یادگیری ماشین

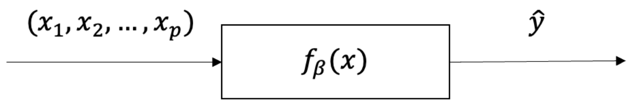
نیم‌سال دوم ۱۴۰۲-۱۴۰۱

مدرس: دکتر سید ابوالفضل مطهری

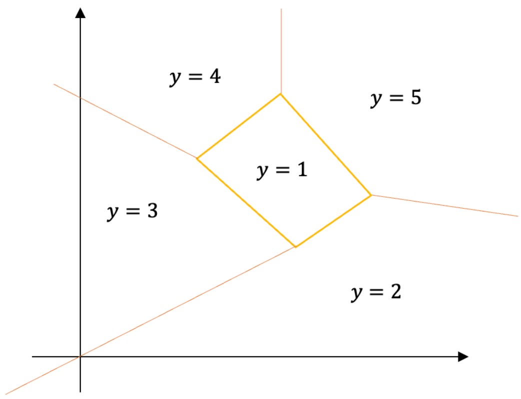
دانشکده‌ی مهندسی کامپیوتر

درس‌نامه ششم

در مسأله دسته‌بندی می‌خواهیم از روی ورودی یکی از k دسته مختلف را تخمین بزنیم.



که در آن $\hat{y} = \{۱, ۲, \ldots, k\}$ می‌باشد. آن چیز که مشخص است، تابع $f_\beta(x)$ فضای ورودی را به k قسمت افراز می‌نماید و هر افراز به یکی از دسته‌ها تخصیص می‌یابد.



همانطور که در مسأله رگرسیون بحث گردید در اینجا هم نیاز به داشتن یک تابع ریسک و یک مدل هستیم تا بتوانیم تعمیم خارج داده را بدست آوریم.

دسته‌بند بیز

فرض نمایید که توزیع مشترک x و y موجود است یعنی $P(x, y)$ را در اختیار داریم. می‌دانیم MAP (Maximum A Posteriori) کمترین خطا را برای تصمیم‌گیری از روی ورودی x در اختیار می‌گذارد.

$$\hat{y} = \arg\max_k \mathbb{P}(y = k|x)$$

برای یک مسأله دو کلاسه کافیست که نسبت احتمالات فوق را با یک مقایسه کنیم:

$$\frac{\mathbb{P}(y = ۱|x)}{\mathbb{P}(y = ۲|x)} \gtrless ۱$$

و اگر از طرفین لگاریتم بگیریم خواهیم داشت:

$$log(\frac{\mathbb{P}(y=1|x)}{\mathbb{P}(y=2|x)}) \gtrless 0$$

دسته‌بند فوق را دسته‌نبد بیز می‌نامیم و این دسته‌بند تابع زیر را کمینه می‌نماید:

$$\mathbb{P}(y \neq \hat{y}) = \mathbb{E}[\mathrm{I}_{y \neq \hat{y}}]$$

از روی رابطه فوق یکی از توابع ریسکی که می‌توانیم از روی داده بدست آوریم به صورت زیر خواهد بود.

$$\mathcal{L}(y, \hat{y}) = \mathrm{I}_{y \neq \hat{y}}$$

(LDA)Linear Discriminant Analysis

یک مسأله k کلاسه را در نظر بگیرید که توزیع مشترک x و y به صورت زیر تعریف شده است:

$$\begin{aligned}
\mathbb{P}(y=1) &= \pi_1; \qquad & \mathbb{P}(x|y=1) &= f_1(x) = \mathcal{N}(\mu_1, \Sigma_1) \\
&\vdots & &\vdots \\
\mathbb{P}(y=k) &= \pi_k; \qquad & \mathbb{P}(x|y=k) &= f_k(x) = \mathcal{N}(\mu_k, \Sigma_k)
\end{aligned}$$

با توجه به دسته‌بند بیز کافیست رابطه زیر را محاسبه نماییم:

$$\begin{aligned}
\hat{y} &= \arg\max_k \mathbb{P}[y=k|x] \\
&= \arg\max_k \mathbb{P}[x|y=k]\mathbb{P}[y=k] \\
&= \arg\max_k \frac{\pi_k e^{-\frac{(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}{2}}}{|2\pi\Sigma_k|^{\frac{1}{2}}} \\
&= \arg\max_k log(\pi_k) - \frac{1}{2} log(|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) \\
&= \arg\max_k \delta_k(x)
\end{aligned}$$

به $\delta_k(x)$ تابع discriminant می‌گوییم.
در حالت اول فرض نماییم که تنها دو کلاس داریم و $\Sigma = \Sigma_1 = \Sigma_2$ در این حالت:

$$\begin{aligned}
\delta_1(x) &= log(\pi_1) - \frac{1}{2} log(|\Sigma|) - \frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1) \\
\delta_2(x) &= log(\pi_2) - \frac{1}{2} log(|\Sigma|) - \frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2)
\end{aligned}$$

مرز بین دو ناحیه را می‌توانیم از روی تساوی قرار دادن دو عبارت فوق بدست آوریم:

$$\begin{aligned}&\delta_1(x) = \delta_2(x)\\ \Rightarrow &log(\frac{\pi_1}{\pi_2}) + (\mu_1 - \mu_2)^T\Sigma^{-1}x - \frac{1}{2}\mu_1\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 = 0\end{aligned}$$

رابطه فوق نشان می‌دهد که ناحیه تصمیم‌گیری یک ناحیه خطی است. در حقیقت که ابرصفحه وجود دارد که بین کلاسهای ۱ و ۲ تفکیک می‌گذارد. این ابرصفحه دارای بردار نرمال

$$w^T = (\mu_1 - \mu_2)^T\Sigma^{-1}$$

می‌باشد و به صورت زیر قابل بیان است:

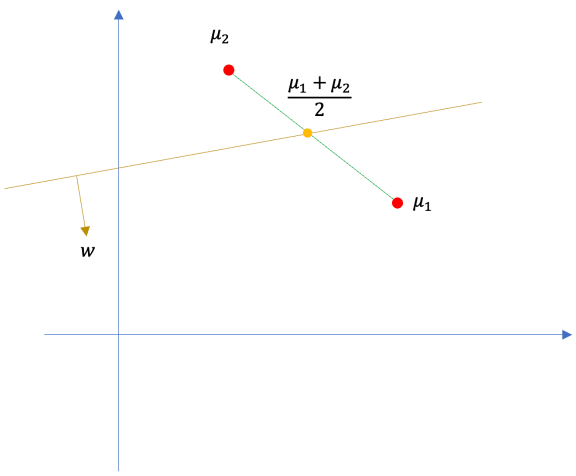
$$w^Tx + b = 0$$

که در آن

$$b = -\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2\Sigma^{-1}\mu_2 + log(\frac{\pi_1}{\pi_2})$$

به طور خاص‌تر اگر $\pi_1 = \pi_2 = \frac{1}{2}$ باشد آنگاه نقطه $\frac{\mu_1+\mu_2}{2}$ بر روی ابرصفحه قرار می‌گیرد. یعنی

$$(\mu_1 - \mu_2)^T\Sigma^{-1}\frac{\mu_1 + \mu_2}{2} = \frac{1}{2}\mu_1\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2$$



اگر $\Sigma = \mathbb{I}$ باشد آنگاه $w = \mu_1 - \mu_2$ خواهد بود و ابرصفحه عمود می‌گردد.
به راحتی می‌توان شرایط فوق را تعمیم داد و برای k دسته دید که نواحی تصمیم‌گیری با یکسری ابرصفحه از یکدیگر جدا می‌گردند.
در یک مسأله یادگیری داده‌ها در دسترس می‌باشند و توزیع را در اختیار نداریم. با توجه به آنکه در مدل فوق توزیعهای هر دسته گوسی مفروض است از روی داده‌ها می‌توانیم پارامترهای توزیع را تخمین بزنیم. در این شرایط

$$\hat{\pi_k} = \frac{N_k}{N}$$

که در آن N_k تعداد مشاهدات از کلاس k می‌باشد.

$$\hat{\mu}_k = \frac{\sum_{y_i=k} x_i}{N_k} = \frac{\sum x_i \mathrm{I}_{y_i=k}}{\sum \mathrm{I}_{y_i=k}}$$

$$\hat{\Sigma} = \frac{\sum_{k=۱}^{k} \sum_{y_i=k} (x_i - \hat{\mu_k})(x_i - \hat{\mu_k})^T}{N-k}$$

مقابله با داده‌های نامتقارن

در پاره‌ای از موارد تعداد نمونه‌های یک کلاس کم می‌باشد و بنابراین دسته‌بند نمی‌تواند به خوبی برای آنها پاسخگو باشد. چند راه حل می‌توان برای این مسأله در نظر گرفت:

۱. **نمونه‌برداری (Resampling)**: در این روش یا به دسته کم از طریق نمونه‌برداری تصادفی از مشاهدات، نمونه‌های تکراری اضافه می‌کنیم و یا اینکه به طور تصادفی از دسته‌های بزرگتر نمونه دور می‌ریزیم. در این روش ممکن است اطلاعات مفید از بین رفته و یا اینکه overfitting رخ دهد.

۲. **تولید نمونه‌های تصادفی**: از روی نمونه‌های با تعداد کم، یک مدل یاد گرفته و داده‌های جدید مصنوعی تولید می‌کنیم. از این روش با نام SMOTE نیز یاد می‌شود که مخفف Synthetic Minority Over-Sampling Technique است.

۳. **روش‌های Ensemble**: از جمله این روش‌ها می‌توان به Bagging و Boosting اشاره کرد که در این درس‌نامه به آن‌ها نمی‌پردازیم.

۴. **یادگیری بر پایه متعادل‌کردن خطا**: خطای کلاسه‌بندی را می‌توان به دو خطای مختلف تفکیک نمود:

$$\mathbb{P}[\hat{y} \neq y] = \mathbb{P}[\hat{y} \neq ۱|y=۱]\mathbb{P}[y=۱] + \mathbb{P}[\hat{y} \neq ۲|y=۲]\mathbb{P}[y=۲]$$

اگر توزیع مشترک y و $\hat{y}$ را در یک جدول نمایش دهیم، داریم:

$\hat{y}$ \ y		π_1	π_2
		1	2
π'_1	1	P_{11}	P_{12}
π'_2	2	P_{21}	P_{22}

در این شرایط می‌توان نوشت:

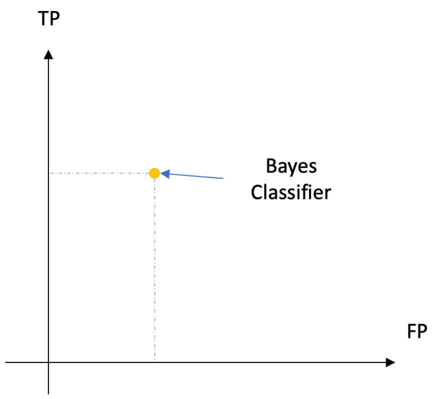
$$\mathbb{P}[\hat{y} \neq ۱|y=۱] = \frac{P_{۲۱}}{P_{۱۱}+P_{۲۱}} = \alpha_۱$$

$$\mathbb{P}[\hat{y} \neq ۲|y=۲] = \frac{P_{۱۲}}{P_{۱۲}+P_{۲۲}} = \alpha_۲$$

اگر $y = ۱$ را با علامت $-$ یعنی بیمارنبودن طرف و $y = ۲$ را با علامت $+$ یعنی بیماربودن طرف نشان دهیم، آنگاه از عبارات زیر استفاده می‌کنیم:

$$\begin{aligned}
&\alpha_۲ : \text{False Negative (FN)} \\
&\alpha_۱ : \text{False Positive (FP)} \\
&۱ - \alpha_۱ : \text{True Negative (TN)} \\
&۱ - \alpha_۲ : \text{True Positive (TP)}
\end{aligned}$$

معمولا تعداد مثبت‌ها در جامعه بیشتر از منفی‌ها می‌باشد و بنابراین تشخیص آن‌ها مهم است. در این شرایط به TP حساسیت (sensitivity) هم می‌گوییم. همچنین به TN که تشخیص گروه بزرگتر است specificity اطلاق می‌گردد. به هر حال داشتن FP و FN بیانگر حساسیت و specificity می‌باشد. می‌توانیم FP و TP را در یک صفحه نمایش دهیم:



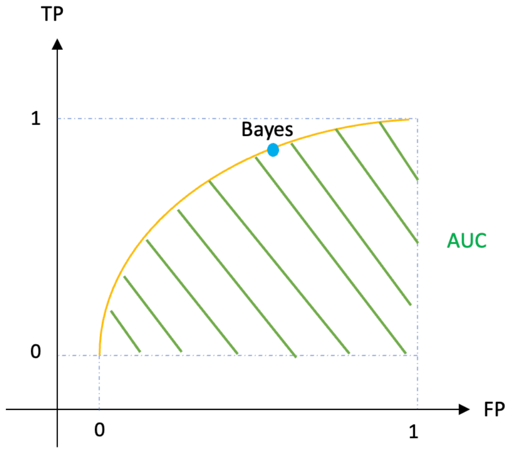
مشاهده کردیم که دسته‌بند Bayes میزان

$$\mathbb{P}[y \neq \hat{y}] = \alpha_۱\pi_۱ + \alpha_۲\pi_۲ = (\text{FP})\pi_۱ + (۱ - \text{TP})\pi_۲$$

را کمینه می‌کند. در عمل دوست داریم $\text{FP} \to ۰$ و $\text{TP} \to ۱$ میل کند. قضیه Neyman-Pearson بیان می‌دارد که برای مصالحه بین $(\alpha_۲, \alpha_۱)$ و یا (FP, TP) کافیست مقدار نسبت احتمالات را با یک آستانه مقایسه نماییم. یعنی:

$$\log \frac{\mathbb{P}(y = ۱|x)}{\mathbb{P}(y = ۲|x)} \gtrless \text{Th}$$

و به ازای آستانه‌های مختلف TP و FPـ های مختلف می‌گیریم. در نتیجه شکل نمودار به صورت زیر خواهد بود:



بنابراین با تغییر در آستانه می‌توانیم منحنی فوق را که ROC (Receiver Operating Characteristics) نامیده می‌شود به دست آوریم. مساحت زیر ROC را AUC (Area Under the Curve) می‌نامیم. در حالت ایده‌آل $\text{AUC} = ۱$ می‌باشد. یعنی در $\text{FP} = ۰$ و $\text{TP} = ۱$ قابل حصول است. همه مطالب فوق را برای شرایطی مطرح نمودیم که توزیع در اختیار است. علی‌القاعده برای هر سیستم دسته‌بندی می‌توانیم از روی داده‌های آزمون ماتریس confusion را بدست آوریم که در حقیقت همان ماتریس توزیع مشترک $(y, \hat{y})$ است. با این تفاوت که به جای فرکانس، تعداد نشان داده می‌شود. از روی این ماتریس می‌توانیم همه پارامترها را تخمین بزنیم. همچنین با تغییر آستانه در LDA می‌توانیم ROC و AUC را نیز بدست آوریم و از کیفیت کار مطلع شویم.

Quadratic Discriminant Analysis (QDA)

روابط LDA را با فرض $\Sigma_۱ = \Sigma_۲ = \cdots = \Sigma_k$ بدست آوردیم. حال اگر این فرض را کنار بگذاریم، آنگاه تابع discriminant به شکل زیر می‌شود.

$$\sigma_k(x) = -\frac{۱}{۲}(x - \mu_k)^T \Sigma_k^{-۱}(x - \mu_k) - \frac{۱}{۲}\log|\Sigma_k| + \log \pi_k$$

در این شرایط نواحی تصمیم‌گیری با توابع درجه دوم از یکدیگر جدا می‌گردند.