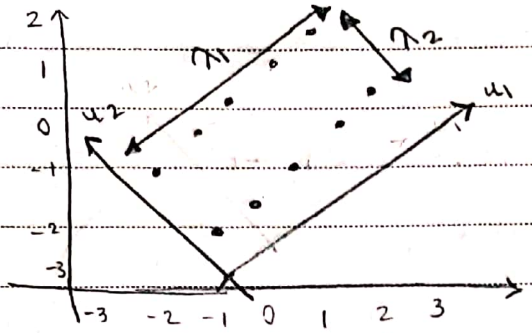
1) الف: u_1 و u_2 به ترتیب component های حاصل از PCA هستند. label داده‌ها در نمودار تأثیری در component های PCA ندارند زیرا



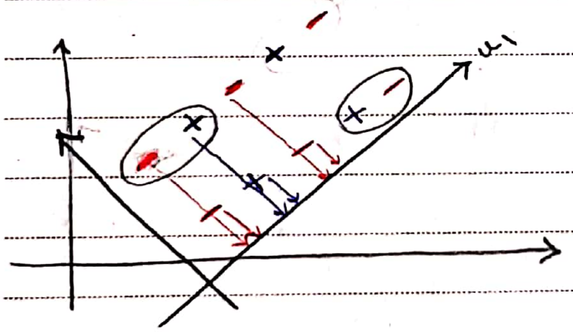
PCA یک روش unsupervised است که در آن واریانس تصویر داده‌ها روی component ها باید بیشینه شود. به عبارت دیگر، معادله زیر d جواب برای $(\vec{u}, \lambda)$ دارد و آن $\vec{u}$ ای انتخاب می‌شود که مقدار ویژه‌اش (λاش) بیشینه باشد. (جمع انرژی y_i از در معادله زیر بیشینه باشد.)

$$S_{d\times d}\, u_{d\times 1} = \lambda u_{d\times 1} \quad \text{و} \quad S = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$$

1) ب: حالت اول:

2D data, 100% error

1D data 0% error

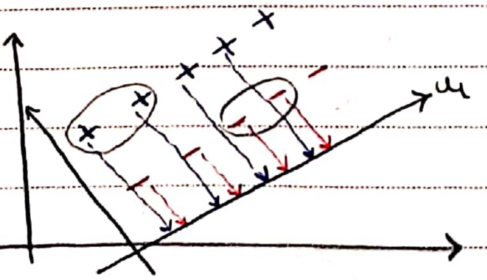


همانطور که دیده می‌شود با این نوع دسته‌بندی در فضای 2 بعدی، دسته‌بند 1NN با دقت 0 درصد (خطای 100 درصد) برازش داده خواهد شد. اما چنانچه همین داده‌ها بر روی محور اول PCA ($\vec{u_1}$) تصویر شوند، دسته‌بند 1NN دقت 100% خواهد داشت.

حالت دوم:

2D data 0% error

1D data 100% error



با این نوع برچسب‌گذاری، دسته‌بند 1NN به دقت 100% می‌رسد، اما در حالتی که داده‌ها روی $\vec{u_1}$ تصویر شوند، نزدیک‌ترین نمونه به هر +، یک نمونه با برچسب − خواهد بود (توجه نکردن PCA به label نمونه‌ها)

Subject:
Year.　　Month.　　Date.　　(　)

(2

$$S_1^2 = \sum_{y_i \in Y_1} (y_i - m_1)^2 = \sum_{y_i \in Y_1} y_i^2 + \underbrace{\cancel{\sum_{Y_1} m_1^2}}_{n m_1^2} - 2 m_1 \sum_{Y_1} y_i$$

$$\Rightarrow \frac{1}{n_1} s_1^2 = \frac{1}{n_1} \sum_{Y_1} y_i^2 + m_1^2 - 2 m_1 \Big(\underbrace{\cancel{\frac{1}{n_1} \sum_{Y_1} y_i}}_{m_1} \Big) =$$

$$\frac{1}{n_1} \sum_{y_i \in Y_1} y_i^2 - m_1^2$$

(m_1 تعریف)

به همین ترتیب برای $\frac{1}{n_2} s_2^2$ هم داریم: $\frac{1}{n_2} s_2^2 = \frac{1}{n_2} \sum_{y_j \in Y_2} y_j^2 - m_2^2$

$$\Longrightarrow (m_1 - m_2)^2 + \frac{1}{n_1} s_1^2 + \frac{1}{n_2} s_2^2 = \frac{1}{n_1} \sum_{Y_1} y_i^2 + \frac{1}{n_2} \sum_{Y_2} y_j^2$$

$$\underbrace{-2 \frac{1}{n_1} \sum_{Y_1} y_i \times \frac{1}{n_2} \sum_{Y_2} y_j}_{-2 m_1 m_2} = \boxed{\frac{1}{n_1} \sum_{y_i \in Y_1} y_i^2 + \frac{1}{n_2} \sum_{y_j \in Y_2} y_j^2 - \frac{2}{n_1 n_2} \sum_{y_i \in Y_1} \sum_{y_j \in Y_2} y_i y_j}$$

ادامه در صفحه بعد

ادامه 2

عبارت صفحه قبل را میتوان بلوند آی تیتر داد که همه termها وارد یک summation شوند:

$$\frac{1}{n_1}\sum_{y_i\in Y_1} y_i^2 = \frac{1}{n_1 n_2}\sum_{y_i\in Y_1} y_i^2 \underbrace{\sum_{y_j\in Y_2} 1}_{n_2} =$$

$$\frac{1}{n_1 n_2}\sum_{y_i\in Y_1}\sum_{y_j\in Y_2} y_i^2$$

به همین ترتیب برای term دوم نیز خواهیم داشت:

$$\frac{1}{n_2}\sum_{y_j\in Y_2} y_j^2 = \frac{1}{n_1 n_2}\sum_{y_i\in Y_1}\sum_{y_j\in Y_2} y_j^2$$

عبارت سوم را هم به همان صورت که بود بازنویسی میکنیم: $\frac{-2}{n_1 n_2}\sum_{y_i\in Y_1}\sum_{y_j\in Y_2} y_i y_j$

$$= \frac{1}{n_1 n_2}\sum_{y_i\in Y_1}\sum_{y_j\in Y_2}\left(y_i^2 + y_j^2 - 2y_i y_j\right) =$$

$$\boxed{\frac{1}{n_1 n_2}\sum_{y_i\in Y_1}\sum_{y_j\in Y_2}(y_i - y_j)^2}$$

همانطور که دیده شد، دو تعریف J ارائه شده در سوال با هم معادلند.

Subject:
Year. Month. Date. ( )

3) الف: فرض کنید M تا component داریم. پارامترها عبارت اند از: $\alpha_1, \dots, \alpha_K$
$\lambda_1, \dots, \lambda_K$

$$P(x_i \mid \theta) = \sum_{K=1}^{M} \alpha_K \, \text{Poisson}(x_i \mid \lambda_K) \quad \text{I}$$

x_i متغیر observed

z_i متغیر hidden

$$P(x_i, z_i \mid \theta) = \prod_{K=1}^{M} \left(\alpha_K \, \text{Poisson}(x_i \mid \lambda_K)\right)^{z_{iK}} \quad \text{II}$$

$z_{iK} = 1$

همانند جزوه درس، z_i یک بردار M تایی است بطوریکه فقط در اندیس K ام مقدار 1 دارد و در بقیه اندیس ها مقدار 0 دارد.

حال عبارت maximum likelihood را می نویسیم:

$$\hat{\theta} = \underset{\theta}{\arg\max}\; E_{Z \mid x, \theta}\left[\sum_{i=1}^{N} \log P(x_i, z_i \mid \theta)\right] \quad \text{II}$$

$$E_{Z \mid x, \theta}\left[\sum_{i=1}^{N} \log \prod_{K=1}^{M} \left(\alpha_K \, \text{Possion}(x_i \mid \lambda_K)\right)^{z_{iK}}\right] =$$

$$\text{w.t.} \; \sum_{K=1}^{M} \alpha_K = 1$$

$$E_{Z \mid x, \theta}\left[\sum_{i=1}^{N} \sum_{K=1}^{M} \left(z_{iK} \log \alpha_K + z_{iK} \log(\text{Poission}(x_i \mid \lambda_K))\right)\right]$$

امید ریاضی خطی است

$$\sum_{i=1}^{N} \sum_{K=1}^{M} \left(E[z_{iK}] \log \alpha_K + E[z_{iK}] \log(\text{Poission}(x_i \mid \lambda_K))\right)$$

حال $E_{Z \mid x, \theta}[z_{iK}]$ را بدست می آوریم: z_{iK} is a boolean variable

$$\Longrightarrow E_{Z \mid x, \theta}[z_{iK}] = P(z_{iK} = 1 \mid x_i, \theta) = \frac{P(x_i \mid z_{iK} = 1, \theta) \cdot P(z_{iK} = 1 \mid \theta)}{P(x_i \mid \theta)}$$

با توجه به I

$$\Longrightarrow E_{Z \mid x, \theta}[z_{iK}] = \frac{\text{Poission}(x_i \mid \lambda_K) \, \alpha_K}{\sum_{K=1}^{M} \alpha_K \, \text{Poission}(x_i \mid \lambda_K)} = \gamma_{iK}$$

حال که $E_{Z \mid x, \theta}[z_{iK}] = \gamma_{iK}$ را بدست آوردیم، آنرا در رابطه Expectation maximization می گذاریم:

PAPCO

Subject:
Year.   Month.   Date.   ( )

$$E_Z[L(\theta)] = \sum_{i=1}^{N}\sum_{k=1}^{M}\Big(\gamma_{ik}\log(\alpha_k) + \gamma_{ik}\log(\text{Poission}(x_i|\lambda_k))\Big)$$

حال باید از $E_Z[L(\theta)]$ نسبت به θ (پارامترهای λ_k و α_k) مشتق بگیریم. از آنجا که

برای α_k قید داریم $(\sum_i \alpha_k = 1)$، باید تابع لاگرانژ را تشکیل دهیم:

$$\mathcal{L} = E_Z[L(\theta)] - \lambda\Big(\sum_{k=1}^{M}\alpha_k - 1\Big)$$

$$\Rightarrow \frac{\partial \mathcal{L}}{\partial \alpha_k} = 0 \Rightarrow \frac{\partial}{\partial \alpha_k}\Big[\sum_{i=1}^{N}\sum_{k=1}^{M}\gamma_{ik}\log(\alpha_k)\Big] - \frac{\partial}{\partial \alpha_k}\Big[\lambda\sum_{k=1}^{M}\alpha_k\Big] = 0 \Rightarrow$$

$$\sum_{i=1}^{N}\frac{\gamma_{ik}}{\alpha_k} - \lambda = 0 \Rightarrow \alpha_k = \sum_{i=1}^{N}\gamma_{ik}/\lambda$$

$$, \sum_{k=1}^{M}\alpha_k = 1 = \frac{\sum_{k=1}^{M}\sum_{i=1}^{N}\gamma_{ik}}{\lambda} \Rightarrow \lambda = \sum_{k=1}^{M}\sum_{i=1}^{N}\gamma_{ik} \quad \Big\} \Rightarrow$$

$$\boxed{\alpha_k = \frac{\sum_{i=1}^{N}\gamma_{ik}}{\sum_{k=1}^{M}\sum_{i=1}^{N}\gamma_{ik}}} = \frac{\sum_{i=1}^{N}\gamma_{ik}}{N}$$

اما برای پارامتر λ_k چون قید نداریم، نیازی به Lagrangian نیست:

$$\frac{\partial E_Z[L(\theta)]}{\partial \lambda_k} = \frac{\partial}{\partial \lambda_k}\Big(\sum_{i=1}^{N}\gamma_{ik}\big(\underset{-1}{-\lambda_k} + \underset{x_i/\lambda}{x_i\log\lambda_k} - \underset{0}{\log x_i!}\big)\Big) = 0$$

$$\Rightarrow -\sum_{i=1}^{N}\gamma_{ik} + \sum_{i=1}^{N}\gamma_{ik}\,x_i/\lambda_k = 0 \Longrightarrow$$

$$\boxed{\lambda_k = \frac{\sum_{i=1}^{N}\gamma_{ik}\,x_i}{\sum_{i=1}^{N}\gamma_{ik}}}$$

n×d d×1

d×n n×d

Subject:

Year. Month. Date. ( )

3) ب) : می دانیم حاصل ضرب ماتریس $X_{n\times d}$ و $P_{i_{d\times 1}}$ ، برابر تصویر داده ها روی

محور i ام PCA خواهد بود. Xp_i projected samples on ith component

$$= X_{n\times d}\cdot P_{i_{d\times 1}}$$

طبق تعریف واریانس برای مجموعه داده Xp_1 :

$$Var(Xp_i) = (Xp_i)^T (Xp_i) = p_i^T X^T X p_i$$

حال باید اثبات کنیم : $\boxed{\lambda_1 = p_1^T X^T X p_1}$

برای این کار ، از تعریف ماتریس کوواریانس استفاده می کنیم :

تعداد component ها (بردارهای ویژه)

$$\left(C_{d\times d} = \sum_{i=1} \lambda_i p_i p_i^T\right) \times p_1 \longrightarrow Cp_1 = \sum \lambda_i p_i \underbrace{p_i^T p_1}$$

(این مقدار برای هر $i \neq 1$ ، صفر است (بردارهای ویژه بر هم عمودند))

$$\Longrightarrow Cp_1 = \lambda_1 p_1 \underbrace{p_1^T p_1}_{\|p_1\|^2=1} \Rightarrow Cp_1 = \lambda_1 p_1$$

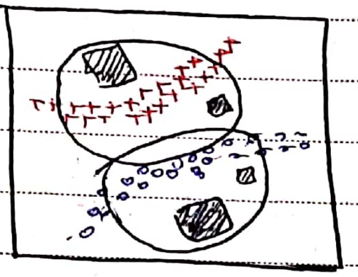
همچنین طبق تعریف $C = X^T X$

$$p_1^T \times (Cp_1 = \lambda_1 p_1) \Rightarrow p_1^T C p_1 = \lambda_1 \underbrace{p_1^T p_1}_{1}$$

$$\Rightarrow \boxed{p_1^T X^T X p_1 = \lambda_1}$$

Subject:
Year. Month. Date. ( )

4) الف: مدل ب مناسب نیست زیرا در جاهایی از فضا که احتمال صفر است (داده‌ای وجود ندارد) این مدل احتمال قابل توجهی میدهد. (در شکل با ▨ نشان داده شده اند.)



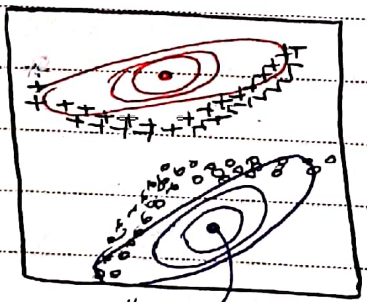
مدل ب

اما مدل الف مناسب‌تر است زیرا مقادیر میانگین و ماتریس کوواریانس مربوط به component ها طوری هستند که در جاهایی که داده‌ای نداریم $E_{z|x_n,\theta}[z_{ik}]$ (امید ریاضی تعلق نمونه n به cluster k) نیز نزدیک صفر است.

4) ب: خروجی ب نمی‌تواند نتیجه اولین گام استفاده از EM با component های گوسی (GMM) باشد.

دلیل آن اینست که مقادیر μ و Σ در الگوریتم GMM طوری update میشوند که میانگین component های گوسی روی فضاهایی با density بالا بنشیند و سپس با استفاده از Σ سایر نقاط در فضاهای اطراف با density کمتر نیز پوشش داده خواهند شد.

اما همانطور که دیده میشود در شکل ب میانگین های توزیع گوسی در فضاهایی افتاده اند که $E_{z|x_n,\theta}[z]$ صفر است. پس بیشتر update نشده اند. در شکل الف clustering و update شدن μ و Σها مناسبتر است.



میانگین در جایی با density صفر

PAPCO