امین نظری

پویش کنید

هزاران نفر را نجات بده!

یکی از مهم‌ترین مسائل در زمینهٔ پزشکی، تشخیص سریع و به‌موقع بیماری‌هاست. زیرا تشـــخیص زودهنگام بیماری می‌تواند، به درمان ســـریع بیمار کمک کند و از میزان مرگ و میر در بیماران بکاهد. در میان بیماری‌های گوناگون، سرطان سینه، به‌عنوان شایع‌ترین سرطان بین زنان و دومین ســـرطان شایع جهان، همواره هزینه‌های زیادی را بر جامعه تحمیل می‌کند. در جهان، سرطان سینه، زندگی ۱۰ درصد از زنان را تحت تأثیر قرار داده اســـت. در سال‌های اخیر و در کشور ما ایران، ســـرطان سینه افزایشی چشمگیر داشته است. همین موضوع انگیزهٔ اصلی نوشتن این مقاله است. بنابراین، سعی می‌کنیم توانایی‌ها و قابلیت‌های فن‌های یادگیری ماشین برای شناسایی سرطان سینه را بررسی کنیم.

معمولاً بیماری سرطان به دو نوع خوش‌خیم و بدخیم تقسیم می‌شود که روش‌های درمان هرکدام متفاوت هستند. بنابراین، تشخیص نوع سرطان اقدامی ضروری برای مقابله با آن است. برای این منظور باید از بافت مشـــکوک به سرطان نمونه‌برداری کرد و با بررسی و آزمایش نمونه‌ها، نوع ســـرطان را تشخیص داد. برای تشخیص از ویژگی‌هایی نظیر اندازهٔ سلول‌ها، بافت سلول، سفتی یا نرمی سلول استفاده می‌شود. اما متأســـفانه دقت تشخیص نوع سرطان توسط پزشکان، حدود ۷۹ درصد است. این در حالی است که روش‌های یادگیری ماشین با دقتی حدود ۹۷ درصد توانایی تشـــخیص نوع سرطان را دارند. این افزایش دقت بسیار مهم است و می‌تواند به جلوگیری از مرگ هزاران نفر منجر شود. در این مقاله ســـعی می‌کنیم با استفاده از روش‌های یادگیری ماشـــین، بیماری سرطان سینه را دسته‌بندی و فرایند انجام پروژه را بررسی کنیم. این فرایند شامل چندین مرحله است:

۱. شـــناخت و پیش‌پردازش داده: در این مرحله، ابتدا داده‌ها را جمع‌آوری و بررسی می‌کنیم. سپس آن‌ها را برای کار با الگوریتم‌های یادگیری ماشین پالایش و به قالب مناسب تبدیل می‌کنیم. در اینجا ما از مجموعه‌دادگان آماده استفاده خواهیم کرد که از وبگاه **uci**[1] قابل دریافت هستند. در این مجموعه دادگان ۵۶۷ نمونه وجود دارد که هر نمونه ۳۰ ویژگی دارد. ســـرطان حدود ۶۲ درصد از بیماران مبتلا به سرطان سینه از نوع خوش‌خیم و مابقی بدخیم است.

۲. ساخت مدل: برای ساخت مدل می‌توان از الگوریتم‌های یادگیری ماشین استفاده کرد. با توجه به اینکه مسئلهٔ ما از نوع مسائل یادگیری با نظارت اســـت، می‌توانیم از الگوریتم‌هایـــی نظیر درخت تصمیم، ماشین بردار پشتیبان و شبکهٔ عصبی چندلایه برای پیش‌بینی نوع بیماری اســـتفاده کنیم. در این روش‌ها، ابتدا با استفاده از داده‌های برچسب‌گذاری‌شده، مدل آموزش داده می‌شود. سپس مدل می‌تواند با دریافت دادهٔ بدون برچسب، آن را برچسب‌گذاری کند.

۳. ارزیابی مدل: پس از ساخت مدل باید آن را با استفاده از داده‌های جدید ارزیابی کرد و میزان دقت مدل را اندازه گرفت. مدل باید دقت کافی (دقتی بیش از دقت انســـان) داشته باشد تا قابلیت استفاده در محیط‌های واقعی را داشته باشد. بنابراین، از بین مدل‌های ساخته‌شده توسط الگوریتم‌های گوناگون، مدلی که بیشترین دقت را داشته باشد، می‌تواند به‌عنوان مدل نهایی انتخاب شـــود و در ساخت برنامه‌های کاربردی مورد استفاده قرار گیرد.

ما برای پیاده‌سازی الگوریتم‌های یادگیری ماشین از زبان برنامه‌نویسی

	دقت	کارایی	صحت
SVM بدون عادی‌سازی داده	۱/۰	۰/۹۰	۰/۹۴
SVM با عادی‌سازی داده	۱/۰	۰/۹۴	۰/۹۷

جدول شماره ۱

پایتون اســـتفاده کرده‌ایم. برای ساخت و ارزیابی مدل باید داده‌ها را به دو گروه داده‌های آموزشـــی و داده‌های آزمایشـــی تقســـیم کرد. داده‌های آموزشـــی برای یادگیری مدل مورد استفاده قرار می‌گیرند و از داده‌های آزمایشی برای آزمون و ارزیابی آن استفاده خواهد شد. بنابراین، داده‌ها را به دو گروه آموزشی و آزمایشی به نسبت ۷۰ به ۳۰ تقسیم می‌کنیم. به عبارت دیگر، از ۷۰ درصد از داده‌ها برای آموزش مدل و از ۳۰ درصد برای آزمایش استفاده خواهیم کرد.

بـــرای ارزیابی عملکرد مدل‌ها از معیارهایـــی نظیر صحت[2]، دقت[3] و کارایی[4] استفاده می‌شود. برای تعریف این موارد، ابتدا دانستن تعریف اصطلاحات مثبت کاذب[5]، منفی کاذب[6]، مثبت واقعی[7] و منفی واقعی[8] نیاز است.

مثبت کاذب (FN): بیانگر تعداد رکوردهایی است که دستهٔ واقعی آن‌ها منفی اســـت و مدل، رتبهٔ (کلاس) آن‌ها را به اشـــتباه مثبت پیش‌بینی کرده است.

مثبت واقعی (TP): بیانگر تعداد رکوردهایی است که دستهٔ واقعی آن‌ها مثبت اســـت و الگوریتم نیز برچسب آن‌ها را به‌درستی مثبت تشخیص داده است.

منفی واقعی (TN): بیانگر تعداد رکوردهایی است که دستهٔ واقعی آن‌ها منفی اســـت و الگوریتم نیز دســـتهٔ آن‌ها را به‌درســـتی منفی تشخیص داده است.

منفی کاذب (FN): بیانگر تعداد رکوردهایی است که دستهٔ واقعی آن‌ها مثبت اســـت و مدل، آن‌ها را به اشتباه منفی پیش‌بینی کرده است.

دقت: دقت مدل مشـــخص می‌کند چند درصد از مواردی که مدل به‌عنوان سرطان بدخیم پیش‌بینی کرده، درست بوده‌اند.

.....

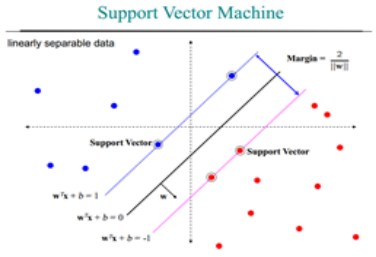
فراخوان: تعیین می‌کند چند درصد از افرادی که ســـرطان بدخیم داشته‌اند، به‌درستی شناسایی شده‌اند.

.....

صحت: کیفیت و کارایی الگوریتم را مشخص می‌کند.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

ما در اینجا از الگوریتم ماشـــین بردار پشتیبان استفاده کرده‌ایم. این الگوریتم سعی می‌کند مرز تصمیم (خوش‌خیم‌بودن یا بدخیم‌بودن ســـرطان) را به‌گونه‌ای بیابد که این مرز از نمونه‌های هر دو کلاس، بیشـــترین حاشیه یا فاصله را داشته باشند. برای مثال، در شکل زیر می‌توان بی‌نهایت خط ترسیم کرد که نمونه‌های هر دو کلاس آبی و قرمز را به‌درستی از هم تفکیک کنند. اما کدام خط مناسب‌ترین خط است؟ SVM می‌گوید، خطی بهترین است که بیشترین فاصله را از نمونه‌های هر دو کلاس داشته باشد.



در جدول ۱ این معیارها برای الگوریتم ماشین بردار پشتیبان آورده شده‌اند.

پی‌نوشت‌ها

1. https://goo.gl/U2Uwz2
2. Accuracy
3. Precision
4. Recall
5. False positive
6. False negative
7. True positive
8. True negative