ارزیابی عملکرد روش‌های دسته‌بندی متن

عملکرد یک سیستم دسته‌بندی متن، از طریق پارامترهای متفاوتی نظیر صحت (Accuracy) یادآوری (Recall) ، دقت (Precision) و امتیاز F1 سنجیده می‌شود. درک این معیارها، به کاربران اجازه می‌دهد تا بفهمند که یک مدل دسته‌بندی توسعه داده شده، تا چه حد در تحلیل داده‌های متنی خوب عمل می‌کند.برای ارزیابی (Evaluation) عملکرد یک سیستم دسته‌بندی داده‌های متنی، می‌توان از یک مجموعه داده تست ثابت (مجموعه‌ای از داده‌های متنی با اندازه از پیش تعیین شده که کلاس (برچسب) هر کدام از نمونه‌های موجود در آن مشخص شده است) یا از روشی به نام (Cross Validation) استفاده کرد. چنین فرآیندی در مرحله ارزیابی، داده‌های آموزشی را به دو زیر مجموعه تقسیم می‌کند؛ زیر مجموعه اول برای آموزش مدل یادگیری ماشین و زیر مجموعه دوم برای تست عملکرد سیستم استفاده می‌شود.

معیار یادآوری

معیار یادآوری(Recall) ، بیان کننده نسبت «تعداد داده‌های متنی درست دسته‌بندی شده» در یک کلاس خاص، به تعداد کل داده‌هایی است که باید در همان کلاس خاص دسته‌بندی شوند. مقدار بالا برای معیار یادآوری، بیانگر تعداد کم داده‌هایی است که به اشتباه، در آن کلاس خاص دسته‌بندی نشده‌اند. استفاده از این معیار، به تنهایی، برای ارزیابی عملکرد سیستم درست نیست و باید در کنار معیار دقت(Precision) مورد استفاده قرار بگیرد. زیرا، به راحتی می‌شود مدل‌های دسته‌بندی متنی طراحی کرد که یادآوری بالایی داشته باشند و این لزوما به معنای دقت (Precision) بالا نیست

$$Recall = \frac{TP}{TP + FN}$$

معیار صحت

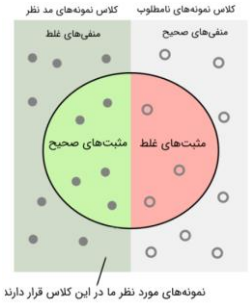
معیار صحت (Accuracy) ، بیان کننده تعداد «پیش‌بینی‌های صحیح انجام شده» توسط دسته‌بند، تقسیم بر، تعداد «کل پیش‌بینی‌های انجام شده» توسط همان دسته‌بند است. با این حال، این معیار به تنهایی، معیار مناسبی برای ارزیابی عملکرد یک دسته‌بند نیست. زمانی که کلاس‌های موجود در داده‌ها نامتوازن (Imbalanced) باشند (یعنی، تعداد داده‌های متعلق به کلاس (برچسب) خاص از کلاس‌های دیگر بسیار بیشتر باشد)، ممکن است سیستم با پدیده خاصی به نام تناقض صحت (Paradox Accuracy)مواجه شود. در نتیجه این تناقض، مدل دسته‌بند به احتمال زیاد عملکرد بسیار خوبی در پیش‌بینی کلاس (برچسب) داده‌ها از خود نشان می‌دهد؛ زیرا، اکثریت داده‌ها تنها به یکی از کلاس‌ها تعلق دارند. در صورتی که چنین

پدیده‌ای رخ دهد، بهتر است که معیارهای دیگری نظیر (فراخوانی | Recall) نرخ یادآوری و دقت (Precision) برای ارزیابی عملکرد سیستم در نظر گرفته شوند.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

معیار دقت

معیار دقت (Precision) ، نسبت تعداد «پیش‌بینی‌های صحیح انجام شده» برای نمونه‌های یک کلاس خاص، به تعداد «کل پیش‌بینی‌ها» برای نمونه‌های همان کلاس خاص را (این تعداد، مجموع تمامی پیش‌بینی‌های صحیح و پیش‌بینی‌های نادرست را شامل می‌شود) ارزیابی می‌کند. مقدار بالا برای معیار دقت، بیانگر تعداد کم داده‌هایی است که به اشتباه، در کلاس خاص دسته‌بندی شده‌اند. شایان توجه است که معیار دقت، فقط برای مواردی ارزیابی می‌شود که در آن‌ها، مدل دسته‌بندی تعلق یک نمونه به یک کلاس خاص را پیش‌بینی کرده باشد. برای برخی از فعالیت‌ها، نظیر ارسال پاسخ‌های خودکار به ایمیل‌ها(Automated Email Responses) ، مدل‌هایی نیاز است که سطح دقت آن‌ها بالا باشد؛ به عبارت دیگر، پاسخ‌ها تنها باید زمانی به کاربران ارسال شوند که مدل دسته‌بندی، با احتمال بالا، پیش‌بینی‌های درستی انجام داده باشد. در هنگام ارزیابی عملکرد یک مدل دسته‌بندی متن، بهتر است که از این معیار در کنار معیار یادآوری (Recall) استفاده شود.



$$Precision = \frac{TP}{TP + FP}$$

در این فرمول، وجود در مخرج باعث می شود که اگر تعداد تشخیص های اشتباه مان بالا باشد، صحت الگوریتم عددی FP نزدیک به صفر نشان دهد و بنابراین کارآیی مدل، زیر سوال برود.

MCC پارامتر دیگری است که برای ارزیابی کارایی الگوریتم‌های یادگیری ماشین از آن استفاده می‌شود. این پارامتر بیان‌گر کیفیت کلاس‌بندی برای یک مجموعه باینری می‌باشد. سنجه‌ای است که بیان‌گر بستگی مابین مقادیر مشاهده شده از کلاس باینری و مقادیر پیش‌بینی شده از آن می‌باشد. مقادیر مورد انتظار برای این کمیت در بازه ۱- و ۱ متغیر می‌باشد. مقدار ۱+، نشان دهنده پیش‌بینی دقیق و بدون خطای الگوریتم یادگیر از کلاس باینری می‌باشد. مقدار ۰، نشان دهنده پیش‌بینی تصادفی الگوریتم یادگیر از کلاس باینری می‌باشد. مقدار ۱-، نشان دهنده عدم تطابق کامل مابین موارد پیش‌بینی شده از کلاس باینری و موارد مشاهده شده از آن می‌باشد.

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

با این توضیحات، معیارهای یادآوری و صحت به جای معیار اولیه دقت، کاربرد وسیع تری در دنیای امروز یادگیری ماشین پیدا کرده است. در اغلب موارد، این دو معیار با هم رشد و حرکت نمی کنند. گاهی ما صحت مدل را با الگوریتم های دقیقتر بالا می بریم، یعنی آنهایی را که مثبت اعلام می کنیم، اکثراً درست هستند و موارد نادرست مثبت ما بسیار کم هستند یعنی صحت الگوریتم ما بسیار بالاست اما ممکن است جنبه یا ویژگی خاصی از داده ها را در نظر نگرفته باشیم و تعداد کل نمونه های مثبت، بسیار بیشتر از نمونه های اعلام شده ما باشد یعنی بازخوانی بسیار پایینی داشته باشیم.

از طرفی ممکن است کمی الگوریتم تشخیصی خود را ساده تر بگیریم تا تعداد مثبت های تشخیصی خود را بالا ببریم، در این صورت میزان اشتباهات ما زیادتر شده، صحت الگوریتم عدد پایین تر و بازخوانی آن، عدد بالاتری را نشان می دهد.

اگر بخواهیم میانگین معمولی دو معیار بازخوانی و صحت را ملاک کار درنظر بگیریم، برای حالت هایی که صحت بالا و بازخوانی پایینی داریم (و یا بالعکس)، میانگین معمولی عددی قابل قبول خواهد بود در صورتی که نباید نمره قبولی بگیرد.

اگر بتوانیم معیاری ترکیبی از این دو معیار برای سنجش الگوریتم های دسته بندی به دست آوریم، تمرکز بر آن معیار به جای بررسی همزمان این دو، مناسبتر خواهد بود...

معیار امتیازF1

این معیار، پارامترهای دقت (Precision) و یادآوری (Recall) را با هم ترکیب می‌کند تا مشخص شود یک مدل دسته‌بند تا چه حد عملکرد خوبی از خود نشان می‌دهد. به این معیار، میانگین متوازن (Harmonic Mean) دو معیار دقت (Precision) و یادآوری (Recall) نیز گفته می‌شود. این معیار، نسبت به معیار صحت (Accuracy) ، تصویر دقیق‌تری از نحوه عملکرد مدل دسته‌بند روی تمامی کلاس‌های موجود در داده‌ها ترسیم می‌کند.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

با توجه به محاسبات انجام گرفته برای معیارهای Precision وRecall ، در این مرحله می‌توان مقدار کمیت وزن‌دار F را محاسبه نمود که پارامتر مناسبی برای ارزیابی کیفیت کلاس‌بندی می‌باشد و همچنین توصیف‌کننده میانگین وزن‌دار مابین دو کمیت Precisionو Recall می‌باشد. برای یک الگوریتم کلاس‌بندی کننده در شرایط ایده‌آل، مقدار این کمیت برابر با ۱ می‌باشد و در بدترین وضعیت برابر با صفر می‌باشد.

گاها در مسایل دنیای واقعی میتوانیم یک حد آستانه پذیرش هم برای این مساله در نظر گرفت به گونه ای که مقادیر بالاتر از آن مقدار خاص مورد تایید قرار میگرند.

Reference: https://blog.faradars.org/text-mining-algorithms/