یادگیری ماشین

نیم‌سال اول ۱۴۰۳-۱۴۰۲

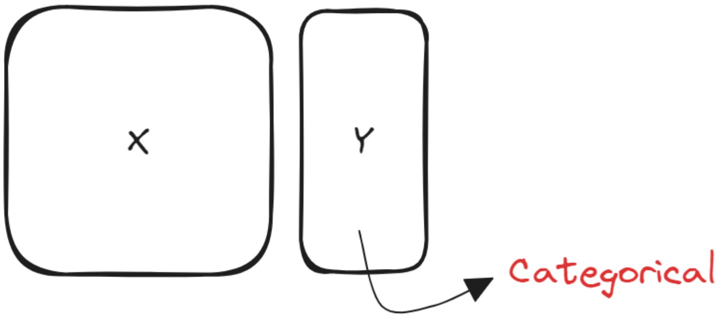
مدرس: دکتر سید ابوالفضل مطهری

دانشکده‌ی مهندسی کامپیوتر

درس‌نامه دوم

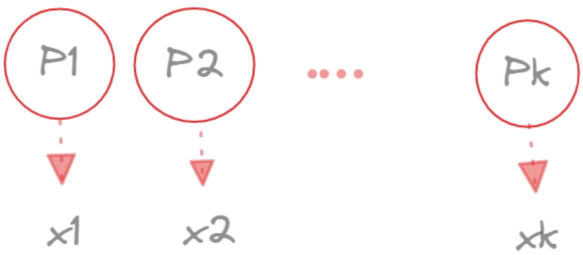
دسته‌بندی (Classification)

در اینجا، داده (که به شکل ماتریس X نشان داده می شود) و وکتور نهایی (Y) را داریم. این بار پاسخ ها به صورت Categorical هستند.



شکل ۱: ماتریس ورودی و خروجی

یکی از راه هایی که برای حل این نوع مسائل داریم، محاسبه‌کردن $P(X)$ اول تا k ام و در نظر گرفتن بزرگ‌ترین آن هاست. به این روش **Maximum Likelihood** می‌گوییم.



شکل ۲: احتمال هر یک از دسته ها

$$h : X \rightarrow \{۱, ۲, ..., k\}$$

$$P_۱ \rightarrow x \xrightarrow{h(x)} P_۲$$

$$E = \mathbb{1}\{Y \neq h(X)\}$$

این فرمول نشان دهنده مقدار ارور ما می باشد که به ازای هر اشتباه ۱ واحد به ارور اضافه می شود. (در اینجا Y و X هر دو متغیر تصادفی هستند.)

$$P(Y, X) = P(X|Y)P(Y)$$

$$P_۱(X) = P(X|Y = ۱)$$

$$P_۲(X) = P(X|Y = ۲)$$

هدف این است که تابع تخمینی را پیدا کنیم که امید ریاضی خطا را کمینه کند.

$$\min_h \ \mathbb{E}[E]$$

که در آن

$$h : X \to \{۱, ۲, ..., k\}$$

تابع فرضیه است.

MAP - Maximum A Posteriori

در آمار بیزی، MAP به معنای حداکثر احتمال پسین است. این تکنیک راهی برای تخمین پارامترهای یک توزیع احتمال پسین است که با استفاده از داده‌های مشاهده شده و توزیع احتمال پیشین، به دست می‌آید.

۱. توزیع احتمال پسین (Posterior Distribution)

$$P(\theta|D) \propto P(D|\theta) \cdot P(\theta)$$

۲. Map Estimation

$$\theta_{MAP} = \arg\max_{\theta} P(\theta|D)$$

نمادهای فوق، معانی زیر را می‌رسانند.

- $P(\theta|D)$: توزیع احتمال پسین پارامترها (θ) با توجه به داده‌های مشاهده شده (D)
- $P(D|\theta)$: احتمال داده‌ها به شرط پارامترها.
- $P(\theta)$: توزیع احتمال پیشین پارامترها.

از MAP می توان برای به دست آوردن یک تخمین نقطه ای از یک کمیت مشاهده نشده بر اساس داده‌های تجربی استفاده کرد.

$$\operatorname{argmax} P(Y|X) = \max \; P(Y=i|X) \;\; if \; i \in \{۱, ۲, ..., k\}$$
$$\operatorname{argmax} P(Y|X) = \frac{P(X|Y=i)P(Y=i)}{P(X)}$$

و چون argmax را روی i می گیریم پس باید صورت کسر را بیشینه کنیم.

$$\operatorname{argmax} P(X|Y=i)P(Y=i)$$

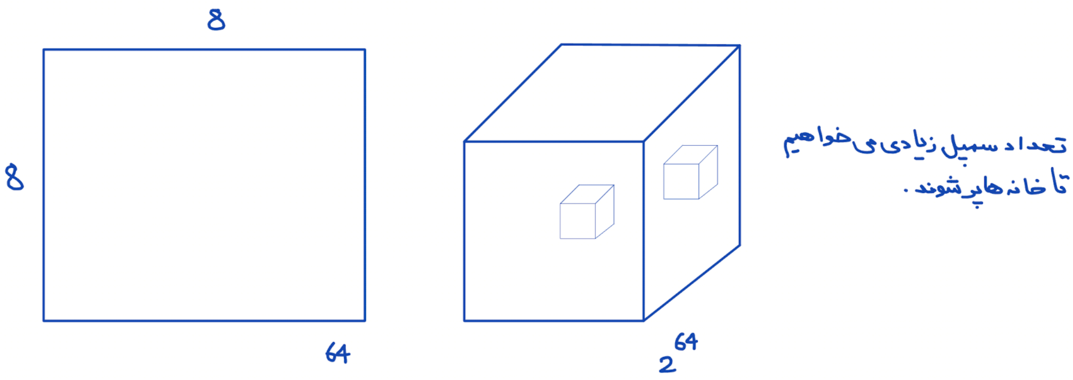
به این روش دسته‌بندی، روش بیز گفته می‌شود و به هر دسته‌بند که به این روش عمل کند، Bayes Classifier می‌گوییم.
اگر P_i ها را نداشته باشیم، می‌توانیم:
۱) از نمونه ها استفاده کرده و P_i ها را تخمین بزنیم. $\{\mathrm{D} = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$
به این روش **Density Estimation** می گوییم.
۲)

$$\begin{pmatrix} h_۱(x) \\ h_۲(x) \\ . \\ . \\ . \\ h_k(x) \end{pmatrix} \left(\xrightarrow{Softmax} \right) \begin{pmatrix} \frac{e_۱^h(x)}{z} \\ \frac{e_۲^h(x)}{z} \\ . \\ . \\ . \\ \frac{e_k^h(x)}{z} \end{pmatrix}$$

$$z = \Sigma e_i^h(x)$$

در Density Estimation اولین کاری که باید انجام دهیم، بدست‌آوردن هیستوگرام از نمونه‌های موجود است. توزیع نمونه‌ها در هر قسمت به صورت Binomial می باشد.

در این جا با چالش هایی رو به رو هستیم که یکی از آن ها نفرین ابعاد یا Curse of Dimensionality می‌باشد.



نفرین ابعاد یا Curse of Dimensionality یک مفهوم در زمینه‌ی آمار و یادگیری ماشین است و به چالش‌ها و مشکلاتی اشاره دارد که در فضاهای با ابعاد بالا (تعداد زیادی از ویژگی‌ها یا متغیرها) به وجود می‌آید. برخی از چالش‌ها و مسائل مرتبط با Curse of Dimensionality عبارتند از:

- تنوع ویژگی‌ها: با افزایش ابعاد فضا، تنوع (تفاوت) بین نقاط داده‌ها نسبت به همدیگر افزایش می‌یابد. این امر ممکن است منجر به افزایش فاصله‌های اقلیدسی بین نقاط شود.
- نیاز به داده‌های بیشتر: با افزایش ابعاد، نیاز به تعداد داده‌های آموزشی بیشتر برای کسب یک مدل قابل اعتماد نیز افزایش می‌یابد. این امر می‌تواند منجر به مشکل کمبود داده در بعضی مسائل شود.
- داده‌های Sparse: در فضاهای با ابعاد بالا، داده‌ها اغلب به صورت sparse می‌شوند؛ به عبارت دیگر، اکثر نقاط در فضا خالی هستند. این موضوع باعث افزایش پیچیدگی در مدل‌سازی و تحلیل داده می‌شود.
- پیداکردن الگوها و روابط معنادار: افزایش ابعاد ممکن است باعث کاهش قدرت تفسیر و تحلیل الگوها و روابط معنادار در داده‌ها شود.

برای مقابله با Curse of Dimensionality، رویکردهایی نظیر انتخاب ویژگی، تجمیع داده، و الگوریتم‌های خاص مخصوص فضاهای با ابعاد بالا مورد استفاده قرار می‌گیرد. همچنین، استفاده از روش‌های کاهش ابعاد مانند تحلیل مؤلفه‌های اصلی (PCA) و تحلیل تفسیری مؤلفه‌ها (ICA) نیز به عنوان راه‌حل‌های معمول به شمار می‌آید.

Naive Bayes

$$P(x_1, x_2, ..., x_p)$$

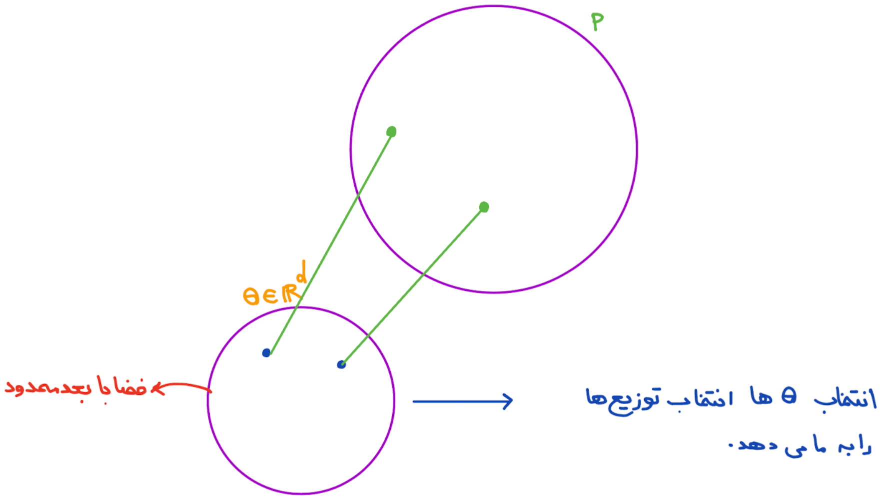
$$\Downarrow$$

$$P(x_1)P(x_2)\ ...\ P(x_p)$$

$$\Downarrow$$

$$P(x_1)P(x_2|x_1)P(x_3|x_2, x_1)\ ...\ P(x_p|x_p - 1, ..., x_1)$$

$$P \begin{cases} \text{Parametric} \\ \text{Non-parametric} \end{cases}$$



اگر دو θ یک توزیع را به ما بدهد، می گوییم این توزیع نسبت به θ ها Identifiable نیست.

Parametric

$$X \sim P_\theta \xrightarrow{Density} \lambda e^{-\lambda x}$$

$$X \sim P_\theta \xrightarrow{Dist} ۱ - e^{-\lambda x}$$

Non Parametric

$$\text{Density } f(x) \quad , \quad \int f''(x)^2 \, dx \leqslant Threshold$$

LDA: Linear Discriminant Analysis

فرض کنیم دو توزیع توام گاوسی با احتمال پیشین π داریم. یعنی:

$$\pi_1, \mathcal{N}(\mu_1, \Sigma) \longrightarrow \theta_1 = (\hat{\pi}_1, \hat{\mu}_1, \hat{\Sigma})$$

$$\pi_2, \mathcal{N}(\mu_2, \Sigma) \longrightarrow \theta_2 = (\hat{\pi}_2, \hat{\mu}_2, \hat{\Sigma})$$

قرار است θ_1 و θ_2 را تخمین بزنیم. می‌توانیم از Maximum Likelihood استفاده کنیم اما با مشکلاتی مواجه می‌شویم.. مشکل اصلی ما این است که Σ، از دو توزیع مختلف آمده است و ۲ تخمین متفاوت برای آن بدست می‌آید و همچنین ابعداد آن به صورت quadratic زیاد می‌شود. بنابراین به دنبال راه حل دیگری می‌رویم.

راه حل

فرض کنیم داده ها به صورت $x \in X \subseteq R^d$ از دو توزیع گاوسی آمده‌اند.

$$f_x = (x|y=1) = \mathcal{N}(\mu_1, \Sigma_1)$$

$$f_x = (x|y=0) = \mathcal{N}(\mu_2, \Sigma_2)$$

$$p(y=1) = \pi_1$$

و هدف ما این است که جهت نامساوی را در عبارت زیر بدست آوریم:

$$p(y=0|x) \quad \boxed{؟} \quad p(y=1|x)$$

سعی می‌کنیم نامساوی بالا را حل کنیم.

$$g(x) = \frac{p(y=1|x)}{p(y=0|x)} \lesseqslant 1$$

$$\longrightarrow \frac{\frac{p(x|y=1)p(y=1)}{p(x)}}{\frac{p(x|y=0)p(y=0)}{p(x)}} \lesseqslant 1$$

$$\longrightarrow \frac{\pi_1 \frac{1}{\sqrt{(2\pi)^d|\Sigma_1|}} \exp\left(-\frac{1}{2}((x-\mu_1)^T {\Sigma_1}^{-1}(x-\mu_1))\right)}{\pi_0 \frac{1}{\sqrt{(2\pi)^d|\Sigma_0|}} \exp\left(-\frac{1}{2}((x-\mu_0)^T {\Sigma_0}^{-1}(x-\mu_0))\right)} \lesseqslant 1$$

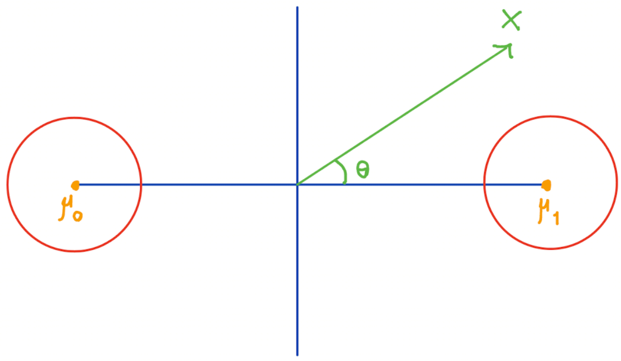
از دو طرف لگاریتم گرفته و به صورت زیر رابطه را بازنویسی می‌کنیم.

$$g(x) = \log(\frac{\pi_1}{\pi_0}) + \frac{1}{2}\log(\frac{\Sigma_0}{\Sigma_1}) + \frac{1}{2}((x-\mu_0)^T {\Sigma_0}^{-1}(x-\mu_0)) - \frac{1}{2}((x-\mu_1)^T {\Sigma_1}^{-1}(x-\mu_1)) \lesseqslant 0$$

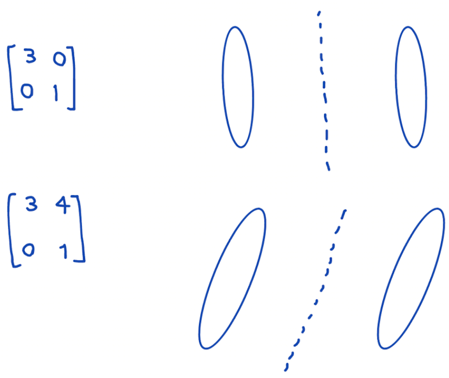
تابع g(x) یک تابع درجه دو است. Σ_0 و Σ_1 را یکی درنظر می‌گیریم تا تابع به شکل خطی بدست آید.

$$g(x) = \log(\frac{\pi_1}{\pi_0}) + \left(x - \frac{\mu_1 - \mu_0}{2}\right)^T \Sigma^{-1}(\mu_1 - \mu_0)$$

برای مثال اگر $\pi_1 = \frac{1}{2}$ و $\Sigma = I_d$



در شکل بالا مکان خط جدا کننده به مقدار π ربط دارد مثلا اگر $\pi_1 = \frac{1}{3}$ باشد خط به گروه ۱ نزدیکتر می‌شود. همچنین شکل توزیع‌ها، جهت خط و زاویه‌دار بودن آن نیز به میزان Σ بستگی دارد و با تغییر آن دوران خواهیم داشت. به عنوان مثال اگر $\Sigma = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$ باشد، توزیع ها به شکل بیضی درمی‌آیند. و اگر $\Sigma = \begin{pmatrix} 3 & 4 \\ 0 & 1 \end{pmatrix}$ باشد شکل دسته‌بند ما به صورت زیر خواهد بود.



حالا فرض کنیم توزیع دو کلاس گاوسی با Σ برابر هستند. پارامتر های زیر را می‌خواهیم با استفاده از MLE تخمین بزنیم.

$$\hat{\mu_0} = \frac{1}{n_0} \sum_{y_i=0} x_i$$

$$\hat{\mu_1} = \frac{1}{n_1} \sum_{y_i=1} x_i$$

$$\hat{\Sigma} = \frac{۱}{n - ۲} \sum_{i=۰,۱} \sum_{j} (x_j{}^i - \hat{\mu_i})(x_j{}^i - \hat{\mu_i})^T$$

$$\hat{\pi} = \frac{n_۱}{n_۰}$$

اینجا می‌توان Σها را یکی در نظر نگرفت و به صورت

$$\Sigma_۱ = \Sigma + y_۱ I \, , \Sigma_۰ = \Sigma + y_۰ I$$

نوشت. در این صورت دیگر تابع خطی نمی‌شود ولی با این وجود، به سختی درجه دو هم نمی‌باشد.