بسمه تعالی

دانشگاه صنعتی اصفهان

پاسخنامه تکلیف دوم

مبانی ریاضی علوم داده

پاییز ۱۴۰۴

۱

(الف)

روش PCA یک روش خطی برای کاهش بُعد است که بر مبنای بیشینه‌سازی واریانس داده‌ها عمل می‌کند. فرض کنید داده‌ها به صورت ماتریس $X \in \mathbb{R}^{d\times n}$ (هر ستون یک نمونه) داده شده‌اند و داده‌ها مرکز شده‌اند. در PCA به دنبال زیرفضای خطی kـبعدی هستیم که تصویر داده‌ها روی آن، بیشترین واریانس ممکن را حفظ کند (یا به‌طور معادل، خطای بازسازی کمینه شود).

اگر تجزیه مقادیر تکین داده‌ها به صورت

$$X = U\Sigma V^T$$

باشد، آنگاه نمایش kـبعدی داده‌ها با نگه‌داشتن k مقدار تکین اول به صورت

$$X_k = U_k\Sigma_k V_k^T$$

به دست می‌آید. ستون‌های U_k جهت‌های اصلی داده هستند و این روش از نظر بهینگی در نرم فربنیوس و طیفی تضمین‌شده است. از نقاط قوت PCA می‌توان به تفسیرپذیری بالا و حذف نویز خطی اشاره کرد؛ اما این روش نیازمند محاسبه SVD بوده و به خود داده‌ها وابسته است.

(ب)

در روش Random Projection (و در قالب کلی‌تر، Compressive Sensing) کاهش بُعد با استفاده از یک تصویرسازی تصادفی انجام می‌شود. ایده‌ی اصلی آن است که داده‌های با بُعد بالا را با ضرب در یک ماتریس تصادفی به فضای کم‌بعد منتقل کنیم، بدون آن‌که فاصله‌ها (یا ساختار هندسی کلی داده‌ها) به‌طور معنادار تخریب شوند.

به طور دقیق، یک ماتریس تصادفی

$$R \in \mathbb{R}^{k\times d}$$

(مثلاً با درایه‌های گاوسی یا رادماخر) انتخاب می‌شود و نگاشت به صورت

$$y = Rx$$

تعریف می‌گردد. طبق لم جانسون–لیندن‌اشتراوس، اگر

$$k = O\left(\frac{\log n}{\varepsilon^2}\right),$$

آنگاه با احتمال بالا فاصله‌های زوج‌نقاط تا خطای نسبی ε حفظ می‌شوند.

در فشرده‌سازی یا Compressive Sensing نیز ایده مشابه است، با این تفاوت که فرض می‌شود داده (یا سیگنال) در یک پایه‌ی مناسب تُنُک است و می‌توان آن را از تعداد کمی اندازه‌گیری تصادفی با حل یک

مسئله بهینه‌سازی بازیابی کرد. نکته‌ی کلیدی این است که در Random Projection طول بردار تصویرشده حفظ نمی‌شود، اما روابط نسبی بین نقاط (مانند فاصله‌ها و زوایا) تقریباً ثابت می‌مانند.

(پ)

مقایسه‌ی دو روش به صورت زیر است:

- PCA یک روش **داده‌محور** است و جهت‌های تصویرسازی مستقیماً از ساختار داده استخراج می‌شوند، در حالی که Random Projection کاملاً **داده‌ناوابسته** و تصادفی است.
- PCA بهینه‌ترین تقریب کم‌رنک را (در معنای نرم فربنیوس و طیفی) ارائه می‌دهد، اما محاسبات آن پرهزینه‌تر است؛ در مقابل Random Projection بسیار سریع و مقیاس‌پذیر است.
- در PCA تفسیرپذیری جهت‌ها بالا است، ولی در Random Projection جهت‌ها فاقد معنای تفسیری مستقیم‌اند.
- PCA برای فشرده‌سازی و نویززدایی مناسب‌تر است، در حالی که Random Projection و Compressive Sensing برای حفظ فاصله‌ها، محاسبات سریع و کار با داده‌های بسیار بزرگ کاربرد بیشتری دارند.

۲

(الف)

اگر $A \in \mathbb{R}^{d\times n}$ با مقادیر تکین $\sigma_1 \geq \cdots \geq \sigma_r > 0$ و $r = \operatorname{rank}(A)$ باشد، داریم:

$$\|A\|_F^2 = \sum_{i=1}^{r} \sigma_i^2, \qquad \|A\|_2 = \sigma_1.$$

چون $\sigma_i \leq \sigma_1$ برای همه i، نتیجه می‌شود:

$$\|A\|_F^2 = \sum_{i=1}^{r} \sigma_i^2 \leq \sum_{i=1}^{r} \sigma_1^2 = r\sigma_1^2 = r\,\|A\|_2^2 \quad \Rightarrow \quad \|A\|_F \leq \sqrt{r}\;\|A\|_2\,.$$

(ب)

از مرتب بودن مقادیر تکین داریم $\sigma_1 \geq \cdots \geq \sigma_k$، بنابراین:

$$\sum_{i=1}^{k} \sigma_i^2 \geq k\sigma_k^2.$$

از طرف دیگر $\sum_{i=1}^{k} \sigma_i^2 \leq \sum_{i=1}^{r} \sigma_i^2 = \|A\|_F^2$، پس:

$$k\sigma_k^2 \leq \|A\|_F^2 \quad \Rightarrow \quad \sigma_k \leq \frac{\|A\|_F}{\sqrt{k}}.$$

(پ)

تقریب کم‌رنکِ بهینه با برش SVD داده می‌شود:

$$A_k = \sum_{i=1}^{k} \sigma_i\, u_i v_i^T, \qquad \operatorname{rank}(A_k) = k.$$

طبق قضیه اکهارت–یانگ:

$$\|A - A_k\|_2 = \sigma_{k+1}.$$

همچنین از نامساوی بخش (ب) نتیجه می‌شود $\sigma_{k+1} \leq \|A\|_F / \sqrt{k}$، لذا با انتخاب $B := A_k$ داریم:

$$\|A - B\|_2 \leq \frac{\|A\|_F}{\sqrt{k}}.$$

۳

(الف)

اگر A مربعی و وارون‌پذیر باشد، با نوشتن $A = U\Sigma V^T$ داریم:

$$A^{-1} = V\Sigma^{-1}U^T.$$

از طرفی شبه‌وارون مور–پنروز در این حالت برابر است با:

$$A^* = V\Sigma^{-1}U^T,$$

پس:

$$A^* = A^{-1}.$$

(ب)

برای گزاره‌ها:

(۱)

$$\left\|v_1^T A^*\right\|_2 = \min_{\|x\|=1} \frac{1}{\|Ax\|_2}.$$

از آن‌جا که $A^* = \sum_{i=1}^{r} \frac{1}{\sigma_i} v_i u_i^T$ داریم:

$$v_1^T A^* = \frac{1}{\sigma_1} u_1^T \quad \Rightarrow \quad \left\|v_1^T A^*\right\|_2 = \frac{1}{\sigma_1}.$$

و نیز $\max_{\|x\|=1} \|Ax\|_2 = \|A\|_2 = \sigma_1$، پس

$$\min_{\|x\|=1} \frac{1}{\|Ax\|_2} = \frac{1}{\sigma_1}.$$

بنابراین گزاره درست است.

(۲) $A^*AA^* = A^*$ (درست؛ از شرایط مور–پنروز)

(۳) $A = AA^*A$ (درست؛ از شرایط مور–پنروز)

(۴) $AA^* = I$ (نادرست در حالت کلی)

مثال:

$$A = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \Rightarrow A^* = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad AA^* = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \neq I_2.$$

(۵) $A^* = (A^TA)^*A^T$ (درست)

۴

(الف)

با فرض $\sigma_1 \geq \cdots \geq \sigma_r > 0$ و قید $\sum_{i=1}^{r} c_i^2 = 1$ داریم:

$$\sum_{i=1}^{r} c_i^2 \sigma_i^2 \leq \sigma_1^2 \sum_{i=1}^{r} c_i^2 = \sigma_1^2.$$

برابری وقتی رخ می‌دهد که $c_1 = \pm 1$ و $c_{i\neq 1} = 0$. بنابراین:

$$\max = \sigma_1^2.$$

(ب)

اکنون قیود $\sum c_i^2 = 1$ و $\sum c_i = 0$ برقرار است. به دلیل قید جمع صفر، لااقل دو مؤلفه با علامت مخالف لازم است. بهترین حالت آن است که این دو مؤلفه روی بزرگ‌ترین σها قرار گیرند:

$$c_1 = \frac{1}{\sqrt{2}}, \qquad c_2 = -\frac{1}{\sqrt{2}}, \qquad c_3 = \cdots = c_r = 0,$$

که در نتیجه:

$$\sum_{i=1}^{r} c_i^2 \sigma_i^2 = \frac{1}{2}\sigma_1^2 + \frac{1}{2}\sigma_2^2 \quad \Rightarrow \quad \max = \frac{\sigma_1^2 + \sigma_2^2}{2}.$$

۵

(الف)

محاسبه دقیق Ax زمان $O(nd)$ دارد. طبق نتیجه سؤال ۲، برای هر k می‌توان تقریب رتبه‌ـk از A (مثلاً برش (SVD را به صورت A_k در نظر گرفت به طوری که:

$$\|A - A_k\|_2 \leq \frac{\|A\|_F}{\sqrt{k}}.$$

اکنون خروجی را

$$y := A_k x$$

می‌گیریم. آنگاه:

$$\|y - Ax\|_2 = \|(A_k - A)x\|_2 \leq \|A - A_k\|_2 \, \|x\|_2 \leq \frac{\|A\|_F}{\sqrt{k}} \, \|x\|_2 .$$

برای اینکه $\|y - Ax\|_2 \leq \varepsilon \, \|A\|_F \, \|x\|_2$ برقرار باشد کافی است:

$$\frac{1}{\sqrt{k}} \leq \varepsilon \quad \Rightarrow \quad k \geq \frac{1}{\varepsilon^2}.$$

پس انتخاب حرفه‌ای:

$$\boxed{k = \left\lceil \frac{1}{\varepsilon^2} \right\rceil}$$

کفایت می‌کند.

(ب)

اگر $A_k = U_k \Sigma_k V_k^T$ باشد:

$$y = A_k x = U_k \Sigma_k (V_k^T x).$$

محاسبه $V_k^T x$ زمان $O(nk)$ و سپس ضرب در U_k زمان $O(dk)$ دارد، لذا کل زمان:

$$O((d+n)k).$$

۶

(الف)

برای

$$A = \begin{bmatrix} 1 & 2 \\ -1 & 2 \\ 1 & -2 \\ -1 & -2 \end{bmatrix}$$

داریم:

$$A^T A = \begin{bmatrix} 4 & 0 \\ 0 & 16 \end{bmatrix}.$$

پاور متد روی $A^T A$ با $v^{(0)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$:

$$v^{(t+1)} = \frac{A^T A\, v^{(t)}}{\|A^T A\, v^{(t)}\|}.$$

گام ۱:

$$w^{(1)} = \begin{bmatrix} 4 \\ 16 \end{bmatrix}, \quad v^{(1)} = \frac{1}{\sqrt{4^2 + 16^2}} \begin{bmatrix} 4 \\ 16 \end{bmatrix} = \begin{bmatrix} 0.2425356 \\ 0.9701425 \end{bmatrix}.$$

گام ۲:

$$w^{(2)} = \begin{bmatrix} 0.9701424 \\ 15.52228 \end{bmatrix}, \quad v^{(2)} \approx \begin{bmatrix} 0.0623783 \\ 0.9980526 \end{bmatrix}.$$

گام ۳:

$$w^{(3)} = \begin{bmatrix} 0.2495132 \\ 15.96884 \end{bmatrix}, \quad v^{(3)} \approx \begin{bmatrix} 0.0156231 \\ 0.9998780 \end{bmatrix}.$$

پس:

$$v_1 \approx \begin{bmatrix} 0.0156 \\ 0.9999 \end{bmatrix}.$$

(ب)

چون $A^T A = \mathrm{diag}(4, 16)$، مقادیر تکین:

$$\sigma_1 = 4, \qquad \sigma_2 = 2.$$

بردار تکین راست اول متناظر با بزرگ‌ترین مقدار ویژه 16 برابر است با:

$$v_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

و

$$u_1 = \frac{Av_1}{\sigma_1} = \frac{1}{4}\begin{bmatrix} 2 \\ 2 \\ -2 \\ -2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}.$$

تقریب پاورمتد پس از ۳ تکرار با v_1 دقیق هم‌راستا و بسیار نزدیک است.

(پ)

در تفسیر توصیه‌گر، v_1 الگوی غالب در فضای آیتم‌ها (رستوران‌ها) و u_1 میزان هم‌راستایی کاربران با آن الگو را نشان می‌دهد. اختلاف/نسبت σ_1 و σ_2 شدت «تک‌بُعدی بودن» ساختار غالب را مشخص می‌کند.

۷

(الف)

اگر

$$A = \begin{bmatrix} B & 0 \\ 0 & C \end{bmatrix},$$

آنگاه:

$$A^T A = \begin{bmatrix} B^T B & 0 \\ 0 & C^T C \end{bmatrix}.$$

بنابراین مقادیر ویژه $A^T A$ اجتماع مقادیر ویژه $B^T B$ و $C^T C$ است. چون مقادیر تکین ریشه‌ی مربع مقادیر ویژه‌اند، نتیجه می‌شود مقادیر تکین A دقیقاً اجتماع مقادیر تکین B و C (با احتساب تکرار) خواهد بود.

(ب)

برای بردارهای تکین راست نیز: اگر v_B بردار ویژه $B^T B$ باشد، آنگاه

$$\begin{bmatrix} v_B \\ 0 \end{bmatrix}$$

بردار ویژه $A^T A$ است و مشابه آن برای C. با تعریف $u = \frac{Av}{\sigma}$، بردارهای تکین چپ نیز به همین روش از بردارهای تکین بلوک‌ها ساخته می‌شوند.

۸

(الف)

فرض کنید $x_1, \dots, x_n \in \mathbb{R}^d$ و $X \in \mathbb{R}^{d\times n}$ ماتریس مختصات باشد (ستون i برابر x_i) و ماتریس فاصله‌ها $D = [d_{ij}]$ با $d_{ij} = \|x_i - x_j\|$ داده شده باشد. همچنین فرض کنید مرکز هندسی در مبدأ است، یعنی:

$$X\mathbf{1} = 0.$$

اگر $G = X^T X$ باشد، آنگاه:

$$d_{ij}^2 = \|x_i\|^2 + \|x_j\|^2 - 2G_{ij}.$$

با «دو بار مرکزسازی» به رابطه استاندارد می‌رسیم:

$$G = -\frac{1}{2}\, JD^{(2)}J, \qquad J = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T, \qquad D^{(2)} = [d_{ij}^2].$$

(ب)

الگوریتم بازسازی مختصات MDS) کلاسیک):

- ابتدا $D^{(2)} = [d_{ij}^2]$ را تشکیل دهید.
- سپس $J = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ را بسازید.
- ماتریس گرام را حساب کنید: $G = -\frac{1}{2}\, JD^{(2)}J$.
- تجزیه ویژه انجام دهید: $G = V\Lambda V^T$.
- با انتخاب d مقدار ویژه مثبت اول:

$$X = \Lambda_d^{1/2} V_d^T.$$

(در صورت وجود نویز، مقادیر ویژه منفی کوچک معمولاً صفر در نظر گرفته می‌شوند.)