خروجی-ورودی نباشند.
- الف-*اصطلاح منظم سازی*با هدف جستجوی مدلی برای توصیف رابطه‌بین متغیرهای ورودی و خروجی به تابع هدف اضافه می شود،به طوری که ابرصفحه تا حد امکان مسطح نگه داشته شود.
- *متغیرهای‌اسلک*می توان برای در نظر گرفتن خطاهای آموزش، که حاشیه نرم‌نامیده می شود، زمانی که خروجی خارج از ناحیه غیر حساس θ یافت می شود،معرفی کرد. با معرفی متغیرهای slack، تحمل برای عبارت باقیمانده‌که بزرگتر از θ باشد، ایجاد می شود.
- الف-*تابع هسته*به ما اجازه می دهد در فضایی با ابعاد بالاتر کار کنیم،

فضای‌ویژگی. یک تابع هسته که در فضای ورودی اعمال می شود، معادل‌یک ضرب نقطه ای در فضای ویژگی است که در آن معیارهای شباهت‌محاسبه می شوند. این امر بدون نیاز به نگاشت صریح داده های‌ورودی از فضای ورودی به یک فضای ویژگی توسط یک تابع نگاشت،محقق می شود.

بادر دست داشتن تمام این مفاهیم، اکنون می توانیم مدلی با ضخامت‌مشخص، که به عنوان لوله ای که توسط تابع زیان insensitive-β معرفی می شود، برازش دهیم، در حالی که عبارت منظم سازی، مسطح‌بودن این ابرصفحه را در فضای ویژگی تعریف شده توسط تابع هسته‌کنترل می کند.**شکل ۶**این مفاهیم اساسی SVM را نشان می دهد.

ترفندو انتخاب هسته.:می تواند توابع هدف غیرخطی را ثبت کند که ورودی های‌چند متغیره را به خروجی نگاشت می کنند. به طور دقیق تر، ترفندهسته به این معنی است که یک هسته SVR

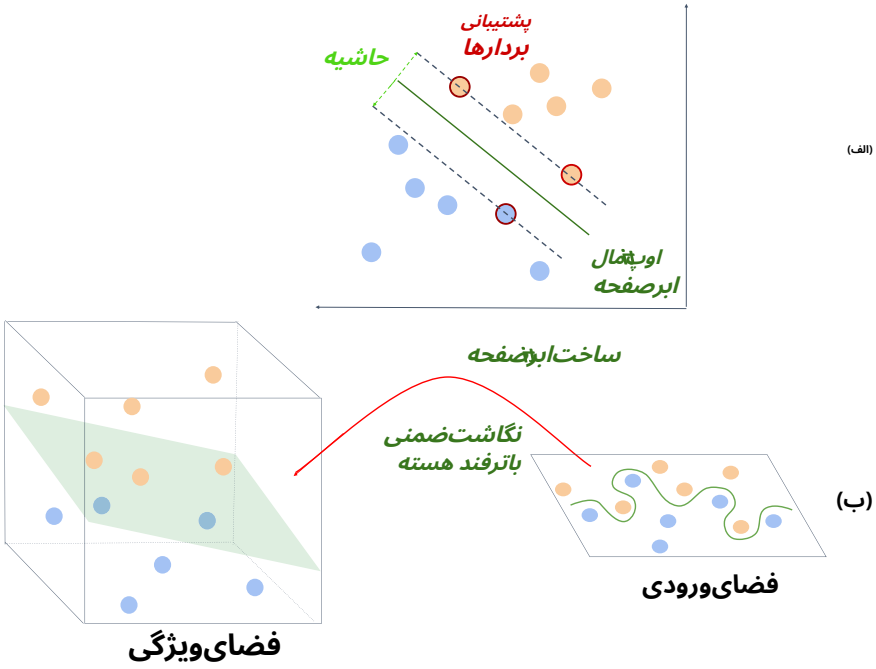
(6)

اعمال‌شده بر مجموعه ای از ورودی ها در فضای ورودی معادل محاسبه‌ضرب نقطه ای به عنوان معیار شباهت در برخی از فضای ویژگی هااست. این امر بدون نیاز به انجام صریح پیش نگاشت ورودی هاحاصل می شود،**ایکس**من، با یک تابع نگاشت. یک تابع هسته‌محاسبه شده در فضای ورودی، اگر و تنها اگر یک تابع متقارن مثبت‌معین باشد، معادل یک ضرب نقطه ای در یک فضای ویژگی است.[۵۱،۵۲]

انتخاب‌هسته تابع پایه شعاعی،

(7)

اغلب‌به این دلیل ساخته می شود که می توان آن را به یک فضای ویژگی با ابعاد نامتناهی‌گسترش داد. اگرچه تابع پایه شعاعی طیف وسیعی را پوشش می دهد



شکل۶تصویرسازی اصول ماشین بردار پشتیبان (SVM). (الف)تصویرسازی از یک حالت ساده که در آن یک ابرصفحه، دو گروه را مستقیماً در فضای ورودی ها از هم جدا می کند. (ب)نمایش انجام طبقه بندی غیرخطی با نگاشت ضمنی ورودی ها به فضاهای ویژگی با ابعاد بالا که در آن نقاط داده می توانند توسط یک ابرصفحه از هم جدا شوند.

ازاثرات احتمالی، منجر به تفسیر دشوارتر مدل نهایی می شود. در عمل،انتخاب تابع هسته بر اساس کارایی محاسباتی است. سایر هسته های‌محبوب شامل هسته های خطی و چندجمله ای هستند.[۴۷]

شبکه های‌عصبی

پیشینه.شبکه های عصبی مجموعه ای از نورون ها و لبه ها را تشکیل می دهندکه منشأ آنها از تحلیل مدار گرفته شده است. می توان وزن های‌مختلفی را به هر لبه متصل کننده نورون ها اعمال کرد. در هر نورون،یک تابع فعال سازی به یک سیگنال ورودی وزن دار اعمال می شودتا یک سیگنال خروجی تولید شود. اغلب از یک تابع سیگموئیدی‌استفاده می شود که شامل یک فیلتر پایین گذر مرتبه اول ازیک تابع پله واحد است. چنین تابع سیگموئیدی مزایای خروجی کران دارو مشتق پذیری پیوسته را دارد که در مرحله انتشار رو به عقب برای‌تنظیم وزن ها (پارامترهای مدل) مورد نیاز است، به مراحل تعریف شده‌در بخش «شبکه عصبی بازگشتی» مراجعه کنید.

نورون هاهمانطور که در شکل نشان داده شده است، به لایه ورودی،لایه(های) پنهان و لایه خروجی تقسیم می شوند.**شکل ۷ الف** لایه های‌پنهان، لایه انتزاع مورد نیاز برای رفتن از لایه ورودی به لایه خروجی‌را انجام می دهند. تعداد لایه های پنهان مشخص می کند که آیاسیستم یک سیستم یادگیری سطحی (با یک یا چند لایه پنهان) یا یادگیری‌عمیق (با لایه های پنهان زیاد) است. بین تعداد لایه های پنهان‌و زمان مورد نیاز برای آموزش مدل، یک بده بستان ذاتی وجود دارد.به همین دلیل، اگرچه مفهوم اصلی تعبیه شده در شبکه عصبی، مفهوم‌جدیدی نیست، اما به دلیل پیشرفت های اخیر در قدرت محاسباتی،کاربردهای جدیدی پیدا کرده است.

اساسی ترین‌نوع آن به عنوان شناخته می شودپیشخورشبکه عصبی، زیرااطلاعات فقط از لایه ورودی به لایه(های) پنهان و در نهایت به لایه خروجی‌منتشر می شود. وضعیت فعلی سیستم توسط هیچ وضعیت گذشته ای‌تعریف نمی شود؛ از این رو، یک سیستم بدون حافظه را نشان می دهد.

درادامه، نمونه هایی از شبکه های عصبی شرح داده شده است: شبکه های‌عصبی بازگشتی، شبکه های حافظه کوتاه مدت بلندمدت و شبکه های‌بازگشتی دروازه دار. شبکه های عصبی قابل توجه دیگری که خارج‌از محدوده این مقاله هستند اما مطالعه بیشتر در مورد آنها را توصیه‌می کنیم، عبارتند از شبکه های عصبی کانولوشنی،[۵۳]شبکه های رمزگذار-رمزگشا،[۵۴،۵۵]و مدل های تولیدی.[۵۶]

شبکه‌عصبی بازگشتی.شبکه های عصبی بازگشتی، دسته ای از شبکه های عصبی‌هستند که به مجموعه داده های سری زمانی اختصاص دارند، زیرا رابطه‌ترتیبی ذاتی مشاهده شده در داده های یک نقطه زمانی با نقطه زمانی‌دیگر را در نظر می گیرند. این شبکه ها در چیزی که در این زمینه به عنوان ...شناخته می شود، موفقیت کسب کرده اند.*داده های ترتیبی*، جایی‌که ترتیب یا توالی زمانی سیگنال نقشی ایفا می کند، یعنی در پردازش‌زبان طبیعی و پیش بینی سری های زمانی. این روش که ارتباط نزدیکی‌با حوزه تحقیقاتی ما دارد، در پیش بینی نتایج حاصل از پرونده های‌الکترونیکی سلامت کاربرد پیدا کرده است، جایی که غنای آن ذاتاً از ساختار همبستگی توالی داده ها ناشی می شود تا اقدامات سریع وحتی پیش بینی شده ای را که باید توسط کادر پزشکی انجام شود، توصیه‌کند.[۵۷]طرح مجدد سوال برای حل یک معمای مدل سازی در حوزه فارماکومتری،تنها زمانی شروع به ظهور کرد که این مقاله در حال نگارش بود.تانگ*و همکاران*یکی از تلاش های نادر در مورد نحوه استفاده از ML (اینجا:RNNها) برای توصیف PK رمیفنتانیل را ارائه داده و نتایج را با روش‌استاندارد طلایی فارماکومتریک NONMEM مقایسه کرده اند.[۵۸] اگرچه‌از مدل های غیراستاندارد PK برای مقایسه استفاده شد و تعمیم پذیری‌نتایج می تواند مورد چالش قرار گیرد، تانگ*و همکاران*... سهم‌ارزشمندی در نشان دادن کاربردهای RNNها در فارماکومتری داشته‌باشد.

شکل‌پایه یک RNN در شکل نشان داده شده است.**شکل ۷ب**که درآن هر حالت فعلی (در زمان)*تی*) با ترکیبی از حالت قبلی سیستم و ورودی‌فعلی تعریف می شود، که مشابه مفهوم سیستم های دینامیکی‌کلاسیک است. وزن های هر لبه را می توان تعیین کرد که چقدربراید به عقب نگاه کرد، مشابه یک ثابت زمانی. برخلاف شبکه عصبی‌پیش خور، یک وزن یکسان در بلوک واحد نورونی منفرد در تمام‌مراحل زمانی گسسته قبلی به اشتراک گذاشته می شود.

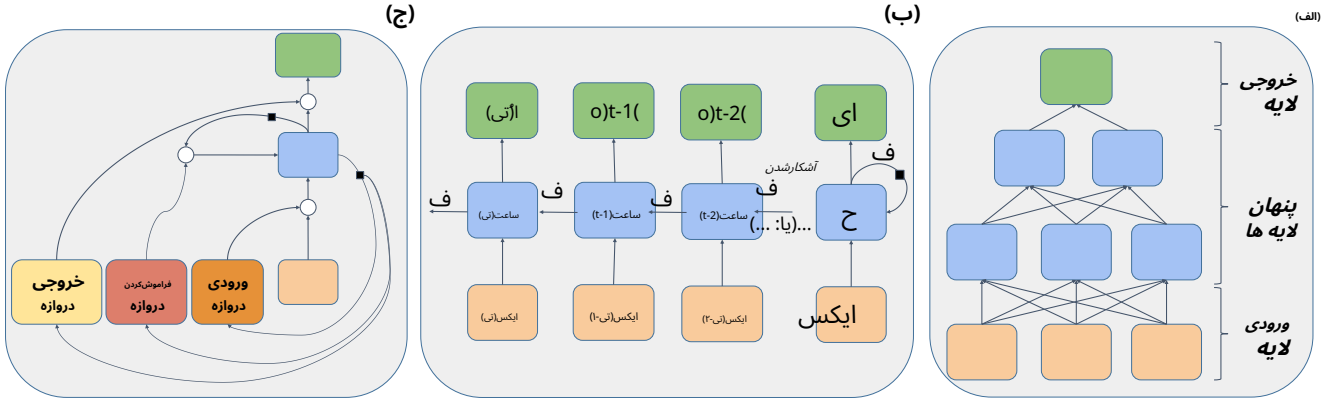
درهسته RNN، از یک توالی ورودی تشکیل شده است که توسط ... تعریف می شود.ایکس(تی)، یک دنباله خروجی مطابق تعریف زیر(ت)، یک توالی حالت پنهان یاسیستمی که به صورت زیر تعریف می شود(ت)و همچنین زیرمدول های زنجیری از واحدهای‌تکراری.

مراحل‌مورد نیاز برای آموزش یک مدل RNN به شرح زیر است:

۱.معماری شبکه را تعریف کنید و مدل را با وزن ها و بایاس های تصادفی‌مقداردهی اولیه کنید.

۲.برای محاسبه خروجی تخمینی، یک انتشار رو به جلو انجام دهید.

۳.خطا را در لایه خروجی محاسبه کنید.



شکل۷شبکه های عصبی (الف)مبانی شبکه های عصبی پیش خور. (ب)آشکار شدن شبکه های عصبی بازگشتی. (ج)بسط شبکه های عصبی بازگشتی با واحدهای دروازه ای. مربع‌سیاه نشان دهنده تأخیر یک گام زمانی گسسته است.

۴.با استفاده از یک رویکرد بهینه سازی، یک الگوریتم انتشار معکوس برای به روزرسانی‌وزن ها انجام دهید.
۵.مراحل ۲ تا ۴ را برای تعداد دوره ها (یا تکرارها) تکرار کنید تا مقدار تابع‌زیان به حداقل برسد.

افزونه های‌این*وانیل*شبکه های عصبی بازگشتی (RNN) برای رسیدگی‌به مشکلات مربوط به گرادیان ناپایدار (مثلاً مشکل گرادیان ناپدیدشونده و معادل جدی تر آن، ناپایداری ناشی از گرادیان انفجاری) توسعه‌داده شدند. این مشکلات در اصل به دلیل ضرب هایی (تحت تأثیرخطاهای عددی) هستند که در انتشار معکوس در رابطه با تخمین‌خطا نسبت به پارامترها در امتداد هر لایه از شبکه عصبی ایجادمی شوند. به عبارت دیگر، گرادیان ناپدید شونده باعث می شود اطلاعاتی‌که باید از یک نقطه زمانی دورتر از زمان فعلی ثبت شوند، ضعیف‌شوند و بنابراین، مدل را برای ثبت حافظه ذخیره شده ارزشمندبا تأخیر زمانی طولانی تر، ضعیف می کنند. در یک رویداد کمترایج که حداقل یک مشتق جزئی الزام پایداری را نقض کند، و به‌ماتریس حالت داشتن حداقل یک مقدار ویژه > 1 تبدیل شود، این امرمنجر به یک مشکل گرادیان انفجاری می شود، یک مشکل شناخته‌شده در سیستم پویای سنتی برای زمان گسسته. وظیفه مورداستفاده برای رسیدگی به این مشکل اساسی، در دو افزونه شناخته‌شده RNN (حافظه کوتاه مدت طولانی (LSTM) و شبکه بازگشتی‌دروازه دار (GRU)) بیشتر توضیح داده خواهد شد.

انواع‌و توسعه های مختلفی در تحقیقات RNN وجود داشته است،هر روش جدید برای پرداختن به یک مسئله متفاوت به کار می رودکه در نهایت منجر به توسعه مدل های قوی تر می شود. به عنوان‌مثال، برای دور زدن مشکل گرادیان ناپایدار، برش گرادیان با مجبورکردن گرادیان به یک آستانه در مرجع پیشنهاد شده است.[۵۹،۶۰] اماتاکنون پذیرفته شده ترین روش، گنجاندن*واحدهای دروازه ای*.

حافظه‌کوتاه مدت بلند و شبکه بازگشتی دروازه دارLSTM
بخشی‌از خانواده بزرگتری از RNN های دروازه دار است که با معرفی واحدهای‌دروازه ای، اطلاعات را حفظ و فراموش می کنند. به طور خاص‌تر، همانطور که در نشان داده شده است، می توان سه واحد دروازه‌ای را در سیستم گنجاند.**شکل ۷ج**اول، کپی کردن یا پاک کردن مستقیم‌کل حالت می تواند توسط گیت فراموشی کنترل شود. رویکردمشابهی نیز توسط گیت ورودی برای تصمیم گیری در مورد اینکه‌آیا سیگنال ورودی فعلی را به عنوان بخشی از به روزرسانی حالت‌در نظر بگیرد یا خیر، انجام می شود. مقدار اطلاعاتی که باید از سیگنال‌حالت قبلی و از سیگنال ورودی اختلال حفظ شود، در هر گام زمانی‌آموخته می شود.[۶۱]سیستم باید با حفظ اطلاعات، وابستگی های‌زمانی بلندمدت را یاد بگیرد، اما باید گاهی اوقات نیز یادبگیرد که اطلاعات را از وضعیت فعلی خود پاک کند.[۶۲] درنتیجه، حل مسائل گرادیان ناپدید شونده و انفجاری. در نهایت، می توان‌یک گیت خروجی، هرچند کمتر رایج، را به عنوان یک مکانیسم گیتینگ‌برای تصمیم گیری در مورد اینکه کدام سیگنال خروجی به سیستم بازگردانده‌شود، معرفی کرد.
یک‌تفسیر ساده تر و در نتیجه، پیاده سازی سریع تر آموزش را می توان در GRU یافت. GRUها همان مشکل گرادیان های ناپایدار را برطرف می کنند و یک‌افزونه جدید به این خانواده از افزونه های RNN هستند. تفاوت اصلی بین LSTM و GRU این است که دومی گیت خروجی را حذف کرده و از گیت های ریست‌و به روزرسانی ساده تری استفاده می کند.[۶۳]با این حال، در تئوری، LSTM باید عملکرد بهتری داشته باشد زیرا می تواند اطلاعات را از فاصله زمانی/تاخیر طولانی تر، وزن دهی یا وزن دهی کند.

نمونه هایی‌از کاربردهای یادگیری ماشینی تحت نظارت در داروشناسی‌بالینی
مدل های‌موجود در فارماکولوژی بالینی معمولاً با ترجمه اصول فیزیولوژیکی‌و دارویی به سیستم های معادلات دیفرانسیل و استفاده ازالگوریتم های حداکثرسازی انتظار برای تخمین پارامترهای مدل ایجادشده اند. این رویکرد با انگیزه مکانیکی در بسیاری از کاربردها مفیدبوده و یک جزء تثبیت شده از برنامه های توسعه دارو است. به طوربالقوه به دلیل موفقیت این رویکردهای تثبیت شده، تاکنون تنها چندنمونه از اعمال روش های یادگیری ماشینی در مسائل فارماکولوژی‌بالینی وجود دارد. ریو*و همکاران*یک شبکه عصبی عمیق رابر روی یک پایگاه داده بزرگ و منظم که شامل ۱۹۲۲۸۴ تداخل دارو بادارو بود، آموزش داد تا تداخلات دارو با دارو و دارو و غذا را برای نسخه ها،توصیه های غذایی و مولکول های جدید پیش بینی کند.[۶۴] ترکیب‌مجموعه داده های حاصل از مطالعات متعدد برای ایجاد پایگاه های‌داده بزرگ، پتانسیل استفاده از یادگیری ماشینی را برای پرداختن‌به سوالات گسترده داروشناسی بالینی افزایش می دهد.

همچنین‌از یادگیری ماشینی برای ایجاد پل بین کشف دارو و توسعه بالینی استفاده شده‌است. به عنوان مثال، هامان*و همکاران*توانستند با استفاده از روش درخت تصمیم گیری،میزان بروز عوارض جانبی را از ساختار شیمیایی یک مولکول پیش بینی کنند.[۶۵]به طور مشابه، لنکستر و سوبی SVM ها را برای پیش بینی خطر de Pointes Torsades از ... پیاده سازی کردند.*در شرایط آزمایشگاهی (in vitro)*داده ها[۶۶]

درحوزه ایمنی شخصی سازی شده، یادگیری ماشین توسط دانهاور مورداستفاده قرار گرفته است.*و همکاران*شخصی سازی ایمنی در زمینه هیپربیلیروبینمی‌در نوزادان.[۶۷]نویسندگان از لاسو و جنگل های تصادفی برای‌پیش بینی از مجموعه داده های بالینی استفاده کردند. علاوه بر این، یادگیری‌تقویتی توسط گاودا مورد استفاده قرار گرفت.*و همکاران*برای شخصی سازی‌مدیریت دارویی کم خونی.[۶۸]از رویکرد مشابهی برای توسعه‌یک سیستم «حلقه بسته» برای کنترل گلوکز با ترکیب یک مدل ریاضی،یک حسگر گلوکز و یک مدل یادگیری تقویتی استفاده شد.[۶۹]چاوادا*وهمکاران*و هنینگ*و همکاران*امکان سنجی بازخورد بیزی برای تنظیم دوز آنتی بیوتیک هارا بررسی کرد.[۷۰،۷۱]حوزه مراقبت های بهداشتی شخصی سازی شده‌می تواند از استفاده از مدل های یادگیری ماشین که تنظیمات‌دوز را در زمان واقعی توصیه می کنند، بسیار بهره مند شود. در یک‌مطالعه اخیر، یک الگوریتم کنترل از نوع یادگیری ماشین با مدل های ساختاری‌موجود PK/PD که برای متخصصان داروشناسی آشنا هستند، ادغام‌شد و مشخص شد که سیستم کنترل حلقه بسته حاصل، عملکرد بهتری‌نسبت به پمپ با کمک حسگر دارد.[۶۹]

غذاهای‌اصلی
- روش های یادگیری نظارت شده، مدل ها را بر اساس جفت های خروجی-ورودی برچسب گذاری شده ی‌مجموعه داده های آموزشی استنتاج می کنند.
- معیارهای عملکرد برای ارزیابی مدل های طبقه بندی و رگرسیون استفاده می شوندتا از بیش برازش مجموعه داده های آموزشی جلوگیری شود.
- روش های یادگیری نظارت شده ی زیادی با بده بستان های مختلف‌بین تفسیرپذیری و عملکرد وجود دارند.
.شکل‌خاصی از شبکه عصبی است که یک سیستم پویا را در زمان گسسته‌نمایش می دهد RNN •
- نمونه هایی از کاربردهای این روش های یادگیری تحت نظارت در زیست شناسی‌محاسباتی و به ویژه داروشناسی بالینی در حال پدیدارشدن هستند.

بحث
دراین آموزش، ما برخی از روش های اساسی یادگیری ماشینی را معرفی کرده ایم‌که احتمالاً برای فارماکولوژی بالینی جالب خواهند بود.

وجامعه فارماکومتری. مقدمه مختصر ما با طیف وسیعی از منابع مرتبطتکمیل شده است. ما با ذکر مثال هایی مرتبط با توسعه دارو، زمینه‌را فراهم کرده ایم. در پایان، خلاصه ای از وضعیت فعلی حوزه های‌یادگیری ماشینی و فارماکولوژی بالینی و همچنین چشم اندازی‌از چگونگی ادغام بیشتر این حوزه ها در آینده ارائه می دهیم.روش های آماری پیشرفته برای فارماکومتری دانان جدید نیستند؛در واقع، چنین روش هایی مدتی است که برای توصیف پدیده هایPK و PD استفاده می شوند. به عنوان مثال، روش های بیزی‌یک جزء جاافتاده از رویکردهای فارماکومتری هستند.[۷۲،۷۳]

بنابراین،به نظر می رسد که با تثبیت و برجسته تر شدن رویکردهای آماری‌و یادگیری ماشینی در صنعت داروسازی، متخصصان فارماکومتری‌از جمله کسانی خواهند بود که از این روش ها بهره می برند.علاوه بر این، فرصت های جدیدی برای بررسی سایر سوالات بالینی،مانند طبقه بندی بیمار بر اساس ویژگی های پایه با ابعاد بالا، ممکن‌است در فارماکولوژی بالینی با استفاده از رویکردهای یادگیری ماشینی‌امکان پذیر شود.

چندین‌نمونه از رویکردهای یادگیری ماشینی که در سوالات فارماکولوژی‌بالینی به کار گرفته شده اند، شامل ادغام تکنیک های مدل سازی«کلاسیک»، مانند تعیین یک مدل ساختاری مبتنی بر درک‌مکانیسمی، و رویکردهای یادگیری ماشینی است.[۶۹-۷۱] رویکردهای‌فارماکومتری کلاسیک مبتنی بر اصول فارماکولوژیکی هستند که‌فرضیه های حاصل از درک فیزیولوژی و خواص دارو را منعکس می کنند.بعید است که این مدل ها در آینده نزدیک به طور کامل با رویکردهای‌یادگیری ماشینی جایگزین شوند. با این حال، هنگامی که مجموعه‌داده ها و مسائل پیچیده تر هستند، تأثیرات و روابط ناشناخته زیادی‌وجود دارد و تمرکز بر درون یابی و ارزیابی سریع است، فارماکومتری‌ممکن است از به کارگیری روش هایی از نوع یادگیری ماشینی‌بهره مند شود. در ادامه، انتظار داریم که تلفیق این درک با مدل های‌یادگیری ماشینی بتواند به مدل های بسیار مؤثری در آینده منجرشود. یک مقاله اخیر با دیدگاهی جدید، جزئیات بیشتری در مورد کاربردهای‌یادگیری ماشینی در فارماکولوژی بالینی ارائه می دهد.[۷۴]

درعصر کلان داده، فرصت های جدید زیادی برای یادگیری ماشینی در داروشناسی بالینی‌وجود دارد. به عنوان مثال، داده های تولید شده از دستگاه های پوشیدنی، چالش های‌جدیدی را در مورد چگونگی پیوند آنها با داده های PK در آینده ایجاد می کنند.علاوه بر این، دسترسی به داده های دنیای واقعی می تواند شواهد محکمی برای‌متغیرهای کمکی، مجموعه داده های کنترل مکمل و مدل های تقویت کننده ای که روی‌مجموعه داده های کوچک آموزش دیده اند، فراهم کند.

درررویکردهای فارماکومتری، یک مدل پیش بینی کننده معمولاً با ادغام‌یک مدل ساختاری و داده های مرتبط ایجاد می شود. مدل ساختاری‌به طور قابل توجهی فضای راه حل را محدود می کند و بنابراین،داده های نسبتاً کمی برای برازش مدل مورد نیاز است. برعکس،در شبکه های عصبی، ساختار مدل از پیش مشخص نشده است‌و بنابراین، داده های نسبتاً بیشتری برای ساخت یک مدل پیش بینی کننده‌مورد نیاز است. همچنین لازم به ذکر است که ما هنوزدر مرحله ابتدایی درک هستیم که در آن نقطه ادغام داده های بزرگ‌تربا این روش های جدید یادگیری ماشین می تواند برای عملکرد درمقایسه با روش های سنتی تر مفید باشد. چالش زیر[۷۵]پیش بینی سری های‌زمانی نشان می دهد که ترکیب روش های آماری کلاسیک ویادگیری ماشینی دقیق ترین پیش بینی را ایجاد می کند و بنابراین، آن‌را به عنوان راهی برای پیشرفت پیشنهاد می دهد. یکی از محرک های‌اصلی موفقیت رویکردهای فارماکومتری این است که مدل هاشامل درک کاملی از فرآیندهای جذب، توزیع، متابولیسم و حذف‌دارو هستند. مدل های ایجاد شده بسیار پیش بینی کننده هستندو بنابراین، کاربرد گسترده ای در پشتیبانی از دارو پیدا می کنند.

توسعه.به دلیل این موفقیت، علیرغم ورود یادگیری ماشینی، انتظار نمی رودکه اهمیت و فعالیت رویکردهای کلاسیک فارماکومتری کاهش‌یابد. در مقابل، می توان آنها را با دانش و بینشی که توسط روش هاو مدل های یادگیری ماشینی استخراج شده است، تقویت و بهبودبخشید.

یک‌چالش مداوم برای اعضای جامعه داروشناسی بالینی که مایل به استفاده‌از روش های یادگیری ماشین هستند، شیوع ذاتی داده های طولی است.تاکنون، روش های یادگیری ماشین زیادی وجود دارند که برای پیش بینی‌به ویژگی های پایه متکی هستند، اما نمونه های نسبتاً کمی وجود داردکه در آن ها از داده های طولی استفاده شده باشد.

درمجموع، ما انتظار داریم که هرگز یک رویکرد جهانی و یکسان برای همه وجودنداشته باشد که مدل سازان از حوزه های مختلف به آن روی آورند. ما خاطرنشان‌می کنیم که حوزه های زیادی از هم افزایی بالقوه وجود دارد که در آن هاحوزه های مدل سازی در حوزه توسعه دارو با هم همپوشانی دارند. جامعه‌داروسازی بالینی به پایه گذاری تحلیل های خود بر اساس اصول دارویی‌ادامه خواهد داد و به تدریج عناصر جدید یادگیری ماشین را به گردش کارخود اضافه خواهد کرد و مدل های خود را بیشتر تقویت خواهد کرد. علاوه براین، جامعه داروسازی بالینی قادر خواهد بود طیف وسیعی از سوالاتی را که‌می تواند با استفاده از رویکردهای یادگیری ماشین به آن ها پاسخ دهد، افزایش‌دهد.

تأمین‌مالی
.تأمین‌مالی شده است F. Hoffmann-La Roche Ltd است. این مطالعه توسط Roche دریافت کننده ی‌بورسیه ی پسادکتری IID

تضادمنافع
همه‌نویسندگان هیچ گونه تضاد منافعی برای این اثر اعلام نکردند.

۱.کاماچو، دی ام، کالینز، کی ام، پاورز، آر کی، کاستلو، جی سی و کالینز، جی جی.یادگیری ماشینی نسل بعدی برای شبکه های بیولوژیکی.*سلول* ۱۵۸۱،۱۷۳–۱۵۹۲ (۲۰۱۸).

۲.شن، دی.، وو، جی. و سوک، اچ.-آی. یادگیری عمیق در تحلیل تصاویر پزشکی.*سال. کشیش. زیست پزشکی. مهندسی*.۱۹،۲۲۱–۲۴۸ (۲۰۱۷).

۳.راجکومار، آ.، دین، جی. و کوهان، آی. یادگیری ماشین در پزشکی. *مجله‌پزشکی انگلستان (N. Engl. J. Med)*.۳۸۰،۱۳۴۷–۱۳۵۸ (۲۰۱۹).

۴.کلاین، اس. سی. نمایش رویدادها در شبکه های عصبی و اتوماتای متناهی (نیروی‌هوایی پروژه رند، سانتا مونیکا، کالیفرنیا، ۱۹۵۱) <ADA596138/ https://apps.dtic.mil/docs/citations>.

5.بریمن، ل. مدل سازی آماری: دو فرهنگ (همراه با توضیحات و پاسخ نویسنده). *آمار.علم*.۱۶،۱۹۹–۲۳۱ (۲۰۰۱).

۶.ریبیرو، ام تی، سینگ، اس. و گسترین، سی. «چرا باید به شما اعتماد کنم؟»: توضیح‌پیش بینی های هر طبقه بندی کننده.*مجموعه مقالات بیست و دومین کنفرانس‌بین المللی ACM SIGKDD در زمینه کشف دانش و داده کاوی، سانفرانسیسکو،کالیفرنیا* ۱۳ تا ۱۷ آگوست ۲۰۱۶

۷.استرومبل، ای. و کونونکو، آی. توضیح مدل های پیش بینی و پیش بینی های منفردبا مشارکت ویژگی ها.*دانش. سیستم اطلاعاتی*. ۴۱،۶۴۷–۶۶۵ (۲۰۱۴).

۸.لیپتون، زد. سی. افسانه تفسیرپذیری مدل.*صف ACM*،۱۶، ۳۱–۵۷ (۲۰۱۸).

۹.هاستی، ت.، تیبشیرانی، ر. و فریدمن، ج.*عناصر یادگیری آماری: داده کاوی، استنتاج‌و پیش بینی*(انتشارات اشپرینگر نیویورک، نیویورک، نیویورک، ۲۰۰۸).

۱۰.جرز، جی.ام.*و همکاران*جایگذاری داده های گمشده با استفاده از روش های آماری و یادگیری‌ماشین در یک مسئله واقعی سرطان سینه.*مصنوع. هوشمند. پزشکی*. ۵۰،۱۰۵–۱۱۵ (۲۰۱۰).

۱۱.دوروگوش، ای وی، ارشوف، وی. و گولین، ای. کت بوست: تقویت گرادیان با پشتیبانی‌از ویژگی های دسته بندی شده.*پیش چاپ arXiv arXiv:1810.11363* (۲۰۱۸).

نظریه‌تصحیح سوگیری انتخاب نمونه. در).Zeugmann, T.)eds
Cortes, C., Mohri, M., Riley, M., Rostamizadeh, A., Freund, Y., Györfi, L., Turán, G. &
.12*نظریه یادگیری الگوریتمی*. 38–53 (اشپرینگر، برلین، هایدلبرگ، 2008).

۱۳.لی، بی. کی، لسلر، جی. و استوارت، ای. ای. بهبود وزن دهی امتیاز گرایش با استفاده‌از یادگیری ماشین.*آمار. پزشکی*.۲۹،۳۳۷–۳۴۶ (۲۰۱۰).

۱۴.نیوبی، دی.، فریتاس، ای.ای. و غفوریان، تی. مقابله با مجموعه داده های کلاس نامتعادل‌در مدل های جذب خوراکی.*مجله شیمی. مدل اطلاعات*.۵۳، ۴۶۱–۴۷۴ (۲۰۱۳).

۱۵.هو، ال.-اچ.، هوانگ، ام.-دبلیو.، کی، اس.-دبلیو. و تسای، سی.-اف. تأثیر تابع فاصله‌بر طبقه بندی k-نزدیک ترین همسایه برای مجموعه داده های پزشکی. *اسپرینگرپلاس*۵،۱۳۰۴ (۲۰۱۶).

۱۶.بروزان، آی.، وات، اس. و فرتی، وی. ادغام معیارهای شباهت توالی مبتنی بر هم ترازی‌و بدون هم ترازی برای طبقه بندی توالی های زیستی.*بیوانفورماتیک* ۳۱،۱۳۹۶–۱۴۰۴ (۲۰۱۵).

۱۷.کولیسون، ای.ای.، بیلی، پی.، چانگ، دی.کی. و بیانکین، ای.وی. زیرگروه های مولکولی‌سرطان پانکراس.*نات. Rev. Gastroenterol. هپاتول*. ۱۶،۲۰۷–۲۲۰ (۲۰۱۹).

۱۸.لوید، اس. کوانتیزاسیون کمترین مربعات در PCM.*نظریه انتقال اطلاعات IEEE*.۲۸،۱۲۹–۱۳۷ (۱۹۸۲).

.خوشه‌بندی مبتنی بر چگالی Kriegel, H.-P., Kröger, P., Sander, J., Zimek, A.
.19*وایلی اینتردیسک. حداقل دانش داده های بازیابی شده. دیسکاو*.۱، ۲۳۱–۲۴۰ (۲۰۱۱).

20.استر، م.، کریگل، اچ.-پی.، ساندر، جی. و شو، XA الگوریتم مبتنی بر چگالی برای کشف‌خوشه ها، یک الگوریتم مبتنی بر چگالی برای کشف خوشه ها در پایگاه‌های‌داده مکانی بزرگ با نویز. *مجموعه مقالات دومین کنفرانس بین المللی‌کشف دانش و داده کاوی*.)AAAI Press, Portland, OR, 1996(226-231

۲۱.زیمک، آ.، شوبرت، ای. و کریگل، اچ.-پی. بررسی تشخیص داده های پرت بدون نظارت‌در داده های عددی با ابعاد بالا.*داده های آماری. تحلیلی. حداقل*۵،۳۶۳–۳۸۷(۲۰۱۲).

۲۲.پیرسون، ک. درباره خطوط و صفحاتی که بیشترین برازش را به سیستم های نقاط‌در فضا دارند.*مجله فلسفی و مجله علمی لندن، ادینبورگ و دوبلین*۲، ۵۵۹–۵۷۲ (۱۹۰۱).

23. van der Maaten, L. & Hinton, G. تجسم داده ها با استفاده از t-SNE. *جی.ماخ. پژوهش های یادگیری*.۹،۲۵۷۹–۲۶۰۵ (۲۰۰۸).

۲۴.بخت، ای.*و همکاران*کاهش ابعاد برای مصورسازی داده های تک سلولی با استفاده‌از UMAP.*نات. بیوتکنولوژی*.۳۷،۳۸–۴۴ (۲۰۱۹).

۲۵.نگوین، ال اچ و هولمز، اس. ده نکته سریع برای کاهش ابعاد مؤثر. *PLoS Comput. Biol*.۱۵،.)2019(e1006907

۲۶.وانگ، سی.، ماچیراجو، آر. و هوانگ، کی. طبقه بندی بیماران سرطان سینه با استفاده‌از روش خوشه بندی اجماعی منظم مولکولی.*روش ها*۶۷،۳۰۴–۳۱۲ (۲۰۱۴) .

۲۷.کوپر، جی اس، بایوم، ام ال و سامرز، ای سی. بینش های اخیر در اپیدمیولوژی بیماری های‌خودایمنی: بهبود تخمین شیوع و درک خوشه بندی بیماری ها.*جی. خودایمنی*.۳۳،۱۹۷–۲۰۷ (۲۰۰۹).

۲۸.والش، دی. و ریبیکی، ال. خوشه بندی علائم در سرطان پیشرفته. *پشتیبانی. مراقبت.سرطان*۱۴،۸۳۱–۸۳۶ (۲۰۰۶).

۲۹.کنسرسیوم ژنوتیپ-بیان بافت (GTEx)*و همکاران*آنالیز آزمایشی ژنوتیپ-بیان بافت (GTEx): تنظیم ژن چند بافتی در انسان.*علم*۳۴۸،۶۴۸–۶۶۰ (۲۰۱۵).

30.ژانگ، جی. دی، برنتنیس، ان.، راث، ای. و ابلینگ، ام. داده کاوی شبکه ای از ژن های‌پاسخ اولیه را به عنوان امضای اجماعی سمیت ناشی از دارو در شرایط آزمایشگاهی(in vitro) و درون بدن (in vivo) آشکار می کند. *فارماکوژنومیک‌جی*.۱۴،۲۰۸–۲۱۶ (۲۰۱۴).

۳۱.جما، آ.*و همکاران*خوشه بندی داروهای ضد سرطان در سرطان ریه بر اساس پروفایل های‌بیان ژن و پایگاه داده حساسیت.*سرطان BMC*۶،۱۷۴ (۲۰۰۶).

۳۲.کخ، ام.ای و والدمن، اچ. خوشه بندی شباهت ساختار پروتئین و ساختار محصول طبیعی‌به عنوان اصول راهنما در کشف دارو.*کشف مواد مخدر. امروز*۱۰،۴۷۱–۴۸۳(۲۰۰۵).

۳۳.روتلینگر، م. و اشنایدر، گ. کاهش ابعاد غیرخطی و نگاشت کتابخانه های ترکیبات برای‌کشف دارو.*نمودار. مدل. مجله مول*.۳۴،۱۰۸–۱۱۷ (۲۰۱۲).

۳۴.عزت، آ.، وو، م.، لی، ایکس.-ال. و کو، سی.-کی. پیش بینی برهمکنش دارو-هدف بااستفاده از یادگیری جمعی و کاهش ابعاد. *روش ها*۱۲۹،۸۱–۸۸ (۲۰۱۷).

۳۵.نستروف، ی.*سخنرانی هایی در مورد بهینه سازی محدب*.جلد۱۳۷ (اشپرینگر، بین‌الملل، ۲۰۱۸).

۳۶.کاولی، جی سی و تالبوت، ان ال سی، درباره بیش برازش در انتخاب مدل و سوگیری‌انتخاب بعدی در ارزیابی عملکرد.*جی. ماخ. یادگیری. پژوهش*.۱۱،۲۰۷۹–۲۱۰۷ (۲۰۱۰).

۳۷.میچل، تی ام*یادگیری ماشین*((شرکت مک گرا-هیل، نیویورک، ۱۹۹۷)

۳۸.بلسون، دبلیو. ای. تطبیق و پیش بینی بر اساس اصل طبقه بندی بیولوژیکی.*آمار JR .جامعه آماری. گروه C . آمار کاربردی*.۸،۶۵–۷۵ (۱۹۵۹).

۳۹.کریشر، جی. پی. کتابشناسی مشروح از کاربردهای تحلیل تصمیم گیری در مراقبت های‌بهداشتی.*عملیات. تحقیق*.۲۸،۹۷–۱۱۳ (۱۹۸۰).

۴۰.شورتلیف، ای اچ، بوکانان، بی جی و فیگنباوم، ای ای. مهندسی دانش برای تصمیم گیری‌پزشکی: مروری بر ابزارهای تصمیم گیری بالینی مبتنی بر کامپیوتر.*مجموعه مقالات IEEE*۶۷،۱۲۰۷–۱۲۲۴ (۱۹۷۹).

۴۱.باخ، پی اچ و بریجز، جی دبلیو. رویکرد درخت تصمیم گیری برای کاربرد مطالعات متابولیسم‌و سینتیک دارو در آزمایش های سم شناسی و داروشناسی درون تنی‌و برون تنی.*قوس. توکسیکول. مکمل*.۸،۱۷۳–۱۸۸ (۱۹۸۵).

۴۲.جردن، تی جی و رایشمن، ال بی. دوزاژ یک بار در روز در مقابل دوز دو بار در روز تئوفیلین:رویکردی برای تحلیل تصمیم گیری جهت ارزیابی سطح خونی تئوفیلین‌و میزان پذیرش.*جناب کشیش. تنفس. دیس*. ۱۴۰،۱۵۷۳–۱۵۷۷ (۱۹۸۹).

۴۳.بریمن، ل.، فریدمن، ج.، اولشن، ر. و استون، س. درخت های طبقه بندی و رگرسیون.*گروه بین المللی وادزورث*۳۷، ۲۳۷–۲۵۱ (۱۹۸۴).

۴۴.کوینلان، جی آر. القای درخت های تصمیم گیری.*ماخ. یاد بگیر*.۱،۸۱–۱۰۶ (۱۹۸۶).

۴۵.بریمن، ل. جنگل های تصادفی.*ماخ. یاد بگیر*.۴۵،۵–۳۲ (۲۰۰۱).

۴۶.سگال، معیارهای یادگیری ماشین MR و رگرسیون جنگل تصادفی. گزارش فنی، (مرکز بیوانفورماتیک و آمار زیستی مولکولی، دانشگاه کالیفرنیا، سانفرانسیسکو، کالیفرنیا،۲۰۰۳).

47. Ben-Hur, A., Ong, CS, Sonnenburg, S., Schölkopf, B. & Rätsch, ز.ماشین های بردار پشتیبان و هسته ها برای زیست شناسی محاسباتی.*PLoS Comput. Biol*.۴،.)2008(e1000173

۴۸.نیومن، آ.*و همکاران*شمارش دقیق زیرمجموعه های سلولی از پروفایل های بیان بافت.*روش های طبیعی*۱۲،۴۵۳–۴۵۷ (۲۰۱۵).

۴۹.واپنیک، وی. و لرنر، ای. تشخیص الگو با استفاده از روش پرتره تعمیم یافته. *کنترل‌اتوماتیک. رم*.۲۴،۷۷۴–۷۸۰ (۱۹۶۳).

۵۰.اسمولا، ای جی و شولکوپف، بی. آموزشی در مورد رگرسیون بردار پشتیبان.*آمار. محاسبات*.۱۴،۱۹۹–۲۲۲ (۲۰۰۴).

۵۱.مرسر، جی. توابع از نوع مثبت و منفی و ارتباط آنها با نظریه معادلات انتگرال. *فلسفه.ترجمه. آر. جامعه شناسی. لندن. بی. زیست. علمی*.۲۰۹، ۴۱۵–۴۴۶، (۱۹۰۹).

52. Schölkopf, B. & Smola, AJ*یادگیری با هسته ها: ماشین های بردار پشتیبان، منظم سازی،بهینه سازی و فراتر از آن*(انتشارات MIT، کمبریج، ماساچوست، ۲۰۰۲).

۵۳.کریژفسکی، آ.، ساتسکور، آی. و هینتون، جی ای. طبقه بندی ایمیج نت با شبکه های‌عصبی کانولوشن عمیق.*پیشرفت ها در سیستم های پردازش اطلاعات‌عصبی*۲۵،۱۰۹۷–۱۱۰۵ (۲۰۱۲).

۵۴.چو، ک.*و همکاران*یادگیری نمایش عبارات با استفاده از رمزگذار-رمزگشای RNN برای‌ترجمه ماشینی آماری.*پیش چاپ arXiv arXiv:1406.1078*(۲۰۱۴).

۵۵.چو، ک.، ون مرینبوئر، ب.، باهداناو، د. و بنجیو، ی. درباره ی ویژگی های ترجمه ی‌ماشینی عصبی: رویکردهای رمزگذار-رمزگشا.*arXiv:1409.1259 arXiv preprint*(۲۰۱۴).

۵۶.کینگما، دی پی، محمد، اس.، رزنده، دی جی و ولینگ، ام. یادگیری نیمه نظارتی با مدل های‌مولد عمیق. NIPS'14: مجموعه مقالات بیست و هفتمین کنفرانس بین المللی‌سیستم های پردازش اطلاعات عصبی،۲،۳۵۸۱–۳۵۸۹ (۲۰۱۴).

۵۷.چوی، ای.*و همکاران*استفاده از مدل های شبکه عصبی بازگشتی برای تشخیص زودهنگام‌شروع نارسایی قلبی.*دانشگاهی OUP*(<24/2/361/2631499/ ۲۰۱۶(>https://academic.oup.com/jamia/article

۵۸.تانگ، جی.-تی.، کائو، وای.، شیائو، جی.-وای. و گوئو، کیو.-ال. پیش بینی غلظت‌پلاسمایی ریمیفنتانیل بر اساس شبکه عصبی المن.*دانشگاه جی. سنت جنوبی*.۲۰،۳۱۸۷–۳۱۹۲ (۲۰۱۳).

59.پاسکانو، ر.، میکولوف، ت. و بنجیو، ی. در مورد دشواری آموزش شبکه های عصبی‌بازگشتی. کنفرانس بین المللی یادگیری ماشین، ادینبورگ، 26 ژوئن - 1ژوئیه 1310 - 1318 (2013).

.درزمینه آکوستیک، گفتار و پردازش سیگنال، کیوتو، 25-30 مارس، 8624-8628 (2012) IEEE پیشرفت ها در بهینه سازی شبکه های بازگشتی. کنفرانس بین المللی 60. Bengio, Y., Boulanger-Lewandowski, N. & Pascanu, R.

۶۱.هوخرایتر، س. و اشمیدهوبر، ج. حافظه کوتاه مدت بلندمدت.*محاسبات عصبی*.۹، ۱۷۳۵–۱۷۸۰ (۱۹۹۷).
۶۲.گودفلو، آی.، بنگیو، وای. و کورویل، ای.*یادگیری عمیق*(انتشارات MIT، کمبریج، MA، ۲۰۱۶) <http://www.deeplearningbook.org>
63.چانگ، جی.، گولسهره، سی.، چو، کی. و بنگیو، وای. ارزیابی تجربی شبکه های عصبی‌بازگشتی دروازه دار در مدل سازی توالی. کارگاه آموزشی یادگیری عمیق در NIPS، مونترال، 8 تا 13 دسامبر (2014).
۶۴.ریو، جی وای، کیم، اچ یو و لی، اس وای. یادگیری عمیق پیش بینی تداخلات دارو و غذارا بهبود می بخشد.*مجموعه مقالات علمی، آکادمی ملی علوم*۱۱۵، ۲۰۱۸. (E4304–E4311
65. Hammann, F., Gutmann, H., Vogt, N., Helma, C. و Drewe, جی.پیش بینی عوارض جانبی داروها با استفاده از مدل سازی درخت تصمیم گیری.*کلینیک. فارماکول. درمان*.۸۸، ۵۲–۵۹ (۲۰۱۰).
66.لنکستر، ام سی و سوبی، ای. پیش بینی بهبود یافته ی تورساد دو پوینت های ناشی‌از دارو از طریق شبیه سازی دینامیک و الگوریتم های یادگیری ماشین. *کلینیک.فارماکول. درمان*.۱۰۰، ۳۷۱–۳۷۹ (۲۰۱۶).
۶۷.دانهاور، آی.*و همکاران*پیش بینی زودهنگام بهبود یافته هیپربیلیروبینمی نوزادی مرتبط‌با علائم بالینی با یادگیری ماشینی.*اطفال. پژوهش*. ۱۲۲،۸۶ (۲۰۱۹).
۶۸.گاودا، ای.ای.*و همکاران*شخصی سازی مدیریت دارویی کم خونی با استفاده از یادگیری‌تقویتی.*شبکه های عصبی*۱۸، ۸۲۶–۸۳۴ (۲۰۰۵).

۶۹.بنهامو، پ.-ی.*و همکاران*تحویل انسولین با حلقه بسته در بزرگسالان مبتلا به دیابت‌نوع 1 در شرایط واقعی: یک کارآزمایی متقاطع تصادفی کنترل شده چند مرکزی12 هفته ای با برچسب باز.*لنست دیگ سلامت*۱.، 2019(e17–e25)

70.چاوادا، آر.، قوش، ن.، ساندارادورا، آی.، مالی، ام. و ون هال، تأسیس‌یک AUC توسط fS۲۴-آستانه نفروتوکسیسیتی گامی به سوی دوز وانکومایسین‌فردی برای باکتریمی استافیلوکوکوس اورئوس مقاوم به متی سیلین‌است.*ضد میکروب. عوامل شیمی درمانی*.۶۱.، 2017(e02535–16)

۷۱.هنینگ، س.، هولتهاوس، ف. و استاتز، س. ای. مقایسه روش های تنظیم دوز برای‌توبرامایسین یک بار در روز در بیماران کودک و نوجوان مبتلا به فیبروز کیستیک.*کلینیک. فارماکوکینتیک*.۵۴، ۴۰۹–۴۲۱ (۲۰۱۵).

۷۲.دانسیریکول، سی.، موریس، آر.جی.، تت، اس.ای. و دافول، اس.بی. رویکرد بیزی برای‌مدل سازی فارماکوکینتیک جمعیتی سیرولیموس.
مجله‌پزشکی بالینی بریتانیا (Br. J. Clin. Pharmacol.).۶۲،۴۲۰–۴۳۴ (۲۰۰۶).
۷۳.لان، دی جی، بست، ان.، توماس، ای.، ویکفیلد، جی. و اشپیگل هالتر، دی. تحلیل بیزی‌مدل های PK/PD جمعیتی: مفاهیم کلی و نرم افزار.*مجله فارماکوکینت. فارماکودین*.۲۹،۲۷۱–۳۰۷ (۲۰۰۲).

۷۴.هاچینسون، ل.*و همکاران*مدل ها و ماشین ها: چگونه یادگیری عمیق، داروشناسی بالینی‌را به سطح بعدی خواهد برد..*Pharmacomet. System. Pharmacol. CPT*.۸،۱۳۱–۱۳۴ (۲۰۱۹).
۷۵.ماکریداکیس، س.، اسپیلیوتیس، ای. و آسیماکوپولوس، و. رقابت M4: نتایج، یافته ها،نتیجه گیری و مسیر پیش رو.*پیش بینی بین المللی ر*.۳۴،۸۰۲–۸۰۸ (۲۰۱۸).