

تمرین پنجم

سوال اول)

(الف)

برای دو روش خوشه‌بندی ذکر شده ابتدا نحوه حل مسئله شامل روش محاسبه و ابرپارامترهای هر یک و در نتیجه فرضیات در نظر گرفته شده درباره داده را توضیح می‌دهیم و سپس تاثیرهای این فرضیات را شرح می‌دهیم.

الگوریتم Kmeans:

فرضیات این الگوریتم راجع به مجموعه داده شامل این است که:

۱. خوشه‌های نهایی دارای اندازه تقریباً یکسان هستند.
 ۲. خوشه‌های نهایی محدب و isotropic هستند یعنی شعاع این خوشه‌ها از هر جهت تقریباً با هم برابر هستند. این فرض باعث می‌شود خوشه‌های در فضای دوبعدی به شکل دایره باشند.
 ۳. خوشه‌های نهایی دارای واریانس تقریباً برابری هستند.
 ۴. تعداد مشخصی از خوشه‌ها را برای مجموعه داده از ابتدا فرض می‌کند.
- این الگوریتم یک مسئله بهینه‌سازی است که سعی می‌کند حاصل جمع intra را که به شکل پایین محاسبه می‌شود را در تمام خوشه‌های کمینه کند. C خوشه‌ها را نشان می‌دهد.

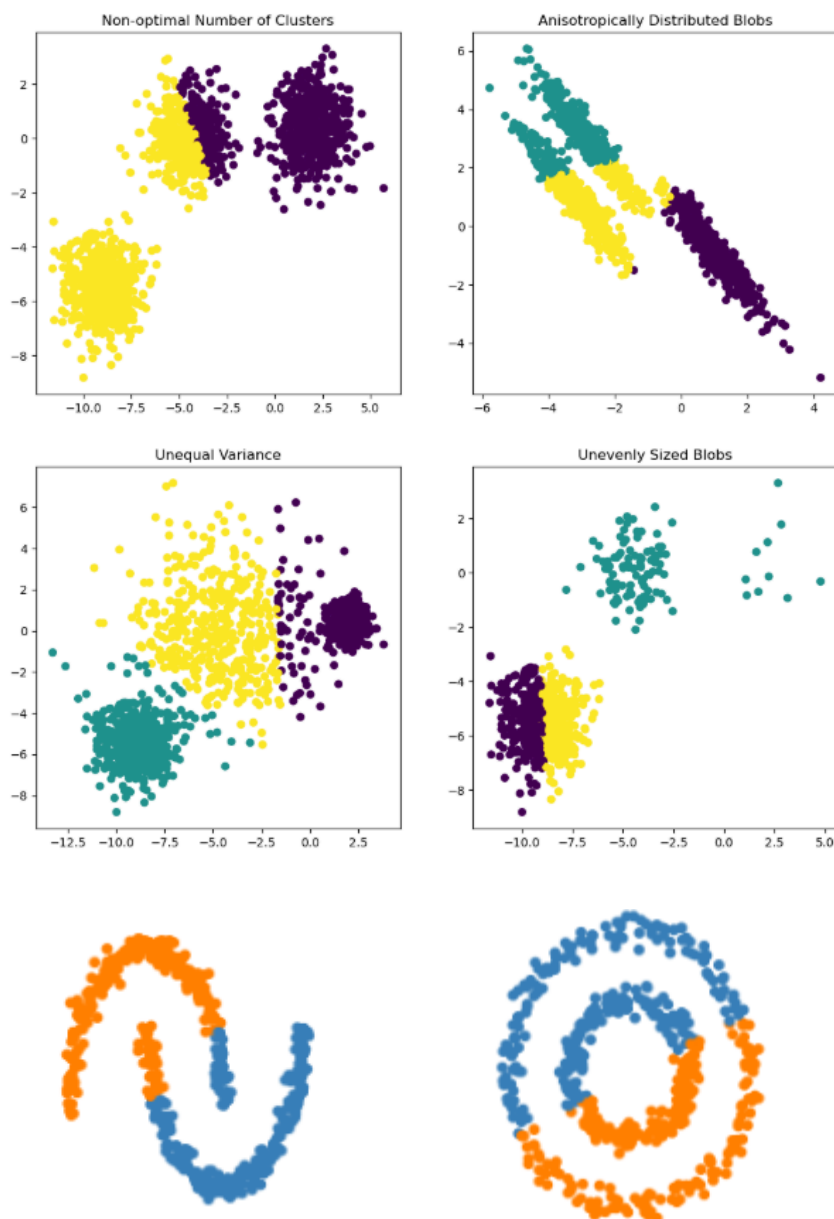
$$\sum_{i=0}^n \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

تنها ابر پارامتر این الگوریتم k یا همان تعداد خوشه‌ها می‌باشد. و در الگوریتم استاندارد kmeans معیار فاصله، فاصله اقلیدوسی می‌باشد.

به طور کلی در هر الگوریتم خوشه‌بندی درباره داده موردنظر فرضیاتی رخ می‌دهد و خروجی نهایی برای داده‌هایی با آن فرضیات بهتر خواهد بود.

اگر هر کدام از فرض‌هایی که ذکر شد درباره داده موثق نباشد این الگوریتم در پیدا کردن خوشه‌ها شکست خواهد خورد و خوشه‌بندی‌ها معنادار نخواهد بود.

در شکل پایین kmeans روی مجموعه داده‌هایی که هر کدام از این فرض‌ها را رعایت نکرده پیاده شده است.



دو شکل پایین و بالا سمت راست مربوط به isotropic نبودن، شکل وسط سمت راست مربوط به اندازه غیربرابر خوشه‌ها، شکل وسط چپ مربوط به میزان واریانس مختلف (پخش) مختلف خوشه‌ها و شکل بالا چپ مربوط به تعداد خوشه‌های اشتباه فرض شده است.

الگوریتم DBSCAN:

همینطور که از اسم این الگوریتم مشخص است، روش محاسبه خوشه‌ها بر اساس نواحی با چگالی بیشتر (dense) در داده است. این الگوریتم centroid based نیست یعنی خوشه‌ها را بر اساس مرکز خوشه‌ها پیدا نمی‌کند. (مانند kmeans) این مورد که DBSCAN فرض می‌کند داده صرفاً با نواحی کم چگالی از هم جدا می‌شوند باعث می‌شود بتواند در مجموعه داده‌ای که خوشه‌ها اندازه یکسان ندارند و isotropic نیستند بسیار بهتر عمل کند. این روش به طور هیچ فرضی درباره توزیع داده‌ها ندارد.

فرضیات این الگوریتم راجع به مجموعه داده شامل این است که:

۱. تمام خوشه‌ها دارای چگالی تقریباً یکسان هستند.
۲. اگر نمونه‌ای در هیچ کدام از خوشه‌ها نباشد، نویز است.
۳. خوشه‌ها صرفاً توسط نواحی کم چگالی از هم جدا می‌شوند و داده‌ها صرفاً بر اساس چگالی با هم ارتباط و وابستگی دارند.

این روش دارای دو ابر پارامتر $min\ samples$ و eps است و مشخص می‌کند که حداقل تعداد داده‌ای که باید در شعاع یک داده باشد تا core point باشد. border points نقاطی هستند که خود دارای $min\ samples$ نقطه در شعاع eps نیستند و خوشه را نمی‌توانند گسترش بدهند. همچنین اگر نقطه‌ای نه core بود نه border، نویز است. کلاً دو نقطه در یک خوشه قرار می‌گیرند اگر زنجیره‌ای از نقاط core آن‌ها را به هم متصل کنند.

به عنوان مثال تاثیرهای آنان به این شکل است که هر چه $min\ samples$ بیشتر باشد خوشه‌ها دارای چگالی بیشتر هستند و برعکس. همینطور eps کمتر باعث می‌شود تا خوشه‌ها به هم نزدیک‌تر باشند. (tightly packed) کلاً جایشگت‌های این دو ابرپارامتر به صورت مستقیم خوشه‌های نهایی را هم از نظر اندازه و تعداد آن‌ها تحت تاثیر قرار می‌دهد.

(ب)



این دو مجموعه داده داده دارای دو خوشه اند که شکل آن‌ها محدب و isotropic نیست. در نتیجه الگوریتمی مانند kmeans جواب نمی‌دهد. (در بخش الف توضیح داده شد.) برای خوشه‌بندی درست می‌توانیم از DBSCAN استفاده کنیم زیرا فرضی در این باب انجام نمی‌دهد.



این مجموعه داده دارای ۳ خوشه است که به خوبی از هم جدا شده‌اند و دارای شکل isotropic و محدب هستند. در نتیجه kmeans با $k = 3$ برای این مسئله عالی است. البته DBSCAN نیز با انتخاب بهینه دو ابرپارامترش می‌تواند خوب عمل کند ولی ممکن است دو خوشه بالایی را یک خوشه در نظر بگیرد.

(ج)

به طور کلی ابعاد بالای داده برای هر الگوریتم خوشه‌بندی distance based چالش ایجاد می‌کند. دو الگوریتم DBSCAN و Kmeans به عنوان مثال روی فاصله اقلیدوسی نمونه‌ها از هم تمرکز می‌کنند. ما می‌دانیم در ابعاد بسیار بالا معنای این فاصله از بین می‌رود و درواقع curse of dimensionality رخ می‌دهد و فواصل حساب شده دیگر معنی دار نیستند. در نتیجه این موضوع خوشه‌های پیدا شده را به خوبی نمی‌توان از هم جدا کرد. همینطور چون در ابعاد بالا sparsity بالا می‌رود در نتیجه به طور کلی چگالی داده نیز در کل مجموعه داده پایین می‌آید و روش DBSCAN به عنوان مثال دیگر نمی‌تواند به خوبی عمل کند زیرا دیگر ناحیه با چگالی بالا نداریم. درواقع در این الگوریتم‌ها فاصله دیگر قدرت discriminative ندارند در نتیجه خوشه‌های بدست آمده نیز نمی‌توانند معنی‌دار باشند.

(د)

با توجه به این حساسیت بالا می‌توان چند نتیجه‌گیری درباره داده کرد. می‌دانیم ϵ شعاعی را تعریف می‌کند که در آن باید همسایه‌ها را حساب کرد.

- ممکن است داده دارای چگالی متغیر باشد. یعنی در یک ناحیه تعداد بسیار زیادی نمونه وجود دارد و در کنارش این چگالی خیلی کمتر باشد. تغییر ϵ در اینجا می‌تواند در اندازه و تعداد خوشه‌ها بسیار تغییر ایجاد کند.
- ممکن است خوشه‌ها و خود داده به طور کلی بسیار در هم تنیده باشند (tightly coupled) و هرگونه تغییراتی در شعاع موردنظر به صورت زنجیروار بر روی اینکه یک نقطه core باشد یا خیر تاثیر بگذارد.
- همینطور ممکن است داده‌ها به خوبی از هم از طریق نواحی کم‌چگالی قابل جداسازی نباشند.

(ه)

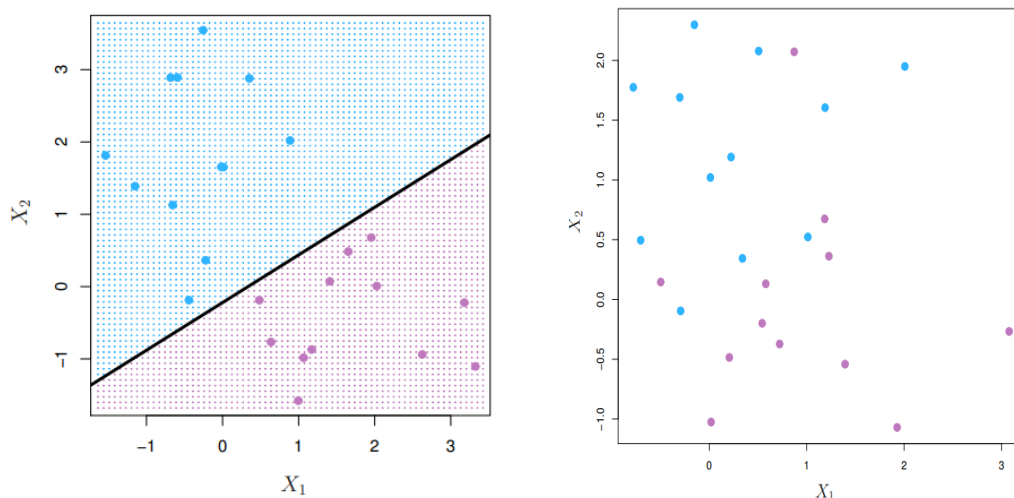
در اینجا از روشی با عنوان elbow criterion استفاده می‌کنیم که مقدار بهینه k را برای ما با توجه جمع sse ها مشخص می‌کند. هدف اصلی در این روش این است که نقطه‌ای را بیابیم که افزودن تعداد خوشه‌ها دیگر به صورت چشمگیر کیفیت خوشه‌ها را افزایش ندهد. بدیهی است که هرچه تعداد خوشه‌ها بالا برود حاصل جمع sse ها روی خوشه‌های مختلف کاهش می‌یابد زیرا خوشه‌ها کوچک‌تر می‌شوند در نتیجه تعداد داده‌های کمتری نیز در هر یک از آن‌ها خواهد بود. اما اگر از جایی تعداد خوشه‌ها بیشتر بشود ما ممکن است یک خوشه را به چندین خوشه بشکنیم و در نتیجه فرقی بین داده‌ها فرض کرده‌ایم که در واقع وجود ندارد. این روش می‌گوید نمودار جمع sse ها را نسبت به k رسم کنیم و نقطه‌ای بهینه می‌شود که شیب بین دو نقطه شروع به تقریباً ثابت و کم می‌شود. یعنی دیگر اضافه کردن خوشه‌ها کمک چندانی از نظر intra به ما نمی‌کند. با توجه به توضیحات ارائه شده، در اینجا این مقدار برابر با $k = 3$ می‌باشد.

سوال دوم)

(الف)

منظور از داده‌های غیر قابل جداسازی (non-separable) این است که نتوان به صورت خطی و با یک ابرصفحه داده‌ها را به طوری جدا کرد که هیچ خطایی نداشته باشیم. یعنی کلاس‌ها در هم آمیخته شده باشند و decision boundary غیر خطی باشد.

روش پایه‌ای SVM یا درواقع Hard SVM یا Maximal Margin Classifier فرض می‌کند که داده به صورت کامل و بدون هیچ خطایی قابل جداسازی هستند و ابرصفحه‌ای وجود دارد که به طول کاملاً درست داده‌ها را از هم جدا می‌کند. یعنی هیچ tolerance ای درقبال missclassification ندارند. این موضوع باعث می‌شود تا اگر داده غیر قابل جداسازی باشد نتوان مسئله را حل کرد زیرا درواقع دنبال ابرصفحه‌ای می‌گردیم که وجود ندارد. در پایین مجموعه داده غیر قابل جداسازی و قابل جداسازی در فضای دوبعدی نشان داده شده است.



(ب)

در SVM، margin، نشان‌دهنده فاصله ابرصفحه تا نزدیک‌ترین نقطه از دو کلاس یا support vectors است. (در یک مسئله دو کلاسه) هر چقدر margin بیشتر باشد مرز شفاف‌تر و بهتری بین داده‌ها وجود دارد و هر چه این مقدار کمتر باشد نشان‌دهنده نزدیک‌تر بودن داده‌ها از دو کلاس به یکدیگر است.

همانطور که می‌دانیم پایه و اساس تمام روش‌های SVM این است که margin را بیشینه کنند. زیرا:

- درواقع margin یک متریک برای میزان confidence ما از پیش‌بینی‌هایی است که انجام می‌دهیم زیرا هر چه فاصله یک نمونه از ابرصفحه پیدا شده بیشتر باشد ما با قطعیت بیشتری می‌توانیم راجع به کلاس آن نمونه حرف بزنیم.

- هر چه margin بیشتر باشد احتمال overfitting نیز بیشتر می‌شود و شدت حساسیت مدل به تغییرات کوچک در داده و نویز کمتر می‌شود.
- پس بیشینه کردن margin به مدل قدرت generalization می‌بخشد.

assumption: observations are i.i.d
 x_1, \dots, x_n

$$f(x; \alpha) = \begin{cases} \alpha x^{\alpha-1} & 0 < x < 1 \\ 0 & \text{o.w} \end{cases}$$

- ابتدا تابع Likelihood را می نویسیم

$$\begin{aligned} \mathcal{L}(\alpha) &= \prod_{i=1}^n f(x_i; \alpha) \\ &= \prod_{i=1}^n \alpha x_i^{\alpha-1} \end{aligned}$$

از طرفین لگاریتم می گیریم:

$$\log(\mathcal{L}(\alpha)) = \log \left[\prod_{i=1}^n \alpha x_i^{\alpha-1} \right]$$

$$\begin{aligned} * \log(ab) &= \log a + \log b * \\ &= \sum_{i=1}^n \log(\alpha x_i^{\alpha-1}) \\ &= \sum_{i=1}^n \log \alpha + \log x_i^{\alpha-1} \\ &= n \log \alpha + \sum_{i=1}^n (\alpha-1) \log x_i \end{aligned}$$

دنبال α ای هستیم که $\log(\mathcal{L}(\alpha))$ را بیشینه کند پس یک مسئله بهینه سازی تک متغیره داریم.

$$\log(\mathcal{L}(\alpha)) = L(\alpha)$$

$$\Rightarrow 0 = \frac{dL(\alpha)}{d\alpha}$$

$$\begin{aligned} * \frac{d}{d\alpha} \log \alpha &= \frac{1}{\alpha} * \\ &= n \frac{1}{\alpha} + \sum_{i=1}^n \log x_i \end{aligned}$$

$$\Rightarrow \alpha = \frac{n}{\sum_{i=1}^n \log x_i}$$

subject

$$A = \underset{m \times n}{U} \underset{m \times m}{\Sigma} \underset{m \times n}{V^T} \Rightarrow A = \underset{r \times r}{U} \underset{r \times r}{\Sigma} \underset{r \times r}{V}$$

حال باید مقادیر ویژه و بردارهای ویژه $A^T A$ را بیابیم.

دستی حسابان I-A راباید حساب لینم:

$$= (15 - \lambda)^2 (1 - \lambda) - 8(15 - \lambda)$$

$$-1200 + 1331$$

$$= \Omega_0 + \varepsilon \lambda$$

$$= (15 - \lambda)(1 - \lambda) - (15 \times 2 - 1 \times 2 \times 1)$$

$$= 15\Delta r - 149\lambda + 11\lambda^2 - \lambda^2 - 20\lambda + 24\lambda^2 - 15\Delta r + 102\lambda$$

$$= -\lambda^2 + 5\lambda - 25$$

$$= \lambda(-\lambda^2 + 9\lambda - 25) = -\lambda(\lambda - 25)(\lambda - 9)$$

$$\lambda_1 = 0, \lambda_2 = 25, \lambda_3 = 9$$

$$A \rightarrow I = A \rightarrow Ax = 0 \rightarrow A = \begin{bmatrix} 1 & 0 & r \\ 0 & 1 & -r \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_r \\ x_f \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

A₂

ref
of A

$$x_1 + 7x_4 = 0$$

$$m_r - \gamma m_r = U$$

$$n_c = 7$$

$$\Rightarrow \vec{n} = \begin{bmatrix} -x+ \\ x+ \\ t \end{bmatrix}$$

s.a.m

for $t=1$ $x = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}$ eigenvector for $\lambda=0$

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

$$(A - 1\lambda I)x = 0$$

$$\Rightarrow \begin{bmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$x_1 = x_2 = t$$

$$x_3 = 0 \quad \text{for } t=1 \quad x = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$

eigenvector for $\lambda=2$

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

$$(A - 2I)x = 0$$

$$\Rightarrow \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$x_1 - x_3 = 0$$

$$x_2 + 2x_3 = 0$$

$$x_3 = t$$

$$\Rightarrow \vec{x} = \begin{bmatrix} t \\ -2t \\ t \end{bmatrix} \quad \text{for } t=1 \quad x = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$$

eigen vector
for $\lambda=9$

حال ماتریس $\sum_{1 \times 3}$ را می سازیم که یک ماتریس قطری است و در اینجا
آن مقادیر یک ماتریس ATA می باشند (بلاکه)

البته به صورت مرتب شده (descending) $\sigma_1 = 0, \sigma_2 = 5, \sigma_3 = 3$

$$\Sigma = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix}$$

در نظر داریم ماتریس های V و U متعامد هستند یعنی بردارهای تک‌جهت
نسبت به هم ۳ تان باید ایک باشند پس:

$$\alpha_1 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \\ \frac{1}{2} \end{bmatrix} \quad \alpha_2 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \\ 0 \end{bmatrix} \quad \alpha_3 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}$$

ستون های V در واقع همان بردارهای ویژه $A^T A$ هستند

$$V = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 0 & \frac{\sqrt{2}}{2} & \frac{1}{2} \end{bmatrix}$$

از طرفی می دانیم رابطه زیر برای U برقرار است.

$$U = \frac{1}{\sigma_i} A \alpha_i$$

$$\begin{aligned} A \alpha_1 &= \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix} \\ A \alpha_2 &= \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix} \end{aligned} \Rightarrow U = \frac{1}{\sigma_i} \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 2 & 2 & 2 \\ 2 & 2 & -2 \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix} \begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \frac{1}{2} \end{bmatrix}$$

transpose of

✓

سوال پنجم)

(الف)

با فرض اینکه $A_{m \times n}$ دارای m سطر و n ستون است با تجزیه آن به ۳ ماتریس زیر می‌رسیم:

• ماتریس $U_{m \times r}$

• ماتریس $S_{r \times r}$

• ماتریس $V_{n \times r}$

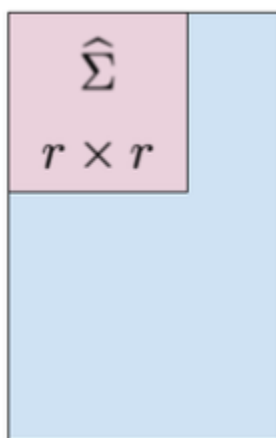
این مدل SVD فقط singular value های غیر صفر را در نظر می‌گیرد که تعداد آن‌را با r نشان می‌دهیم و $r \leq n$. برای همین است که ماتریس S مربعی شده است. (compact SVD)
اگر full SVD داشته باشیم ماتریس‌های بدست آمده به صورت زیر خواهد بود:

• ماتریس $U_{m \times m}$

• ماتریس $S_{m \times n}$

• ماتریس $V_{n \times n}$

در اینجا ماتریس S به طور کلی به صورت زیر است:



S

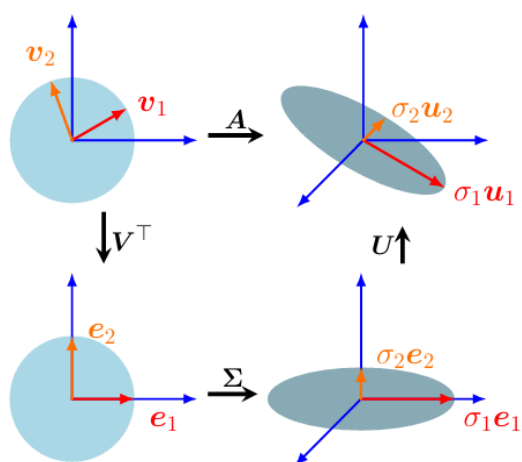
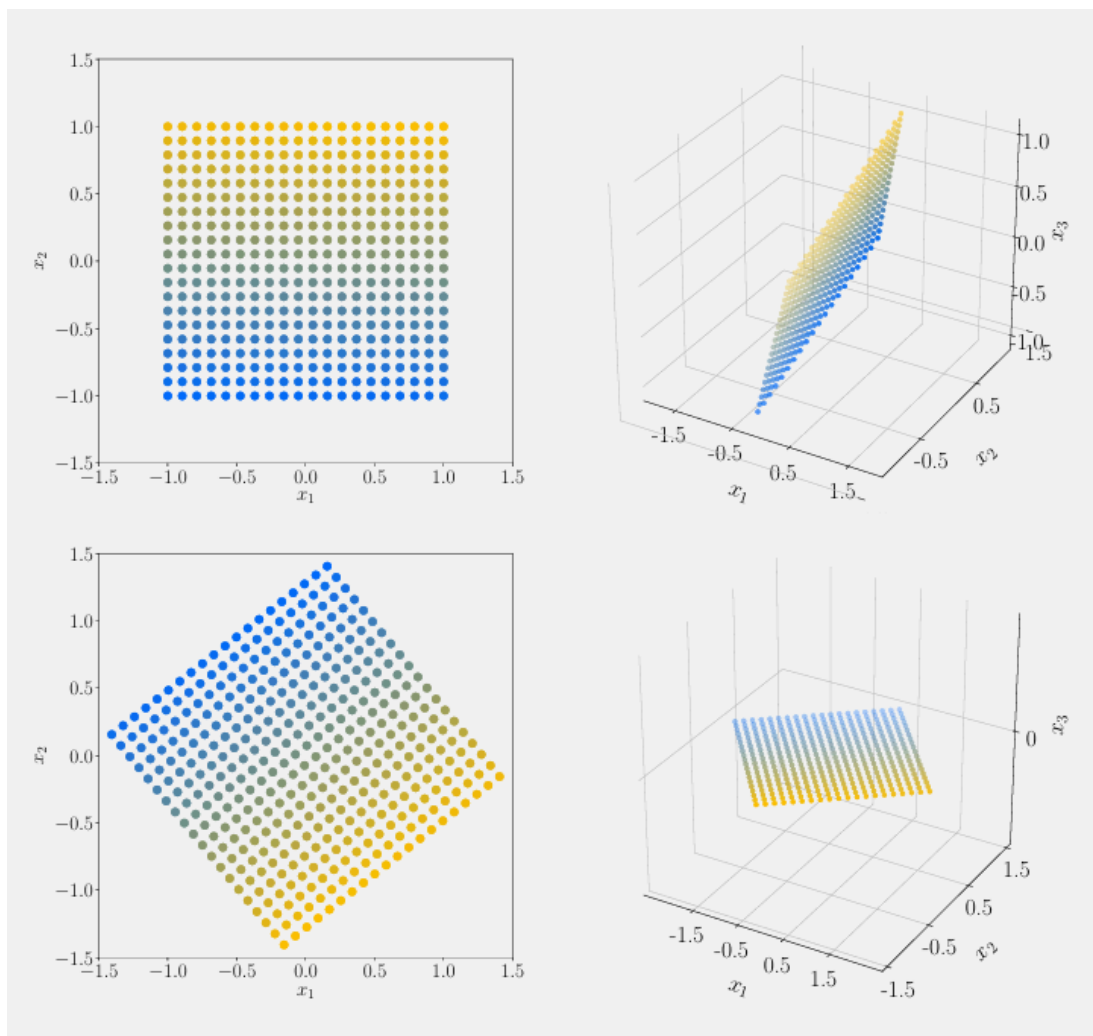
$$\begin{array}{c}
 \boxed{\begin{array}{c} A \\ n \times d \end{array}} = \boxed{\begin{array}{c} \hat{U} \\ n \times r \end{array}} \boxed{\begin{array}{c} \hat{\Sigma} \\ r \times r \end{array}} \boxed{\begin{array}{c} \hat{V}^T \\ r \times d \end{array}} \\
 \begin{array}{ccc} U & \Sigma & V^T \\ n \times n & n \times d & d \times d \end{array}
 \end{array}$$

(ب)

اگر از analogy در اسلاید استفاده کنیم و فرض کنیم در $A_{m \times n}$ دارای m کاربرد و n فیلم هستیم ماتریس U نشان‌دهنده ارتباط بین کاربر و ژانر (concept) است و در واقع هر ستون U نشان‌دهنده هر کدام از آن ژانرها می‌باشد و مقادیر آن نشان‌دهنده میزان آن شخص به آن ژانر. مانند action یا romance. همچنین ماتریس V نشان‌دهنده ارتباط بین هر فیلم و ژانر می‌باشد یعنی هر کدام از ستون‌های آن نشان‌دهنده میزان وجود آن ژانر در هر کدام از فیلم‌ها می‌باشد. همچنین مقادیر در S نیز نشان‌دهنده اهمیت هر کدام از این ژانرها در کل داده موجود است.

اما اگر بخواهیم به صورت جبری بررسی کنیم و ماتریس A را شامل یک تبدیل از فضای R^m به R^n بدانیم این‌گونه است که ستون‌های V نشان‌دهنده بردارهای پایه‌ای فضای R^n است که A بردارها را به آن نگاشت می‌کند (فضای ستونی A). در نتیجه V^T مانند یک عملگر برای basis change می‌ماند. همین‌طور ستون‌های U نشان‌دهنده basis فضای R^m ورودی است یعنی هر کدام از ستون‌های آن هم راستا با فضای سطری A است. و همین‌طور مقادیر در S یکسری re scaling انجام می‌دهد.

تبدیل از R^2 به R^3 با ماتریس $A_{3 \times 2}$ (به صورت پات ساعتگرد)



men: [171, 171, 171, 171, 171]

women: [175, 170, 171, 172, 172]

$$f(x_i; \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \quad (1)$$

$$L(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

$$\ln(L(\mu, \sigma^2)) = \sum_{i=1}^n \left[\ln\left(\frac{1}{\sigma \sqrt{2\pi}}\right) - \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

$$\log a b = \log a + \log b$$

$$L(\mu, \sigma^2) = \log(L(\mu, \sigma^2))$$

$$= -n \ln(\sigma \sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

$$* \quad 0 = \frac{dL(\mu, \sigma^2)}{d\mu} = \frac{d}{d\mu} \left(-n \ln(\sigma \sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right)$$

$$= -\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)$$

$$= -\frac{1}{\sigma^2} \sum_{i=1}^n x_i + n$$

$$\Rightarrow \sum_{i=1}^n x_i = n\mu \rightarrow \mu = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\hat{\mu}_m = \frac{171 + 171 + 171 + 171 + 171}{5} = 171$$

$$\hat{\mu}_w = \frac{175 + 170 + 171 + 172 + 172}{5} = 172$$

$$0 = \frac{dL(\mu, \sigma^2)}{d\sigma} = \frac{d}{d\sigma} \left(-n \ln \sigma - \frac{n}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right)$$

$$= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (x_i - \mu)^2$$

$$\Rightarrow -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (x_i - \mu)^2 = 0 \Rightarrow \sigma = \frac{1}{\sqrt{n}} \sqrt{\sum_{i=1}^n (x_i - \mu)^2}$$

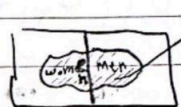
$$\hat{\sigma}_{\text{of men}} : \frac{1}{5} \sqrt{\sum_i (x_i - 174.8)^2} = 4.94 \quad \hat{\sigma}_m$$

$$\hat{\sigma}_{\text{of women}} : \frac{1}{5} \sqrt{\sum_i (x_i - 149.4)^2} = 3.95 \quad \hat{\sigma}_w$$

$$f_{\text{men}}(x; \mu, \sigma^2) = N(174.8, 24.52)$$

$$f_{\text{women}}(x; \mu, \sigma^2) = N(149.4, 15.64)$$

(ب) قاتل پسر، استفاده می‌کنیم:



قاتل احتمال دل

احتمال دیدن فردی با قد 149:

$$P(149) = P(149 | \text{man}) P(\text{man}) + P(149 | \text{woman}) P(\text{woman})$$

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

از مدل هایی که درست آوردیم استفاده می‌کنیم

احتمال مرد بودن کسی با قد 149 دارد:

$$\frac{P(149 | \text{man}) P(\text{man})}{P(149 | \text{man}) P(\text{man}) + P(149 | \text{woman}) P(\text{woman})}$$

$P(149 | \text{man}) \rightarrow$ یعنی می‌دانیم پسر مرد است پس از مدتی

برای قد مردان درست آوردیم یا با استفاده کنیم. همین طور برای زن ها.

$$P(149 | \text{man}) = 0.04$$

$$P(149 | \text{woman}) = 0.04$$

$$P(\text{man}) = P(\text{woman}) = \frac{1}{2}$$

$$\Rightarrow \frac{0.04 \cdot \frac{1}{2}}{0.04 \cdot \frac{1}{2} + 0.04 \cdot \frac{1}{2}} = \frac{0.02}{0.04} = \frac{1}{2}$$

سوال هفتم

Subject:

Date: / /

$$\text{Bayes theorem: } P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

مثال: بر مبنای فرضی که ویژگی‌ها یا تفرقه به C از یک ویژگی مستقل هستند

$$P(x_1, x_2) = P(x_1)P(x_2)$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

پس ای داشته باشیم $X = (x_1, \dots, x_n)$

$$P(X|C) = P(x_1|C) \cdot P(x_2|x_1, C) \cdot \dots \cdot P(x_n|x_1, \dots, x_{n-1}, C)$$

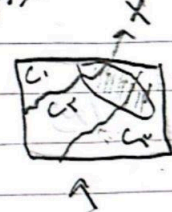
زیرا x_i های $P(x_i|x_1, \dots, x_{i-1}, C) = P(x_i|C)$ چون از هم مستقل هستند

دستی اطلاعاتی درباره x_i به معنی دهت و نقطه می باشد

$$\Rightarrow P(X|C) = \prod_{i=1}^n P(x_i|C)$$

حال باقیانند می‌داریم:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$



$$\Rightarrow \frac{\prod_{i=1}^n P(x_i|C) \cdot P(C)}{P(X)}$$

از طرفی $P(X) = \sum_{C'} P(X|C')P(C')$
 قاعدن احتمال کل \downarrow \downarrow \downarrow
 دیدن x با درستی \downarrow \downarrow \downarrow
 در C \downarrow \downarrow \downarrow

$$P(X|C') \Rightarrow P(X|C') = \prod_{i=1}^n P(x_i|C')$$

$$P(X) = \sum_{C'} P(C') \prod_{i=1}^n P(x_i|C')$$

$$\text{Final formula: } P(C|X) = \frac{P(C) \prod_{i=1}^n P(x_i|C)}{\sum_{C'} P(C') \prod_{i=1}^n P(x_i|C')}$$

در زبان inference چون می‌فهمیم احتمال بالایی را بیاییم به بیش‌ترین احتمال
موقع را دارد و صرفاً باید مقایسه انجام شود. می‌توانیم مترجم را در نظر بگیریم

$$P(c|X) \propto P(c) \prod_{i=1}^n P(x_i|c)$$