رگرسیون لجستیک

استاد راهنما : دکتر وحید رضایی تبار

ارائه دهنده : حسام افشار

فهرست مطالب

- مقدمه
- رگرسیون لجستیک ساده
- ماتریس درهم ریختگی[1]
- تنظیم مدل (به روش ریدج[2] و لاسو[3])
- ارائه یک پروژه عملی

1-Confusion Matrix
2-Ridge
3-Lasso

مقدمه

اغلب برای بیان شدت رابطه خطی بین دو متغیر کمی از ضریب همبستگی استفاده می کنیم. همچنین برای نمایش مدل رابطه بین آن دو نیز از مدل رگرسیونی کمک می گیریم. در این میان یک الگو برای پیش بینی متغیر وابسته براساس متغیر مستقل ایجاد می شود. ولی باید توجه داشت که در مدل ایجاد شده، هر دو متغیر مستقل و وابسته، کمی هستند. همچنین شرط پیوسته بودن این مقدارها نیز در روش رگرسیون نهفته است. ولی ممکن است بخواهیم رابطه بین یک متغیر مستقل (با مقدارهای پیوسته یا گسسته) را با یک متغیر وابسته با مقدارهای کیفی بسنجیم. در این حالت روش عادی رگرسیون خطی جوابگو نخواهد بود و باید از«رگرسیون لجستیک» استفاده کرد. همچنین رگرسیون لجستیک یک روش یادگیری با نظارت است چون داده ها برچسب گذاری شده اند.

رگرسیون لجستیک ساده

در رگرسیون ساده با فرض نرمال بودن داده ها متغیر تبیینی با امید ریاضی رابطه خطی داشت و از امید ریاضی شرطی که همان میانگین بود به عنوان برآوردی از متغیر وابسته استفاده می شد. اما در اینجا فرض می شود متغیر های پاسخ ما دارای توزیع چند جمله ای هستند و تابعی از امید ریاضی با متغیر تبیینی در رابطه خطی است:

$$\ln\left(\frac{P}{1-P}\right) = X\beta \;\rightarrow\; E(Y|X=x) = \hat{P} = \frac{e^{X\hat{\beta}}}{1+e^{X\hat{\beta}}}$$

به این صورت احتمال رخداد متغیر هدف برآورد میشود و اگر این احتمال برای رخداد موفقیت بیشتر بود آن مشاهده را به دسته موفقیت و اگر احتمال برای شکست بیشتر بود آن را به دسته شکست نسبت میدهیم.

رگرسیون لجستیک ساده

همینطور برای حالاتی که بیشتر از دو حالت در متغیر هدف داشته باشیم این احتمال ها برای هر حالت محاسبه می شود و مشاهدات را به رده ای که بیشترین احتمال را داشته باشد نسبت میدهیم.

برای بدست آوردن پارامترهای بهینه یعنی β ها میتوان از روش برآورد بیشینه درستنمایی[1] استفاده کرد. اگر n تعداد نمونه ها باشد و نمونه ها را به شکل $D = (\vec{x}_1, y_1), \dots, (\vec{x}_n, y_n)$ نمایش دهیم آنگاه لگاریتم تابع درستنمایی به شکل زیر است:

$$l(D, \beta) = \log\left(\prod_{i=1}^{n} P(y_i = 1|\vec{x}_i, \beta)^{y_i} \times P(y_i = 1|\vec{x}_i, \beta)^{1-y_i}\right)$$

1-maximum likelihood estimation

رگرسیون لجستیک ساده

حال اگر بخواهیم از لگاریتم تابع درستنمایی نسبت به β مشتق بگیریم تا ضرایب بدست بیایند چون معادلات خطی نیستند فرم بسته برای ضرایب بدست نمی آید. برای همین باید از یک روش عددی مانند نیتون-رافسون[1] این مقادیر را بدست آورد.

ابتدا یادآوری می کنیم که می توان از بسط تیلور[2] برای تقریب تابع f حول نقطه ای مانند x_0 به صورت زیر استفاده کرد.

$$f(x) \approx f(x_0) + (x - x_0)f'(x_0)$$

که اگر x_1 ریشه تابع مورد نظر باشد آنگاه:

$$x_1 \approx x_0 - \frac{f(x_0)}{f'(x_0)}$$

1-Newton-Raphson method
2-Taylor series

رگرسیون لجستیک ساده

برای بهتر شدن می توان این روند را تکرار کرد یعنی تعریف کنیم:

$$x_2 \approx x_1 - \frac{f(x_1)}{f\prime(x_1)} \quad , \quad x_3 \approx x_2 - \frac{f(x_2)}{f(x_2)}$$

و این روند را تا جایی ادامه دهیم که بسیار به ریشه نزدیک شویم. حال ما میخواهیم ریشه مشتق تابع درستنمایی یعنی $l'(D,\beta)$ را نسبت به β را پیدا کنیم پس روش به صورت زیر تعیریف می شود:

$$\beta^{(new)} = \beta^{(old)} - \frac{l'(D,\beta)}{l''(D,\beta)} = \beta^{(old)} + (X'WX)^{-1}.X'(Y-\mu)$$

که در آن μ بردار $\hat{p}_i$ ها و ماتریس W یک ماتریس قطری است که مقادیر قطر اصلی آن $\hat{p}_i(1-\hat{p}_i)$ است و بقیه مقادیر آن صفر است. این روند را تا همگرا شدن β ادامه می دهیم.

ماتریس درهم ریختگی

در رگرسیون هایی که متغیر های هدف مقادیر پیوسته داشتند از میانگین مربعات خطا برای سنجش مدل استفاده میکردیم اما اینجا چون مسئله دسته بندی است باید دید چند درصد از مشاهدات درست دسته بندی شده اند.

برای این منظورماتریس درهم ریختگی مطرح میشود که با استفاده از آن میتوان معیار هایی که نشان دهنده دقت مدل است را استخراج کرد و بسته به نوع مسئله از آن استفاده کرد.

ماتریس درهم ریختگی

		برچسب پیش‌بینی شده	
		مثبت	منفی
برچسب شناخته شده	مثبت	TP	FN
	منفی	FP	TN

ماتریس درهم ریختگی

که با توجه به ماتریس درهم ریختگی داریم:

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

$$Sensitivity(Recall) = \frac{TP}{TP+FN}$$

$$Precision(PPV) = \frac{TP}{TP+FP}$$

$$F - Measure = \frac{2\times TP}{2\times TP+FP+FN}$$

ماتریس درهم ریختگی

دقت[1] : متداول ترین، اساسی ترین و ساده ترین معیار اندازه گیری کیفیت یک دسته بند است و عبارت است از نسبت داده هایی که درست دسته بندی شده اند به کل داده ها. این پارامتر در واقع نشان گر میزان الگوهایی است که درست تشخیص داده شده اند.

حساسیت[2] : به معنی نسبتی از موارد مثبت است که آزمایش آن ها را به درستی به عنوان نمونه مثبت تشخیص داده است.

صحت[3] : بیان کننده این است که چند درصد از الگوهایی که مدل آن ها را مثبت تشخیص داده، در واقعیت هم مثبت هستند.

معیار اف[4] : میانگین هارمونیک بین صحت و حساسیت.

1- Accuracy 3- Precision
2- Sensitivity 4- F-Measure

تنظیم مدل

برای جلوگیری از بیش برازشی در مدل‌های خطی مانند رگرسیون خطی یا رگرسیون لجستیک جریمه ای به تابع هزینه اضافه می‌شود تا از افزایش زیاد پارامترها جلوگیری شود. تابع هزینه را در رگرسیون لجستیک با منفی لگاریتم درست نمایی تعریف می‌کنیم تا کمینه کردن آن به بیشینه کردن تابع درست نمایی بیانجامد. به این کار تنظیم مدل گفته می‌شود. دو راه متداول تنظیم مدل‌های خطی روش‌های ریدج و لاسو هستند.

در تنظیم مدل به روش ریدج تابع هزینه را به این شکل تغییر می‌دهیم:

$$-l(D,\beta) + \lambda||\beta||_2^2 =$$

$$-\sum_{i=1}^{n}(y_i \log P(y_i = 1|x_i,\beta) + (1-y_i)\log p(y_i = 0|x_i,\beta)) + \lambda \sum_{k=0}^{m} {\beta_k}^2$$

تنظیم مدل

در روش ریدج با اضافه کردن جریمه به تابع هزینه باعث میشود بعضی ازضرایب مقدارشان بسیار نزدیک به صفر میشود و از پیچیدگی مدل جلوگیری میکند.

و در تنظیم مدل به روش لاسو تابع هزینه را به این شکل تغییر می‌دهیم:

$$-l(D,\beta) + \lambda||\beta||_1 =$$

$$-\sum_{i=1}^{n}(y_i \log P(y_i = 1|x_i,\beta) + (1-y_i)\log p(y_i = 0|x_i,\beta)) + \lambda\sum_{k=0}^{m}|\beta_k|$$

این روش باعث می‌شود که بسیاری از پارامترهای مدل نهائی صفر شوند و مدل به اصطلاح خلوت شود.

همچنین بهترین λ برای هر دو روش را میتوان با استفاده از روش اعتبارسنجی متقابل[1] پیدا کرد.

1- Cross-Validation

پروژه عملی

سپرده های مدت دار منبع اصلی درآمد بانک ها هستند. سپرده مدت دار یک سرمایه گذاری نقدی است که در یک موسسه مالی نگهداری می شود. پول شما برای نرخ بهره توافق شده در مدت زمان یا مدت معینی سرمایه گذاری می شود. این بانک برای فروش سپرده های مدت دار به مشتریان خود برنامه های توسعه ای مختلفی از جمله بازاریابی ایمیلی، تبلیغات، بازاریابی تلفنی و بازاریابی دیجیتال دارد.

کمپین های بازاریابی تلفنی همچنان یکی از موثرترین راه ها برای ارتباط با مردم هستند. با این حال، آنها نیاز به سرمایه گذاری هنگفت دارند زیرا مراکز تماس بزرگ برای اجرای واقعی این کمپین ها استخدام می شوند. از این رو، شناسایی مشتریانی که بیشترین احتمال را برای تبدیل از قبل دارند بسیار مهم است تا بتوان آنها را به طور خاص از طریق تماس مورد هدف قرار داد.

داده ها مربوط به کمپین های بازاریابی مستقیم (تماس های تلفنی) یک موسسه بانکی پرتغالی است. هدف طبقه بندی این است که پیش بینی کند آیا مشتری برای یک سپرده مدت دار مشترک می شود یا خیر. اغلب برای دسترسی به این که آیا محصول (سپرده مدت دار بانکی) توسط مشتری مشترک می شود یا خیر، بیش از یک تماس با یک مشتری مورد نیاز بود.

پروژه عملی

داده ها شامل 45211 سطر و 17 ستون است که 16 ستون اول متغیرهای وابسته و 1 ستون آخر متغیر هدف است. شرح مفصل ستونها به صورت زیر است:

age	سن افراد (مقداری عددی)
job	نوع شغل افراد که شامل 12 نوع شغل مختلف است (مقدار اسمی: مدیر،ناشناس، بیکار، مدیریت، خدمت خانگی، کارآفرین، کارگر، کار آزاد، بازنشسته، تکنسین، خدمات)
marital	وضعیت تاهل (مقدار اسمی: متاهل ، طلاق گرفته، مجرد)
education	سطح تحصیلات (مقدار اسمی: ناشناس ابتدایی، دبیرستان، دانشگاهی)
default	آیا در پرداخت بدهی کوتاهی داشته یا نه؟ (به صورت دودویی: بله، خیر)
balance	میانگین تراز سالانه، به یورو(مقدار عددی)

پروژه عملی

housing	وام مسکن دارد؟ (دودویی : بله، خیر)
loan	وام شخصی دارد؟ (دودویی : بله، خیر)
contact	نوع ارتباط (مقدار اسمی: تلفنی، تلفن همراه، ناشناس)
day	آخرین تماس چه روزی از ماه بوده (مقداری عددی)
month	آخرین تماس چه روزی از ماه بوده (مقدار اسمی: ماه های سال میلادی)
duration	مدت آخرین تماس، بر حسب ثانیه (مقدار عددی)
campaign	تعداد تماس گرفته شده با این مشتری در طول این کمپین (مقدار عددی)
pdays	تعداد روزهایی که پس از آخرین تماس با مشتری از کمپین قبلی گذشته است (مقدار عددی، -1 به این معنی است که مشتری قبلاً با مشتری تماس گرفته نشده است)
previous	تعداد تماس انجام شده قبل از این کمپین و برای این مشتری (مقدار عددی)
poutcome	نتیجه کمپین بازاریابی قبلی (مقدار اسمی: ناشناس، سایر، شکست، موفقیت)
y	آیا مشتری سپرده مدت دار را ثبت کرده است؟ (دودویی: بله، نه)

پروژه عملی

برای انجام تحلیل آماری روی داده ها از نرم افزار پایتون استفاده میکنیم. ابتدا داده ها را در پایتون میخوانیم. پنج سطر اول داده ها را در پایین مشاهده میکنیم. همانطور که میبینیم بعضی از ستون ها که از داده های گسسته هستند مقدار عددی ندارند و باید به آنها مقدار های عددی داد.

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
58.0	management	married	tertiary	no	2143.0	yes	no	unknown	5.0	may	261.0	1.0	-1.0	NaN	unknown	no
44.0	technician	single	secondary	no	NaN	yes	no	unknown	5.0	may	151.0	1.0	-1.0	0.0	unknown	NaN
33.0	entrepreneur	married	secondary	no	2.0	NaN	yes	unknown	5.0	may	76.0	1.0	-1.0	0.0	unknown	no
47.0	blue-collar	married	unknown	NaN	1506.0	yes	no	unknown	5.0	may	92.0	1.0	-1.0	0.0	unknown	no
33.0	unknown	single	unknown	no	1.0	no	no	unknown	5.0	NaN	198.0	1.0	-1.0	NaN	unknown	no

پروژه عملی

همچنین باید بررسی کرد داده ها دارای داده گمشده هستند یا خیر. با توجه به خروجی بدست آمده مشاهده میشود که به طور تقریبی در هر ستون تعداد 10 درصد آنها داده گمشده هستند.

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
NaN	4627	4601	4578	4551	4520	4617	4577	4435	4571	4410	4554	4399	4554	4627	4500	4539	4523

بنابراین ابتدا به داده ها مقدار عددی نسبت میدهیم سپس داده های گمشده را با استفاده از یک روش تکرار شونده برآورد میکنیم

پروژه عملی

عملیات های فوق را با پایتون انجام داده و همانطور که در خروجی های بدست آمده مشاهده میکنیم داده های گسسته به حالت عددی درآمده اند و همه داده های گمشده برآورد شده اند.

خروجی جدول پنج داده اول بعد از برآورد داده های گمشده:

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	58.0	0.0	0.0	0.0	0.0	2143.000000	0.0	0.0	0.0	5.0	0.0	261.0	1.0	-1.0	0.0	0.0	0.0
1	44.0	1.0	1.0	1.0	0.0	1360.860891	0.0	0.0	0.0	5.0	0.0	151.0	1.0	-1.0	0.0	0.0	0.0
2	33.0	2.0	0.0	1.0	0.0	2.000000	0.0	1.0	0.0	5.0	0.0	76.0	1.0	-1.0	0.0	0.0	0.0
3	47.0	3.0	0.0	2.0	0.0	1506.000000	0.0	0.0	0.0	5.0	0.0	92.0	1.0	-1.0	0.0	0.0	0.0
4	33.0	4.0	1.0	2.0	0.0	1.000000	1.0	0.0	0.0	5.0	1.0	198.0	1.0	-1.0	0.0	0.0	0.0

پروژه عملی

خروجی بررسی داده ها از نظر وجود داده گمشده:

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
NaN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

حال این داده ها را به صورت تصادفی به دو دسته آموزش[1] و آزمون[2] تقسیم میکنیم. این داده های آزمون در فرآیند تولید و انتخاب مدل دخالتی ندارند و صرفاً برای ارزیابی نهایی مدل استفاده می شود. بنابراین این داده های آزمون را ذخیره کرده و کنار میگذاریم که در آینده برای مقایسه مدل های مختلف که برازش میدهیم (رگرسیون لجستیک ، درخت تصمیم و...) استفاده کنیم.

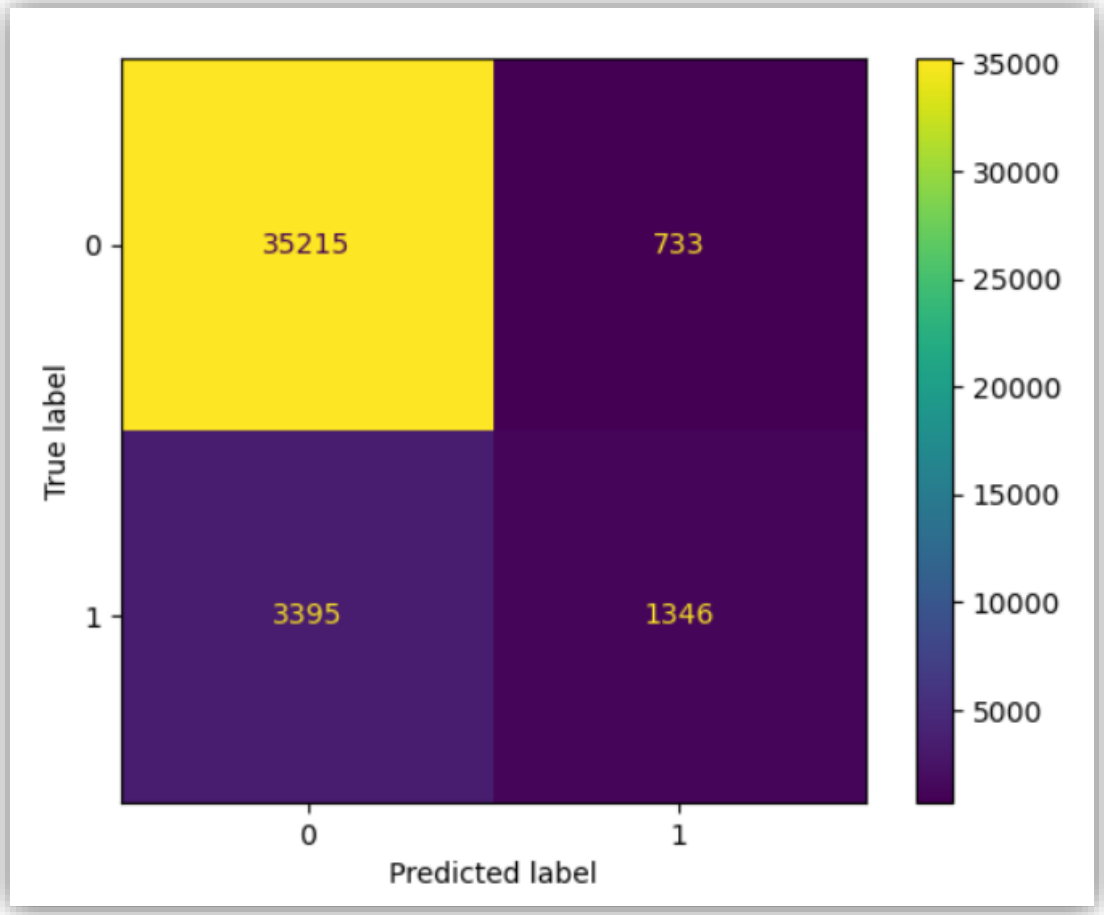
1- Train
2- Test

پروژه عملی

حال با استفاده از داده های آموزش و روش اعتبارسنجی مناسب ترین مدل رگرسیونی را پیدا میکنیم و برازش میدهیم. خروجی ضرایب رگرسیون لجستیک به صورت زیر است:

Feature	coefficients	Feature	coefficients
age	-0.000997	**contact**	0.360612
job	0.016519	**day**	-0.005803
marital	0.108348	**month**	0.085558
education	-0.148838	**duration**	0.003911
default	-0.017843	**campaign**	-0.115572
balance	0.000025	**pdays**	-0.001881
housing	0.655032	**previous**	-0.008566
loan	-0.299464	**poutcome**	0.826489

پروژه عملی



با مدل بدست آمده داده های آموزش را پیشبینی میکنیم و ماتریس درهم ریختگی برای داده های آموزش به صورت مقابل بدست می آید که دقت مدل براساس آن 90 درصد است. حال اگر با همین مدل داده های آزمون را پیشبینی کنیم نشان میدهد که دقت مدل در پیشبینی داده های جدید که در برازش مدل هیچ نقشی نداشته اند به چه صورت است.

پروژه عملی

دقت مدل در پیشبینی داده های جدید هم 89 درصد است که به نظر دقت بالایی است اما ماتریس درهم ریختگی علاوه بر دقت اطلاعات مهم تر دیگری به ما میدهد. مثلاً میتوان دید مدل در پیشبینی دسته 1 یعنی افرادی که سپرده بلند مدت را قبول کرده اند خیلی خوب نبوده. باید توجه داشت که در این مدل حد آستانه 0.5 است یعنی اگر احتمال برآورد شده بالای 0.5 بود داده برای دسته 1 و در غیر این صورت برای دسته 0 است. با تغییر این آستانه یعنی مثلاً وزن بیشتری به یک دسته دادن این خروجی میتواند متفاوت باشد که بسته به هدف مسئله میتوان آن را تغییر داد. مثلاً ممکن است برای بانک بیشتر این مهم باشد که همه افردای که از دسته 1 بوده اند را درست پیشبینی شود یا نه ممکن است بیشتر دقت کلی مدل در پیشبینی هر دو دسته مهم باشد. ولی در این مسئله چون بانک باید هزینه زیادی صرف جذب مشتری میکرد این اهمیت داشت که نسبت افرادی که به درستی به عنوان دسته 1 پیش بینی شده اند به کل افرادی که به عنوان دسته 1 پیش بینی شده اند بالا باشد. در واقع یعنی مهم بود مدل صحت بالایی داشته باشد که مدل ما از این نظر مناسب بوده است.

فایل داده ها و کد تحلیل های انجام شده با پایتون در لینک زیر قرار داده شده:

https://github.com/hesamafshar/Classification/tree/main/Banking%20Dataset%20-%20Marketing%20Targets