تمرین سری هفتم ML

آرمین سعادت
۹۶۱۰۵۸۲۹

مسئله ۱.

الف)

first - iteration: $\mu_1 = 0 \xrightarrow{\text{خوشه ۱}} \{1\} \Rightarrow \mu_1 = \frac{1}{1} = 1$

$\mu_2 = 3 \xrightarrow{\text{خوشه ۲}} \{2, 8, 9\} \Rightarrow \mu_2 = \frac{2+8+9}{3} = 6.\bar{3}$

Second - iteration: $\mu_1 = 1 \xrightarrow{\text{خوشه ۱}} \{1, 2\} \Rightarrow \mu_1 = \frac{1+2}{2} = 1.5$

$\mu_2 = 6.\bar{3} \xrightarrow{\text{خوشه ۲}} \{8, 9\} \Rightarrow \mu_2 = \frac{8+9}{2} = 8.5$

third - iteration: $\mu_1 = 1.5 \xrightarrow{\text{خوشه ۱}} \{1, 2\}$

$\mu_2 = 8.5 \xrightarrow{\text{خوشه ۲}} \{8, 9\}$ $\Rightarrow$ همگرا شده است.

مقدار تابع هزینه:

$$J = \sum_{n=1}^{4} \sum_{k=1}^{2} r_{nk} \| x_n - \mu_k \|^2 = (1-1.5)^2 + (2-1.5)^2 + (8-8.5)^2 + (9-8.5)^2$$

$$= 4 \times \left(\frac{1}{2}\right)^2 = 1$$

ب)

مرحله ۱: هر خوشه را به نزدیک‌ترین متا-خوشه تخصیص می‌کنیم. معیار نزدیک بودن خوشه i به متا-خوشه j، $\| v_j - \mu_i \|$ است. پس برای هر خوشه i، تعیین می‌کنیم که μ_i به کدام v_j نزدیک‌تر است و به همان متا-خوشه آن را تخصیص می‌کنیم.

مرحله ۲: در این مرحله مراکز متا-خوشه‌ها به‌روزرسانی می‌شود. برای این کار هر v_j، برابر با میانگین μ_iهای که به آن تخصیص یافته‌اند می‌شود. یعنی مرکز هر متا-خوشه، برابر میانگین مراکز خوشه‌هایی است که به آن متا-خوشه تخصیص یافته‌اند. رابطه‌ی ریاضی این به‌روزرسانی به شکل مقابل است:

$$V_m = \frac{\sum_{i=1}^{K} y_{im} \mu_i}{\sum_{i=1}^{K} y_{im}}$$

مرحله ۳: در واقع همان مرحله E در حالت K-means عادی است. در این مرحله هر داده را به نزدیک‌ترین خوشه تخصیص می‌کنیم. ملاک نزدیک بودن هم برای داده n و خوشه K به این شکل است:

$$\|x_n-\mu_K\|$$

مرحله ۴: در واقع همان مرحله‌ی M در حالت K-means عادی است. مرکز هر خوشه را برابر با میانگین داده‌هایی که به آن خوشه تخصیص یافته است قرار می‌دهیم. یعنی مرکز هر خوشه از رابطه زیر بدست می‌آید:

$$\mu_K = \frac{\sum_{n=1}^{N} r_{nK} x_n}{\sum_{n=1}^{N} r_{nK}}$$

مسئله ۲.

الف:

$$P(F) = P(F) \ , \ P(A) = P(A), \ P(S) = \sum_F \sum_A P(S|F,A)P(F)P(A)$$

$$P(H) = \sum_S P(H|S)P(S) \ , \ P(N) = \sum_S P(N|S)P(S)$$

$$P(F,A,S,H,N) = P(N|F,A,S,H)\,P(H|A,S,F)\,P(S|F,A)\,P(F|A)\,P(A)$$

رابطه بالا از chain-rule به دست می‌آید. از طرفی طبق مدل نیز داریم:

$$\left.\begin{array}{l} P(N|F,A,S,H) = P(N|S) \\ P(H|A,S,F) = P(H|S) \\ P(F|A) = P(F) \end{array}\right\} \ P(F,A,S,H,N) = P(N|S)P(H|S)P(S|F,A)P(F)P(A)$$

ب:

$$P(A=1|F,S,H,N) = \frac{P(A=1,F,S,H,N)}{P(F,S,H,N)} = \frac{P(A=1,F,S,H,N)}{P(A=0,F,S,H,N)+P(A=1,F,S,H,N)}$$

با استفاده از رابطه به دست آمده در قسمت الف (توسط chain-rule و مدل بیزی) داریم:

$$P(A,F,S,H,N) = P(A)P(F)P(S|F,A)P(N|S)P(H|S)$$

بنابراین به دست می‌آید

$$P(A=1|F,S,H,N) = \frac{P(A=1)P(F)P(S|F,A=1)P(N|S)P(H|S)}{P(F)P(N|S)P(H|S)\left[P(A=0)P(S|F,A=0)+P(A=1)P(S|F,A=1)\right]}$$

$$= \frac{P(A=1)P(S|F,A=1)}{P(A=0)P(S|F,A=0)+P(A=1)P(S|F,A=1)}$$

د)

طبق تعریف مدل‌های Markov blanket ، باید فرزندان A ، والدین A ، و والدین فرزندان A را انتخاب کنیم. A والد ندارد. فرزند A می‌شود S. و والد فرزند A می‌شود F. پس جواب می‌شود $\underline{F, S}$.

این جواب منطقی است، چرا که در قسمت (ب) به این نتیجه رسیدیم که برای پیدا کردن $P(A|F,S,H,N)$ تنها به متغیرهای F و S نیاز داریم. (علاوه بر A). پس جواب همان F، S است.

ت)

در همان بخش (ب) به جواب ساده شده مورد نظر رسیدیم:

$$P(A=1|F,S,H,N) = \frac{P(A=1)P(S|F,A=1)}{\sum_A P(A)P(S|F,A)}$$

ث)

متغیر unobserved همان A است. پس پارامترهای که به A ربط دارند، نیازمند محاسبه گام E هستند.

از بخش‌های قبلی داشتیم که:

$$P(F,A,S,H,N) = P(F)P(A)P(S|F,A)P(H|S)P(N|S)$$

همانطور که مشاهده می‌شود در ۲ عبارت متغیر A ظاهر شده است:

$P(A), P(S|F,A)$

بنابراین پارامترهای $P(A=1)$ ، $P(S=1|F=1,A=1)$ ، $P(S=1|F=1,A=0)$ ،

$P(S=1|F=0,A=1)$ و $P(S=1|F=0,A=0)$ نیازمند محاسبه گام E هستند.

(ج

محاسبه تخمین $P(A=1)$:

$$P(A=1) = \underset{P(A=1)}{\arg\max} \sum_{i=1}^{\text{۵}} \sum_{z} Q(z) \log P(A)$$

$$= \log P(A=1) + \log P(A=1) + 0 + \text{۰/۸} \log P(A=1) + \text{۰/۴} \log P(A=1)$$

$$+ 0 + 0 + \log P(A=0) + \text{۰/۲} \log P(A=0) + \text{۰/۶} \log P(A=0)$$

$$= \text{۳/۲} \log P(A=1) + \text{۱/۸} \log P(A=0) = \text{۳/۲} \log(P(A=1)) + \text{۱/۸} \log(1-P(A=1)) = E$$

$$\frac{dE}{dP(A=1)} = 0 \Rightarrow \frac{\text{۳/۲}}{P(A=1)} + \frac{-\text{۱/۸}}{1-P(A=1)} = 0 \Rightarrow P(A=1) = \text{۰/۶۴}$$

محاسبه تخمین $P(S=1|F=0,A=1)$:

$$P(S=1|F=0,A=1) = \arg\max \sum \sum_{z} Q(z) \log P(S=1|F=0,A=1)$$

$$= \log P(S=1|F=0,A=1) + \log P(S=0|F=0,A=1)$$

$$+ \text{۰/۸} \log P(S=0|F=0,A=1) + \text{۰/۴} \log P(S=1|F=0,A=1)$$

$$= \text{۱/۴} \log P(S=1|F=0,A=1) + \text{۱/۸} \log(1-P(S=1|F=0,A=1)) = E$$

$$\frac{dE}{dP(S=1|F=0,A=1)} = \frac{\text{۱/۴}}{P(S=1|F=0,A=1)} + \frac{-\text{۱/۸}}{1-P(S=1|F=0,A=1)} = 0 \Rightarrow P(S=1|F=0,A=1) = \frac{\text{۷}}{\text{۱۶}}$$

مسئله ۳.

الف)

برای بیشینه کردن کران پایین داریم:

$$\max \sum_z Q(z)\log\frac{P(x,z|\theta)}{Q(z)} = \max\left\{\sum_z Q(z)\log\frac{P(z|x,\theta)}{Q(z)} + \sum_z Q(z)\log P(x|\theta)\right\}$$

$$\sum_z Q(z)\log P(x|\theta) \longrightarrow \text{Const}$$

$\Rightarrow$ کافی است مکس باز عبارت را حساب کنیم $\longrightarrow$ می‌توانیم مینیمم معکوس عبارت را حساب کنیم.

$$\Rightarrow \min \sum_z Q(z)\log\frac{Q(z)}{P(z|x,\theta)} = \min \sum q_k \log\frac{q_k}{z_k}$$

با استفاده از ضرایب لاگرانژ داریم:

$$\sum q_k\log\frac{q_k}{r_k} + \beta\left(\sum q_k - 1\right) - \sum \alpha_k q_k = L(\theta,q,\beta,\alpha)$$

$$\frac{\partial L}{\partial q_k}=0 \Rightarrow \log\frac{q_k}{r_k}+1+\beta-\alpha_k=0 \Rightarrow q_k = r_k e^{\alpha_k-1-\beta}$$

$$\Rightarrow L(\theta,q,\beta,\alpha) = \sum_k r_k e^{\alpha_k-1-\beta} - \beta$$

اگر $r_k=0$، پس q_kام مساوی صفر خواهیم داشت: $q_k=r_k$

درصورتی که $r_k\neq 0$، طبق لاگرانژ نتیجه می‌شود که $\alpha_k=0$ (زیرا $q_k\neq 0$).

پس اگر $\alpha_k=0$ داریم:

$$q_k = r_k e^{-1-\beta} \Rightarrow e^{-1-\beta}=1 \Rightarrow -1-\beta=0 \Rightarrow \boxed{\beta=-1}$$

که نتیجه می‌شود $\boxed{q_k=r_k}$.

ب) :

طبق رابطه‌ی اسلاید ۲۸ داریم:

$$E[\log P(Z, x|\theta)] = \sum_{i=1}^{N} \sum_{j=1}^{K} \gamma_{ji} \Big(\underbrace{\log \frac{e^{-\lambda_j} \lambda_j^{m_i}}{m_i!}}_{P(m_i; \lambda_j)} + \log \pi_j \Big) = J$$

$$\frac{\partial J}{\partial \lambda_K} = 0 \Rightarrow \frac{\partial}{\partial \lambda_K} \sum_{i=1}^{N} \gamma_{Ki} \Big(\log \frac{e^{-\lambda_K} \lambda_K^{m_i}}{m_i!} + \log \pi_K \Big) = 0$$

$$\frac{\partial}{\partial \lambda_K} \log \frac{e^{-\lambda_K} \lambda_K^{m_i}}{m_i!} = \frac{\partial}{\partial \lambda_K} \log e^{-\lambda_K} + \frac{\partial}{\partial \lambda_K} \log \frac{\lambda_K^{m_i}}{m_i!}$$

$$= -1 + \frac{m_i}{\lambda_K} \Rightarrow \sum_{i=1}^{N} \gamma_{Ki} \Big(-1 + \frac{m_i}{\lambda_K} \Big) = 0 \Rightarrow \boxed{\lambda_K = \frac{\sum_{i=1}^{N} \gamma_{Ki} m_i}{\sum_{i=1}^{N} \gamma_{Ki}}}$$

$$\frac{\partial \, E[\log P(Z, x|\theta)] + \lambda \big(1 - \sum_{i=1}^{K} \pi_i\big)}{\partial \pi_K} = 0$$

$$\Rightarrow \frac{1}{\pi_K} \sum \gamma_{Ki} - \lambda = 0 \Rightarrow \left\{ \begin{array}{l} \pi_K = \dfrac{\sum_{i=1}^{N} \gamma_{Ki}}{\lambda} \\ \sum_{i=1}^{K} \pi_K = 1 \Rightarrow N = \lambda \end{array} \right\} \boxed{\pi_K = \frac{\sum_{i=1}^{N} \gamma_{Ki}}{N}}$$

مسئله ۴.

الف)

خب این سؤال عملاً در اسلایدها جواب داده شده است.

فرض می‌کنیم تا مرحله‌ی $i-1$ را پیش رفته‌ایم. یعنی v_1 تا v_{i-1} جواب‌های مراحل ۱ تا $i-1$ هستند.

$$\max_{v_i} \frac{1}{N}\sum_{n=1}^{N}\left(v_i^T x^{(n)} - v_i^T \bar{x}\right)^2 = v_i^T S v_i \quad , \quad v_i^T v_j = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{o.w.} \end{cases}$$

$$\mathcal{L}(v_i, \lambda_i, \alpha) = v_i^T S v_i - \sum_{j=1}^{i-1} \alpha_j v_i^T v_j + \lambda_i (1 - v_i^T v_i)$$

$$\Rightarrow \frac{\partial \mathcal{L}}{\partial v_i} = 2 S v_i - \sum_{j=1}^{i-1} \alpha_j v_j - 2\lambda_i v_i = 0$$

$$\Rightarrow \forall k \in \{1, i-1\} : \quad v_k^T \left[2 S v_i - \sum \alpha_j v_j - 2\lambda_i v_i \right] = 0$$

$$\left.\begin{aligned} &\Rightarrow 2 v_k^T S v_i = \alpha_k v_k^T v_k = \alpha_k \\ &v_k^T S = \lambda_k v_k^T \end{aligned}\right\} \quad 2 v_k^T S v_i = 2\lambda_k \underbrace{v_k^T v_i}_{0} = 0 = \alpha_k$$

بردار ویژه

$$\Rightarrow \forall k \in \{1, i-1\} \quad \alpha_k = 0$$

$$\frac{\partial \mathcal{L}}{\partial v_i} = 2 S v_i - \sum_{j=1}^{i-1} \alpha_j v_j - 2\lambda_i v_i = 0 \quad , \quad \alpha_j = 0$$

$$\Rightarrow 2 S v_i = 2 \lambda_i v_i \Rightarrow S v_i = \lambda_i v_i \Rightarrow$$ v_i بردار ویژه است

$$\max v_i^T S v_i = \max v_i^T \lambda_i v_i = \max \lambda_i$$

بنابراین iامین محور پیدا شده در الگوریتم PCA، بردار ویژه متناظر با iامین بزرگترین مقدار ویژه است.

ب)

$$K(x,x') = x^T x' \quad , \quad S = \frac{1}{N}\, x.x^T$$

S ← ماتریس کوواریانس

به ازای بردار ویژه K (مثلاً b) داریم:

$$\left.\begin{aligned} Kb &= N\lambda b \\ K &= x^T x \end{aligned}\right\} \quad x^T x b = N\lambda b \xrightarrow[\text{ضرب می‌کنیم.}]{\text{از چپ در } x} x x^T x b = N\lambda x b \rightarrow (1)$$

$$\left.\begin{aligned} S = \frac{1}{N} x x^T \Rightarrow x x^T &= NS \\ (1) & \end{aligned}\right\} \quad NSxb = N\lambda x b \Rightarrow Sxb = \lambda x b$$

بنابراین xb بردار ویژه S است. پس داریم:

$$S.(xb) = \lambda (xb)$$

از طرفی با توجه به ماتریس کرنل داریم:

$$V_i = \sum b_j \phi(x_j) = xb$$

$$\Rightarrow S.(xb) = \lambda(xb) \Rightarrow \boxed{S V_i = \lambda V_i}$$

بنابراین با استفاده از کرنل خطی، مسئله PCA حل شد (به ازای یک جواب دلخواه).
بنابراین این ۲ روش با یکدیگر معادل هستند.

نکته مهم: در روش حل فرض شد که میانگین داده‌ها صفر است بنابراین عبارت $x - \bar{x}$ به x تبدیل شد.
این فرض از کلیت مسئله کم نمی‌کند چرا که می‌توان داده‌ها را شیفت داد تا میانگین صفر شود و راه‌حل درست است.

ب)

خب ابتدا ماتریس داده ها را تجزیه SVD می کنیم.

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

$$\Rightarrow X^T X = V \Sigma^T U^T U \Sigma V^T \quad , \quad U^T U = I$$

$$\Rightarrow X^T X = V \Sigma^T \Sigma V^T = V \Sigma' V^T$$

از تجزیه SVD می دانیم که درایه‌های قطر اصلی ماتریس Σ در واقع مقادیر ویژه ماتریس X هستند. پس ماتریس Σ' که $\Sigma^T \Sigma$ است، روی قطر اصلی اش توان دوم مقادیر ویژه X را دارد.

از روش PCA، مطلوب مقادیر ویژه ماتریس $\frac{1}{N} X^T X$ می باشد. (و در ادامه بردار ویژه متناظر آن‌ها)

پس اگر درایه های قطر اصلی Σ را s_i بنامیم، مقادیر ویژه ماتریس $\frac{1}{N} X^T X$ برابر با $\frac{s_i^2}{N}$ می شود.

پس از روی تجزیه SVD و با در اختیار داشتن Σ، می توانیم به مقادیر ویژه دسترسی داشته باشیم.

بردار ویژه ها نیز ستون های V در تجزیه SVD است.

ت)

ماتریس داده‌ها را $X_{n \times d}$ در نظر می گیریم.

روش SVD از اردر زمانی $O(nd \min(n,d))$ است.

روش کوواریانس از مرتبه‌ی $O(nd^2)$ است.

پس اگر ابعاد داده ها یعنی d بیشتر از n باشد، روش SVD برابر $O(n^2 d)$ می شود که چون $n < d$، از $O(nd^2)$ بهتر است.

پس در این حالت استفاده از روش SVD بهتر است.