یادگیری ماشین — دکتر منصور رزقی

تمرین دوم — دانشکده علوم ریاضی ، گروه علوم کامپیوتر، گرایش داده‌کاوی

سبحان عیوضی پور کفشدوزی — ۴۰۴۵۲۵۲۱۰۰۴

تمرین ۱:

سوال ۲.۳ در تمرین ۱ سری اول تکرار شده است. بنابراین، ۴ سوال دیگر را مورد بررسی قرار می‌دهیم.

سوال ۲.۶:

می‌دانیم:

$$P(H|e_1, e_2) = \frac{P(H)P(e_1, e_2|H)}{p(e_1, e_2)} \propto P(H)P(e_1, e_2|H) \quad (۱)$$

همچنین، می‌دانیم:

$$P(e_1, e_2) = \Sigma_h P(h)P(e_1|h)P(e_2|h) \quad (۲)$$

- اگر بدانیم $E_1 \perp E_2|H$ برقرار است، می‌توانیم با داشتن $P(e_1|H)$ و $P(e_2|H)$، مقدار $P(e_1, e_2|H)$ را محاسبه کنیم.

$$P(e_1, e_2|H) = P(e_1|H)P(e_2|H) \quad (۳)$$

- اگر هیچ فرضی درباره استقلال شرطی نداشته باشیم، نمی‌توانیم از فرمول (۳) استفاده کنیم.

- (الف) در این بخش، هیچ اطلاعاتی درباره استقلال شرطی نداریم. گزینه‌ها را بررسی می‌کنیم:

- (i): اگرچه $P(e_1|H)$ و $P(e_2|H)$ را در اختیار داریم، اما چون هیچ فرضی درباره استقلال شرطی نداریم، $P(e_1, e_2|H)$ را نمی‌توانیم تعیین کنیم. پس داشتن مقادیر ارائه‌شده کافی نیست.
- (ii): همه مقادیر موجود در تساوی (۱) را در اختیار داریم و این برای محاسبه کافی است.
- (iii): مشابه حالت (i)، چون هیچ فرضی درباره استقلال شرطی نداریم، عبارت $P(e_1, e_2|H)$ به شکل مستقیم، قابل محاسبه نیست. همچنین، چون $P(e_1, e_2)$ را نداریم، نمی‌توانیم نرمالایز کنیم. پس مقادیر ارائه شده، کافی نیستند.

نتیجه : **فقط مجموعه مقادیر بخش (ii)، برای محاسبه کافی می‌باشد.**

- (ب) در این بخش، فرض استقلال شرطی برقرار است. به بررسی گزینه‌ها می‌پردازیم:

- (i): چون فرض استقلال شرطی برقرار است، می‌توانیم از فرمول (۳) استفاده کنیم و $P(e_1, e_2|H)$ را با ضرب $P(e_1|H)$ و $P(e_2|H)$ محاسبه کنیم. پس مقادیر ارائه‌شده برای محاسبه (۱) کافی هستند.
- (ii): مشابه بخش (الف)، همه مقادیر موجود در تساوی (۱) را در اختیار داریم و این برای محاسبه کافی است.

- (iii): مشابه بخش (i)، چون فرض استقلال شرطی برقرار است، می‌توانیم از فرمول (۳) استفاده کنیم و $P(e_1, e_2|H)$ را با ضرب $P(e_1|H)$ و $P(e_2|H)$ محاسبه کنیم. همچنین، جهت محاسبه $P(e_1, e_2)$، می‌توانیم از مقادیر موجود استفاده کرده و فرمول (۲) را به کار ببریم. پس مقادیر ارائه شده در این بخش نیز کافی هستند.

نتیجه: **مجموعه مقادیر تمامی بخش‌های (i)، (ii) و (iii) برای محاسبه کافی هستند.**

سوال ۲.۷:

فرض می‌کنیم، X_1 و X_2، متغیرهای تصادفی سکه‌های عادلانه مستقل هستند.

$$P(X_1 = 0) = P(X_1 = 1) = \frac{1}{2}$$

$$P(X_2 = 0) = P(X_2 = 1) = \frac{1}{2}$$

$$P(X_1 = a, X_2 = b) = P(X_1 = a)P(X_2 = b) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}, \quad a, b \in \{0, 1\}$$

تعریف می‌کنیم: $X_3 = X_1 \oplus X_2$. بنابراین:

$$X_3 = \begin{cases} 0 & \text{اگر } X_1 = X_2 \text{، } \\ 1 & \text{اگر } X_1 \neq X_2 \text{، } \end{cases}$$

طبق تعریف، X_1 و X_2، مستقل هستند. مستقل بودن X_1 و X_3 را بررسی می‌کنیم.
باید نشان دهیم این تساوی برقرار است:

$$P(X_1 = a, X_3 = c) = P(X_1 = a)P(X_3 = c) \quad , (a, c \in \{0, 1\})$$

ابتدا $P(X_3 = 0)$ را محاسبه می‌کنیم:

$$P(X_3 = 0) = P(X_1 = 0, X_2 = 0) + P(X_1 = 1, X_2 = 1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

مشابهاً $P(X_3 = 1) = \frac{1}{2}$ برقرار است. عبارت $P(X_1 = 0, X_3 = 0)$ را بررسی می‌کنیم. اگر $X_1 = 0$ و $X_3 = 0$ باشد، $X_2 = 0$ دارای احتمال $\frac{1}{4}$ است. همچنین:

$$P(X_1 = 0)P(X_3 = 0) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

عبارت $P(X_1 = 0, X_3 = 1)$ را بررسی می‌کنیم. اگر $X_1 = 0$ و $X_3 = 1$ باشد، $X_2 = 1$ دارای احتمال $\frac{1}{4}$ است. همچنین:

$$P(X_1 = 0)P(X_3 = 1) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

به همین ترتیب، $P(X_1 = 1)P(X_3 = 0) = \frac{1}{4}$ و $P(X_1 = 1)P(X_3 = 1) = \frac{1}{4}$ برقرار است و در نتیجه، X_1 و X_3 مستقل هستند. به طریق مشابه، می‌توان نشان داد که X_2 و X_3 هستند. در ادامه، بررسی می‌کنیم که آیا استقلال متقابل برقرار است یا خیر. یعنی برقراری این تساوی را چک می‌کنیم:

$$P(X_1 = 1, X_2 = 1, X_3 = 1) = P(X_1 = 1)P(X_2 = 1)P(X_3 = 1)$$

سمت راست، محاسبه می‌شود:

$$P(X_1 = 1)P(X_2 = 1)P(X_3 = 1) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$$

وقتی $X_1 = 1$و $X_2 = 1$است، طبق تعریف XOR، داریم: $X_3 = 0$. پس امکان ندارد به صورت همزمان $X_1 = 1$، $X_2 = 1$و $X_3 = 1$ رخ دهد. در نتیجه:

$$P(X_1 = 1, X_2 = 1, X_3 = 1) = 0$$

بنابراین:

$$P(X_1 = 1, X_2 = 1, X_3 = 1) \neq P(X_1 = 1)P(X_2 = 1)P(X_3 = 1)$$

مشاهده می‌شود که علیرغم برقراری استقلال بین هرجفت از متغیرها، بین سه متغیر، استقلال متقابل برقرار نیست.

سوال ۲.۱۲:

فرض کنید $D = \{x_1, \ldots, x_N\}$ داده‌های مشاهده‌شده باشند که مستقل و یکنواخت توزیع شده‌اند. همچنین درباره توزیع تجربی می‌دانیم:

$$\hat{p}_{emp}(x) = \frac{1}{N}\Sigma_{i=1}^{N}\, \mathbf{I}[x_i = x]$$

همگرایی KL بین توزیع تجربی و مدل فرضی $q(x|\theta)$ را می‌نویسیم:

$$KL(\hat{p}_{emp}||q) = \Sigma_x \hat{p}_{emp}(x) log\ \hat{p}_{emp}(x) - \Sigma_x \hat{p}_{emp}(x) log\ q_\theta(x)$$

بخش اول عبارت حاصل شده، به θ وابسته نیست؛ اما بخش دوم به θ وابسته است. بنابراین مینیمم‌سازی همگرایی KL، معادل مینیمم‌سازی بخش دوم است؛ یعنی:

$$\begin{aligned} min_\theta\ KL(\hat{p}_{emp}||q) &\iff min_\theta[-\Sigma_x \hat{p}_{emp}(x) log\ q_\theta(x)] \\ &\iff max_\theta[\Sigma_x \hat{p}_{emp}(x) log\ q_\theta(x)] \end{aligned}$$

عبارت داخل کروشه را ساده می‌کنیم:

$$\Sigma_x \hat{p}_{emp}(x) log\ q_\theta(x) = \Sigma_x (\frac{1}{N}\Sigma_{i=1}^{N} \mathbf{I}[x_i = x]) log\ q_\theta(x)$$

$$= \frac{1}{N}\Sigma_{i=1}^{N}\Sigma_x\, \mathbf{I}[x_i = x] log\ p_\theta(x) = \frac{1}{N}\Sigma_{i=1}^{N}\, log\ p_\theta(x_i)$$

مینیمم‌سازی همگرایی KL را ادامه می‌دهیم:

$$\begin{aligned} min_\theta\ KL(\hat{p}_{emp}||q) &\iff max_\theta[\Sigma_x \hat{p}_{emp}(x) log\ q_\theta(x)] \\ &\iff max_\theta\ \frac{1}{N}\Sigma_{i=1}^{N}\, log\ p_\theta(x_i) \\ &\iff max_\theta\ \Sigma_{i=1}^{N}\, log\ p_\theta(x_i) \\ &\iff max_\theta \prod_{i=1}^{N} p_\theta(x_i) \end{aligned}$$

عبارت حاصل‌شده، معادل تابع درست‌نمایی بیشینه است. اگر فرض کنیم، مقدار پارامتر بیشینه‌کننده تابع درست نمایی، برابر $\hat{\theta}$ است، در نتیجه، می‌توان نوشت:

$$argmin_\theta\ KL(\hat{p}_{emp}||q) = argmax_\theta \prod_{i=1}^{N} p_\theta(x_i) = \hat{\theta}$$

بدین ترتیب، حکم ثابت گردید. □

سوال ۲.۱۵:

فرمول $mutual\ \ information$ را می‌نویسیم:

$$I(X,Y) = \Sigma_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$$

ابتدا نشان می‌دهیم:

$$I(X,Y) = H(X) - H(X|Y)$$

طبق فرمول احتمال توأم:

$$p(x,y) = p(x|y)p(y)$$

جای‌گذاری می‌کنیم:

$$I(X,Y) = \Sigma_{x,y} p(x,y) log \frac{p(x|y)p(y)}{p(x)p(y)} = \Sigma_{x,y} p(x,y) log \frac{p(x|y)}{p(x)}$$

$$= \Sigma_{x,y} p(x,y) log\, p(x|y) - \Sigma_{x,y} p(x,y) log\, p(x)$$

ابتدا عبارت $\Sigma_{x,y} p(x,y) log\, p(x|y)$ را حساب می‌کنیم:

$$\Sigma_{x,y} p(x,y) log\, p(x|y) = \Sigma_y p(y) \Sigma_x p(x|y) log\, p(x|y) = -H(X|Y)$$

سپس عبارت $\Sigma_{x,y} p(x,y) log\, p(x)$ را حساب می‌کنیم:

$$\Sigma_{x,y} p(x,y) log\, p(x) = \Sigma_x log\, p(x) \Sigma_y p(x,y) = \Sigma_x p(x) log\, p(x) = -H(X)$$

مقادیر حاصل‌شده را جای‌گذاری می‌کنیم:

$$I(X,Y) = \Sigma_{x,y} p(x,y) log\, p(x|y) - \Sigma_{x,y} p(x,y) log\, p(x) = H(X) - H(X|Y)$$

در ادامه، به شکل مشابه، نشان می‌دهیم:

$$I(X,Y) = H(Y) - H(Y|X)$$

طبق فرمول احتمال توأم:

$$p(x,y) = p(y|x)p(x)$$

جای‌گذاری می‌کنیم:

$$I(X,Y) = \Sigma_{x,y} p(x,y) log \frac{p(y|x)p(x)}{p(x)p(y)} = \Sigma_{x,y} p(x,y) log \frac{p(y|x)}{p(y)}$$

$$= \Sigma_{x,y} p(x,y) log\, p(y|x) - \Sigma_{x,y} p(x,y) log\, p(y)$$

ابتدا عبارت $\Sigma_{x,y} p(x,y) log\, p(y|x)$ را حساب می‌کنیم:

$$\Sigma_{x,y} p(x,y) log\, p(y|x) = \Sigma_x p(x) \Sigma_y p(y|x) log\, p(y|x) = -H(Y|X)$$

سپس عبارت $\Sigma_{x,y} p(x,y) log\, p(y)$ را حساب می‌کنیم:

$$\Sigma_{x,y} p(x,y) log\, p(y) = \Sigma_y log\, p(y) \Sigma_x p(x,y) = \Sigma_y p(y) log\, p(y) = -H(Y)$$

مقادیر حاصل‌شده را جای‌گذاری می‌کنیم:

$$I(X,Y) = \Sigma_{x,y} p(x,y) log\, p(y|x) - \Sigma_{x,y} p(x,y) log\, p(y) = H(Y) - H(Y|X)$$

در نتیجه:

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

بدین ترتیب، حکم ثابت گردید. □

تمرین ۲:

- (الف): تابع سیگموید را در نظر می‌گیریم:

$$\sigma(a) = \frac{1}{1+e^{-a}} = (1+e^{-a})^{-1}$$

چون:

$$\frac{d}{da}(1+e^{-a}) = -e^{-a}$$

بنابراین، مشتق تابع سیگموید برابر است با:

$$\frac{d\sigma(a)}{da} = -1(1+e^{-a})^{-2}.(-e^{-a}) = \frac{e^{-a}}{(1+e^{-a})^2}$$

عبارت حاصل‌شده را بر حسب $\sigma(a)$می‌نویسیم؛ چون:

$$\sigma(a) = \frac{1}{1+e^{-a}}\ , \qquad 1-\sigma(a) = 1 - \frac{1}{1+e^{-a}} = \frac{1+e^{-a}-1}{1+e^{-a}} = \frac{e^{-a}}{1+e^{-a}}$$

بنابراین:

$$\frac{d\sigma(a)}{da} = \frac{e^{-a}}{(1+e^{-a})^2} = \frac{1}{1+e^{-a}}.\frac{e^{-a}}{1+e^{-a}} = \sigma(a)(1-\sigma(a))$$

به نتیجه مطلوب دست یافتیم. □

- (ب): برای رگرسیون لجستیک، این تساوی برقرار است:

$$p(y=1|x,w) = \sigma(w^T x)$$

قرار می‌دهیم: $a_n = w^T x_n$. مقدار $log-likelihood$ برای داده (x_n, y_n) برابر است با:

$$l(w) = \Sigma_{n=1}^{N}[y_n log\sigma(a_n) + (1-y_n)log(1-\sigma(a_n))]$$

قرار می‌دهیم:

$$l_n = y_n log\sigma(a_n) + (1-y_n)log(1-\sigma(a_n))$$

بنابراین:

$$l(w) = \Sigma_{n=1}^{N} l_n$$

می‌خواهیم $\nabla_w l(w)$ را حساب کنیم. طبق $chain\ rule$ می‌توان نوشت:

$$\nabla_w l(w) = \frac{\partial l}{\partial a}.\frac{\partial a}{\partial w}$$

ابتدا $\frac{\partial l}{\partial a}$ را حساب می‌کنیم. تک نقطه l_n را در نظر بگیرید. با مشتق‌گیری خواهیم داشت:

$$\frac{\partial l_n}{\partial a_n} = y_n \frac{1}{\sigma(a_n)}\sigma'(a_n) - (1-y_n)\frac{1}{1-\sigma(a_n)}\sigma'(a_n)$$

عبارت معادل $\sigma'(a_n)$را جای‌گذاری می‌کنیم:

$$\frac{\partial l_n}{\partial a_n} = y_n(1-\sigma(a_n)) - (1-y_n)\sigma(a_n) = y_n - y_n\sigma(a_n) - \sigma(a_n) + y_n\sigma(a_n) = y_n - \sigma(a_n)$$

همچنین طبق تعریف $a_n = w^T x_n$، می‌توان نوشت:

$$\frac{\partial a_n}{\partial w} = x_n$$

در نتیجه، با جای‌گذاری نتایج در گرادیان، برای یک نمونه خواهیم داشت:

$$\nabla_w l_n = (y_n - \sigma(a_n))x_n$$

پس برای کل مجموعه داده، حاصل گرادیان برابر است با:

$$\nabla_w l(w) = \Sigma_{n=1}^{N}(y_n - p_n)x_n, \qquad p_n = \sigma(a_n) = \sigma(w^T x_n)$$

عبارت حاصل‌شده را به فرم ماتریسی نیز بازنویسی می‌کنیم:

$$\nabla_w l(w) = X^T(y - p), \qquad X \in \mathbf{R}^{N\times d}, \quad y, p \in \mathbf{R}^N$$

بدین ترتیب، خواسته مسئله محاسبه گردید.

تمرین ۳:

جهت درک و تحلیل بهتر پرسش‌های مطرح‌شده، نیاز به انجام برخی محاسبات است. این محاسبات را پیش از بررسی پرسش‌ها انجام می‌دهیم. مسئله مینیمم‌سازی مطرح‌شده به شکل زیر است:

$$J(w) = -l(w, \mathcal{D}_{train}) + \lambda||w||_2^2$$

$$l(w, \mathcal{D}) = \frac{1}{|\mathcal{D}|}\Sigma_{i\in\mathcal{D}} log\ \sigma(y_i x_i^T w), \quad y_i \in \{-1, +1\}$$

جهت سهولت در محاسبه، مسئله را به گونه‌ای می‌نویسیم که به جای $y_i \in \{-1, +1\}$، داشته باشیم $y_i' \in \{0, 1\}$. نحوه نوشتن $l(w, \mathcal{D})$ بدین صورت تغییر می‌کند:

$$J(w) = -l(w, \mathcal{D}_{train}) + \lambda||w||_2^2$$

$$l(w, D) = \frac{1}{|\mathcal{D}|}\Sigma_{i\in\mathcal{D}} l_i$$

$$l_i(w) = y_i log\ p_i + (1 - y_i)log(1 - p_i)$$

$$p_i = \sigma(w^T x_i)$$

با توجه به گرادیان $log - likelihood$ که در تمرین دوم محاسبه شد، گرادیان و ماتریس $Hessian$ متناظر با این مسئله را محاسبه می‌کنیم:

$$\nabla_w J(w) = \frac{1}{|\mathcal{D}|}\Sigma_{i\in\mathcal{D}}(p_i - y_i)x_i + 2\lambda w, \qquad p_i = \sigma(w^T x_i)$$

$$\nabla_w^2 J(w) = \frac{1}{|\mathcal{D}|}\Sigma_{i\in\mathcal{D}}\frac{\partial}{\partial w}[(p_i - y_i)x_i] + 2\lambda = \frac{1}{|\mathcal{D}|}\Sigma_{i\in\mathcal{D}}(\frac{\partial p_i}{\partial w})x_i^T + 2\lambda$$

طبق آنچه در تمرین دوم بخش (الف) محاسبه شد:

$$\nabla_w^2 J(w) = \frac{1}{|\mathcal{D}|}\Sigma_{i\in\mathcal{D}}[p_i(1 - p_i)x_i]x_i^T + 2\lambda = \frac{1}{|\mathcal{D}|}\Sigma_{i\in\mathcal{D}}p_i(1 - p_i)x_i x_i^T + 2\lambda$$

عبارت حاصل‌شده را به فرم ماتریسی می‌نویسیم. با تعریف ماتریس قطری S خواهیم داشت:

$$\nabla_w^2 J(w) = X^T S X + 2\lambda I, \qquad S_{ii} = \frac{p_i(1-p_i)}{|\mathcal{D}|}$$

در ادامه، نشان می‌دهیم که ماتریس $Hessian$ حاصل‌شده، معین مثبت است؛ یعنی به ازای هر $v \neq 0$، نا مساوی $v^T \nabla_w^2 J(w) v > 0$ برقرار است. مثبت معین بودن $2\lambda I$ به شرط $\lambda > 0$مشخص است. کافی است، مثبت معین بودن $X^T S X$ را نشان دهیم. تغییر متغیر $y = Xv$ را در نظر بگیرید. با دنبال کردن محاسبات، خواهیم داشت:

$$v^T X^T S X v = y^T S y = \frac{1}{\mathcal{D}} \Sigma_{i \in \mathcal{D}} p_i(1-p_i) y_i^2 > 0$$

نا مساوی بالا واضحاً برقرار است. چون $0 < p_i < 1$ است، قطعاً عبارت حاصل‌شده مثبت خواهد بود. چون v بردار دلخواه غیرصفر است، ماتریس $X^T S X$ مثبت معین است و در نتیجه، ماتریس $\nabla_w^2 J(w)$ مثبت معین است. از برقرار بودن این نامساوی، می‌توان نتیجه گرفت که این مسئله اکیداً محدب است و بنابراین، دارای دقیقاً یک جواب مینیمم محلی است که سراسری نیز هست.
حال، به پرسش‌های مطرح‌شده پاسخ می‌دهیم:

- (آ): نادرست - طبق توضیحات ارائه شده، فقط یک جواب بهینه محلی دارد. حتی، اگر $\lambda = 0$ باشد، تابع هزینه لجستیک، محدب است(نه اکیداً محدب)، اما باز هم چندین جواب بهینه محلی ندارد.

- (ب): نادرست - شرط بهینگی بر اساس گرادیان به صورت $\nabla_w J(\hat{w}) = 0$ حاصل می‌شود:

$$\nabla_w J(w) = \frac{1}{|\mathcal{D}|} \Sigma_{i \in \mathcal{D}} (p_i - y_i) x_i + 2\lambda w = 0$$

$$\implies 2\lambda \hat{w} = \frac{1}{|\mathcal{D}|} \Sigma_{i \in \mathcal{D}} (y_i - p_i) x_i$$

  این یک سیستم خطی است که هیچ مکانیزمی جهت تولید صفرهای زیاد و ایجاد پراکندگی ندارد.(اگر به جای منظم‌سازی $ridge$، از منظم‌سازی $lasso$ استفاده شده بود، طبق شهود هندسی‌ـاحتمالاتی ارائه‌شده در تمرین سری اول، پاسخ مثبت می‌بود.)

- (ج): درست - اگر داده آموزشی، خطی قابل‌تفکیک باشد، یعنی w^*ای وجود دارد که به ازای هر i، نا مساوی $y_i(w^{*T} x_i) > 0$ برقرار باشد. تابع هزینه لجستیک $L_i(w) = log(1 + e^{-y_i w^T x_i})$ را در نظر بگیرید. اگر به دنبال $scale$ کردن w به * باشیم به طوری که $\alpha \longrightarrow \infty$ برقرار باشد، خواهیم داشت:

$$lim_{\alpha \longrightarrow \infty} L_i(\alpha w^*) = lim_{\alpha \longrightarrow \infty} log(1 + e^{-y_i \alpha w^{*T} x_i}) = log(1+0) = 0$$

  بنابراین، تابع هزینه می‌تواند به صفر میل کند. در نتیجه، هیچ مینیمم‌کننده متناهی‌ای وجود نخواهد داشت و وزن‌ها می‌توانند به بی‌نهایت میل کنند. درحالی که اگر $\lambda > 0$ برقرار بود، از این موضوع، جلوگیری می‌شد.

- (د): نادرست - درباره نقش ابرپارامتر λ باید توجه کرد:
  - اگر $\lambda = 0$ باشد، مدل تا حد امکان تلاش می‌کند داده تمرین را برازش کند.
  - اگر $\lambda > 0$ باشد، منظم‌سازی تلاش می‌کند وزن‌ها را به صفر سوق داده و از انعطاف‌پذیری مدل بکاهد.
  - اگر $\lambda \longrightarrow \infty$ باشد، وزن‌ها برابر صفر می‌شوند و مدل برای هر ورودی، یکسان پیش‌بینی خواهد کرد.

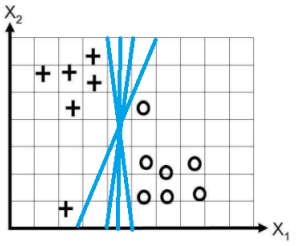
  وقتی $\lambda = 0$ باشد، به برآورد بیشینه درست‌نمایی MLE دست خواهیم یافت و MLE، بیشینه درست‌نمایی آموزشی ممکن را می‌دهد. وقتی $\lambda > 0$ باشد و به مرور افزایش یابد، از MLE دور می‌شویم؛ در نتیجه، به جای اینکه بیشتر بر بیشینه‌سازی $l(\hat{w}, \mathcal{D}_{train})$ تمرکز کنیم، بر کوچک کردن $||w||_2^2$ متمرکز می‌شویم.
  بدین ترتیب، با افزایش λ، مقدار $l(\hat{w}, \mathcal{D}_{train})$ کاهش می‌یابد.

- (ه): نادرست - درباره کارکرد ابرپارامتر λ در بخش (د) صحبت کردیم. انتخاب λ بهینه، کمک می‌کند تعادل بین واریانس و بایاس برقرار شده و از وقوع کم‌برازش یا بیش‌برازش جلوگیری شود. اگر λ بیش از اندازه بزرگ شود، مدل بیش از حد، ساده خواهد شد و روی مجموعه تست، بد عمل خواهد کرد. تصور کنید اگر اینطور نبود، پس کاری می‌کردیم که $\lambda \longrightarrow \infty$ رخ دهد؛ اما این موضوع، باعث می‌گردد مدل برای هر ورودی، به شکل یکسانی پیش‌بینی کند و در نتیجه، ضعیف عمل خواهد کرد؛ بنابراین، می‌توان دریافت با افزایش λ، مقدار $l(\hat{w}, \mathcal{D}_{test})$ ابتدا افزایش و سپس کاهش می‌یابد و یک روند غیریکنوا شاهد خواهیم بود.

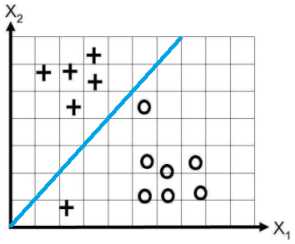
تمرین ۴:

بر اساس مدل مطرح‌شده، مرز تصمیم به صورت $w_0 + w_1x_1 + w_2x_2 = 0$ می‌باشد.

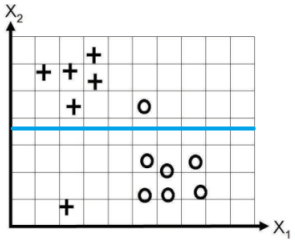
- (۱) بدون منظم‌سازی MLE سعی می‌کند کلاس‌ها را کامل جدا کند و مرز تصمیم از فضای خالی بین کلاس‌ها عبور کند. در این حالت، خط مد نظر یگانه نیست و تعداد خطوط متعددی می‌توانند داده ها را جدا نموده و خطای تمرین ۰ داشته باشند.



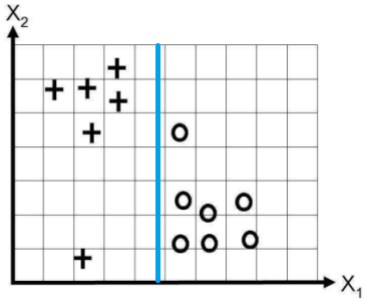
- (۲) وقتی $\lambda \longrightarrow \infty$ رخ دهد، باعث می‌شود $w_0 \longrightarrow 0$ اتفاق بیفتد. در نتیجه مرز تصمیم به صورت $w_1x_1 + w_2x_2 = 0$ در خواهد آمد که از مرکز مختصات می‌گذرد. w_1و w_2 باید طوری تنظیم شوند که خطای دسته‌بندی مینیمم شود. در شکل زیر، یک خطای دسته‌بندی مشاهده می‌کنید.



- (۳) مشابه حالت قبل، اما این بار $w_1 \longrightarrow 0$رخ می‌دهد و مرز تصمیم، $w_0 + w_2x_2 = 0$ خواهد شد که معادل $x_2 = \frac{-w_0}{w_2}$ است. پس یک خط افقی خواهیم داشت که مستقل از x_1 است. در شکل زیر، مدل دارای دو خطای تصمیم‌گیری است.



- (۴) مشابه حالت قبل، اما این بار $w_2 \longrightarrow 0$رخ می‌دهد و مرز تصمیم، $w_0 + w_1 x_1 = 0$ خواهد بود که معادل $x_1 = \frac{-w_0}{w_1}$ است. پس یک خط عمودی خواهیم داشت که مستقل از x_2 است. در شکل زیر، مدل فاقد خطا است.



تمرین ۵:

برای این مسئله، فرض‌های زیر را در نظر می‌گیریم:

- کلاس $Y \in \{1, \dots, C\}$، دارای چنین $prior$ای است: $P(Y = c) = \pi_c$
- ویژگی دودویی $X_j \in \{0, 1\}$، دارای احتمال $P(X_j = 1|Y = c) = \theta_{jc}$ است.
- فرض $Naive\ Bayes$ نیز برقرار است؛ یعنی به ازای $j \neq k$، $X_j \perp X_k$ برقرار است.

می‌خواهیم $I(X_j, Y)$ را محاسبه کنیم. فرمول MI را می‌نویسیم:

$$I(X_j, Y) = \Sigma_{x_j \in \{0,\dots,1\}} \Sigma_{c=1}^{C} P(X_j = x_j, Y = c)\ log \frac{P(X_j = x_j, Y = c)}{P(X_j = x_j)P(Y = c)}$$

بر اساس فرض‌های ارائه شده، مقادیر احتمال $joint$ و $marginal$ را محاسبه می‌کنیم:

$$P(X_j = 1, Y = c) = \pi_c \theta_{jc}, \qquad P(X_j = 0, Y = c) = \pi_c(1 - \theta_{jc})$$

$$P(X_j = 1) = \Sigma_{c=1}^{C} \pi_c \theta_{jc} = \phi_j, \qquad P(X_j = 0) = 1 - \phi_j$$

مقادیر محاسبه‌شده را در فرمول MI جای‌گذاری می‌کنیم:

$$I(X_j, Y) = \Sigma_{c=1}^{C} [\pi_c \theta_{jc} log \frac{\pi_c \theta_{jc}}{\pi_c \phi_j} + \pi_c(1 - \theta_{jc}) log \frac{\pi_c(1 - \theta_{jc})}{\pi_c(1 - \phi_j)}]$$

لگاریتم‌ها را ساده می‌کنیم:

$$I(X_j, Y) = \Sigma_{c=1}^{C} [\pi_c \theta_{jc} log \frac{\theta_{jc}}{\phi_j} + \pi_c(1 - \theta_{jc}) log \frac{(1 - \theta_{jc})}{(1 - \phi_j)}]$$

بدین ترتیب، عبارت مربوطه محاسبه گردیده و به نتیجه مدنظر دست یافتیم.

تمرین ۶:

این مسئله، یک دسته‌بندی دودویی با لیبل‌های نویزی است. همچنین، برای هر داده تمرینی x_n، به جای ارائه یک تعریف محکم از لیبل‌ها به شکل $y_n \in \{0, 1\}$، صرفاً از ارائه احتمال تعلق به یک کلاس صحبت می‌کنیم؛ بنابراین، می‌توان گفت که به جای $hard\ label$، با $soft\ label$ مواجه هستیم.

$$\pi_n = P(y_n = 1|x_n)$$

بر اساس این تعریف:

- کلاس ۱ دارای احتمال π است.
- کلاس ۰ دارای احتمال $1 - \pi$ است.

اگر مدل چنین پیش‌بینی‌ای داشته باشد:

$$p_n = P(y_n = 1|x_n; w)$$

توزیع درست‌نمایی هر داده به صورت زیر خواهد بود:

$$l_n = p_n^{\pi_n}(1 - p_n)^{\pi_n}$$

زیرا برای لیبل‌های نویزی، مدل‌سازی توزیع لیبل مشاهده‌شده، یک توزیع برنولی با پارامتر π_n خواهد بود. بنابراین، $likelihood$ برای کل مجموعه داده، بدین صورت خواهد بود:

$$\mathcal{L}_n = \Sigma_{n=1}^{N} p_n^{\pi_n}(1 - p_n)^{\pi_n}, \qquad p_n = P(y_n = 1|x_n; w)$$

در نتیجه، $log - likelihood$ برای کل مجموعه داده، بدین صورت خواهد بود:

$$log\ \mathcal{L}_n = \prod_{n=1}^{N} \pi_n log\ p_n + (1 - \pi_n) log(1 - p_n), \qquad p_n = P(y_n = 1|x_n; w)$$

نتیجه حاصل‌شده، همان رابطه $cross - entropy$ است؛ با این تفاوت که برای لیبل‌های نویزی لحاظ شده است. عبارت فوق، نتیجه نهایی مد نظر ماست. حال، بر حسب نوع مدل انتخابی، می‌توان مقدار محاسبه‌شده برای p_n را جای‌گذاری نمود؛ به عنوان مثال، اگر $p_n = P(y_n = 1x_n; w) = \sigma(w^T x_n)$ باشد، عبارت$log - likelihood$ بدین صورت خواهد بود:

$$log\ \mathcal{L}_n = \prod_{n=1}^{N} \pi_n log\ \sigma(w^T x_n) + (1 - \pi_n) log(1 - \sigma(w^T x_n))$$

تمرین ۷:

- (۱) تابع هدف $primal$ برای SVM بدین صورت است:

$$min_{w,b,\xi}\ \frac{1}{2}||w||^2 + C\ \Sigma_{i=1}^{N} \xi_i$$
$$s.t\ \ y_i(w^T x_i + b) \geq 1 - \xi_i,$$
$$\xi_i \geq 0, \quad i = 1, \ldots, N$$

عبارت $\frac{1}{2}||w||^2$ میزان $margin$ را بیشینه می‌کند. برای حالتی که نقاط به صورت خطی جداپذیر باشند، یعنی $\xi_i = 0$، نامساوی‌ها به $y_i(w^T x_i + b) \geq 1$ تبدیل می‌شوند. فاصله یک نقطه از مرز تصمیم‌گیری برابر $\frac{|w^T x + b|}{||w||}$ است. برای بردارهای پشتیبان که $|w^T x + b| = 1$ برقرار است، $margin$ برابر $\frac{1}{||w||}$ می‌باشد. بنابراین، بیشینه‌سازی $||w||^2$ معادل کمینه‌سازی $\frac{1}{||w||}$ است.

نقش پارامترها در این مسئله، بدین صورت است:

- w ، بردار نرمال به ابرصفحه است و جهت‌گیری آن را تعیین می‌کند.
- b عبارت بایاس است که باعث شیفت ابرصفحه از مرکز مختصات می‌گردد.
- ξ_i ها متغیرهای $slack$ هستند که اجازه $misclassification$ می‌دهند. (حالت $soft\ \ margin$)
- C یک ابرپارامتر است که تعادل بین بیشینه‌سازی $margin$ و خطای کلاس‌بندی را برقرار می‌سازد.

- (۲) تابع لاگرانژ این مسئله، بدین صورت است:

$$\mathcal{L}(w,b,\xi,\alpha,\mu) = \frac{1}{2}||w||^2 + C\ \Sigma_{i=1}^{N}\xi_i - \Sigma_{i=1}^{N}\ \alpha_i[y_i(w^T x_i + b) - 1 + \xi_i] - \Sigma_{i=1}^{N}\mu_i\xi_i$$

$$\alpha_i, \mu_i \geq 0, \quad i = 1, \ldots, N$$

شروط KKT را می‌نویسیم:

- شرط ۱:

$$\nabla_w \mathcal{L} = w - \Sigma_{1=1}^{N}\alpha_i y_i x_i = 0 \implies w = \Sigma_{1=1}^{N}\alpha_i y_i x_i$$

- شرط ۲:

$$\frac{\partial \mathcal{L}}{\partial b} = -\Sigma_{i=1}^{N}\alpha_i y_i = 0 \implies \Sigma_{i=1}^{N}\alpha_i y_i = 0$$

- شرط ۳:

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - \mu_i \implies \alpha_i = C - \mu_i \implies 0 \leq \alpha_i \leq C \quad \text{چون } \mu_i \geq 0 \text{ ،}$$

جای‌گذاری می‌کنیم:

$$\mathcal{L}(w = \Sigma_{1=1}^{N}\alpha_i y_i x_i, b, \xi, \alpha, \mu) = \frac{1}{2}||\Sigma_{1=1}^{N}\alpha_i y_i x_i||^2 + C\ \Sigma_{i=1}^{N}\xi_i$$

$$-\Sigma_{i=1}^{N}\ \alpha_i[y_i((\Sigma_{1=1}^{N}\alpha_i y_i x_i)^T x_i + b) - 1 + \xi_i] - \Sigma_{i=1}^{N}\mu_i\xi_i$$
$$= \frac{1}{2}\Sigma_{i=1}^{N}\Sigma_{j=1}^{d}\alpha_i\alpha_j y_i y_j x_i^T x_j - \Sigma_{i=1}^{N}\Sigma_{j=1}^{d}\alpha_i\alpha_j y_i y_j x_j^T x_i$$
$$-b\Sigma_{i=1}^{N}\alpha_i y_i + \Sigma_{i=1}^{N}(\alpha_i - \alpha_i\xi_i) - \Sigma_{i=1}^{N}\mu_i\xi_i + C\ \Sigma_{i=1}^{N}\xi_i$$

به خاطر شرط دوم KKT و تقارن $x_j^T x_i = x_i^T x_j$، می‌توان ساده کرد:

$$\mathcal{L}(w,b,\xi,\alpha,\mu) = -\frac{1}{2}\Sigma_{i=1}^{N}\Sigma_{j=1}^{d}\alpha_i\alpha_j y_i y_j x_i^T x_j + \Sigma_{i=1}^{N}(\alpha_i - \alpha_i\xi_i) - \Sigma_{i=1}^{N}\mu_i\xi_i + C\ \Sigma_{i=1}^{N}\xi_i$$

$$= -\frac{1}{2}\Sigma_{i=1}^{N}\Sigma_{j=1}^{d}\alpha_i\alpha_j y_i y_j x_i^T x_j + \Sigma_{i=1}^{N}\alpha_i - \Sigma_{i=1}^{N}\alpha_i\xi_i + \Sigma_{i=1}^{N}\xi_i(C - \mu_i)$$

با لحاظ کردن شرط سوم KKT، خواهیم داشت:

$$\mathcal{L}(w,b,\xi,\alpha,\mu) = -\frac{1}{2}\Sigma_{i=1}^{N}\Sigma_{j=1}^{d}\alpha_i\alpha_j y_i y_j x_i^T x_j + \Sigma_{i=1}^{N}\alpha_i - \Sigma_{i=1}^{N}\alpha_i\xi_i + \Sigma_{i=1}^{N}\alpha_i\xi_i$$

$$= -\frac{1}{2}\Sigma_{i=1}^{N}\Sigma_{j=1}^{d}\alpha_i\alpha_j y_i y_j x_i^T x_j + \Sigma_{i=1}^{N}\alpha_i$$

بدین ترتیب، مسئله دوگان استخراج می‌گردد.

$$max_\alpha \ -\frac{1}{2}\Sigma_{i=1}^{N}\Sigma_{j=1}^{d}\alpha_i\alpha_j y_i y_j x_i^T x_j + \Sigma_{i=1}^{N}\alpha_i$$

$$s.t \quad \Sigma_{i=1}^{N}\alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, N$$

نسخه دوگان برای محاسبات سودمندتر است؛ زیرا در مسئله دوگان، متغیرها فقط به صورت یک ضرب داخلی $x_i^T x_j$ ظاهر می‌شوند. این مسئله، این اجازه را به ما می‌دهد که ضرب داخلی را با یک تابع کرنل $K(x_i, x_j) = \phi(x_i)^T\phi(x_j)$ جایگزین کنیم که بدون نیاز به محاسبه صریح نگاشت ϕ که ممکن است به فضای با ابعاد بسیار بالا برود، سودمند است. همچنین، در جواب بهینه α، تعداد کمی از α_iها غیرصفر هستند و فقط بردارهای پشتیبان در محاسبات دخیل هستند. این نحوه نمایش باعث می‌گردد که سرعت در تست افزایش یابد. به علاوه، در مسئله اصلی، تعداد متغیرها برابر $d+1$ متغیر و در مسئله دوگان برابر N است. در نتیجه، در حل مسائل با ابعاد بسیاربزرگ، این نسخه کارآمدتر به نظر می‌رسد.

- (۳) شروط نوشته شده در نسخه دوگان را شرح می‌دهیم:
  - $\Sigma_{i=1}^{N}\alpha_i y_i = 0$: بیانگر شرکت متوازن کلاس‌هاست.
  - $0 \leq \alpha_i \leq C$: بر اساس مقادیر α_i می‌توان گفت:
    - اگر $\alpha_i = 0$، یعنی نقطه‌ها درست دسته‌بندی شده‌اند. (نه بردار پشتیبان)
    - اگر $0 < \alpha_i < C$، یعنی نقطه بر margin قرار دارد. $(y_i(w^T x_i + b) = 1)$
    - اگر $\alpha_i = C$، یعنی نقطه اشتباه دسته‌بندی شده یا درون margin قرار گرفته است.

  همچنین درباره بردارهای پیشتیبان، این دو نکته حائز اهمیت است:
  - نقاط با $\alpha_i > 0$، بردارهای پشتیبان هستند.
  - تابع تصمیم را می‌توان بدین صورت نوشت:

$$f(x) = \Sigma_{i \in SV}\alpha_i y_i K(x_i, x) + b$$

- (۴) علت ارائه این فرمول‌بندی برای مسئله دوکلاسه، این است که با رسم یک ابرصفحه تفکیک‌کننده، می‌توان تفسیر هندسی طبیعی‌تری ارائه کرد. همچنین مسئله بهینه‌سازی ساده‌تری داریم و از نظر تئوری، اثبات آسان‌تر می‌گردد. در ادامه، سه روش تعمیم SVM، به مسائل چندکلاسه را توضیح می‌دهیم.

  - یکی در برابر بقیه: در این روش برای هر کلاس، یک SVM جداگانه آموزش می‌دهیم. اگر تعداد کلاس‌ها k باشد، k مدل SVM می‌سازیم. هر SVM تلاش می‌کند یک کلاس را از بقیه کلاس‌ها جدا کند. به ازای نمونه جدید، برای هر SVM ، یک مقدار امتیاز یا فاصله لحاظ می‌گردد و بر این اساس، مدلی که بیشترین فاصله از مرز یا بیشترین احتمال را بدهد، برنده است. مزیت این روش، این است که ساده، سریع و مناسب برای تعداد کلاس‌های زیاد است. معایب این روش این است که اگر داده نامتوازن باشد، ممکن است یکی از مدل‌ها ضعیف شود. همچنین، مرز تصمیم نهایی، کاملاً یکپارچه نیست.
  - یکی در برابر یکی: در این روش برای هر جفت کلاس، یک SVM می‌سازیم. بنابراین اگر تعداد کلاس‌ها k باشد، تعداد SVM ها برابر $\frac{k(k-1)}{2}$ می‌باشد. به ازای نمونه جدید، هر SVM رای می‌دهد که نمونه به کدام یک از جفت کلاس‌های متناظر تعلق دارد. در پایان، کلاسی که بیشترین رای را بیاورد، انتخاب می‌شود. مزیت این روش، این است که معمولاً دقیق‌تر است و هر SVM روی بخش کوچکی از داده کار می‌کند. معایب این روش این است که اگر k زیاد باشد، هزینه محاسباتی بالا رفته و مدل پیچیده می‌شود. همچنین، ترکیب تصمیمات SVM های متعدد، می‌تواند دشوار باشد.

- چندکلاسه یکپارچه: برخلاف دو روش قبل که بر اساس ترکیب چند SVM دودویی هستند، در این روش، یک مسئله بهینه‌سازی چندکلاسه تعریف می‌شود و یک SVM واحد برای همه کلاس‌ها ساخته می‌شود. برای هر کلاس، یک بردار وزن جداگانه تعریف می‌کنیم و مدل، تلاش می‌کند به ازای نمونه جدید، امتیاز کلاس صحیح را نسبت به بقیه کلاس‌ها بیشتر در نظر بگیرد. از مزایای این روش، این است که از نظر تئوری، مرز تصمیم یکپارچه و قوی‌ای دارد و دقیق‌تر است. همچنین، ارائه یک مدل واحد برای دسته‌بندی، ساختار تمیزتری ارائه می‌دهد. از معایب این روش، می‌توان به پیچیدگی حل مسئله بهینه‌سازی، پیاده‌سازی پیچیده‌تر نسبت به دو روش قبلی و کندی در دیتاست‌های بسیاربزرگ اشاره کرد.

تمرین ۸:

- (۱) قانون مطرح‌شده بدین صورت است:

$$w^{(t+1)} = w^{(t)} + \eta.y_i x_i, \qquad b^{(t+1)} = b^{(t)} + \eta.y_i$$

علت اینکه این قانون، فقط هنگام دسته‌بندی اشتباه اعمال می‌گردد، این است:

- اگر $y_i(w^T x_i + b) > 0$ باشد، نمونه درست دسته‌بندی شده است و به آپدیت نیازی نیست.
- در غیر این صورت، دسته‌بندی اشتباه رخ داده است و به‌روزرسانی، مرز تصمیم را به سوی دسته‌بندی درست متمایل می‌کند.

نقش ابرپارامترها بدین صورت است:

- w بردار نرمال به مرز تصمیم است و جهت‌گیری آن را مشخص می‌کند.
- b عبارت بایاس است و باعث شیفت مرز تصمیم از مرکز مختصات می‌گردد.

تابع تصمیم این الگوریتم نیز بدین صورت است:

$$f(x) = sign(w^T x + b)$$

بدین ترتیب، یک مرز تصمیم خطی به شکل $w^T x + b = 0$ ساخته می‌شود.

- (۲)

- پرسپترون همیشه نمونه‌هایی را که اشتباه طبقه‌بندی شده‌اند، انتخاب می‌کند و وزن‌ها را طوری به‌روزرسانی می‌کند که پیش‌بینی روی آن نمونه بهتر شود ، زاویه وزن‌ها با بردار ویژگی نمونه درست کوچکتر شود و وزن‌ها در جهت درست هل داده شوند. حالا اگر داده‌ها خطی تفکیک‌پذیر باشند، یعنی یک ابرصفحه وجود دارد که تمام نقاط مثبت را از منفی جدا می‌کند. پس یک بردار وزن بهینه هست که همیشه پیش‌بینی را درست انجام می‌دهد. هر بار که پرسپترون اشتباه می‌کند، وزن‌ها را نزدیک‌تر به این بردار بهینه می‌کند. در نتیجه، تعداد خطاها محدود است، وزن‌ها از خطی که جواب درست است، دور نمی‌شوند و در نهایت پرسپترون به یک برداری می‌رسد که همه نقاط را درست طبقه‌بندی می‌کند. بنابراین پرسپترون تضمینی همگراست و تعداد خطاهایش از یک حد مشخص بیشتر نمی‌شود.
- اگر داده‌ها خطی تفکیک‌پذیر نباشند، هیچ وزن بهینه‌ای وجود ندارد که همه داده‌ها را درست طبقه‌بندی کند. پرسپترون دائم بین چند نقطه اشتباه گیر می‌کند. به‌روزرسانی‌های مکرر باعث می‌شود وزن‌ها دائماً تغییر جهت بدهند. نه زاویه با یک جهت درست کاهش می‌یابد و نه وزن‌ها محدود باقی می‌مانند. بدین ترتیب وزن‌ها ممکن است تا بی‌نهایت بزرگ شود. در نتیجه، الگوریتم ممکن است تا ابد بین خطاهای مختلف رفت‌وبرگشت کند و همگرا نشود. معمولاً رفتار این گونه است که الگوریتم دائماً وزن‌ها را آپدیت می‌کند ولی هیچ‌وقت به بردار ثابتی نمی‌رسد؛ چون هیچ مرز خطی واحدی نیست که همه داده‌ها را درست جدا کند.

- (۳)

  - الگوریتم کلاسیک، نرخ یادگیری ثابت است اما در نسخه دیگر، قابل تنظیم است که این موجب همگرایی نرم‌تر می‌گردد.
  - تعداد تکرار در الگوریتم کلاسیک، نامحدود و در نسخه دیگر محدود است که این موجب جلوگیری از ایجاد حلقه‌های بی‌نهایت روی داده‌های تفکیک‌ناپذیر می‌گردد و بر پایداری مدل در برابر این‌گونه داده‌ها موثر است.
  - افزودن قابلیت شافل کردن در هر دوره، موجب همگرایی سریعتر و کاهش چرخه‌ها می‌گردد.
  - افزودن منظم‌سازی باعث جلوگیری از بیش‌برازش و بهبود تعمیم‌پذیری می‌گردد و دقت را بالاتر می‌برد.
  - همچنین افزودن قابلیت توقف زودهنگام در نسخه $sklearn$ که اختیاری است، می‌تواند در تشخیص خودکار همگرایی و تضمین توقف موثر باشد و اطمینان مدل را بالاتر ببرد.

  نکته: **پاسخ پرسش این بخش، به جهت مقایسه دو نسخه کلاسیک و $sklearn$ الگوریتم پرسپترون، پاسخ نکته تمرین عملی ۴ نیز محسوب می‌شود.**

- (۴) مدل خطی، نمی‌تواند مرزهای تصمیم غیرخطی یاد بگیرد و قدرت نمایش محدودی برای الگوهای پیچیده دارد. دو روش برای افزایش قدرت مدل معرفی می‌کنیم.

  - مهندسی ویژگی: در این حالت، ورودی را به فضای دارای ابعاد بالاتر می‌بریم(مشابه خصوصیات چندجمله‌ای‌ها). در این حالت، همچنان داریم از مدل پرسپترون ساده استفاده می‌کنیم اما به خاطر مشکل احتمالی نفرین ابعاد و طراحی دستی فیچرها، با ضعف مواجه هستیم.
  - پرسپترون چندلایه (شبکه عصبی): ایده بدین صورت است که چندین پرسپترون را با تابع فعالسازی غیرخطی کنار هم می‌چینیم و از طریق لایه‌های پنهان ایجاد شده، ورودی را به خروجی مدنظر می‌رسانیم.

    با این روش می‌توان تقریباً هر تابعی را تقریب زد اما تعداد پارامترها بیشتر است؛ همچنین پدیده پس‌انتشار را هم باید لحاظ کنیم.