بخش ۴:

.۱

رگرسیون لاجیستیک با دو کلاس:

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

$$p(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}$$

برای تخمین پارامترهای $\beta 0, \beta 1, \ldots, \beta p$ از روش بیشینه درست نمایی استفاده می کنیم.

رگرسیون لاجیستیک با پیش از دو کلاس(K تا):

می توان برای هر Kکلاس، K-1 کلاس دیگر را در یک کلاس جدید در نظر گرفت و اینگونه این مسئله به حل K تا رگرسیون لاجیستیک با دو کلاس تبدیل می‌شود.

.۲

در قسمت ۴ داده ها متوازن نیستند. کلاس با برچسب ۱، ۷۵٪ از مشاهدات را تشکیل داده است در حالی که کلاس با برچسب ۰ تنها ۲۵٪ از مشاهدات را. همین مسئله موجب می‌شود تا نتوان به نتیجه‌ی مدل اعتماد کرد.مثلا فرض کنید مدلی طراحی کرده‌ایم که همواره برچسب ۱ را پیش‌بینی کند، دقت این مدل بر روی داده‌های نامتوازن قسمت۴، ۷۵٪ خواهد بود(دقتی بالا برای مدلی کاملا الکی).در صورتی که در قسمت ۵ که داده ها را متوازن کرده‌ایم نتایج قابل اعتماد تری خواهیم داشت.

۳. در قسمت ۱ ما نتیجه‌ی خوبی را بدست آوردیم با در نظر گرفتن همه‌ی فیچرها .در قسمت ۶ با در نظر گرفتن فیچرهایی که ارتباط بیشتری با تارگت دارند هم نتیجه‌ی بهتری گرفته‌ایم و هم به علت کمتر بودن فیچر ها با سرعت بیشتری به نتایج رسیدیم.

.۴

انتخاب زیرمجموعه

انتخاب بهترین زیرمجموعه (Best Subset Selection)

در این روش، کمترین مربعات خطا را بر روی تمام زیرمجموعه‌های ممکن از متغیرها اعمال می‌کنیم. یعنی ابتدا بر روی تمام p مدلی که هر مدل تنها شامل یک متغیر است، سپس بر روی $\binom{p}{2} = \frac{p(p-1)}{2}$ مدلی که هر کدام شامل دو متغیر هستند و ... . در کل در این روش، 2^p مدل را آموزش می‌دهیم. سپس از بین تمام این مدل‌ها، بهترین مدل را انتخاب می‌کنیم.

الگوریتم ۱: انتخاب بهترین زیرمجموعه

۱. مدل تهی M_0 را در نظر بگیرید که شامل هیچ متغیری نیست. این مدل میانگین نمونه را برای هر مشاهده پیش‌بینی می‌کند.
۲. به ازای $k = 1.2. \dots . p$:
   i. تمام $\binom{p}{k}$ مدل با k متغیر را آموزش دهید.
   ii. از بین این $\binom{p}{k}$ مدل، بهترین مدل (مدل با کمترین RSS یا بیشترین R^2) را انتخاب کرده و M_k بنامید.
۳. بهترین مدل از بین $M_0. M_1. \dots . M_p$ را با توجه به خطای ارزیابی (cross-validated error)، $C_p(AIC)$، BIC یا R^2 تغییریافته بیابید.

در مرحله سوم باید انتخاب بهترین مدل با دقت صورت گیرد. زیرا با افزایش تعداد متغیرها، میزان RSS به طور یکنواخت کاهش و میزان R^2 به طور یکنواخت افزایش می‌یابد. علاوه بر این، باید معیار انتخاب بر اساس خطای تست باشد، نه خطای آموزشی. بنابراین در این مرحله باید از معیارهای دیگری از جمله خطای ارزیابی، $C_p(AIC)$، BIC یا R^2 تغییریافته استفاده کرد.

روش پسرو backward stepwise selection

در این روش ابتدا با مدلی شامل تمام متغیرها آغاز می کنیم و در هر مرحله یکی از متغیرها (متغیر با کمترین فایده) را از مدل حذف می‌کنیم. این روند تاحذف تمام متغیرها ادامه می‌یابد.

الگوریتم ۲: روش پسرو

۱. مدل M_p شامل تمام p متغیر را در نظر بگیرید.
۲. به ازای k=p,p-1,…,1:
   i. تمام k مدلی که یک متغیر کمتر از M_k دارند (یعنی $k-1$ متغیر دارند) را در نظر بگیرید.
   ii. از بین آن‌ها بهترین مدل (از نظر معیار RSS یا R^2) را انتخاب کرده و آن را M_{k-1} بنامید.
۳. بهترین مدل از بین $M_0. M_1. \dots . M_p$ را با توجه به خطای ارزیابی (cross-validated error)، $C_p(AIC)$، BIC یا R^2 تغییریافته بیابید

مشابه روش پیشرو، در این روش نیز باید $1 + \frac{p(p-1)}{2}$ حالت را بررسی کنیم. این روش نیز تضمینی برای یافتن بهترین زیرمجموعه از متغیرها ندارد. این روش تنها زمانی قابل استفاده است که $p < n$ و در غیر این صورت (وقتی که p خیلی بزرگتر از n باشد)، تنها روش معتبر برای انتخاب زیرمجموعه‌ای از متغیرها، روش پیشرو است.

۵.

در هر دو روش تضمینی برای یافتن بهترین مجموعه از متغیرهاوجود ندارد.

راه حل: استفاده از روش : روش ترکیبی (Hybrid)

در روش ترکیبی که ترکیبی از دو روش پیشرو و پسرو است، مشابه روش پیشرو متغیرها را به مدل اضافه می‌کنیم، اما پس از اضافه کردن هر متغیر به مدل، ممکن است یکی از متغیرهایی که دیگر حضور آن در مدل فایده‌ای ندارد، از مدل حذف شود. این ترکیب بیشتر به دلیل تقلید رفتار روش بهترین زیرمجموعه ولی با پیچیدگی محاسباتی مشابه روش‌های پیشرو و پسرو است.

۶.

آنالیز افتراقی خطی Linear Discriminant Analysis

دراین روش به جای آنکه مستقیما $\Pr(Y = k|X = x)$ را مدل کنیم، ابتدا احتمال توزیع متغیر ورودی به شرط کلاس را به-دست می آوریم و از روی آن و با استفاده از دانش پیشین، $\Pr(Y = k|X = x)$ را مدل می‌کنیم. چرا گاهی از این روش به جای لاجیستیک استفاده می کنیم:

- وقتی کلاس ها به خوبی از هم قابل تفکیک هستند، تخمین های روش لاجیستیک نااستوار است.
- وقتی تعداد داده ها کم است و توزیع متغیرها در هر کلاس تقریبا نرمال است، روش linear discriminant بهتر و استوارتر از روش لاجیستیک عمل می کند.
- روش linear discriminant برای حالت چندکلاسه معروف تر است.

استفاده از قضیه بیز برای کلاس بندی

فرض کنید که K تا کلاس داریم. فرض کنید π_k توزیع پیشین کلاس kام باشد. یعنی احتمال اینکه یک داده‌ی تصادفی عضو کلاس kام باشد. فرض کنید $f_k(x) = \Pr\,(X = x|Y = k)$ توزیع متغیر ورودی برای داده های کلاس kام باشد. آنگاه طبق قضیه بیز خواهیم داشت:

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \qquad (3.1)$$

تخمین π_k از روی نمونه های آموزشی بسیار ساده است. کافی است محاسبه کنیم که چه کسری از داده ای آموزشی برچسب کلاس kام را دارند. اما تخمین $f_k(x)$ به این سادگی ها نیست؛ مگر اینکه فرض کنیم دارای توزیع ساده ای باشد. به $\Pr(Y = k|X = x)$ احتمال پسین می گوییم. به ازای هر داده ی جدید این احتمال پسین را برای هر کدام از K کلاس بدست می آوریم و نهایتا داده را به کلاسی نسبت می دهیم که احتمال پسین بیشتری داشته باشد. این قانون تصمیم گیری دارای کمترین نرخ خطا است.

حالت تک متغیره

ابتدا فرض کنیم فقط یک متغیر ورودی داریم و فرض کنیم $f_k(x)$ برای تمام کلاس ها دارای توزیع نرمال به فرم زیر است:

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)$$

اگر فرض کنیم که واریانس تمام کلاس ها با هم برابر است یعنی $\sigma_1 = \sigma_2 = \cdots = \sigma_k = \sigma$، با جایگذاری توزیع نرمال فوق در رابطه (3.1) خواهیم داشت:

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

حال با لگاریتم گرفتن و ساده کردن برخی جملات به فرمول زیر می رسیم که به آن discriminant score گویند:

$$\delta_k(x) = \frac{x\mu_k}{\sigma^2} - \frac{{\mu_k}^2}{2\sigma^2} + \log(\pi_k) \qquad \textit{(4.1)}$$

هر داده $X{=}x$ای را به کلاس kامی نسبت می دهیم که دارای مقدار $\delta_k(x)$ بزرگتری باشد.

همیشه مقادیر دقیق پارامترهای میانگین μ_k و واریانس σ را نداریم و باید این مقادیر را از روی نمونه های آموزشی به صورت زیر تخمین بزنیم:

$$\hat{\pi}_k = \frac{n_k}{n}$$

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:\, y_i = k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:\, y_i=k} (x_i - \hat{\mu}_k)^2$$

که n تعداد کل نمونه های آموزشی و n_k تعداد نمونه های آموزشی با برچسب کلاس kام است. با جایگذاری این مقادیر تخمینی در فرمول (4.1) ، داده x را به کلاسی نسبت می دهیم که مقدار $\hat{\delta}_k(x)$ بزرگتری داشته باشد.

$$\hat{\delta}_k(x) = \frac{x\hat{\mu}_k}{\hat{\sigma}^2} - \frac{{\hat{\mu}_k}^2}{2\hat{\sigma}^2} + \log\left(\hat{\pi}_k\right)$$

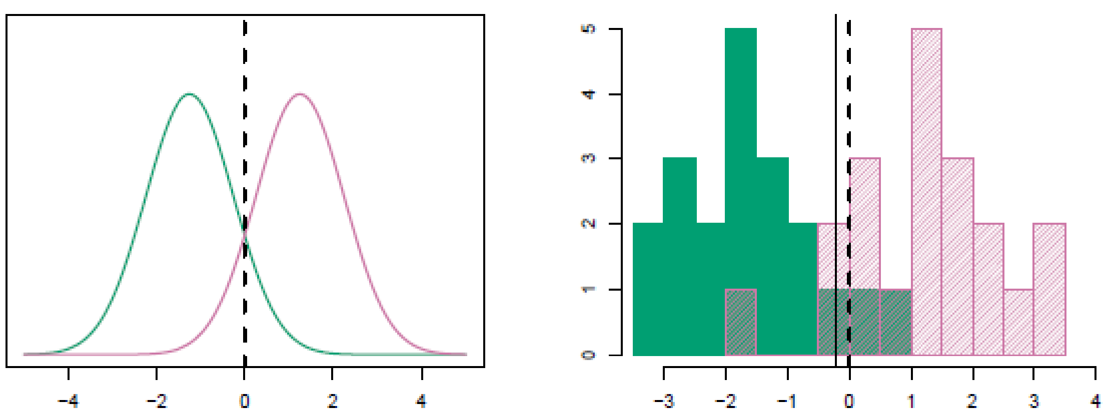
اگرفرض کنیم $K=2$ و $\pi_1 = \pi_2$ باشد، داده را به کلاس ۱ نسبت می دهیم اگر

$$2x(\mu_1 - \mu_2) > {\mu_1}^2 - {\mu_2}^2$$

باشد. بنابراین مرز تصمیم گیری نقطه زیر است:

$$x = \frac{{\mu_1}^2 - {\mu_2}^2}{2x(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}$$

به عنوان مثال داده هایی که در سمت چپ شکل زیر مشاهده می کنید از دو توزیع با واریانش های یکسان و برابر ۱ و میانگین های $\mu_1 = -1.25$ و $\mu_2 = 1.25$ آمده اند. اگر $\pi_1 = \pi_2 = 0.5$ مرز تصمیم، x=0 خواهد بود که با خط نقطه چین در شکل نشان داد هشده است. بنابراین داده های با x<0 به کلاس سبز و داده های با x>0 به کلاس بنفش تعلق خواهند گرفت.



قسمت سمت راست شکل بالا، ۲۰ نمونه آموزشی تولید شده توسط توزیع های احتمالاتی هر کلاس را نشان می دهد. ما مقادیر میانگین و واریانس و توزیع پیشین را برحسب نمونه های آموزشی استخراج کردیم و نهایتا مرز تصمیم گیری به صورت $\frac{\hat{\mu}_1+\hat{\mu}_2}{2}$ را با خط مشکی صاف در نمودار نشان داده ایم که کمی با خط کلاس بند واقعی $\frac{\mu_1+\mu_2}{2}$ که با نقطه چین نشان داده شده است، متفاوت است.

LDA و PCA هر دو روش هایی برای تبدیلات خطی هستند.LDA تحت نظارت است در حالیکه PCA بدون نظارت است.

PCA یک رویکرد کلی برای denoising و کاهش ابعاد است و به اطلاعاتی مانند برچسب کلاس در یادگیری تحت نظارت نیاز ندارد(PCA برچسب های کلاس را نادیده می گیرد). بنابراین می توان از آن در یادگیری بدون نظارت استفاده کرد. از LDA برای ایجاد فضای چند بعدی استفاده می شود. PCA برای فروپاشی فضای چند بعدی استفاده می شود.

LDA در مورد کلاسهای توزیع شده نرمال با کوواریانسهای برابر فرضیات را ارائه می دهد.