



تمرین سری سوم

امیر حسین محمدی

۹۹۲۰۱۰۸۱



(آ) (۳ نمره) آیا الگوریتم K-means با معیار فاصله اقلیدسی، حالت خاصی از الگوریتم EM است که در آن از k تابع گوسی با واریانس یکسان برای هربعد استفاده می‌کنیم؟



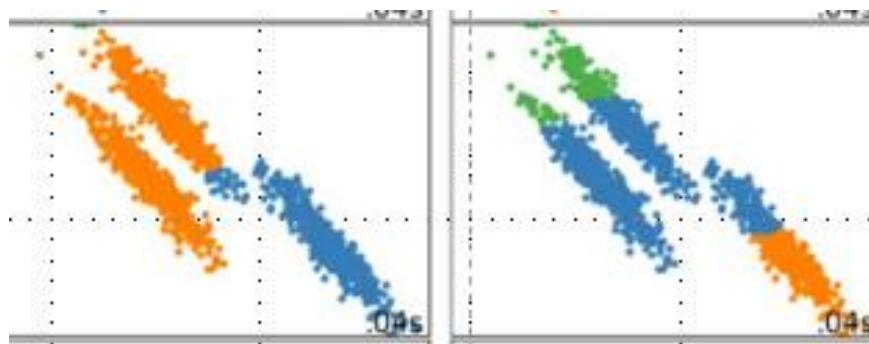
بله درست است و در واقع در `kmeans`، اندازه گیری فاصله اقلیدسی تضمین می کند که مناطق اطراف یک مرکز خوشه متشکل از نقاط نزدیک به آن مرکز (که یک خوشه است) گروهی شکل باشد. همچنین، این مدل اندازه گیری از ایجاد خوشه های با اندازه دلخواه جلوگیری می کند. بنابر این هر خوشه یک ماتریس کواریانس مشابه دارد و این باعث می شود که هر خوشه به فضای گروهی به خود بگیرد (بنابراین `kmeans` ترکیبی از k گوسی با واریانس یکسان برای هر بعد هستند و زیر مجموعه ای از الگوریتم EM است) که این مسئله سبب می شود که الگوریتم `kmeans` به ازای توزیع های مختلف نتایج مطلوب نداشته باشد.



(ب) (۴ نمره) درخوشه‌بندی سلسله مراتبی با کدامیک از معیارهای شباهت خوشه‌ای امکان دارد که داده‌ای در یک خوشه به داده‌ای از خوشه دیگر، نزدیکتر از داده‌ای درخوشه خودش باشد.



این مسئله هم در حالت خوشه بندی single link وجود دارد هم در حالت خوشه بندی complete link وجود دارد. در حالت single link ممکن است حالتی پیش بیاید که مثلاً تو تا توزیع (خوشه) به موازات هم قرار دارند و قسمتی از نقاط یک خوشه به نقاط خوشه دیگر نزدیک تر باشد تا نقاط داده های خودش. این مسئله در حالت complete link نیز وجود دارد زیرا ما در complete محدودیتی داریم تحت عنوان اینکه اندازه گیری بین دو خوشه بر اساس بیشترین فاصله بین دو نقطه ی آنها انجام می شود بنابراین در این حالت می تواند و قسمتی از نقاط یک خوشه به نقاط خوشه دیگر نزدیک تر باشد تا نقاط داده های خودش





(ج) (۲ نمره) به یک مجموعه از راس‌ها p -cluster گفته می‌شود اگر حداقل p درصد از یال‌های این راس‌ها به راس‌های داخل این مجموعه متصل باشند. اگر ما خوشه‌ها را p -cluster های گراف در نظر بگیریم، آیا نتیجه خوشه‌بندی با این تعریف مطلوب است؟

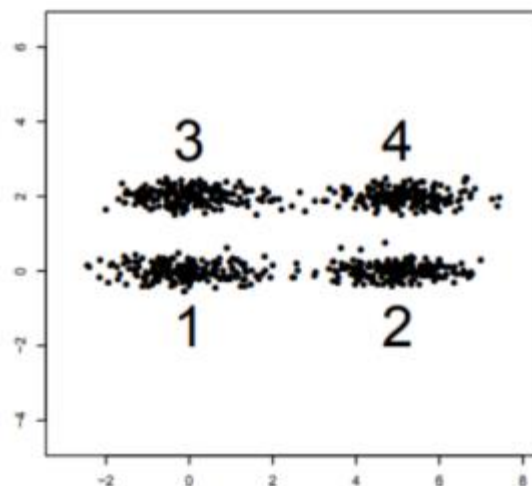


همانطور که گفته شد p cluster به مجموعه ای از راس ها گفته می شود که حداقل p درصد از یال های این راس ها به راس های داخل این کلاستر (مجموعه) متصل باشد. با توجه به فرضیات مسئله اگر ما تعداد یال ها را افزایش دهیم این مسئله می تواند باعث شود که برای مجموعه ای راس ها دیگر p درصد از یال به گره های همان مجموعه نروند و به گره های مجموعه های دیگر متصل باشند بنابراین در این حالت ممکن است راس های موجود در یک کلاستر به کلاستر های دیگر اختصاص یابد و این مسئله بر خلاف روند خوشه بندی است. بنابراین این روش تا به اندازه ای می تواند مانند خوشه بندی خوب عمل کند و به صورت کلی خوب نیست.

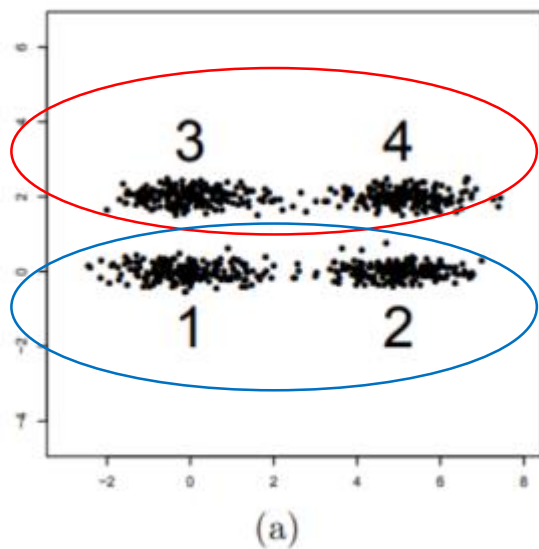


۱.۲. (۶ نمره) با توجه به مجموعه داده‌های شکل ۱ به سوالات زیر پاسخ دهید:

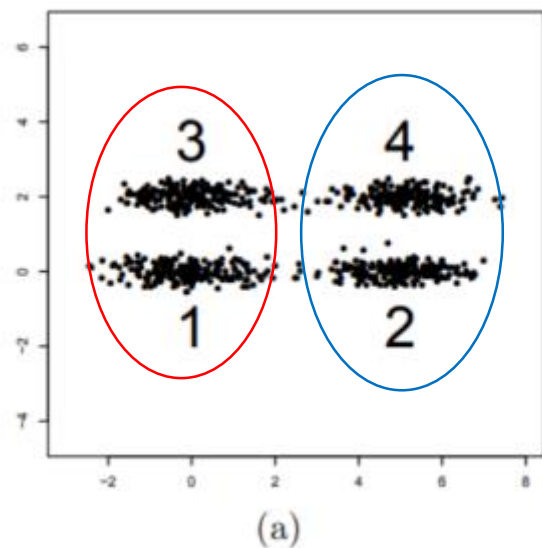
(آ) (۳ نمره) در مجموعه داده a اگر با استفاده از خوشه‌بندی سلسله مراتبی با $K = 2$ خوشه‌بندی صورت گیرد، با استفاده از هرکدام از معیارهای شباهت خوشه‌ای $single_link$ و $complete_link$ و $average_link$ ، ۴ دسته مشخص شده به چه خوشه‌ای تعلق می‌گیرند؟



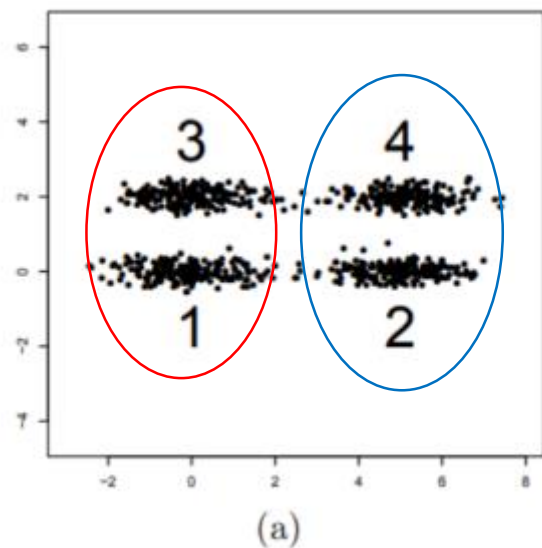
(a)



در این حالت با توجه به الگوریتم single link، به طور مثال اگر توزیع شماره ۳ را به عنوان یک کلاستر در نظر بگیریم این توزیع با نزدیک ترین نمونه موجود در توزیع چهار فاصله ی کمتری دارد (زیرا نمونه هایی نیز بین ۳ و ۴ وجود دارد) تا توزیع ۱ و ۲ (برای بقیه ی حالات هم به همین صورت) بنابراین توزیع ۳ و ۴ در یک خوشه و توزیع ۱ و ۲ در یک خوشه قرار می گیرند.



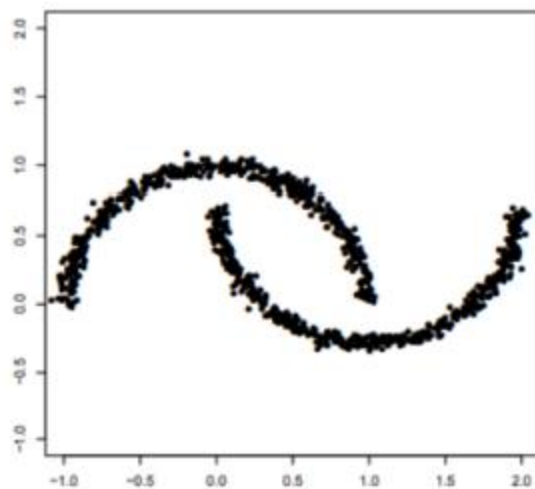
در این حالت با توجه به الگوریتم complete link، ما دورترین نقاط فواصل را در نظر می گیریم و سپس نقطه ی منیمم را انتخاب می کنیم. همانطور که از شکل مشخص است. مثلاً وقتی 3 را یک کلاستر در نظر بگیریم فاصله ی کلاستر 3 تا دورترین نقطه ی کلاستر 1 کمتر است (تا کلاستر 4 و 2) بنابراین توزیع 3 و 1 در یک کلاستر قرار می گیرند. (همین روند برای 4 و 2)



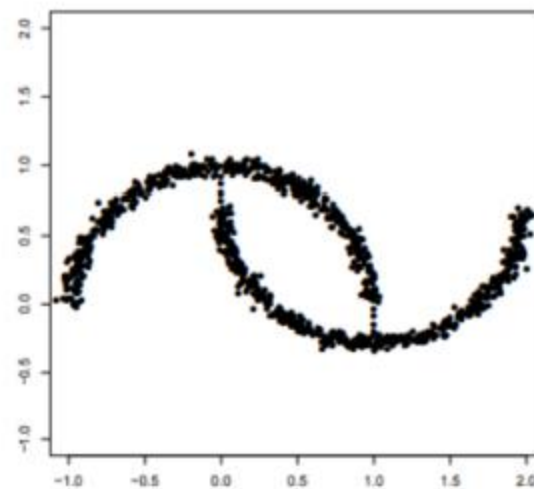
در این حالت با توجه به الگوریتم **average link**، ما میانگین فواصل نقاط را در نظر می گیریم و سپس نقطه ی منیمم را انتخاب می کنیم. همانطور که از شکل مشخص است. مثلاً وقتی 3 را یک کلاستر در نظر بگیریم فاصله ی کلاستر 3 تا میانگین نقاط کلاستر 1 کمتر است (تا کلاستر 4 و 2) بنابراین توزیع 3 و 1 در یک کلاستر قرار می گیرند. (همین روند برای 4 و 2)



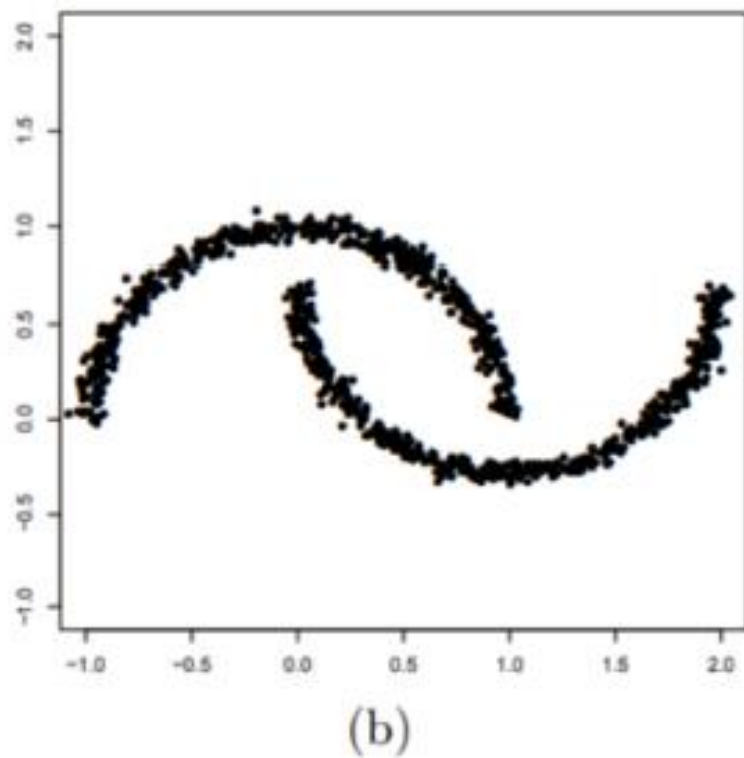
(ب) (۳ نمره) کدامیک از سه معیار فاصله در صورت وجود میتوانند داده‌ها در دو شکل c و b را با موفقیت جدا کند؟ پاسخ خود را به صورت خلاصه توضیح دهید.



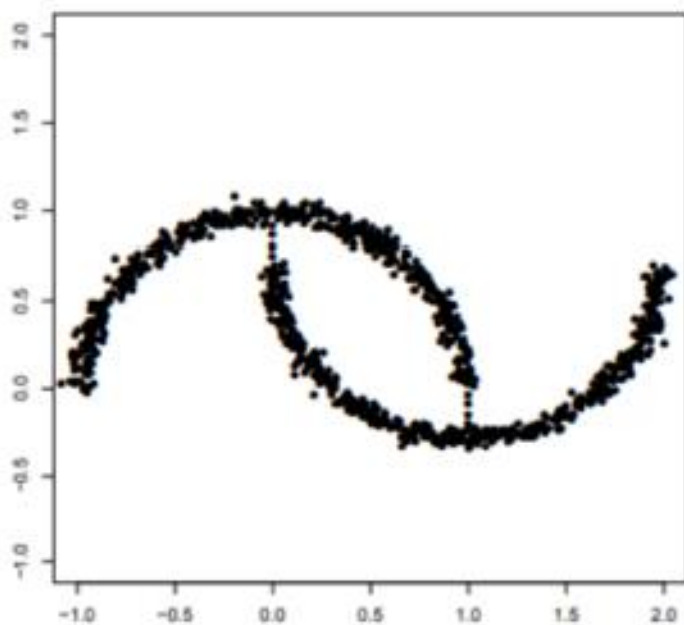
(b)



(c)



در این حالت با توجه به الگوریتم **single link**، که بر اساس نزدیک ترین فواصل و مینیمم آنها خوشه بندی انجام می دهد. می توانیم **complete link** را به عنوان بهترین گزینه در نظر بگیریم. (زیرا همانطور که شکل پیداست دو توزیع از یکدیگر جدا هستند و نمونه های هر دسته کمترین فاصله را نسبت به یک دیگر دارند تا خوشه ی دیگر)



(c)

در این حالت با توجه به الگوریتم single link، که بر اساس نزدیک ترین فواصل و مینیمم آنها خوشه بندی انجام می دهد. چون دو دسته نسبت به حالت قبل که از یکدیگر از هم جدا بودند در این جا جدا نیستند می تواند کل دو توزیع در یک خوشه قرار بگیرند در صورتی که واقعا در یک خوشه نیستند. بنابراین برای این حالت می توانیم از الگوریتم complete link استفاده کنیم که بر اساس دورترین فواصل خوشه بندی انجام می دهد



۱.۳. (۵ نمره) برای ترکیب دو خوشه در خوشه‌بندی سلسله‌مراتبی روشی به نام **Ward** هست که با محاسبه تابع هزینه مورد استفاده در خوشه‌بندی k-means یعنی یافتن کمینه مربع مجذور فاصله، دو خوشه را ترکیب می‌کند. تابع هزینه مورد استفاده، تابع زیر است:

$$\text{cost}(T) = \sum_{x \in S} \min_{t \in T} \|x - t\|^2$$


ثابت کنید برای هر دو خوشه C و C' تابع هزینه ترکیب این دو خوشه، معادله زیر حاکم است:

$$\text{cost}(C \cup C') = \text{cost}(C) + \text{cost}(C') + \frac{|c||c'|}{|c|+|c'|} \|\text{mean}(C) - \text{mean}(C')\|^2$$



فرض می کنیم:

- $\mu = \text{mean}(C)$,
- $\mu' = \text{mean}(C')$,
- $\mu^- = \text{mean}(C \cup C')$

طبق فرضیات مسئله 

$$\text{cost}(T) = \sum_{x \in S} \min_{t \in T} \|x - t\|^2$$
$$\text{cost}(C \cup C') = \text{cost}(C) + \text{cost}(C') + \frac{|c||c'|}{|c|+|c'|} \|\text{mean}(C) - \text{mean}(C')\|^2$$

$$\begin{aligned} \text{cost}(C \cup C') - \text{cost}(C) - \text{cost}(C') &= \sum_{x \in C \cup C'} \|x - \mu^-\|^2 - \sum_{x \in C} \|x - \mu\|^2 - \sum_{x \in C'} \|x - \mu'\|^2 \\ &= \sum_{x \in C} (\|x - \mu^-\|^2 - \|x - \mu\|^2) + \sum_{x \in C'} (\|x - \mu^-\|^2 - \|x - \mu'\|^2) \\ &= |C| \cdot \|\mu - \mu^-\|^2 + |C'| \cdot \|\mu' - \mu^-\|^2 \\ &= |c| \cdot \left\| \frac{|c'|}{|c| + |c'|} (\mu' - \mu^-) \right\|^2 + |c'| \cdot \left\| \frac{|c|}{|c| + |c'|} (\mu - \mu^-) \right\|^2 \\ &= \frac{|c| \cdot |c'|}{|c| + |c'|} \|\mu - \mu'\|^2 \end{aligned}$$



۱.۴. (۶ نمره) جدول فاصله برای ۶ شی زیر را در نظر بگیرید:

	A	B	C	D	E	F
A	0					
B	0.12	0				
C	0.51	0.25	0			
D	0.84	0.16	0.14	0		
E	0.28	0.77	0.70	0.45	0	
F	0.34	0.61	0.93	0.20	0.67	0

جدول ۱: داده‌های مسئله ۵

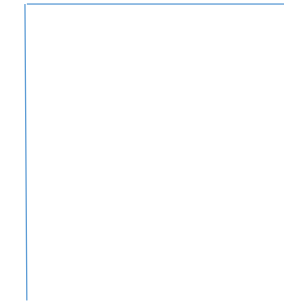
(آ) نمودار درختی برای خوشه‌بندی^۱ به روش Single-linkage را رسم کنید.

(ب) نمودار درختی برای خوشه‌بندی به روش Complete-linkage را رسم کنید.

(ج) دو مقدار از جدول بالا را چنان تغییر دهید که نمودارهای دو سوال قبل مشابه شوند.



	A	B	C	D	E	F
A	0					
B	0.12	0				
C	0.51	0.25	0			
D	0.84	0.16	0.14	0		
E	0.28	0.77	0.70	0.45	0	
F	0.34	0.61	0.93	0.20	0.67	0



A

B

$$\text{Min}(\text{dist}(A,B),C)=\text{Min}(\text{dist}(A,C),\text{dist}(B,C))=\text{Min}(0.51,0.25)=0.25$$

$$\text{Min}(\text{dist}(A,B),D)=\text{Min}(\text{dist}(A,D),\text{dist}(B,D))=\text{Min}(0.84,0.16)=0.16$$

$$\text{Min}(\text{dist}(A,B),E)=\text{Min}(\text{dist}(A,E),\text{dist}(B,E))=\text{Min}(0.28,0.77)=0.28$$

$$\text{Min}(\text{dist}(A,B),F)=\text{Min}(\text{dist}(A,F),\text{dist}(B,F))=\text{Min}(0.34,0.61)=0.34$$

	AB	C	D	E	F
AB	0				
C	0.25	0			
D	0.16	0.14	0		
E	0.28	0.70	0.45	0	
F	0.34	0.93	0.20	0.67	0



	AB	C	D	E	F
AB	0				
C	0.25	0			
D	0.16	0.14	0		
E	0.28	0.70	0.45	0	
F	0.34	0.93	0.20	0.67	0



$$\text{Min}(\text{dist}(C,D),AB)=\text{Min}(\text{dist}(C,AB),\text{dist}(D,AB))=\text{Min}(0.25,0.16)=0.16$$

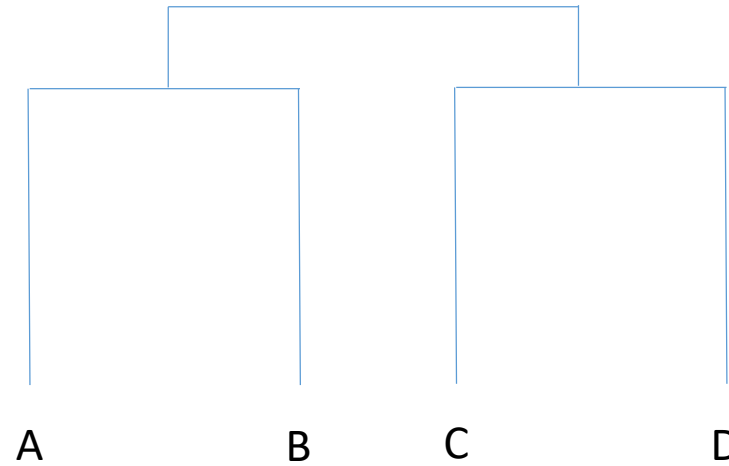
$$\text{Min}(\text{dist}(C,D),E)=\text{Min}(\text{dist}(C,E),\text{dist}(D,E))=\text{Min}(0.70,0.45)=0.45$$

$$\text{Min}(\text{dist}(C,D),F)=\text{Min}(\text{dist}(C,F),\text{dist}(D,F))=\text{Min}(0.93,0.20)=0.20$$

	AB	CD	E	F
AB	0			
CD	0.16	0		
E	0.28	0.45	0	
F	0.34	0.20	0.67	0



	AB	CD	E	F
AB	0			
CD	0.16	0		
E	0.28	0.45	0	
F	0.34	0.20	0.67	0

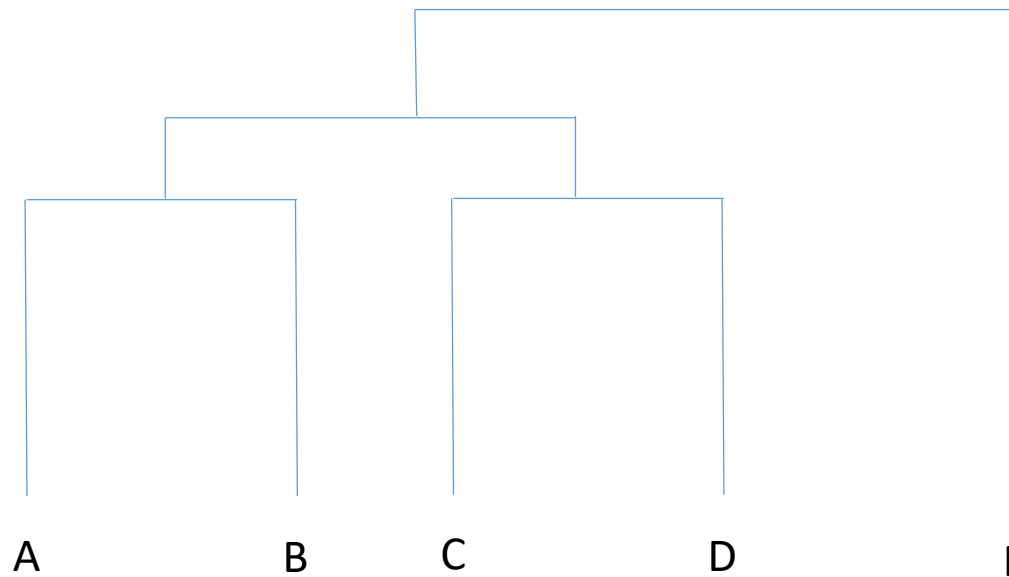


$\text{Min}(\text{dist}(\text{AB}, \text{CD}), \text{E}) = \text{Min}(\text{dist}(\text{AB}, \text{E}), \text{dist}(\text{CD}, \text{E})) = \text{Min}(0.28, 0.45) = 0.28$

$\text{Min}(\text{dist}(\text{AB}, \text{CD}), \text{F}) = \text{Min}(\text{dist}(\text{AB}, \text{F}), \text{dist}(\text{CD}, \text{F})) = \text{Min}(0.34, 0.20) = 0.20$

	ABCD	E	F
ABCD	0		
E	0.28	0	
F	0.20	0.67	0

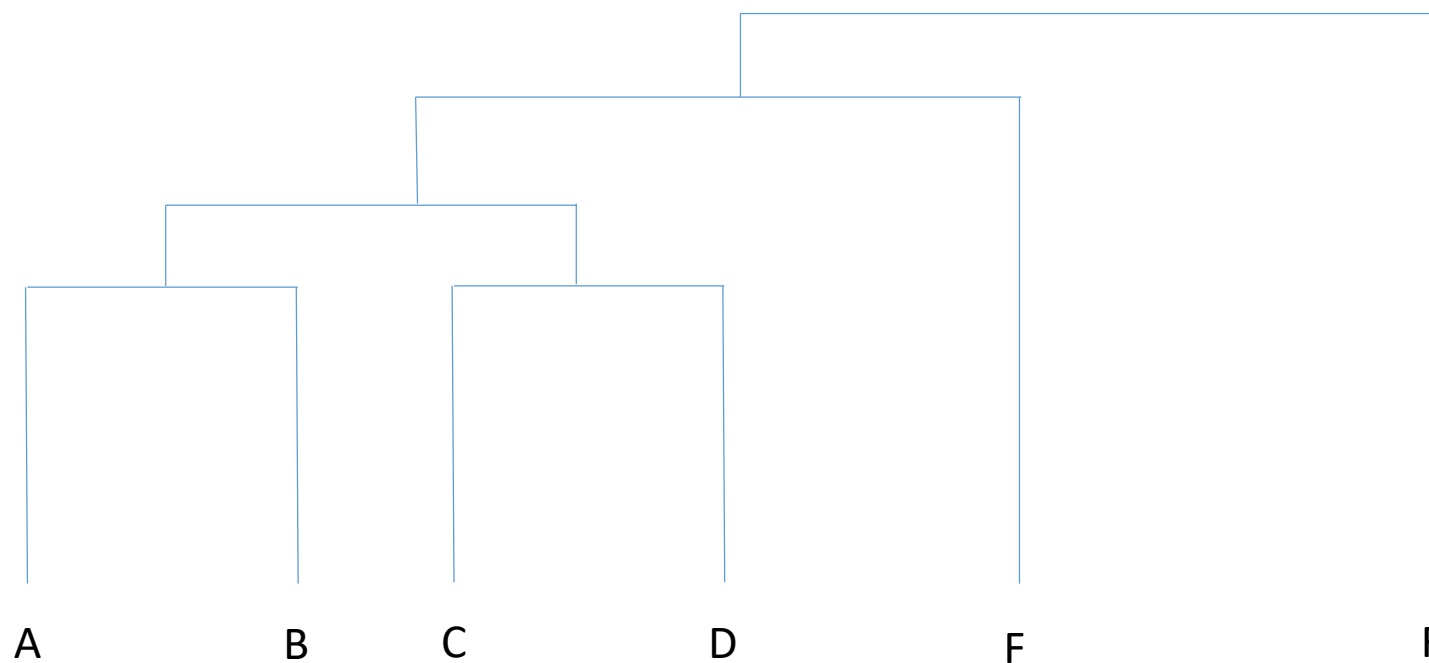
	ABCD	E	F
ABCD	0		
E	0.28	0	
F	0.20	0.67	0



$$\text{Min}(\text{dist}(\text{ABCD}, \text{F}), \text{E}) = \text{Min}(\text{dist}(\text{ABCD}, \text{E}), \text{dist}(\text{F}, \text{E})) = \text{Min}(0.28, 0.67) = 0.28$$

	ABCDF	E
ABCDF	0	
E	0.28	0

	ABCDF	E
ABCDF	0	
E	0.28	0





	A	B	C	D	E	F
A	0					
B	0.12	0				
C	0.51	0.25	0			
D	0.84	0.16	0.14	0		
E	0.28	0.77	0.70	0.45	0	
F	0.34	0.61	0.93	0.20	0.67	0



A

B

$$\text{Max}(\text{dist}(A,B),C)=\text{Mac}(\text{dist}(A,C),\text{dist}(B,C))=\text{Max}(0.51,0.25)=0.51$$

$$\text{Max}(\text{dist}(A,B),D)=\text{Max}(\text{dist}(A,D),\text{dist}(B,D))=\text{Max}(0.84,0.16)=0.84$$

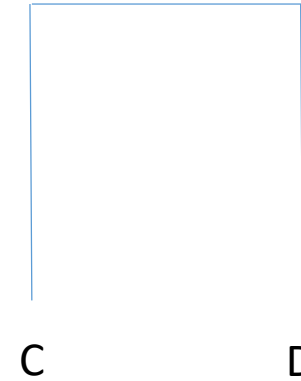
$$\text{Max}(\text{dist}(A,B),E)=\text{Max}(\text{dist}(A,E),\text{dist}(B,E))=\text{Max}(0.28,0.77)=0.77$$

$$\text{Max}(\text{dist}(A,B),F)=\text{Max}(\text{dist}(A,F),\text{dist}(B,F))=\text{Max}(0.34,0.61)=0.61$$

	AB	C	D	E	F
AB	0				
C	0.51	0			
D	0.84	0.14	0		
E	0.77	0.70	0.45	0	
F	0.61	0.93	0.20	0.67	0



	AB	C	D	E	F
AB	0				
C	0.51	0			
D	0.84	0.14	0		
E	0.77	0.70	0.45	0	
F	0.61	0.93	0.20	0.67	0



$\text{Max}(\text{dist}(C,D),AB)=\text{Max}(\text{dist}(C,AB),\text{dist}(D,AB))=\text{Max}(0.51,0.84)=0.84$

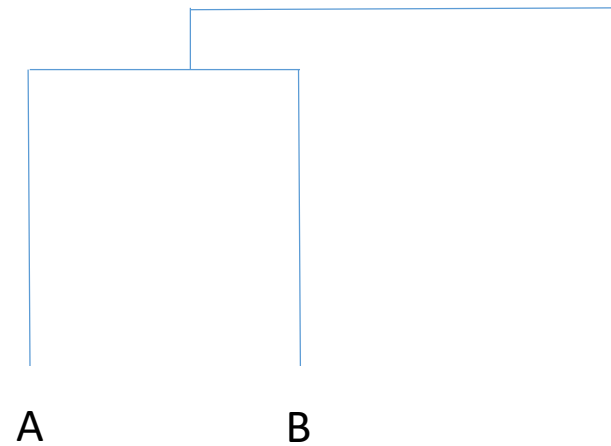
$\text{Max}(\text{dist}(C,D),E)=\text{Max}(\text{dist}(C,E),\text{dist}(D,E))=\text{Max}(0.70,0.45)=0.70$

$\text{Max}(\text{dist}(C,D),F)=\text{Max}(\text{dist}(C,F),\text{dist}(D,F))=\text{Max}(0.93,0.20)=0.93$

	AB	CD	E	F
AB	0			
CD	0.84	0		
E	0.77	0.70	0	
F	0.61	0.93	0.67	0



	AB	CD	E	F
AB	0			
CD	0.84	0		
E	0.77	0.70	0	
F	0.61	0.93	0.67	0



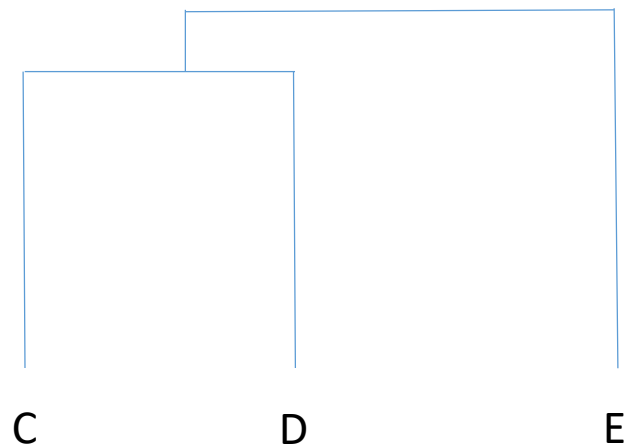
$\text{Max}(\text{dist}(\text{AB}, \text{F}), \text{CD}) = \text{Max}(\text{dist}(\text{AB}, \text{CD}), \text{dist}(\text{F}, \text{CD})) = \text{Max}(0.84, 0.93) = 0.93$

$\text{Max}(\text{dist}(\text{AB}, \text{F}), \text{E}) = \text{Max}(\text{dist}(\text{AB}, \text{E}), \text{dist}(\text{F}, \text{E})) = \text{Max}(0.77, 0.67) = 0.77$

	ABF	CD	E
ABF	0		
CD	0.93	0	
E	0.77	0.70	0



	ABF	CD	E
ABF	0		
CD	0.93	0	
E	0.77	0.70	0

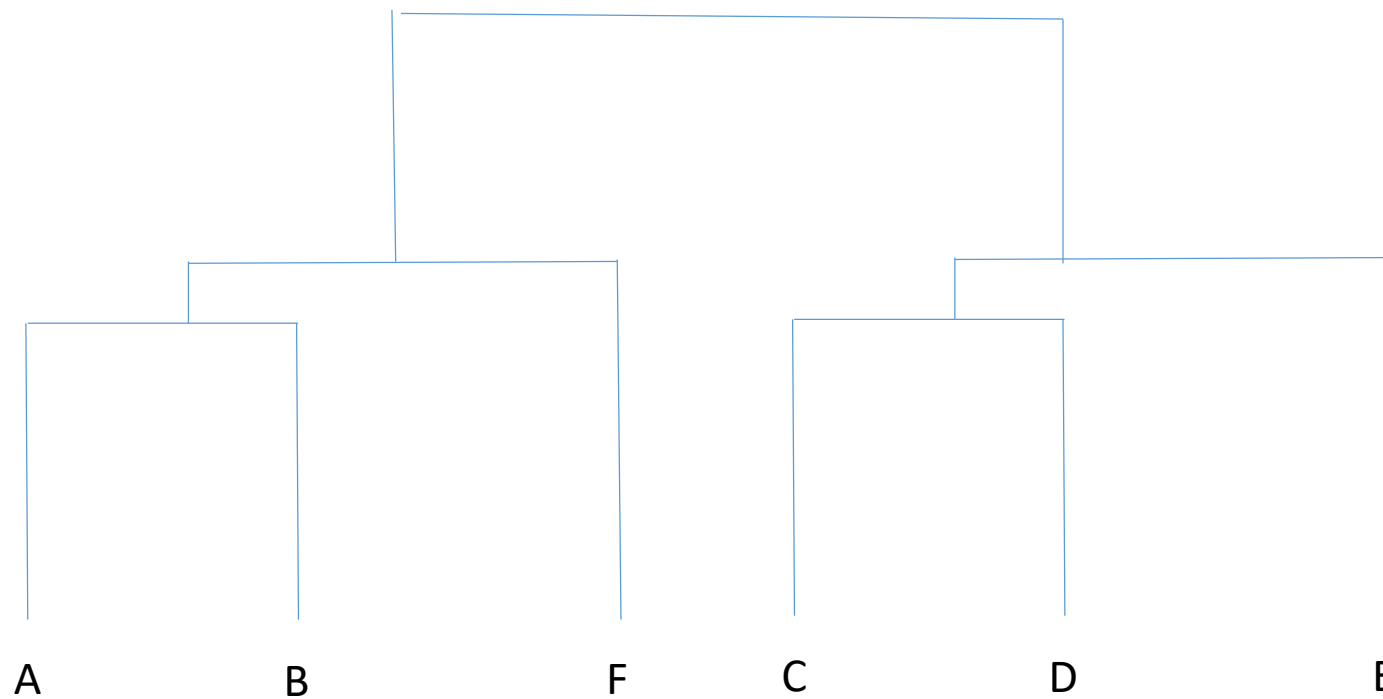


$\text{Max}(\text{dist}(\text{CD}, \text{E}), \text{ABF}) = \text{Max}(\text{dist}(\text{CD}, \text{ABF}), \text{dist}(\text{E}, \text{ABF})) = \text{Max}(0.93, 0.77) = 0.93$

	ABF	CDE
ABF	0	
CDE	0.93	0



	ABF	CDE
ABF	0	
CDE	0.93	0





برای اینکه هر دو دندوگرام شبیه یکدیگر شوند باید دو مقدار بر اساس منطق زیر تغییر کند.

❖ همانطور که در حل ماتریس به روش complete دیده شده پس از انتخاب AB (مشابه حالت single)، حالت AB, F انتخاب می شوند (چون مینیمم مقدار را داشتند و ۰.۶۱ بود)، در حالی که ما می خواهیم AB و CD انتخاب شوند بنابراین باید میزان فاصله ی بین این دو را کاهش دهیم و می توانیم مقدار A, D را به یک عدد کمتر ۰.۵۱ دهیم. بعد از این ما می خواهیم که مقدار بین ABCD و F مینیمم شود (نمیخواهیم E انتخاب شود) بنابراین می توان مقدار بین C, F را از ۰.۹۳ به یک عدد مثلا کمتر 0.63 کاهش دهیم.

	A	B	C	D	E	F
A	0					
B	0.12	0				
C	0.51	0.25	0			
D	0.51	0.16	0.14	0		
E	0.28	0.77	0.70	0.45	0	
F	0.34	0.61	0.63	0.20	0.67	0



۲.۱. با مراجعه به بخش ۲ مقاله [Ben-Dor et al.](#) معیار TNoM در انتخاب ویژگی تک متغیره را توضیح دهید.



نمرات TNoM و INFO، که در این مقاله به آن اشاره شده است، دو روش طبیعی برای تعیین کمیت (information level) رتبه ی یک بردار بر اساس همگن ترین تقسیم بندی آن است. میزان TNoM score مربوط به پارتیشنی در بردار V است که آن را به بهترین وجه به دو homogeneous prefix و homogeneous suffix تقسیم می کند. TNoM score برای رتبه بندی بردار V به صورت زیر تعریف می شود:

$$TNoM(v) = \min_{x, y=v} \min([\#_-(x) + \#_+(y)], [\#_+(x) + \#_-(y)])$$

که در رابطه ی بالا $\#_s(x)$ تعداد باری است که S در بردار v ظاهر شده است. در این روش ابتدا برای هر پارتیشین x, y از بردار v، ابتدا یک دسته بندی در نظر می گیریم به این صورت که لیبل x برای دسته های مثبت است و لیبل y برای دسته های منفی، در این حالت تعداد حالات misclassification به صورت $\#_-(x)$ و $\#_+(y)$ در نظر گرفته می شود. سپس یک دسته بندی برعکس در نظر گرفته می شود که تعداد misclassification ها به صورت $\#_+(x)$ و $\#_-(y)$ در نظر گرفته می شوند. در نهایت، پارتیشنی را که بهترین دسته بندی برای آن انجام می گیرد و کمترین misclassification را دارد برمی گردانیم. برای مثال برای رنک بردار زیر:

$$v = \langle +, +, +, -, +, +, +, -, -, -, +, -, -, +, - \rangle.$$

بهترین پارتیشن به صورت زیر خواهد بود:

$$v = \langle +, +, +, -, +, +, + \rangle; \langle -, -, -, +, -, -, +, - \rangle$$

بنابراین TNoM بردار V به صورت زیر محاسبه می شود:

$$TNoM(v) = 1 + 2 = 3.$$



توجه داشته باشید که پارتیشن ۷ برابر با انتخاب سطح بیان آستانه و از شمارش تعداد دسته بندی های غلط ناشی می شود برای همین به آن و از این رو نام Threshold Number of Misclassification یا TNOM می گویند.

TNOM Score، rule هایی را تشخیص نمی دهد که باعث خطاهای k one-sided می شود.



۲.۲. روش‌های انتخاب ویژگی چه مزیتی نسبت به روش‌های استخراج ویژگی (Feature extraction e.g., PCA) دارند.



در feature selection با به دنبال ویژگی هایی هستیم که بیشترین اطلاعات را در مورد target های مسئله ی ما می دهند. در واقع ما دنبال ویژگی هایی هستیم که بیشترین همبستگی و correlation را با ستون target دارند. برای مثال ما ۱۰ ویژگی داریم که ۹ تا از آنها بخش اعظمی (۹۰ درصد) از target را تولید می کند و ما اگر مثلاً همان ۱۰ ویژگی را در نظر بگیریم بازهم به صورت تقریبی همان (۹۰ درصد) از target را تولید می کند. در واقع ما در feature selection ما به نوعی با کاهش ابعاد نیز مواجه هستیم و با انتخاب بهترین ویژگی ها، داده های مسئله را کاهش ابعاد می دهیم.

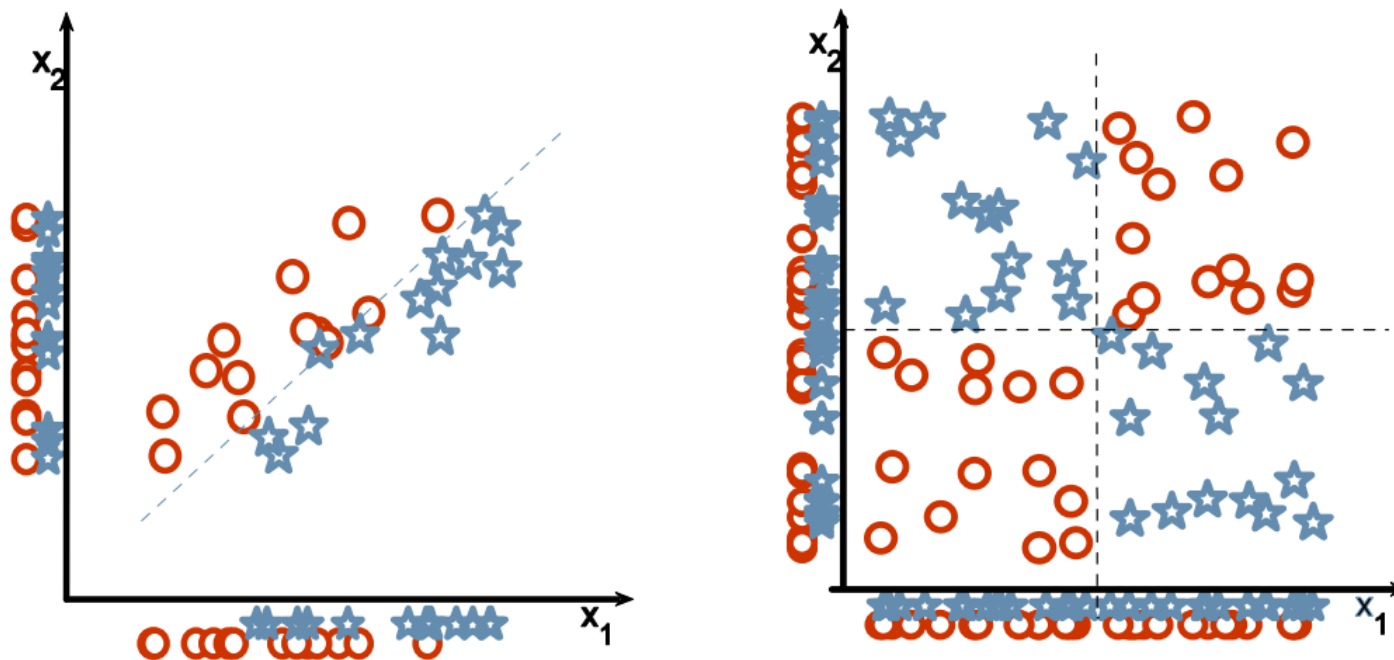
استخراج ویژگی یا feature extraction زمانی استفاده می شود که ما دید جامعی در مورد ویژگی ها نداریم. در روش های استخراج ویژگی ما اغلب وقتی کاهش ابعاد می دهیم، داده های کاهش بعد یافته اغلب قابل برگشت نیست زیرا برخی از اطلاعات در روند کاهش ابعاد از بین می روند و این از بین رفتن اطلاعات می تواند برای استنتاج و نتیجه گیری های ما مصر باشد زیرا ممکن است اطلاعاتی در روند کاهش ابعاد حذف شود بسیار مهم هستند و ما به اشتباه آنها را حذف کرده ایم. این در حالی است که در روش های feature selection ما داده هایی را حذف می کنیم که بیشترین بازنمایی را از target های ما دارد. برای استخراج ویژگی می توانید Feature Extract را بر روی داده های داده شده اعمال کنیم و سپس Feature Selection را با توجه به Target Variable اعمال کنیم تا زیرمجموعه را انتخاب کنید که می تواند به ساخت یک مدل خوب با نتایج خوب کمک کند.



۲.۳. مثالی از عدم کارایی هر یک از دو معیار همبستگی پیرسون و Mutual information در انتخاب ویژگی تک متغیره بنزید.



مشکل روش های پیرسون و mutual را به صورت تک متغیره و وابسه از هم نگاشت می کنیم هر کدام از معیار های پیرسون و mutual صفر می شود و به ما کمکی نمی کنند. در واقع وقتی ما در روش پیرسون داده ها را نگاشت می کنیم وابستگی خطی بین داده ها را از بین می بریم و این مسئله باعث می شود که ما به correlation صفر برسیم در حالی که واقعا داده ها در ابعاد اصلی تفکیک پذیر هستند (زیرا ما در روش پیرسون به دنبال روابط خطی هستیم). در روش mutual هم به همین صورت چون روش mutual بر اساس وابستگی داده ها تصمیم گیری می کند وقتی ما داده ها به یک بعد نگاشت می کنیم این وابستگی از بین می روند و طبق رابطه mutual صورت کسر صفر می شود و به این ترتیب معیار mutual نیز صفر می شود و اطلاعاتی به ما نمی دهد. همونطور که در شکل زیر می بینیم.





۳.۱. ماتریس X که نمونه‌ها در سطرهاى آن قرار گرفته اند در نظر بگیرید. نشان دهید یافتن راستاهایی که داده‌های بازسازی شده روی آنها فاصله‌ی کمی با داده‌های اصلی دارند معادل یافتن راستاهای با واریانس زیاد است. به عبارت دیگر نشان دهید رابطه

$$\underset{w.r.t. D^T D = I}{\operatorname{argmin}_D} \|X - X D D^T\|_F$$

نتیجه می‌دهد

$$(X^T X) D = \lambda D$$



❖ فرض می کنیم که ما دو ماتریس X و D را داریم که X حاوی داده های ما و D ماتریس است که داده های ما را به یک فضای با ابعاد کمتر نگاشت می کند.

❖ همچنین ما یک ماتریس Z خواهیم داشت که project داده های ما در فضای جدید است که به صورت زیر تعریف می شود:

$$Z = D^T X$$

❖ همچنین ما برای بازسازی مجدد داده های X از داده های کاهش بعد یافته خواهیم داشت:

$$X^{\sim} = DZ$$

❖ می دانیم پس از بازیابی مقادیر کاهش بعد یافته همواره با یک خطای reconstruction مواجه خواهیم بود که به صورت زیر تعریف می شود:

$$||X - X^{\sim}||$$



❖ بنابراین با توجه به فرضیات مسئله ما به دنبال مینیمم سازی حداکثری میزان خطای reconstruction هستیم.

$$\min_{D \in R^{d \times k}} \sum_{i=1}^n ||X_i - DD^T X_i||^2$$

❖ فرض می کنیم که داده های ما به صورت تجربی از توزیع یکنواخت پیروی می کنند بنابراین با توجه به فرضیات برای محاسبه ی امید ریاضی خواهیم داشت:

$$E[f(X)] = \frac{1}{n} \sum_{i=1}^n f(X_i)$$

❖ همچنین با فرض center شدن داده ها برای محاسبه ی واریانس داریم:

$$var[f(X)] + (E[f(X)])^2 = E[f(X)^2] = \frac{1}{n} \sum_{i=1}^n f(X_i)^2$$



می دانیم که خطای reconstruction به دلیل وجود اختلاف در میزان واریانس داده های اصلی و داده های اهش بعد یافته ی بازسازی شده است. بنابراین ما دنیال این هستیم که با حداکثر کردن میزان واریانس داده های بازسازی شده این خطا را میزان حداقلی برسانیم. بنابراین مینم کردن خطای بازسازی با ماکسیمم کردن میزان واریانس داد های بازسازی شده رابطه ی مسنقیم دارد.



واریانس داده های
واقعی $||X||$

اختلاف واریانس
داده های واقعی و
بازسازی شده

$$||X - D^T X|| = ||(I - D^T)X||$$

واریانس داده های
بازسازی شده $||D^T X||$

❖ در ادامه مقدار واریانس داده های باز سازی شده در یک D ضرب می شود که این مسئله تاثیری در طول بردار نخواهد داشت و صرفا سبب Rotate آن می شود، بنابراین بر اساس امید راضی و رابطه ی بین داده های اصلی و داده های بازسازی شده و اختلاف آنها خواهیم داشت:

$$E[||X^2||] = E[||D^T X||^2] + E[||X - DD^T X||^2]$$



❖ بنابراین همانطور که گفته شد ما به دنبال ماکسیم کردن رابطه ی زیر هستیم:

$$E \left[||D^T X||^2 \right]$$

❖ بنابراین برای ماکسیم کردن واریانس داده های project شده داریم:

$$\max_{||D||=1} E[(D^T X)] = \max_{||D||=1} \frac{1}{n} \sum_{i=1}^n (D^T X_i)^2 = \max_{||D||=1} \frac{1}{n} ||D^T X||^2 = \max_{||D||=1} \frac{1}{n} D^T \left(\frac{1}{n} X X^T \right) D$$

❖ برای محاسبه رابطه ی زیر بر اساس ضریب لاگرانژ و شرط $D^T D = 1$ خواهیم داشت:

$$\max_{||D||=1} \frac{1}{n} D^T \left(\frac{1}{n} X X^T \right) D$$

محاسبه ی مشتق

$$\begin{aligned} \xrightarrow{\text{blue arrow}} \max D^T X X^T - \lambda D^T D = 1 & \xrightarrow{\text{blue arrow}} \frac{\partial}{\partial D} = 0, (X X^T - \lambda I) D = 0 & \xrightarrow{\text{blue arrow}} (X X^T) D = \lambda D \end{aligned}$$



۳.۲. ماتریس X که نمونه‌ها در ستون‌های آن قرار گرفته‌اند با یک کرنل به یک فضای غیرخطی برده‌ایم و اکنون ماتریس $K = \phi^T(X)\phi(X)$ را در اختیار داریم. از PCA می‌دانیم بردار v که داده‌ها در جهت آن بیشترین واریانس را دارند در رابطه زیر صدق می‌کند

$$\phi(X)\phi^T(X)v = \lambda v$$

با استفاده از K بازتاب داده‌ها بر بردار v را بیابید.

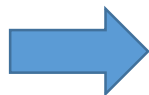


با توجه به فرضیات مسئله داریم



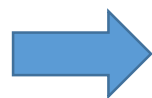
$$K = \phi^T(X)\phi(X)$$
$$\phi(X)\phi^T(X)v = \lambda v$$

فرض می کنیم



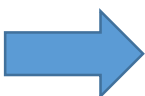
$$\sum_{i=1}^m \phi(\mathbf{x}_i) = 0$$

همچنین برای بردارهای ویژه نیز خواهیم داشت



$$\mathbf{C}\mathbf{v}_j = \lambda_j \mathbf{v}_j, j = 1, \dots, N$$

بنابراین پس از بازنویسی رابطه ی PCA با توجه به فرضیات مسئله خواهیم داشت:



$$\frac{1}{m} \sum_{i=1}^m \phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^T \mathbf{v}_j = \lambda_j \mathbf{v}_j, j = 1, \dots, N$$

1

بنابراین پس از بازنویسی رابطه بردار های ویژه با توجه به فرضیات مسئله خواهیم داشت:



$$\mathbf{v}_j = \sum_{i=1}^m a_{ji} \phi(\mathbf{x}_i)$$

2



بنابراین بر اساس رابطهی
۱ و ۲ خواهیم داشت:



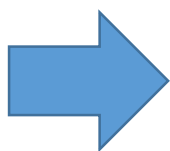
$$\frac{1}{m} \sum_{i=1}^m \phi(\mathbf{x}_i) \phi(\mathbf{x}_i)^T \left(\sum_{l=1}^m a_{jl} \phi(\mathbf{x}_l) \right) = \lambda_j \sum_{l=1}^m a_{jl} \phi(\mathbf{x}_l)$$

فرم ساده سازی شده
عبارت بالا



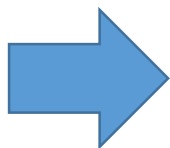
$$\frac{1}{m} \sum_{i=1}^m \phi(\mathbf{x}_i) \left(\sum_{l=1}^m a_{jl} K(\mathbf{x}_i, \mathbf{x}_l) \right) = \lambda_j \sum_{l=1}^m a_{jl} \phi(\mathbf{x}_l)$$

ضرب طرفین در $\phi(\mathbf{x}_k)^T$



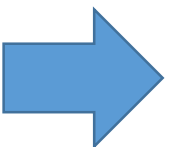
$$\frac{1}{m} \sum_{i=1}^m \phi(\mathbf{x}_k)^T \phi(\mathbf{x}_i) \left(\sum_{l=1}^m a_{jl} K(\mathbf{x}_i, \mathbf{x}_l) \right) = \lambda_j \sum_{l=1}^m a_{jl} \phi(\mathbf{x}_k)^T \phi(\mathbf{x}_l)$$

تبدیل به فرم کرنل



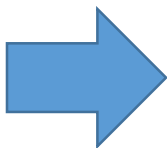
$$\frac{1}{m} \sum_{i=1}^m K(\mathbf{x}_k, \mathbf{x}_i) \left(\sum_{l=1}^m a_{jl} K(\mathbf{x}_i, \mathbf{x}_l) \right) = \lambda_j \sum_{l=1}^m a_{jl} K(\mathbf{x}_k, \mathbf{x}_l), \forall j, k$$

پس از ساده سازی
خواهیم داشت که K
همان کرنل ما است



$$\mathbf{K}^2 \mathbf{a}_j = m \lambda_j \mathbf{K} \mathbf{a}_j$$

از طرفی K را حذف می کنیم و
چون صرفاً این حذف بر روی
مقادیر ویژه صفر تاثیر می گذارد
پس ایرادی ندارد



$$\mathbf{K} \mathbf{a}_j = m \lambda_j \mathbf{a}_j$$



با توجه به نرمال کردن بردار \mathbf{a}_j $\Rightarrow \mathbf{v}_j^T \mathbf{v}_j = 1 \Rightarrow \sum_{k=1}^m \sum_{l=1}^m a_{jl} a_{jk} \phi(\mathbf{x}_l)^T \phi(\mathbf{x}_k) = 1 \Rightarrow \mathbf{a}_j^T \mathbf{K} \mathbf{a}_j = 1$

بنابراین بر اساس رابطه ی ۳ و ۴ داریم $\Rightarrow \lambda_j m \mathbf{a}_j^T \mathbf{a}_j = 1, \forall j$

بنابراین بازتاب داده ها بر بردار \mathbf{V} بر اساس کرنل K به صورت مقابل خواهد بود $\Rightarrow \phi(\mathbf{x})^T \mathbf{v}_j = \sum_{i=1}^m a_{ji} \phi(\mathbf{x})^T \phi(\mathbf{x}_i) = \sum_{i=1}^m a_{ji} K(\mathbf{x}, \mathbf{x}_i)$



۳.۳. Non-negative Matrix Factorization (NMF) یکی از روش‌های کاهش ابعاد است که همزمان قابلیت خوشه‌بندی داده‌ها را نیز دارد. NMF با داشتن ماتریس X که داده‌های آن در ستون‌ها قرار گرفته و ویژگی‌هایشان مقادیر مثبت دارند مساله زیر را حل می‌کند

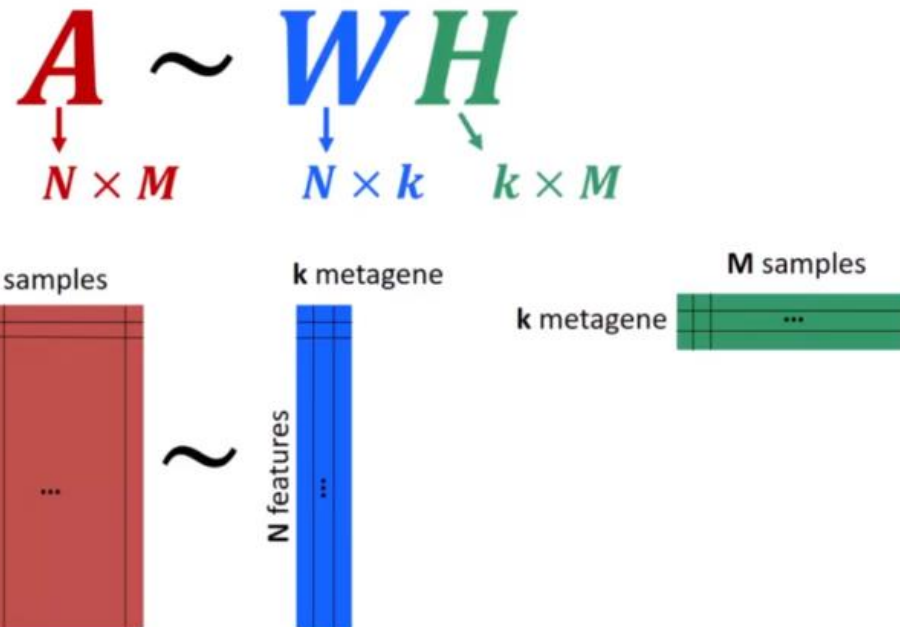
$$\operatorname{argmin}_{W \geq 0, H \geq 0} \|X - WH\|_F$$

اگر ابعاد ماتریس X $m \times n$ بوده و ابعاد ماتریس W $m \times p$ باشد که p بعد نهایی است، کدام ماتریس داده‌های تبدیل‌یافته را در خود جای می‌دهد.



❖ فرض می کنیم که یک ماتریس به نام A داریم با ابعاد، m در n که در آن m سطرهای شامل ویژگی هاست و n ستون های شامل نمونه هاست. طبق الگوریتم NMF ما به دنبال تقسیم ماتریس A به حاصل ضرب دو ماتریس با ابعاد بسیار کوچیکتر (ماتریس W با ابعاد N در K و ماتریس H با ابعاد K در M) هستیم (شکل ۱). که در این الگوریتم K بسیار کوچکتر از N است. بنابراین K یک پارامتر است که بهترین مقدار آن باید محاسبه شود. همانطور که در شکل دو مشخص می شود طبق الگوریتم ما به دو نبال پیدا کردن بهترین k (جهت خوشه بندی) از بین مقادیر ممکن هستیم و تا بر اساس آن ماتریس H را که بهترین محاسبه کنیم (k رنک ماتریس H است). بنابراین ماتریس W داده ها تبدیل را در خود دارد.

شکل ۱



شکل ۲

