رگرسیون

۱. فرض کنید تعداد داده‌های آموزش ما N و تعداد ویژگی‌ها M باشد؛ تابع پیش‌بینی یک مدل رگرسیون به‌شکل زیر تعریف شده باشد:

$$f(X) = \sum_{i=0}^{M} w_i x_i$$

و تابع هزینه نیز به‌صورت زیر تعریف می‌شود:

$$L = \sum_{i=1}^{N} \left(f(x_i) - Y_i\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{M} w_j^2$$

به سؤالات زیر پاسخ دهید:

(الف) گرادیان این تابع هزینه را نسبت به پارامترهای w محاسبه کنید.

(ب) رابطه‌ی به‌روزرسانی وزن‌ها در الگوریتم گرادیان کاهشی را بنویسید.

(ج) چرا استفاده از این تابع هزینه برای مسائل دسته‌بندی مناسب نیست؟

(د) گرادیان تابع هزینه‌ی رگرسیون لاجستیک را برای یک مسئله‌ی دسته‌بندی دودویی محاسبه کنید. سپس آن را با گرادیان در حالت رگرسیون مقایسه نمایید.

تابع پیش‌بینی در رگرسیون لاجستیک به‌صورت زیر تعریف می‌شود:

$$f(X) = \frac{1}{1 + e^{-\sum_{i=0}^{M} w_i x_i}}$$

و تابع هزینه‌ی آن به شکل زیر است:

$$L = -\sum_{i=1}^{N} \left[Y_i \log(f(x_i)) + (1 - Y_i)\log(1 - f(x_i))\right] + \frac{\lambda}{2}\sum_{j=1}^{M} w_j^2$$

---

۲. مجموعه‌داده‌ی زیر را در نظر بگیرید:

x_1	x_2	y
۱	۲	۳
۲	۳	۵
۳	۵	۷

مدلی به صورت زیر در نظر بگیرید:

$$y = w_0 + w_1 x_1 + w_2 x_2$$

با استفاده از الگوریتم گرادیان کاهشی، وزن‌ها را در چهار مرحله به‌روزرسانی کنید. (فرض کنید وزن‌های اولیه $\begin{bmatrix} ۰ \\ ۰ \\ ۰ \end{bmatrix}$ باشد.)

---

۳. فرض کنید مدل رگرسیونی به‌صورت زیر تعریف شده باشد:

$$\hat{y} = w_۱ x^۲ + w_۲ x + w_۳$$

و تابع هزینه نیز به‌صورت زیر باشد:

$$f(w_۱, w_۲, w_۳) = \sum_i (\hat{y}_i - y_i)^۲$$

به سؤالات زیر پاسخ دهید:

(الف) گرادیان تابع هزینه را نسبت به پارامترها محاسبه کرده و رابطه‌ی به‌روزرسانی وزن‌ها را در الگوریتم گرادیان کاهشی بنویسید.

(ب) تعریف معیار $R^۲$ (ضریب تعیین) را ارائه کنید.

(ج) توضیح دهید که این معیار در چه شرایطی به مقدار ۱ نزدیک می‌شود و چه زمانی ممکن است مقدار آن منفی شود.

(د) بین معیارهای MSE و $R^۲$ از نظر تفسیر و کاربرد، مقایسه‌ای انجام دهید.

---

۴. فرض کنید مدل رگرسیونی به‌صورت ماتریسی تعریف شده باشد:

$$\hat{y} = Xw \quad \text{و} \quad L(w) = \|Xw - y\|^۲$$

به سؤالات زیر پاسخ دهید:

(الف) گرادیان تابع $L(w)$ را نسبت به بردار پارامترها w محاسبه کنید.

(ب) رابطه‌ی به‌روزرسانی وزن‌ها در الگوریتم گرادیان کاهشی را بنویسید.

(ج) توضیح دهید که افزایش یا کاهش نرخ یادگیری α چه تأثیری بر فرایند آموزش مدل دارد.

---

۵. فرض کنید قصد داریم یک مسئله‌ی رگرسیون چندمتغیره را در نظر بگیریم. تابع هزینه‌ای که باید کمینه شود، به‌صورت زیر تعریف می‌شود:

$$\min_W \quad F(W) = \lambda W^\top W + \|XW - Y\|_۲^۲$$

به سؤالات زیر پاسخ دهید:

(الف) اگر بخواهیم این مسئله را با استفاده از الگوریتم گرادیان کاهشی تصادفی Stochastic Gradient Descent (SGD) حل کنیم، شبه‌کد (pseudo-code) آن را بنویسید.

(ب) فرض کنید تعاریف زیر برقرار هستند:

$$W_1 = \arg\min_W L(W)$$

$$W_2 = \arg\min_W \Big(L(W) + \lambda W^\top W \Big)$$

که در آن $L(W)$ یک تابع نامنفی است. نشان دهید که:

$$\|W_1\|_2 \leq \|W_2\|_2$$

و توضیح دهید این رابطه چه ارتباطی با فرمول‌بندی مسئله دارد.

۶.(الف) فرض کنید هدف، تخمین سن افراد از طریق رگرسیون و بر اساس تصاویر اسکن مغزی آن‌ها باشد. تعداد افراد موجود برابر با ۱۰ نفر است و برای هر فرد، بردار ویژگی‌ای با ۱۵۰۰۰ ویژگی مختلف در دسترس است. با توجه به مطالب آموخته‌شده، در این حالت استفاده از کدام‌یک از روش‌های تنظیم‌سازی **L1** یا تنظیم‌سازی **L2** مناسب‌تر است؟ دلیل خود را بیان کنید.

(ب) آیا ممکن است الگوریتم گرادیان کاهشی هنگام آموزش یک مدل رگرسیون لجستیک در یک کمینه‌ی موضعی گیر کند؟ توضیح دهید.

---

۷. فرض کنید:

$$y = w^T x, \quad x \in \mathbb{R}^L, \quad y \in \mathbb{R}$$

$$X = [x_1, x_2, \dots, x_N], \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_L \end{bmatrix}$$

(الف) اگر رگرسیون فقط روی ویژگی j انجام شود، پارامتر w_j به صورت زیر محاسبه می‌شود:

$$w_j = \frac{X_j^\top y}{X_j^\top X_j}$$

که در آن X_j بردار ستون j از ماتریس داده‌ها است.

(ب) اگر ویژگی‌ها مستقل باشند (یعنی ستون‌های ماتریس داده‌ها مستقل)، آنگاه پارامترهای بهینه حاصل از آموزش رگرسیون روی همه ویژگی‌ها برابر است با پارامترهای بهینه آموزش روی هر ویژگی به صورت مستقل:

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_L \end{bmatrix} \quad \Longrightarrow \quad w_j = \frac{X_j^\top y}{X_j^\top X_j} \quad \forall j$$

(ج) فرض کنید مدل به صورت زیر تعریف شده است:

$$y = w^\top x + w_0$$

و رگرسیون فقط روی ویژگی j انجام می‌شود. پارامترهای $w_\circ$ و w_j را می‌توان با کمینه کردن تابع هزینه (مانند میانگین مربعات خطا) به‌دست آورد که معمولاً با حل معادله نرمال یا استفاده از روش‌های بهینه‌سازی صورت می‌گیرد.

۸.(الف) دو روش برای کاهش واریانس مدل ارائه دهید.

(ب) اگر بعضی از ویژگی‌ها با هم هم‌بسته باشند، چه اثری بر بایاس و واریانس دارند؟ اگر همه‌ی ویژگیهای وابسته به جز یکی از آنها را حذف کنیم، چه تغییری رخ می‌دهد؟

(ج) با ذکر دلیل مشخص کنید کدام‌یک از گزاره‌های زیر درست است و چرا؟

- اگر بایاس زیاد باشد، افزایش داده‌های آموزش باعث کاهش آن می‌شود.
- افزایش پیچیدگی مدل همیشه باعث کاهش خطای آموزش و افزایش خطای تست می‌شود.
- اگر bias زیاد است، اضافه کردن تعداد داده‌های آموزش کمک زیادی به کم کردن bias نمی‌کند.
- کم کردن خطای مدل روی داده‌های آموزش منجر به کاهش خطای مدل روی داده‌های تست می‌شود.
- افزایش پیچیدگی مدل رگرسیون همواره منجر به کاهش خطای مدل روی داده‌ی آموزش و افزایش خطای مدل روی داده‌ی تست می‌شود.

دسته‌بندی

۹.(الف) فرض کنید تابع زیر داده شده است:

$$f(w) = \frac{۱}{n}\sum_{i=۱}^{n} \max(\circ, ۱ - y^{(i)} w^\top x^{(i)}).$$

مشتق تابع $f(w)$ نسبت به w را به‌دست آورید.

(ب) فرض کنید داریم:

$$f(w) = \max(\circ, ۱ - y\, w^\top x).$$

مشتق تابع $f(w)$ نسبت به w را محاسبه کنید.

---

۱۰. در این‌جا یک رویکرد تمایزی برای حل مسئله‌ی طبقه‌بندی که در شکل ۱ نشان داده شده، در نظر گرفته می‌شود. مجموعه داده‌ی آموزشی دوبعدی شامل نقاط با برچسب $y = ۱$ (نشان داده شده با علامت +) و $y = \circ$ (نشان داده شده با علامت o) است.

(الف) قصد داریم مسئله‌ی طبقه‌بندی دودویی را با مدل ساده‌ی رگرسیون لجستیک خطی حل کنیم:

$$P(y=۱|\vec{x},\vec{w}) = g(w_\circ + w_۱ x_۱ + w_۲ x_۲) = \frac{۱}{۱ + \exp(-w_\circ - w_۱ x_۱ - w_۲ x_۲)}.$$

توجه داشته باشید که داده‌های آموزشی را می‌توان با یک مرز خطی و بدون خطا طبقه‌بندی کرد. اکنون مدل رگرسیون لجستیک خطی منظم‌شده‌ای را در نظر بگیرید که در آن تابع هدف زیر بیشینه می‌شود:

$$\sum_{i=۱}^{n} \log(P(y_i|x_i, w_\circ, w_۱, w_۲)) - C w_j^۲$$

که در آن فقط یکی از پارامترها (w_j) منظم می‌شود. فرض کنید C عددی بسیار بزرگ است.

با توجه به داده‌های آموزشی، در صورت منظم‌سازی هر یک از پارامترهای w_j، خطای آموزشی چگونه تغییر می‌کند؟ مشخص کنید که آیا خطا افزایش می‌یابد یا همچنان صفر باقی می‌ماند. برای هر حالت، توضیح مختصری ارائه دهید:

- در صورت منظم‌سازی $w_۲$
- در صورت منظم‌سازی $w_۱$
- در صورت منظم‌سازی $w_۰$

(ب) اگر نوع منظم‌سازی را به نُرم $L_۱$ (قدر مطلق) تغییر دهیم و تنها پارامترهای $w_۱$ و $w_۲$ را منظم کنیم (و نه $w_۰$ را)، آنگاه تابع هدف به شکل زیر خواهد بود:

$$\sum_{i=۱}^{n} \log P(y_i|x_i, w_۰, w_۱, w_۲) - C(|w_۱| + |w_۲|).$$

مجدداً مسئله‌ی طبقه‌بندی را با همین مدل در نظر بگیرید.

(آ) با افزایش مقدار پارامتر منظم‌سازی C، کدام یک از حالات زیر را انتظار دارید؟ (فقط یکی را انتخاب کنید) پاسخ خود را به‌صورت مختصر توضیح دهید.

- ابتدا $w_۱$ صفر می‌شود، سپس $w_۲$.
- ابتدا $w_۲$ صفر می‌شود، سپس $w_۱$.
- هر دو $w_۱$ و $w_۲$ به‌طور هم‌زمان صفر می‌شوند.
- هیچ‌یک دقیقاً صفر نمی‌شوند؛ فقط با افزایش C کوچک‌تر می‌شوند.

(ب) اگر C بسیار بزرگ باشد و تنها $w_۱$ و $w_۲$ با نُرم $L_۱$ منظم شوند (بدون منظم‌سازی $w_۰$)، انتظار دارید مقدار $w_۰$ چه باشد؟ توضیح دهید.
(راهنما: فرض کنید تعداد نمونه‌های هر کلاس برابر است.)

(ج) حال فرض کنید از کلاس + که با $y = ۱$ مشخص می‌شود، نمونه‌های بیشتری به‌دست می‌آید و برچسب‌ها نامتوازن می‌شوند. در این حالت نیز برای مقدار بزرگ C و همان منظم‌سازی $L_۱$، انتظار دارید $w_۰$ چه مقداری بگیرد؟ توضیح دهید.

---

۱۱. (الف) گرادیان تابع هزینه‌ی رگرسیون لاجستیک برای مسئله‌ی دسته‌بندی دوکلاسه را به‌دست آورید و استفاده از آن را با حالت رگرسیون مقایسه کنید.

$$f(X) = \frac{۱}{۱ + e^{-W^\top X}}$$

$$L = -\sum_{i=۱}^{N} \left(Y_i \log(f(X_i)) + (۱ - Y_i)\log(۱ - f(X_i))\right) + \frac{\lambda}{۲}\sum_{j=۱}^{M}(w_j)^۲$$

---

۱۲. فرض کنید $Y \sim \text{Bernoulli}(p)$ باشد؛ یعنی متغیر تصادفی باینری با پارامتر p باشد.

همچنین فرض کنید N مشاهدهٔ مستقل و هم‌توزیع از Y داریم.

(الف) تابع درست‌نمایی این مشاهدات را به‌دست آورید. سپس لگاریتم درست‌نمایی را محاسبه کنید. در نهایت عبارت

$$-\log(\text{Likelihood})$$

را به‌دست آورید.

(ب) فرض کنید رابطهٔ بین p و x را به‌صورت زیر مدل می‌کنیم:

$$\log(\frac{p}{1-p}) = ax + b$$

این رابطه در قسمت قبل، به چه مدل مشهوری منجر می‌شود؟ (فرض کنید برای هر مشاهده از Y، مقدار متناظر x نیز داریم و این مشاهدات نیز مستقل و هم‌توزیع هستند.)

---

۱۳. در مسئله‌ی طبقه‌بندی، معمولاً تابع خطایی که هدف کمینه‌سازی آن است، خطای ۱/۰ می‌باشد:

$$l(f(x), y) = \mathbb{I}\{f(x) \neq y\}$$

که در آن $f(x), y \in \{0, 1\}$ (یعنی طبقه‌بندی دودویی در نظر گرفته شده است).

در این سؤال، اثر استفاده از یک تابع خطای نامتقارن به‌صورت زیر بررسی می‌شود:

$$\ell_{\alpha,\beta}(f(x), y) = \alpha \cdot \mathbb{I}\{f(x) = 1,\ y = 0\} + \beta \cdot \mathbb{I}\{f(x) = 0,\ y = 1\}$$

در این تابع خطا، دو نوع خطا وزن‌های متفاوتی دارند که با ضرایب $\alpha, \beta > 0$ مشخص می‌شوند.

(الف) فرض کنید کلاس $y = 0$ بسیار نادر باشد (یعنی $P(y = 0)$ کوچک باشد). در این حالت، یک دسته‌بند ساده مانند $f(x) = 1$ برای همه‌ی x، ممکن است ریسک قابل‌قبولی داشته باشد. نشان دهید چگونه انتخاب مقادیر مناسب برای α و β در تابع خطای $l_{\alpha,\beta}$ می‌تواند منجر به همین رفتار شود و ریسک این طبقه‌بند را کمینه کند.

(ب) مسئله‌ی طبقه‌بندی زیر را در نظر بگیرید: ابتدا برچسب Y از توزیع $\text{Bernoulli}\left(\frac{1}{2}\right)$ انتخاب می‌شود، یعنی با احتمال $\frac{1}{2}$ مقدار ۱ دارد. سپس، اگر $Y = 1$ باشد، متغیر X از توزیع $\text{Bernoulli}(p)$ نمونه‌برداری می‌شود؛ و اگر $Y = 0$ باشد، X از $\text{Bernoulli}(q)$ با $p > q$ گرفته می‌شود.

در این حالت، طبقه‌بند بهینه‌ی بیز و ریسک (خطر) آن را بیابید.

(ج) تابع خطای معمولی ۱/۰ را در نظر بگیرید و فرض کنید $P(y = 0) = P(y = 1) = \frac{1}{2}$. همچنین، فرض کنید چگالی‌های شرطی کلاس‌ها گوسی با ویژگی‌های زیر باشند: کلاس ۰ دارای میانگین μ_0 و کواریانس Σ_0، و کلاس ۱ دارای میانگین μ_1 و کواریانس Σ_1 است. علاوه بر این، فرض می‌شود $\mu_0 = \mu_1$.

برای حالت زیر، کانتورهای تابع چگالی شرطی برای هر کلاس را رسم کنید و آن‌ها را به‌ترتیب با $p(x|y = 0)$ و $p(x|y = 1)$ برچسب بزنید. همچنین، مرز تصمیم‌گیری بیز را مشخص کنید و نواحی مربوط به پیش‌بینی کلاس‌های ۰ و ۱ را نشان دهید:

$$\Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$

---

۱۴. در بسیاری از مسائل واقعی، داده‌ها ابعادی در حد میلیون دارند، ولی هر نمونه تنها شامل چند صد ویژگی غیرصفر است. در این سؤال، هدف ما بهینه‌سازی الگوریتم گرادیان کاهشی تصادفی (SGD) با تنظیم ℓ_2 برای داده‌های پراکنده است.

در رگرسیون لجستیک با تنظیم ℓ_2، هدف، بیشینه‌سازی تابع زیر است (برای سادگی، پارامتر بایاس w_0 در نظر گرفته نشده است):

$$F(w) = \frac{1}{N} \sum_{j=1}^{N} l(x^{(j)}, y^{(j)}, w) - \frac{\lambda}{2} \sum_{i=1}^{d} w_i^2$$

که در آن:

$$l(x^{(j)}, y^{(j)}, w) = y^{(j)} \left(\sum_{i=1}^{d} w_i x_i^{(j)} \right) - \ln \left(1 + \exp \left(\sum_{i=1}^{d} w_i x_i^{(j)} \right) \right)$$

هنگام انجام گرادیان کاهشی روی یک نمونه تصادفی $(x^{(j)}, y^{(j)})$، تابع هدف به صورت تقریبی برابر است با:

$$F(w) \approx l(x^{(j)}, y^{(j)}, w) - \frac{\lambda}{2} \sum_{i=1}^{d} w_i^2$$

تعریف پراکندگی (Sparsity): فرض کنید داده‌ها در فضای $\mathbb{R}^d$ تعریف شده‌اند. فرض کنید در هر نمونه به طور میانگین فقط s مؤلفه از d مؤلفه غیرصفر هستند، به‌طوری‌که $s \ll d$. در این حالت، داده‌ها را پراکنده می‌نامیم.

سؤال ۱. فرض کنید $\lambda = 0$. قانون به‌روزرسانی الگوریتم گرادیان کاهشی تصادفی (SGD) را برای بردار وزن w با اندازه‌گام η، بر اساس یک نمونه‌ی تصادفی $(x^{(j)}, y^{(j)})$ بنویسید.

سؤال ۲. اگر از ساختار داده چگال (Dense) استفاده شود، میانگین زمان اجرای هر به‌روزرسانی w چقدر خواهد بود؟ در صورت استفاده از ساختار داده پراکنده (Sparse) چطور؟ تفاوت‌ها را توضیح دهید.

سؤال ۳. حال فرض کنید $\lambda > 0$. قانون به‌روزرسانی الگوریتم SGD را برای هر مؤلفه w_i با اندازه‌گام η و با توجه به نمونه $(x^{(j)}, y^{(j)})$ بنویسید.

سؤال ۴. اگر داده‌ها به صورت چگال ذخیره شوند، زمان میانگین اجرای هر به‌روزرسانی در حالت $\lambda > 0$ چقدر خواهد بود؟

سؤال ۵. فرض کنید $w^{(t)}$ بردار وزن پس از به‌روزرسانی tـاُم باشد. حال فرض کنید k به‌روزرسانی متوالی با نمونه‌هایی انجام شود که در همه‌ی آن‌ها ویژگی iـاُم صفر باشد، یعنی:

$$x_i^{(t+1)} = x_i^{(t+2)} = \cdots = x_i^{(t+k)} = 0$$

مقدار $w_i^{(t+k)}$ را به صورت تابعی از $w_i^{(t)}$، k، η و λ بنویسید.

سؤال ۶. با استفاده از پاسخ سؤال قبل، الگوریتمی کارآمد برای پیاده‌سازی SGD با تنظیم ℓ_2 بر روی داده‌های پراکنده طراحی کنید. این الگوریتم باید از ویژگی‌های غیرصفر به‌طور مؤثر استفاده کند. زمان میانگین اجرای هر به‌روزرسانی (بر حسب s و d) چقدر خواهد بود؟ چه زمانی باید هر مؤلفه‌ی w_i را به‌روزرسانی کرد؟

---

۱۵. فرض کنید طبقه‌بند بیز ساده ویژگی‌ها $x_1, x_2, \ldots, x_n$ را شرطی و مستقل نسبت به برچسب y در نظر می‌گیرد، یعنی:

$$P(y \mid X = (x_1, \ldots, x_n)) \propto P(X, y) = P(y) \cdot \prod_{i=1}^{n} P(x_i \mid y).$$

بنابراین، قاعده‌ی پیش‌بینی طبقه به صورت زیر است:

$$\hat{y} = \arg\max_{y} \; P(y) \prod_{i=1}^{n} P(x_i \mid y).$$

حال، طبقه‌بند غیرخطی زیر را در نظر بگیرید که قاعده طبقه‌بندی آن به شکل زیر است:

$$\hat{y} = \arg\max_{y} \; S \left(w_0 + \sum_{i=1}^{n} w_{y,i} x_i \right),$$

که در آن S تابع سیگموید است.

۱. فرض کنید ویژگی‌ها دودویی هستند، یعنی $x_i \in \{۰, ۱\}$. در این حالت، قاعده طبقه‌بندی بیز ساده را طوری بازنویسی کنید که به شکل طبقه‌بند غیرخطی بالا درآید. (نکته: از خاصیت دودویی بودن x_i استفاده کنید.)

۲. وزن‌های $w_۰$ و $w_{y,i}$ برای $i = ۱, \dots, n$ را برحسب احتمال‌های بیز ساده $P(y)$ و $P(x_i \mid y)$ بیان کنید. فرض کنید همه احتمال‌ها غیرصفر هستند.

۳. فرض کنید ویژگی‌ها دیگر دودویی نیستند. مسئله طبقه‌بندی دوکلاسه (با علامت $+$ و $\diamond$) در فضای دوبعدی را با دو طبقه‌بند رگرسیون لجستیک و بیز ساده گاوسی مقایسه کنید. برای بیز ساده گاوسی فرض کنید واریانس هر دو کلاس در هر بعد برابر است.

برای هر یک از سه مجموعه داده (که در مستندات آمده است)، مرز تصمیم رگرسیون لجستیک را با خط پیوسته و مرز تصمیم بیز ساده گاوسی را با خط‌چین رسم کنید. اگر هرکدام از طبقه‌بندها قادر به طبقه‌بندی درست داده نبودند، یک جمله کوتاه علت را در سمت راست شکل بنویسید.

توجه: فقط برای مواردی که داده‌ها قابل طبقه‌بندی درست هستند، مرز تصمیم را رسم کنید.

---

۱۶. دو فرضیه‌ی زیر را برای طبقه‌بندی دودویی با استفاده از تابع سیگموید ϕ_{sig} در نظر بگیرید:

- فرضیه ۱:

$$\phi_{\text{sig}}(w_۱ x^{(۱)} + w_۲ x^{(۲)})$$

- فرضیه ۲:

$$\phi_{\text{sig}}(w_۰ + w_۱ x^{(۱)} + w_۲ x^{(۲)})$$

داده‌های آموزش به صورت زیر هستند:

$$x^{(۱)} = \begin{bmatrix} ۱ \\ ۰ \end{bmatrix}, \quad y^{(۱)} = ۱; \quad x^{(۲)} = \begin{bmatrix} ۱ \\ ۱ \end{bmatrix}, \quad y^{(۲)} = -۱; \quad x^{(۳)} = \begin{bmatrix} ۰ \\ ۰ \end{bmatrix}, \quad y^{(۳)} = ۱.$$

۱. آیا برچسب داده‌ی سوم در یادگیری وزن‌های فرضیه‌ی اول تأثیری دارد؟ در فرضیه‌ی دوم چطور؟ پاسخ خود را توضیح دهید.

۲. آیا پرسپترون یادگرفته شده، حاشیه بین داده‌های آموزش و مرز تصمیم‌گیری را بیشینه کرده است؟ توضیح دهید.

---

۱۷. همان‌طور که می‌دانید، الگوریتم رگرسیون لجستیک معمولاً برای حالت دسته‌بندی باینری به کار می‌رود، اما به سادگی قابل تعمیم به حالت چندکلاسه (با K کلاس) نیز هست.

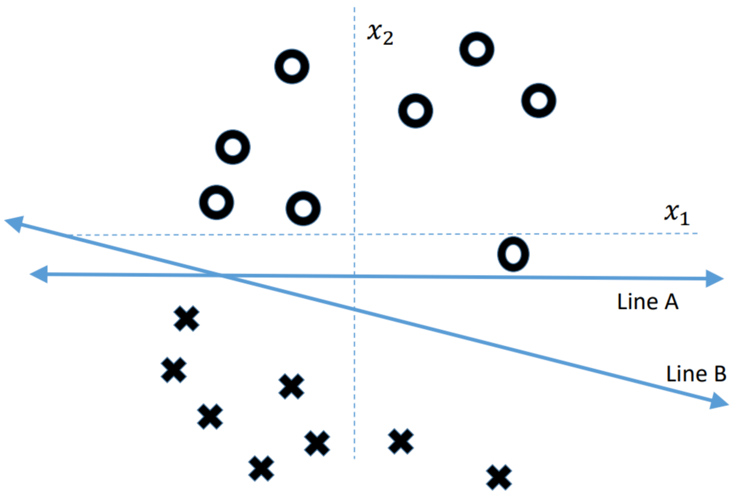
(الف) چه تغییراتی لازم است انجام دهیم تا الگوریتم از حالت باینری به حالت Kـ کلاسه تبدیل شود؟

(ب) شبه‌کد کامل الگوریتم تغییر یافته را با فرض حل مسئله بهینه‌سازی با استفاده از گرادیان کاهشی تصادفی (SGD) بنویسید.

---

۱۸. هر داده ما دو ویژگی (feature) دارد که یکی با محور افقی و دیگری با محور عمودی نشان داده شده است. هر داده به یکی از دو کلاس تعلق دارد، کلاس ۱ همان ”X“ است و کلاس ۱− همان ”O“ است. مدل رگرسیون لجستیک ما به صورت زیر است:

$$P(y = ۱ \mid x, w) = \frac{۱}{۱ + \exp\big(-w_۰ - w_۱ x[۱] - w_۲ x[۲]\big)}$$



(الف) دو خط A, B که در تصویر نشان داده شده است، مربوط به نتیجه آموزش ما روی این داده‌ها هستند. مقدار $w = [w_0, w_1, w_2]$ که خطوط A و B را تولید کنند را بیابید (برای هر خط یک مقدار پیشنهاد کنید). خط A خط افقی است که محور عمودی را در -1 قطع می‌کند و خط B با شیب ۲ محور عمودی را در -2 قطع می‌کند.

(ب) نشان دهید برای مقادیر w ای که در بخش الف یافتید، مقدار $2w$ نیز همان دو خط جداکننده را نشان می‌دهد (همان A, B).

(ج) در درس آموختیم که تابع loss مناسب برای این مسئله

$$J(w) = \sum_{i=1}^{n} \log\left(1 + \exp(-y_i x_i^T w)\right)$$

است. اگر سعی کنیم این loss را کمینه کنیم، w ای که در بخش الف یافتیم، برابر با جواب

$$\hat{w} = \arg\min_{w} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i x_i^T w)\right)$$

نمی‌شود. چرا؟ با توجه به بخش ب جواب دهید. چرا این موضوع یک مشکل ایجاد می‌کند؟

(د) حال فرض کنید که ما به تابع loss، بخش regularizer برابر با $\lambda(|w_1| + |w_2|)$ را اضافه می‌کنیم. برای مقادیر بزرگ λ کدام یک از دو خط A, B انتظار می‌رود loss کمتری داشته باشند؟ (یعنی مقدار J در بالا به علاوه قسمت regularizer کمتر داشته باشد). توجه داشته باشید که این بخش صرفاً مفهومی است و نیازی به محاسبه عددی نیست.

---

۱۹. موارد درست را با ذکر دلیل بیان کنید:

فرض کنید

$$S = \{(x^{(1)}, y^{(1)}), \cdots, (x^{(n)}, y^{(n)})\}$$

مجموعه‌ای شامل n نقطه‌ی خطی‌قابل‌تفکیک در $\mathbb{R}^d$ باشد که توسط یک جداکننده عبوری از مبدأ قابل تفکیک‌اند.

مجموعه‌ی S' از S به صورت زیر ساخته شده است: $S' = \{(cx^{(1)}, y^{(1)}), \cdots, (cx^{(n)}, y^{(n)})\}$

که در آن $c > 1$ یک ثابت است.

فرض کنید الگوریتم پرسپترون را به‌طور جداگانه روی هر دو مجموعه داده اجرا می‌کنیم و می‌دانیم که این الگوریتم روی S همگرا می‌شود.

کدام‌یک از گزاره‌های زیر درست‌اند؟

(الف) کران خطا الگوریتم پرسپترون روی S' بزرگ‌تر از کران خطای آن روی S است.

(ب) الگوریتم پرسپترون زمانی که روی S و S' اجرا شود، یک طبقه‌بند مشابه (تا یک ضریب ثابت) برمی‌گرداند. (یعنی اگر w_S و $w_{S'}$ خروجی‌های الگوریتم پرسپترون برای S و S' باشند، آنگاه $w_S = c_1 w_{S'}$ برای یک ثابت c_1 برقرار است.)

(ج) الگوریتم پرسپترون روی S' همگرا می‌شود.

---

۲۰. یک مدل regression logistic را در نظر بگیرید. می‌دانیم که:

$$y = g(\mathbf{w}^T \mathbf{x})$$

و تابع هزینه آن **binary cross entropy** است که به صورت زیر تعریف می‌شود:

$$J = -\big(y \log(g(z)) + (1 - y) \log(1 - g(z))\big)$$

که در آن:

$$g(z) = \frac{1}{1 + e^{-z}}.$$

حال تابع هزینه روش تغییر داده شده از regression logistic به صورت زیر است:

$$J_{\text{new}} = \frac{e^{-z}}{1 + e^{-z}}$$

که همچنان همان**binary cross entropy** است.

پارامترها و پیش‌بینی‌های مدل یادگرفته شده جدید نسبت به مدل اولیه چه تفاوتی دارند؟ لطفاً به صورت ریاضی توضیح دهید.

---

۲۱. آقای دانیال مسئول باجه سینما شده است و اخیراً تعدادی فیلم با ویژگی‌ها و ژانرشان را دریافت کرده است و می‌خواهد با استفاده از الگوریتم پرسپترون چندکلاسه، مدل‌هایی بسازد که با دریافت ویژگی‌های مرتبط به هر فیلم جدید بتواند ژانر آن را پیش‌بینی کند. او برای سادگی بیشتر صرفاً سه ژانر کمدی، درام و اکشن را در نظر می‌گیرد که محبوبیت بیشتری دارند و برای هر فیلم یک بردار ویژگی در نظر می‌گیرد. داده‌های آموزش به شکل زیر هستند:

$$X = \begin{bmatrix} -4 & 2 & -1 \\ 2 & -3 & 4 \\ 1 & 1 & 1 \end{bmatrix}, \quad Y = \begin{bmatrix} \text{Drama} \\ \text{Action} \\ \text{Comedy} \end{bmatrix}$$

بردارهای وزن مختص به هر کلاس به شرح زیر است:

$$\mathbf{w}_{Comedy} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \mathbf{w}_{Action} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{w}_{Drama} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

الگوریتم پرسپترون چندکلاسه را یک بار به ترتیب نمونه‌های داده شده روی این مجموعه اجرا کنید و در هر مرحله:

(الف) پیش‌بینی ژانر نمونه مربوطه را بنویسید.

(ب) وزن‌های به‌روزرسانی شده $\mathbf{w}$ را برای هر کلاس گزارش دهید.

راهنمای حل:

در هر نمونه $(x^{(i)}, y^{(i)})$،

- پیش‌بینی کلاس به صورت

$$\hat{y} = \arg\max_{c\in\{\text{Drama، Action، Comedy}\}} \mathbf{w}_c^\top x^{(i)}$$

محاسبه می‌شود.

- اگر $\hat{y} \neq y^{(i)}$، وزن‌ها به صورت زیر به‌روزرسانی می‌شوند:

$$\mathbf{w}_{y^{(i)}} \leftarrow \mathbf{w}_{y^{(i)}} + x^{(i)}, \quad \mathbf{w}_{\hat{y}} \leftarrow \mathbf{w}_{\hat{y}} - x^{(i)}.$$

وزن سایر کلاس‌ها بدون تغییر می‌ماند.

- اگر $\hat{y} = y^{(i)}$، هیچ تغییری در وزن‌ها صورت نمی‌گیرد.

---

۲۲. آقای نیان یک کارگردان است که تعدادی فیلم‌نامه در اختیار دارد. او قبل از شروع ساخت فیلم‌ها می‌خواهد پیش‌بینی کند که آیا هر فیلم خاص به سود قابل‌توجهی خواهد رسید یا خیر. برای این منظور، دو منتقد به نام‌های عماد و ایلیا استخدام می‌کند که هر یک از فیلم‌نامه‌ها را به‌صورت مستقل بررسی کرده و به آن‌ها نمره‌ای بین ۱ تا ۵ می‌دهند (فرض کنید نمرات تنها اعداد صحیح هستند). نظرات عماد و ایلیا روی یکدیگر تأثیری ندارند، اما ممکن است کاملاً دقیق نباشند.

در جدول زیر، فیلم‌نامه‌ها و امتیازات داده شده توسط عماد و ایلیا همراه با وضعیت سوددهی آن‌ها آمده است:

فیلم‌نامه	امتیاز عماد	امتیاز ایلیا	سوددهی
نیان و برادران	۱	۱	خیر
نیان در سرزمین عجایب	۳	۲	بله
نیان برگرد!	۴	۵	خیر
حمله به نیان	۳	۴	بله
نیان، کیمیاگر تمام فلزی	۲	۳	بله

اکنون نیان می‌خواهد با استفاده از یادگیری ماشین، یک دسته‌بند بسازد که بتواند پیش‌بینی کند هر یک از فیلم‌هایش سودده خواهند بود یا خیر. در این مرحله، قصد دارد از الگوریتم پرسپترون استفاده کند؛ البته کمی متفاوت: اگرچه مسئله، دوکلاسه است، اما می‌خواهد از نسخه چندکلاسه آن بهره ببرد.

فرض کنید ویژگی‌هایی که نیان استفاده می‌کند (همراه با یک bias) به صورت زیر تعریف شده‌اند:

$$f_۰ = ۰, \quad f_۱ = \text{امتیاز عماد}, \quad f_۲ = ۲ \times \text{امتیاز ایلیا}$$

(آ) فرض کنید نیان پرسپترون را با وزن‌های اولیه زیر آموزش می‌دهد:

$$\mathbf{w}_{\text{بله}} = \begin{bmatrix} ۰ \\ ۰ \\ -۱ \end{bmatrix}, \quad \mathbf{w}_{\text{خیر}} = \begin{bmatrix} ۰ \\ ۰ \\ ۱ \end{bmatrix}$$

i. نیان در اولین مرحله آموزش کدام فیلم‌نامه را انتخاب می‌کند؟ چرا؟

ii. وزن‌های به‌روزرسانی شده برای هر دو کلاس را پس از مرحله اول بنویسید.

(الف) بدون توجه به داده‌های آموزش موجود، فرض کنید نیان می‌خواهد بررسی کند که آیا ویژگی‌های فعلی در شرایط مختلف قدرت تشخیص کافی دارند یا خیر. لطفاً مشخص کنید در کدام یک از سناریوهای زیر، پرسپترون می‌تواند با ویژگی‌های مذکور داده‌ها را با دقت کامل دسته‌بندی کند (دلایل خود را نیز ذکر کنید):

i. منتقدان عالی هستند: اگر مجموع امتیازات آن‌ها بیشتر از ۸ باشد، فیلم حتماً سودده است؛ در غیر این صورت سودده نیست.

ii. منتقدان هنری هستند: در این حالت فیلم سودده است اگر و تنها اگر هر منتقد امتیاز ۲ یا ۳ بدهد.

iii. منتقدان نظراتشان عجیب ولی متفاوت است: در این حالت فیلم سودده است اگر و تنها اگر هر دو منتقد با هم موافق باشند.

(ب) نیان که از امتیازات عماد و ایلیا خسته شده بود، تصمیم می‌گیرد ویژگی‌های جدیدی جایگزین آن‌ها کند. ویژگی‌های جدید (به همراه bias) به صورت زیر تعریف شده‌اند:

$$f_0 = 1,$$

$$E_i = \begin{cases} 1 & \text{اگر امتیاز عماد برابر } i \text{ باشد} \\ 0 & \text{در غیر این صورت} \end{cases} \quad , \quad I_i = \begin{cases} 1 & \text{اگر امتیاز ایلیا برابر } i \text{ باشد} \\ 0 & \text{در غیر این صورت} \end{cases} \quad , \quad i = 1, 2, \ldots, 5$$

مجدداً با توجه به ویژگی‌های جدید به بخش (ب) پاسخ دهید. ممکن است چند سناریو درست باشند.

درخت تصمیم

۲۳. مجموعه داده‌های آموزش زیر در اختیار داریم:

y	x2	x1
+	T	T
+	F	T
−	T	F
−	F	F
+	F	T
+	F	F

(الف) با در نظر گرفتن هر سه حالت پایه و با محاسبه معیار بهره‌وری اطلاعات (**Information Gain**) ، درخت تصمیم مناسب برای دسته‌بندی این داده‌ها را به‌دست آورده و رسم نمایید.

(ب) اگر داده‌های سنجش زیر در اختیار باشد، دقت مدل را محاسبه کنید. (در صورت تساوی احتمال برچسب مثبت و منفی، برچسب مثبت + انتخاب شود.)

y	x2	x1
+	T	F
+	F	T
−	T	F
−	F	F

(ج) قصد داریم با توجه به داده‌های سنجش و به کمک روش کاهش خطا (**Error Reduction**)، درخت تصمیم را هرس کنیم. درخت هرس‌شده نهایی را رسم کرده و دلایل هرس یا عدم هرس هر گره را توضیح دهید.

۲۴. (الف) فرض کنید X یک متغیر تصادفی است که فقط دو مقدار ۰ و ۱ را می‌پذیرد. اگر احتمال وقوع ۰ را p در نظر بگیریم و $H(X)$ را به صورت تابع $h(p)$ تعریف کنیم، ثابت کنید که $h(p)$ یک تابع مقعر است. آیا این گزاره برای متغیر تصادفی با n مقدار نیز صادق است؟ دلیل خود را بیان کنید.

(ب) با استفاده از نتیجه بخش قبل، بیشینه (ماکزیمم) مقدار $H(X)$ را برای یک متغیر تصادفی با n مقدار به دست آورید.

(ج) آیا به طور کلی می‌توان حد بالایی برای $H(X)$ تعیین کرد؟ توضیح دهید.

---

۲۵. متأسفانه تعداد زیادی موجود فضایی ناشناخته به شهر ایوب یورک ($AyoubYork$) حمله کرده‌اند؛ اما همه آن‌ها برای مردم خطرناک نیستند و بعضاً رفتار دوستانه‌ای دارند. برای تشخیص این ویژگی، مجموعه داده‌های آموزشی در اختیار ما قرار گرفته است که شامل ویژگی‌های وزن (که می‌تواند دو مقدار چاق و عادی باشد)، رنگ چشم (دو مقدار آبی و بنفش) و تعداد چشم (که می‌تواند ۳، ۲ یا ۴ باشد) است.

نمای کلی این مجموعه داده به شرح زیر است:

خطرناک؟	تعداد چشم	رنگ چشم	وزن
خیر	۲	آبی	عادی
خیر	۲	بنفش	عادی
خیر	۲	بنفش	عادی
خیر	۳	بنفش	چاق
خیر	۳	بنفش	چاق
بله	۴	آبی	چاق
بله	۴	آبی	عادی
بله	۴	بنفش	عادی
بله	۳	آبی	چاق
بله	۳	آبی	چاق

در تمام بخش‌های زیر، در صورت تساوی بین تعداد برچسب‌ها در یک برگ، موجود فضایی را بی‌خطر در نظر بگیرید.

(الف) درخت تصمیم را با نشان دادن تمام مراحل و محاسبات رسم کنید.

(ب) اگر بخواهیم درخت را یک مرحله زودتر خاتمه دهیم و عمق آن را یک سطح کمتر کنیم، برای هر برگ تعیین کنید چه برچسبی اختصاص داده خواهد شد.

(ج) فرض کنید به دلیل حساسیت بالاتر در شناسایی موجودات خطرناک و حفظ امنیت شهر، جریمه‌ی تشخیص اشتباه نمونه‌های خطرناک ۴ برابر جریمه‌ی اشتباه تشخیص موجودات بی‌آزار باشد. در این صورت، برگ‌های مرحله قبل چگونه برچسب‌گذاری خواهند شد؟

(د) برای بخش‌های (الف) و (ب)، مقادیر دقت (Accuracy)، دقت مثبت کاذب (Precision)، بازیابی (Recall) و **F1 Score** را روی داده‌های آموزشی محاسبه کنید.

---

۲۶. (الف) اشکال تخمین زدن احتمال $P(A|B)$ به صورت تعداد رخداد B چیست؟

(ب) توضیح دهید هر کدام از روش‌های زیر چه تأثیری در بیش‌برازش مدل‌ها دارند:

- کم کردن تعداد برگ‌ها در درخت تصمیم

- محدود کردن حداکثر طول درخت تصمیم

دربارهٔ تأثیر آن‌ها در دقت مدل در داده‌های آموزش چه می‌توان گفت؟

---

۲۷. در هر مورد با ذکر دلیل درست یا غلط بودن گزاره مورد ذکر را بررسی کنید.

(الف) تابع softmax معمولاً در مسائل دسته‌بندی چندکلاسه به‌کار می‌رود، زیرا خروجی آن یک بردار احتمال است که مجموع مؤلفه‌های آن برابر با ۱ بوده و هر مقدار، احتمال تعلق نمونه به یکی از دسته‌ها را نشان می‌دهد.

(ب) بالا بودن واریانس مدل به‌معنای حساسیت زیاد آن نسبت به تغییرات جزئی در داده‌های آموزشی است که منجر به بیش‌برازش می‌شود. برای کاهش واریانس، معمولاً از روش‌هایی مانند ضریب‌زدایی (منظم‌سازی) از نوع L1 یا L2، کاهش پیچیدگی مدل، یا افزایش حجم داده‌های آموزشی استفاده می‌شود.

(ج) در شرایطی که ویژگی‌های ورودی هم‌بسته (وابسته به یکدیگر) هستند، استفاده از رگرسیون Ridge به دلیل دارا بودن جمله‌ی ضریب‌زدا از نوع L2 مناسب‌تر است، زیرا باعث پایداری بیشتر و همگرایی بهتر مدل در مقایسه با Lasso می‌شود.

(د) ضریب‌زدایی از نوع L2 معمولاً موجب کاهش واریانس مدل و در عین حال افزایش اندکی در بایاس می‌شود. این تبادل میان بایاس و واریانس، در بسیاری از موارد باعث بهبود عملکرد مدل در تعمیم به داده‌های جدید خواهد شد.

PCA و سایر

۲۸. در یک مسئله‌ی طبقه‌بندی دوکلاسه با دو ویژگی، از هر کلاس دو داده در اختیار داریم:

$$\omega_۱ : \begin{bmatrix} ۴ \\ ۱ \end{bmatrix}, \quad \begin{bmatrix} ۱ \\ ۰ \end{bmatrix} \qquad \omega_۲ : \begin{bmatrix} ۲ \\ ۵ \end{bmatrix}, \quad \begin{bmatrix} ۳ \\ ۴ \end{bmatrix}$$

(الف) با محاسبه‌ی بردار میانگین و ماتریس کوواریانس، داده‌ها را با استفاده از PCA به فضای یک‌بعدی ببرید و تمایزپذیری کلاس‌ها را بررسی کنید. مسئله را برای هر دو راستا (بردار ویژه) حل کنید.

(ب) در قسمت (الف) برداری پیشنهاد دهید که اگر به همه‌ی داده‌ها اضافه شود، مؤلفه‌ی اصلی اول (بردار ویژه‌ی اول) تغییر نکند.

---

۲۹. تحلیل مؤلفه‌های اصلی (PCA) یک الگوریتم کاهش ابعاد است که ماتریس داده را به‌صورت مجموعه‌ای از مؤلفه‌های اصلی نمایش می‌دهد که نسبت به یکدیگر متعامد (orthogonal) هستند. هر یک از این مؤلفه‌های اصلی، یک منبع از تغییرات موجود در ماتریس داده اولیه را ثبت می‌کند، به‌طوری که می‌توان ابعاد داده را کاهش داد و در عین حال، تا جای ممکن اطلاعات آن را حفظ کرد.

مؤلفه‌های اصلی به‌صورت بردارهای ویژه (eigenvectors) ماتریس کوواریانس داده تعریف می‌شوند؛ بنابراین، ما PCA را با استفاده از تجزیه ویژه (eigendecomposition) پیاده‌سازی خواهیم کرد.

اثر ماتریس M، که با $\mathrm{Tr}(M)$ نشان داده می‌شود، برابر است با مجموع درایه‌های قطری ماتریس M.

اکنون چند قضیه را اثبات می‌کنیم که در مسئله‌ی بعدی مفید خواهند بود. تعریف می‌کنیم:

$$\Sigma := \frac{۱}{n} X^T X$$

که در آن X یک ماتریس داده $n \times d$ است. فرض کنید x_i سطر iـ‌ام ماتریس X است (یعنی x_i یک بردار dـ‌بعدی است). همچنین، فرض کنید $\lambda_۱ \geq \lambda_۲ \geq \cdots \geq \lambda_d$ مقادیر ویژه‌ی ماتریس Σ باشند.

(الف) دو خاصیت زیر را اثبات کنید:

- $\text{Tr}(\Sigma) = \sum_{i=1}^{d} \lambda_i$
- $\text{Tr}(\Sigma) = \frac{1}{n-1} \sum_{i=1}^{n} \|x_i\|_2^2$

نکته ۱: ماتریس‌های متقارن به‌صورت متعامد قطری‌پذیر هستند، یعنی یک ماتریس متقارن A را می‌توان به‌شکل $A = UDU^T$ نوشت که در آن D یک ماتریس قطری حاوی مقادیر ویژه‌ی A در قطر اصلی و U یک ماتریس متعامد است که ستون‌هایش بردارهای ویژه‌ی A هستند.

نکته ۲: برای دو ماتریس A و B داریم:

$$\text{Tr}(AB^T) = \text{Tr}(B^T A)$$

(ب) فرض کنید P یک ماتریس افکنش $d \times k$ است که $P^T P = I_k$. افکنش از $\mathbb{R}^{n\times d}$ به $\mathbb{R}^{n\times k}$ به‌شکل زیر تعریف می‌شود:

$$X \mapsto XP$$

- نشان دهید :

$$\arg \max_{P\in\mathbb{R}^{d\times k}, P^T P=I} \text{Tr}\left(\frac{1}{n} P^T X^T X P\right) = P^*$$

که در آن هر ستون از P^* بردار ویژه‌ی u_i متناظر با λ_i است، برای $i = 1, ..., k$. توجه کنید که u_i و λ_i همان بردارها و مقدارهای ویژه‌ی Σ هستند.

- اثبات کنید که:

$$\text{Tr}\left(\frac{1}{n} P^{*T} X^T X P^*\right) = \sum_{i=1}^{k} \lambda_i$$

با استفاده از نتایج بدست آمده، داریم:

$$1 - \frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_d} = 1 - \frac{\sum_{i=1}^{n} \|P^{*T} x_i\|_2^2}{\sum_{i=1}^{n} \|x_i\|_2^2}$$

این معیار به‌عنوان خطای بازسازی PCA با استفاده از k جهت از d جهت اصلی شناخته می‌شود.

---

۳۰. در این سؤال، از داده‌های موجود در **The Database of Faces (AT&T)** استفاده می‌کنیم که این مجموعه‌داده شامل ۴۰۰ تصویر چهره افراد مختلف است، که هر تصویر ۱۱۲ × ۹۲ پیکسل است. تصاویر به‌صورت سیاه‌وسفید هستند، بنابراین هر پیکسل با یک عدد حقیقی بین ۰ (مشکی) و ۱ (سفید) نمایش داده می‌شود.

ماتریس Σ به‌صورت یک ماتریس ۱۰۳۰۴ × ۱۰۳۰۴ تعریف می‌شود:

$$\Sigma = \frac{1}{n-1} \sum_{i=1}^{n} x_i x_i^T$$

که در آن، بردارهای x_i نقاط موجود در مجموعه‌داده (به صورت بردارهای ستونی) هستند و n تعداد کل تصاویر است. حال، ۵۰ مؤلفه‌ی اصلی برتر (PCA Dimensions) را محاسبه کنید؛ این مؤلفه‌ها، ۵۰ بُعدی هستند که بهترین بازسازی از داده‌ها را انجام می‌دهند.

(الف) مقادیر ویژه $\lambda_1, \lambda_2, \lambda_{10}, \lambda_{30}, \lambda_{50}$ را محاسبه کنید. همچنین، مجموع مقادیر ویژه $\sum_{i=1}^{d} \lambda_i$ را بنویسید. (راهنما: از جواب سؤال قبلی استفاده کنید.)

(ب) از سوال قبل می‌دانیم که خطای بازسازی نسبی هنگام استفاده از k مؤلفه‌ی اصلی برابر است با:

$$1 - \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{d} \lambda_i}$$

این خطای نسبی را برای $k = 1$ تا $k = 50$ رسم کنید. محور افقی (X) برابر k و محور عمودی (Y) مقدار خطای نسبی باشد. نمودار را به صورت واضح و با برچسب‌گذاری مناسب ارائه دهید.

(ج) با استفاده از بخش‌های قبل توضیح دهید که اولین مقدار ویژه (eigenvalue) چه چیزی را نمایش می‌دهد، و چرا فکر می‌کنید λ_1 به‌طور قابل‌توجهی از بقیه بزرگ‌تر است.

---

۳۱. فرض کنید $\mathbf{x}_1, \ldots, \mathbf{x}_m$ بردارهایی در فضای $\mathbb{R}^d$ باشند و $\mathbf{x}$ یک بردار تصادفی با توزیع یکسان روی $\mathbf{x}_1, \ldots, \mathbf{x}_m$ باشد. همچنین فرض کنید $\mathbb{E}[\mathbf{x}] = 0$.

مسئله یافتن بردار واحد $\mathbf{w} \in \mathbb{R}^d$ را در نظر بگیرید به طوری که متغیر تصادفی $\langle \mathbf{w}, \mathbf{x} \rangle$ دارای واریانس ماکزیمم باشد. یعنی می‌خواهیم مسئله زیر را حل کنیم:

$$\underset{\mathbf{w}:\|\mathbf{w}\|=1}{\operatorname{argmax}} \operatorname{Var}\big[\langle \mathbf{w}, \mathbf{x} \rangle\big] = \underset{\mathbf{w}:\|\mathbf{w}\|=1}{\operatorname{argmax}} \frac{1}{m} \sum_{i=1}^{m} \langle \mathbf{w}, \mathbf{x}_i \rangle^2$$

نشان دهید که جواب این مسئله این است که $\mathbf{w}$ را برابر اولین مؤلفه اصلی بردارهای $\mathbf{x}_1, \ldots, \mathbf{x}_m$ قرار دهیم.

---

۳۲. فرض کنید $\mathbf{w}_1$ اولین مؤلفهٔ اصلی داده‌ها است (مطابق سؤال قبل). حال می‌خواهیم یک بردار واحد دیگر $\mathbf{w}_2 \in \mathbb{R}^d$ بیابیم که واریانس تصویر داده‌ها روی آن، یعنی $\langle \mathbf{w}_2, \mathbf{x} \rangle$، را بیشینه کند؛ با این شرط اضافه که $\langle \mathbf{w}_2, \mathbf{x} \rangle$ با $\langle \mathbf{w}_1, \mathbf{x} \rangle$ نامرتبط باشد.

یعنی مسئلهٔ زیر را می‌خواهیم حل کنیم:

$$\underset{\mathbf{w}\in\mathbb{R}^d:\|\mathbf{w}\|=1,\ \mathbb{E}[\langle \mathbf{w}_1, \mathbf{x} \rangle \langle \mathbf{w}, \mathbf{x} \rangle]=0}{\operatorname{argmax}} \qquad \operatorname{Var}[\langle \mathbf{w}, \mathbf{x} \rangle]$$

نشان دهید که بردار $\mathbf{w}$ که این شرط‌ها را ارضا می‌کند و واریانس تصویر را بیشینه می‌کند، همان دومین مؤلفهٔ اصلی داده‌ها است.

راهنمایی:

$$\mathbb{E}[\langle \mathbf{w}_1, \mathbf{x} \rangle \langle \mathbf{w}, \mathbf{x} \rangle] = \mathbf{w}_1^\top \left(\mathbb{E}[\mathbf{x}\mathbf{x}^\top] \right) \mathbf{w} = \mathbf{w}_1^\top A \mathbf{w},$$

که در آن $A = \frac{1}{m} \sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^\top$ ماتریس کوواریانس است.

از آن‌جا که $\mathbf{w}_1$ بردار ویژهٔ A است، شرط نامرتبط بودن معادل این است که:

$$\langle \mathbf{w}_1, \mathbf{w} \rangle = 0$$

بنابراین مسئلهٔ بیشینه‌سازی واریانس با این شرط، معادل یافتن بردار ویژه‌ای است که بیشترین مقدار ویژه را دارد و بر $\mathbf{w}_1$ عمود است؛ که دقیقاً تعریف دومین مؤلفهٔ اصلی است.

---

۳۳. قصد داریم مدلی بسازیم که با گرفتن ورودی X، متغیر خروجی Y را تخمین بزند. فرض کنید M مدل ضعیف به نام‌های $f_1(X), f_2(X), \ldots, f_M(X)$ داریم که روی نمونه‌های bootstrap شده‌ای از دیتاست اصلی آموزش داده شده‌اند و دارای بایاس و واریانس یکسان هستند. مدل نهایی به شکل زیر تعریف می‌شود:

$$f_{\text{ensemble}}(X) = \frac{1}{M}\sum_{i=1}^{M} f_i(X)$$

(الف) بایاس و واریانس مدل $f_{\text{ensemble}}(X)$ را بر حسب بایاس و واریانس مدل‌های ضعیف محاسبه کنید. در این قسمت فرض کنید مدل‌های ضعیف از یکدیگر مستقل هستند. تحلیل کنید که افزایش تعداد مدل‌های ضعیف M چه تأثیری روی بایاس و واریانس مدل نهایی دارد.

(ب) حال فرض کنید مدل‌های ضعیف مستقل نیستند و ضریب همبستگی (correlation) بین هر دو مدل $f_i(X)$ و $f_j(X)$ (برای $i \neq j$) برابر ρ است. حال مانند قسمت الف، بایاس و واریانس مدل نهایی را محاسبه کرده و تأثیر تعداد مدل‌ها M و وابستگی بین آن‌ها ρ را روی این مقادیر بررسی کنید.

(ج) به سؤالات زیر به‌طور خلاصه پاسخ دهید:

- آیا یادگیرنده‌های ضعیف در AdaBoost نیاز به مشتق‌پذیر بودن دارند؟ چرا؟
- بین Boosting و Bagging، کدام یک از نظر محاسباتی گران‌تر می‌باشد؟ چرا؟

۳۴.(الف) یادگیری مبتنی بر نمونه (Instance-Based Learning) چیست؟ این روش چه تفاوتی با یادگیری مدل‌محور (Model-Based Learning) دارد؟ به کمک یک مثال عددی یا هندسی این تفاوت را توضیح دهید.

(ب) در الگوریتم k-NN می‌توان به جای وزن‌دهی مساوی به همسایگان، از وزن‌دهی فاصله‌محور استفاده کرد. فرض کنید وزن نقطه‌ی x_i نسبت به نقطه پرس‌وجو x به صورت زیر تعریف شده است:

$$w_i = \frac{1}{\|x - x_i\|^2 + \epsilon}$$

- نشان دهید که این وزن‌دهی باعث تأکید بیشتر بر نزدیک‌ترین همسایگان می‌شود.
- نقش پارامتر ϵ را بررسی کنید و توضیح دهید در چه شرایطی باید بزرگ یا کوچک باشد.

(ج) چرا الگوریتم‌های مبتنی بر نمونه مانند k-NN برای درون‌یابی (Interpolation) مناسب هستند ولی برای برون‌یابی (Extrapolation) ضعیف عمل می‌کنند؟ پاسخ خود را با تحلیل هندسی یا یک تابع نمونه مانند $f(x) = \sin(x)$ برای نقاط محدود ارائه دهید.

۳۵.(الف) در الگوریتم k-NN برای دسته‌بندی یک نقطه جدید، باید فاصله آن با تمامی n نمونه موجود محاسبه شود. اگر بعد داده‌ها d باشد:

- پیچیدگی زمانی الگوریتم را برای پیش‌بینی یک نقطه جدید بنویسید.
- چه روش‌هایی برای کاهش این پیچیدگی پیشنهاد شده‌اند؟ یکی را توضیح دهید.

(ب) در فضای ویژگی‌ها، استفاده از معیار فاصله بر خروجی الگوریتم‌های مبتنی بر نمونه اثرگذار است. سه معیار مختلف برای محاسبه فاصله بین نمونه‌ها را بنویسید و تأثیر انتخاب آن‌ها بر عملکرد الگوریتم را توضیح دهید.

(ج) نشان دهید که اگر تعداد نمونه‌ها $n \to \infty$ باشد و $k \to \infty$ ولی $k/n \to 0$، آنگاه الگوریتم k-NN به صورت سازگار (consistent) عمل می‌کند، یعنی خطای آن به خطای بهینه (Bayes error) میل می‌کند. این نتیجه چه مفهومی در طراحی الگوریتم دارد؟

تعریف خطای بیز

خطای بیز[۱] حداقل خطای ممکن در یک مسئله‌ی طبقه‌بندی است که حتی بهترین طبقه‌بند نظری نیز نمی‌تواند از آن کمتر خطا کند. این خطا ناشی از ابهام ذاتی داده‌ها است؛ یعنی در مواردی که نمونه‌هایی با ویژگی یکسان ممکن است به کلاس‌های مختلف تعلق داشته باشند.

برای یک مسئله‌ی دوکلاسه با:

- $x \in \mathbb{R}^d$ داده‌ی ورودی
- $y \in \{۰, ۱\}$ برچسب کلاس واقعی
- تابع چگالی احتمال شرطی $P(y|x)$

طبقه‌بند بیز به صورت زیر تعریف می‌شود:

$$h_{\text{Bayes}}(x) = \arg\max_{y\in\{۰,۱\}} P(y|x)$$

و خطای بیز برابر است با:

$$\text{Error Bayes} = \mathbb{E}_x \left[\min\left(P(y=۰|x), P(y=۱|x)\right)\right]$$

تذکر:

- خطای بیز به توزیع داده بستگی دارد و مستقل از الگوریتم یادگیری است.
- اگر داده‌ها کاملاً جداپذیر باشند، (separable) خطای بیز صفر خواهد بود.
- هدف الگوریتم‌های یادگیری، نزدیک شدن به این حد پایینِ خطاست.

موفق باشید.

[۱] Bayes Error در واقع نشان‌دهنده‌ی محدودیت‌های ذاتی یک مسئله‌ی طبقه‌بندی است، نه ناتوانی مدل یادگیری.