مقدمه‌ای بر کاربرد رگرسیون لجستیک در داده‌کاوی

رگرسیون لجستیک با ایده رگرسیون خطی گسترش پیدا کرده است و تنها با این شرط که متغیر وابسته از نوع رده‌ای باشد؛ منظور از رده‌ای بودن، این است که متغیر مشتقی از مشاهدات باشد. برای مثال شما به عنوان یک سرمایه‌گذار در بازار بورس، همواره با سه پیشنهاد روبرو هستید: اول آنکه سهمی را بخرید، دوم آنکه سهم خود را حفظ کنید و سوم آنکه سهم خود را بفروشید. در این‌جا هرکدام از پیشنهادات، در یک کلاس طبقه‌بندی می‌شوند و سه کلاس یا رده موجود است. البته در این بخش سعی داریم تنها متغیر باینری را مطرح کنیم که به معنای وجود دو طبقه می‌باشد.

قبل از ورود به مباحث آماری رگرسیون لجستیک، لازم به ذکر است که کاربردی‌ترین استفاده رگرسیون لجستیک در داده‌کاوی، تئوری انتخاب مصرف‌کننده است. اما تئوری انتخاب مصرف کننده چیست؟

در تئوری انتخاب مصرف کننده، بیان می‌شود که مصرف‌کننده در میان انتخاب‌هایی که دارد، انتخابی را انجام می‌دهد که حداکثر سود را برایش فراهم آورد. برای مثال شما به عنوان مصرف‌کننده یا خریدار یک محصول، اولویت‌هایی دارید که به‌واسطه آنها محصول مورد نظر خود را تهیه می‌کنید. این اولویت‌ها می‌تواند براساس ویژگی‌های فردی، اجتماعی و اقتصادی باشند و این اولویت‌ها باید وزن‌دهی شوند و این دقیقاً همان کاری است که رگرسیون لجستیک انجام می‌دهد و رفتار مصرف‌کننده را پیش‌بینی می‌کند.

مدل رگرسیون لجستیک:

ایده اصلی رگرسیون لجستیک برمبنای الگوریتم نایو بیز است و در آن سعی داریم تا $P(Y|X)$ را محاسبه کنیم، با این شرط که دیگر نیازی به محاسبه $P(X|Y)$ و $P(Y)$ نداریم. در رگرسیون لجستیک، از تابعی تحت عنوان تابع لاجیت[1] استفاده می‌شود. P احتمال تعلق داشتن به یک کلاس است و اگر آن را به شکل خطی (مشابه رابطه ذیل) در نظر بگیریم، الزاماً مقادیر خروجی در بازه صفر تا یک نخواهند بود.

$$p = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i$$

برای رفع این مشکل باید از یک تابع غیرخطی استفاده کنیم که از آن به عنوان تابع لاجیت نام می‌بریم و مقادیر خروجی آن همواره در بازه صفر تا یک قرار خواهند داشت. فرمول مربوط به این تابع به شرح ذیل است:

$$logit(p) = ln\left(\frac{p}{1-p}\right) = \ln(p) - \ln(1-p)$$

[1] Logit Function

اکنون با استفاده از این تابع، تابع پاسخ رگرسیون لجستیک، برای حل مشکل قرار گرفتن در بازه صفر و یک، به شکل زیر خواهد شد:

$$p = \frac{۱}{۱ + e^{-(\beta_. + \beta_۱ x_۱ + \cdots + \beta_i x_i)}}$$

لازم به ذکر است که معمولاً ترجیح می‌دهیم به جای محاسبه احتمال، میزان تعلق داشتن به یک کلاس را مورد بررسی قرار دهیم؛ لذا لازم است مفهوم شانس را نیز تشریح کنیم.

مفهوم شانس:

معمولاً در مکالمات روزمره نیز به‌جای صحبت پیرامون احتمال مثلاً برنده شدن، در مورد شانس[۲] برنده شدن بحث می‌شود. شانس، احتمال تعلق داشتن به یک کلاس را نسبت به احتمال تعلق داشتن به کلاس دیگر می‌سنجد و رابطه آن به شکل زیر است:

$$odds = \frac{p}{۱ - p}$$

برای مثال اگر احتمال برنده شدن در مسابقه اسب‌دوانی $\frac{۲}{۳}$ باشد، شانس برنده شدن در مسابقه، طبق فرمول فوق برابر ۲ است؛ یعنی احتمال برنده شدن در مسابقه اسب‌دوانی، دو برابر احتمال برنده نشدن است.

با نوشتن معادله فوق بر حسب p می‌توان با داشتن شانس، مقدار احتمال p را به شکل زیر محاسبه نمود:

$$p = \frac{odds}{۱ + odds}$$

همچنین می‌توان مدل رگرسیون لجستیک را برحسب شانس نیز بیان نمود، که در ابتدا لازم است رابطه بین شانس و احتمال را به شکل زیر نوشت:

$$odds = e^{\beta_. + \beta_۱ x_۱ + \cdots + \beta_i x_i}$$

سپس با لگاریتم گرفتن از طرفین رابطه، به فرمول استاندارد مدل لجستیک بر حسب شانس می‌رسیم:

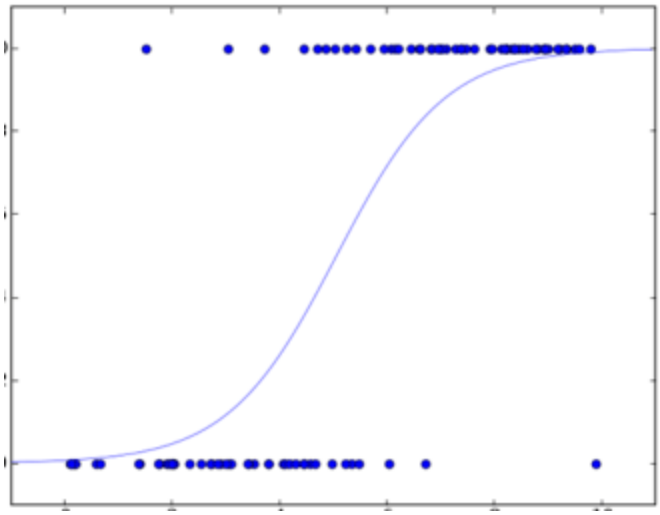
$$\log(odds) = \beta_. + \beta_۱ x_۱ + \cdots + \beta_i x_i$$

مثال۱: فرض کنید بانکی قصد دارد تا کمپینی راه‌اندازی کند و به مشتریان خود وام دهد و متغیرهایی همچون درآمد، سن، تحصیلات، جنسیت و غیره را مدنظر دارد و قصد دارد برمبنای این متغیرها، احتمال پذیرش وام را توسط مشتریان خود بررسی نماید. برای راحتی مسئله در این مثال، تنها یک متغیر (درآمد) را درنظر می‌گیریم. به ازای درآمدهای مختلف برای مثال از صفر تا ۲۵۰ واحد پولی، احتمال دریافت وام را از طریق رابطه زیر محاسبه می‌کنیم:

[۲] Odds

$$p\left(\text{دریافت وام}\middle|\text{درآمد} = x\right) = \frac{۱}{۱ + e^{-(-۶.۳۷+۰.۰۰۹۲x)}}$$

نمودار ذیل، ارتباط بین احتمال دریافت وام و درآمد افراد را به تصویر کشیده است:



همچنین برای مثال وقتی درآمد را صفر در نظر بگیریم، شانس دریافت وام به شکل زیر محاسبه می‌گردد:

$$odd\left(\text{دریافت وام}\right) = e^{-۶.۳۷} = ۰/۰۰۱۷$$

حال از مفهومی به نام مقدار برش استفاده می‌کنیم که برای تصمیم‌گیری در مورد دسته‌بندی داده‌ها به کار می‌رود. اگر مقدار برش در این مثال را ۰.۵ درنظر بگیریم، آنگاه تمامی مقادیر احتمالی محاسبه شده که کمتر از ۰.۵ هستند، به معنای عدم دریافت وام و تمامی مقادیر احتمالی محاسبه شده که بیشتر از ۰.۵ هستند، به معنای دریافت وام از بانک می‌باشند.

در ادامه با مفهوم نسبت شانس آشنا می‌شویم.

نسبت شانس:

در رگرسیون لجستیک، از نسبت شانس[۳] به عنوان آماره‌ای برای بررسی تاثیر متغیرهای مستقل بر روی متغیر خروجی استفاده می‌شود که رابطه آن به شکل زیر است:

$$OR = \frac{\frac{p_۱}{۱ - p_۱}}{\frac{p_.}{۱ - p_.}}$$

[۳] Odds Ratio (OR)

معمولاً یک ابهام بین شانس و نسبت شانس وجود دارد. از آنجایی‌که شانس در واقع نسبت بین احتمال تعلق داشتن به کلاس یک و احتمال تعلق داشتن به کلاس صفر می‌باشد، گاهی به اشتباه نسبت شانس نامیده می‌شود. برای درک هرچه بهتر تفاوت بین این دو مفهوم، مثال زیر را بررسی می‌کنیم:

مثال۲: در یک نمونه بیست نفره که شامل ده فرد با تحصیلات دانشگاهی بوده است، هفت نفر از افرادی که تحصیلات دانشگاهی داشته‌اند و سه نفر از افرادی که تحصیلات دانشگاهی نداشته‌اند، موفق به دریافت وام شده‌اند. باتوجه به توضیحات فوق که در جدول ذیل نیز منعکس شده‌اند، احتمال دریافت وام برای افراد دانشگاهی چند برابر افراد غیردانشگاهی است؟

متغیر مستقل / متغیر وابسته		تحصیلات	
		دانشگاهی	غیردانشگاهی
دریافت وام	بلی	۷	۳
	خیر	۳	۷

شانس دریافت وام برای افراد با تحصیلات دانشگاهی به صورت زیر محاسبه می‌شود:

$$Odds = \frac{۰.۷}{۱ - ۰.۷} = ۲.۳۳$$

همچنین شانس دریافت وام برای افراد با تحصیلات غیردانشگاهی به صورت زیر محاسبه می‌گردد:

$$Odds = \frac{۰.۳}{۱ - ۰.۳} = ۰.۴۳$$

اکنون باتوجه به رابطه نسبت شانس داریم: $OR = \frac{۲.۳۳}{۰.۴۳} = ۵.۴۲$

بدین معنا که احتمال دریافت وام برای فردی که تحصیلات دانشگاهی دارد، بیش از پنج برابر فردی است که تحصیلات دانشگاهی ندارد (در صورت یکسان بودن سایر شرایط).

لازم به ذکر است اگر نسبت شانس بزرگتر از یک باشد، تاثیر متغیر مستقل بر روی متغیر خروجی، مثبت است و اگر نسبت شانس کوچکتر از یک باشد، این تاثیر منفی و در واقع در جهت عکس است. یک بودن نسبت شانس، به معنای عدم تاثیر متغیر مستقل بر روی متغیر وابسته است.

در پایان به چند نکته در مورد الگوریتم رگرسیون لجستیک اشاره می‌کنیم:

- رگرسیون لجستیک در زمینه‌هایی همچون دسته‌بندی مشتریان به عنوان بازگشت یا عدم بازگشت کاربرد دارد.
- زمانی که مجموعه داده بزرگ باشد، جواب‌هایی که الگوریتم نایو بیز و الگوریتم رگرسیون لجستیک ارائه می‌دهند، به یکدیگر نزدیک هستند.
- توصیه می‌شود قبل از استفاده از رگسیون لجستیک، نویزگیری داده‌ها صورت پذیرد و نقاط دورافتاده شناسایی و حذف شوند.
- لازم است ورودی‌هایی که با هم همبستگی دارند به صورت دوبه دو بررسی شوند تا از عدم وجود همبستگی زیاد بین ورودی‌ها اطمینان حاصل شود.

منبع:

www.farabar.net/