به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده مهندسی برق و کامپیوتر

سیستم های هوشمند

تمرین شماره 2

نیما زمان پور

810198407

آذر ماه 1401

فهرست سوالات

تمرین 1

الف) خوشه‌بندی به روش کا-میانگین[1]

به کمک تعریف فاصله داده شده $D_{ij} = dist_{eucl}.(x_i, x_j)^2$ برای نقاط فاصله را در یک ماتریس حساب می‌کنیم.

$$D_{6*6} = \begin{bmatrix} 0 & 1 & 5 & 4 & 9 & 17 \\ 1 & 0 & 2 & 5 & 10 & 20 \\ 5 & 2 & 0 & 12 & 20 & 34 \\ 4 & 5 & 12 & 0 & 1 & 5 \\ 9 & 10 & 20 & 1 & 0 & 2 \\ 17 & 20 & 34 & 5 & 2 & 0 \end{bmatrix}$$

حالا می‌خواهیم الگوریتم K-means را پیاده سازی کنیم. ابتدا نقاط $A = (0,0), F = (4,-1)$ را به عنوان مرکز خوشه اولیه در نظر می‌گیریم. حالا هر دیتاپوینت را به مرکز خوشه نزدیک تر assign می‌کنیم.

با استفاده از ماتریس فاصله میتوان به راحتی این کار را انجام داد.

داده	فاصله از مرکز خوشه C_1	فاصله از مرکز خوشه C_2	خوشه
A	0	17	1
B	1	20	1
C	5	34	1
D	4	5	1
E	9	2	2
F	17	0	2

حالا مرکز خوشه های جدید را بدست می‌آوریم

$$C_1(x_1, x_2) = \left(\frac{0+0-1+2}{4}, \frac{0+1+2+0}{4}\right) = \left(\frac{1}{4}, \frac{3}{4}\right)$$

$$C_2(x_1, x_2) = \left(\frac{3+4}{2}, \frac{0-1}{2}\right) = (3.5, -0.5)$$

[1] K-Means

حالا برای مراکز جدید دیتاپوینت ها را assign می‌کنیم.

داده	فاصله از مرکز خوشه C_1	فاصله از مرکز خوشه C_2	خوشه
A	0.625	12.5	1
B	0.125	14.5	1
C	3.125	26.5	1
D	3.625	2.5	2
E	8.125	0.5	2
F	17.125	0.5	2

داده‌های مربوط به خوشه‌ها تغییر کرده. دوباره میانگین می‌گیریم.

$$C_1(x_1, x_2) = \left(\frac{0+0-1}{3}, \frac{0+1+2}{3}\right) = \left(-\frac{1}{3}, 1\right)$$

$$C_2(x_1, x_2) = \left(\frac{2+3+4}{3}, \frac{0+0-1}{3}\right) = \left(3, -\frac{1}{3}\right)$$

داده	فاصله از مرکز خوشه C_1	فاصله از مرکز خوشه C_2	خوشه
A	1.111	9.111	1
B	0.111	10.777	1
C	1.444	21.444	1
D	6.444	1.111	2
E	12.111	0.111	2
F	22.777	1.444	2

داده‌های متعلق به خوشه ها تغییری نکرد. پس مرکز خوشه ها همین است. و الگوریتم به پایان رسید.

$$C_1(x_1, x_2) = \left(-\frac{1}{3}, 1\right)$$

$$C_2(x_1, x_2) = \left(3, -\frac{1}{3}\right)$$

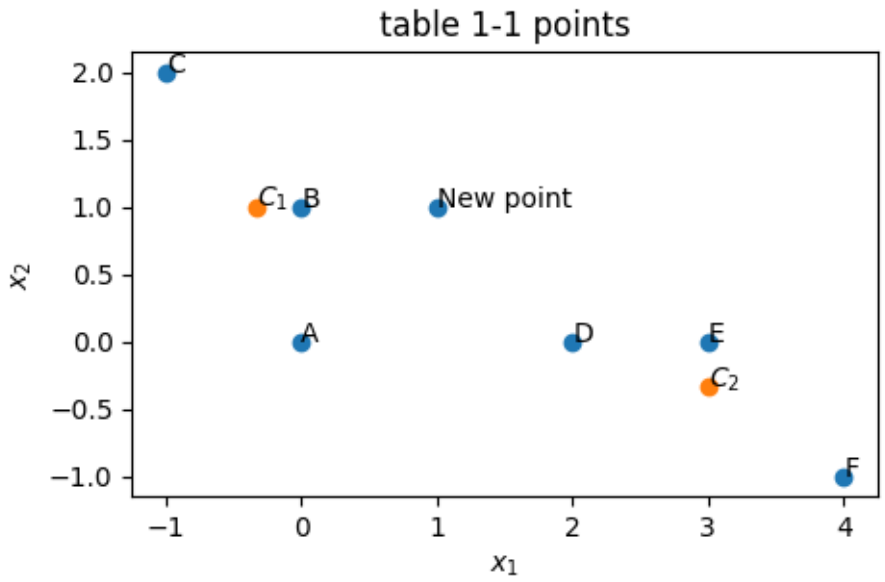

Figure 1 نقاط به همراه مرکز خوشه‌ها

حالا برای نقطه $x^* = (1,1)$ فاصله از مرکز خوشه ها را حساب می‌کنیم.

خوشه	فاصله از مرکز خوشه C_2	فاصله از مرکز خوشه C_1	داده
1	5.777	1.777	(1,1)

داده متعلق به خوشه اول است.

ب) خوشه‌بندی سلسله‌مراتبی

ب.1) پیوند واحد

ابتدا هر نقطه را یک خوشه در نظر می‌گیریم.

مرحله بعد نزدیک ترین 2 نقطه را پیدا می‌کنیم. و آن هارا یک خوشه می‌گیریم.

	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	0.7	0.45	0	
F	0.34	0.61	0.93	0.2	0.67	0

(A,B) یک خوشه شدند. آن هارا در جدول گروه کرده و مقدایر فاصله خوشه را با مقدار Min فاصله از بقیه خوشه ها جایگزین می‌کنیم.

	A,B	C	D	E	F
A,B	0				
C	0.25	0			
D	0.16	0.14	0		
E	0.28	0.7	0.45	0	
F	0.34	0.93	0.2	0.67	0

دوباره روش را اجرا می‌کنیم. کمترین مقدار جدول بین C,D است. (C,D) یک خوشه شدند. آن ها را در یک جدول گروه کرده و مقدایر فاصله خوشه را با مقدار Min فاصله از بقیه خوشه ها جایگزین می‌کنیم.

	A,B	C,D	E	F
A,B	0			

C,D	0.16	0		
E	0.28	0.45	0	
F	0.34	0.2	0.67	0

دوباره روش را اجرا می‌کنیم. کمترین مقدار جدول بین (A,B),(C,D) است. ((C,D),(A,B)) یک خوشه شدند. آن ها را در یک جدول گروه کرده و مقدایر فاصله خوشه را با مقدار Min فاصله از بقیه خوشه ها جایگزین می‌کنیم.

	(A,B),(C,D)	**E**	**F**
(A,B),(C,D)	0		
E	0.28	0	
F	0.2	0.67	0

دوباره روش را اجرا می‌کنیم. کمترین مقدار جدول بین F,((C,D),(A,B)) است. (((C,D),(A,B)),F) یک خوشه شدند. آن ها را در یک جدول گروه کرده و مقدایر فاصله خوشه را با مقدار Min فاصله از بقیه خوشه ها جایگزین می‌کنیم.

	(((C,D),(A,B)),F)	**E**
(((C,D),(A,B)),F)	0	
E	0.28	0

حالا آخرین خوشه ((((C,D),(A,B)),F),E) است.

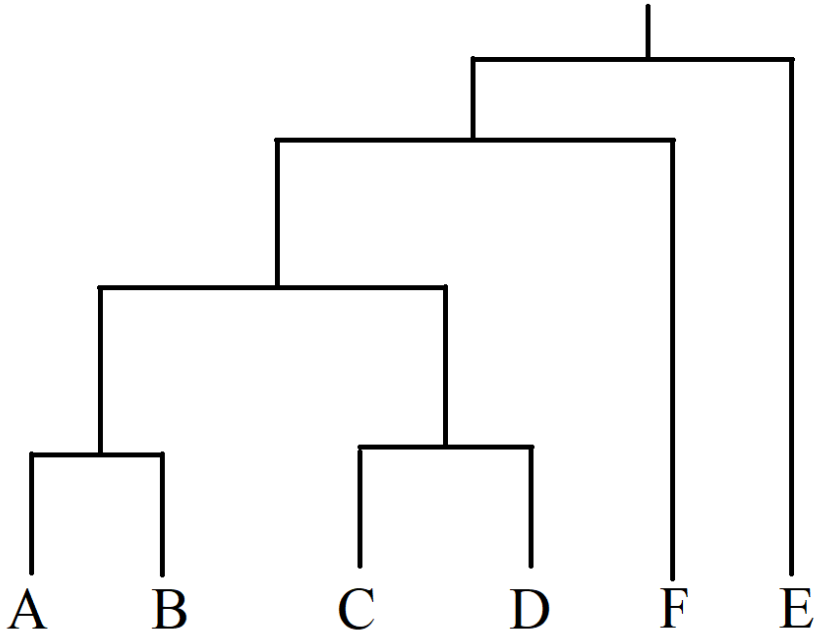

Figure 2 نمودار درخت به روش پیوند واحد

ب.2) پیوند کامل

ابتدا هر نقطه را یک خوشه در نظر می‌گیریم.

مرحله بعد نزدیک ترین 2 نقطه را پیدا می‌کنیم. و آن ها را یک خوشه می‌گیریم.

	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	0.7	0.45	0	
F	0.34	0.61	0.93	0.2	0.67	0

(A,B) یک خوشه شدند. آن هارا در جدول گروه کرده و مقدایر فاصله خوشه را با مقدار Max فاصله از بقیه خوشه ها جایگزین می‌کنیم.

	A,B	C	D	E	F
A,B	0				
C	0.51	0			
D	0.84	0.14	0		
E	0.77	0.7	0.45	0	
F	0.61	0.93	0.2	0.67	0

دوباره روش را اجرا می‌کنیم. کمترین مقدار جدول بین C,D است. (C,D) یک خوشه شدند. آن ها را در یک جدول گروه کرده و مقدایر فاصله خوشه را با مقدار Max فاصله از بقیه خوشه ها جایگزین می‌کنیم.

	A,B	C,D	E	F
A,B	0			
C,D	0.84	0		
E	0.77	0.7	0	
F	0.61	0.93	0.67	0

دوباره روش را اجرا می‌کنیم. کمترین مقدار جدول بین (A,B),F است. ((A,B),F) یک خوشه شدند. آن ها را در یک جدول گروه کرده و مقدایر فاصله خوشه را با مقدار Max فاصله از بقیه خوشه ها جایگزین می‌کنیم.

	((A,B),F)	C,D	E
((A,B),F)	0		
C,D	0.93	0	
E	0.77	0.7	0

دوباره روش را اجرا می‌کنیم. کمترین مقدار جدول بین (C,D),E است. ((C,D),E) یک خوشه شدند. آن ها را در یک جدول گروه کرده و مقدایر فاصله خوشه را با مقدار Max فاصله از بقیه خوشه ها جایگزین می‌کنیم.

	((A,B),F)	**((C,D),E)**
((A,B),F)	0	
((C,D),E)	0.93	0

در آخر نیز دو خوشه را با هم ترکیب می‌کنیم. (((A,B),F),((C,D),E))

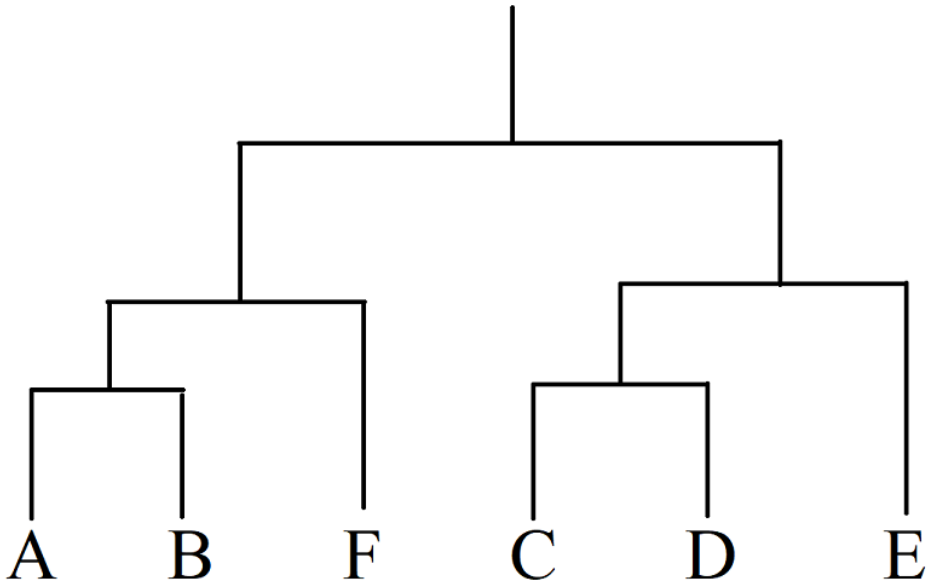

Figure 3 نمودار درخت به روش پیوند کامل

ب.3) مقایسه

باید جوری مقادیر را تغییر دهیم. که جدول سوم قسمت ب.2 منجر به انتخاب ((A,B),(C,D)) شود. پس مقدار Min آن جدول را تغییر می‌دهیم.

	A,B	C,D	E	F
A,B	0			
C,D	0.6	0		
E	0.77	0.7	0	
F	0.61	0.93	0.67	0

حالا دو خوشه (A,B),(C,D) انتخاب شدند. جدول بعدی را می‌کشیم

	(A,B),(C,D)	E	F
(A,B),(C,D)	0		
E	0.77	0	
F	0.93	0.67	0

حالا اینجا خوشه E به خوشه ((A,B),(C,D)) باید متصل شود. که مشابه بخش ب.1 نیست. حالا باید مقدار جدول را تغییر دهیم تا 0.67 دیگر مقدار Min نباشد.

	(A,B),(C,D)	E	F
(A,B),(C,D)	0		
E	0.77	0	
F	0.66	0.67	0

در این صورت اول F مرج می‌شود. و آخر نیز E مرج می‌شود. و همانند قسمت ب.1 می‌شود.

حالا برای مطئن شدن قسمت ب.1 را دوباره حل می‌کنیم.

	A	B	C	D	E	F
A	0					
B	0.12	0				
C	0.51	0.25	0			
D	0.6	0.16	0.14	0		

E	0.28	0.77	0.7	0.45	0	
F	0.34	0.61	0.66	0.2	0.67	0

	AB	**C**	**D**	**E**	**F**
AB	0				
C	0.25	0			
D	0.16	0.14	0		
E	0.28	0.7	0.45	0	
F	0.34	0.6	0.2	0.67	0

حالا C,D را با هم ادغام می‌کنیم.

	AB	**C,D**	**E**	**F**
AB	0			
C,D	0.16	0		
E	0.28	0.45	0	
F	0.34	0.2	0.67	0

میبینیم که جدول همانند جدول سوم بخش ب.1 شده. پس در ادامه حل همان نمودار درختی به دست می‌آید.

حالا جدول جدید را برای روش ب.2 حل می‌کنیم.(خلاصه شده)

	A	**B**	**C**	**D**	**E**	**F**
A	0					
B	0.12	0				

C	0.51	0.25	0			
D	0.6	0.16	0.14	0		
E	0.28	0.77	0.7	0.45	0	
F	0.34	0.61	0.66	0.2	0.67	0

	A,B	**C**	**D**	**E**	**F**
A,B	0				
C	0.51	0			
D	0.6	0.14	0		
E	0.77	0.7	0.45	0	
F	0.61	0.66	0.2	0.67	0

	A,B	**C,D**	**E**	**F**
A,B	0			
C,D	0.6	0		
E	0.77	0.7	0	
F	0.61	0.66	0.67	0

	((A,B),(C,D))	**E**	**F**
((A,B),(C,D))	0		
E	0.77	0	
F	0.66	0.67	0

	(((A,B),(C,D)),F)	E
(((A,B),(C,D)),F)	0	
E	0.77	0

خوشه اصلی: ((((A,B),(C,D)),F),E)

سوال 2

قسمت اول: خوشه بندی کا-میانگین ساده

الف.1) تاثیر تکرار خوشه‌ها

با پیاده سازی الگوریتم برای تعداد خوشه های 5، 10، 20 نمودار های زیر برای تابع هزینه بدست آمد.

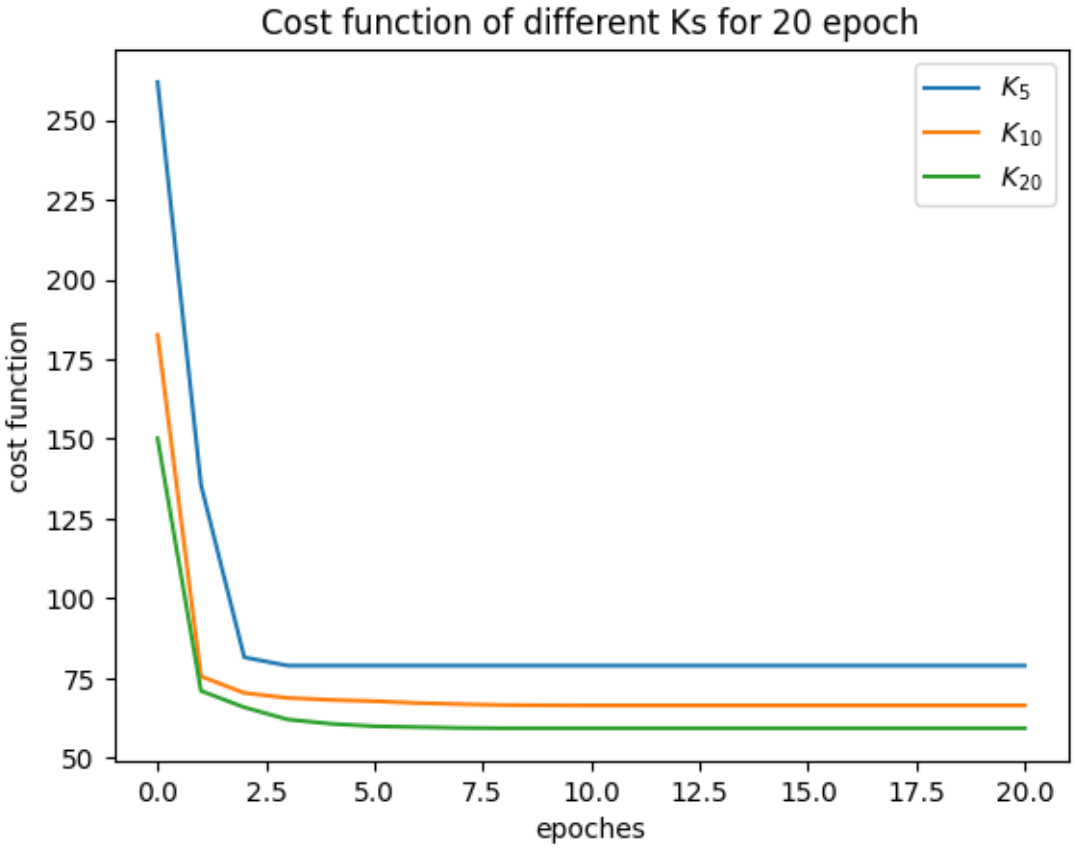

Figure 4 نمودار تابع هزینه مدل به ازای **K** مختلف

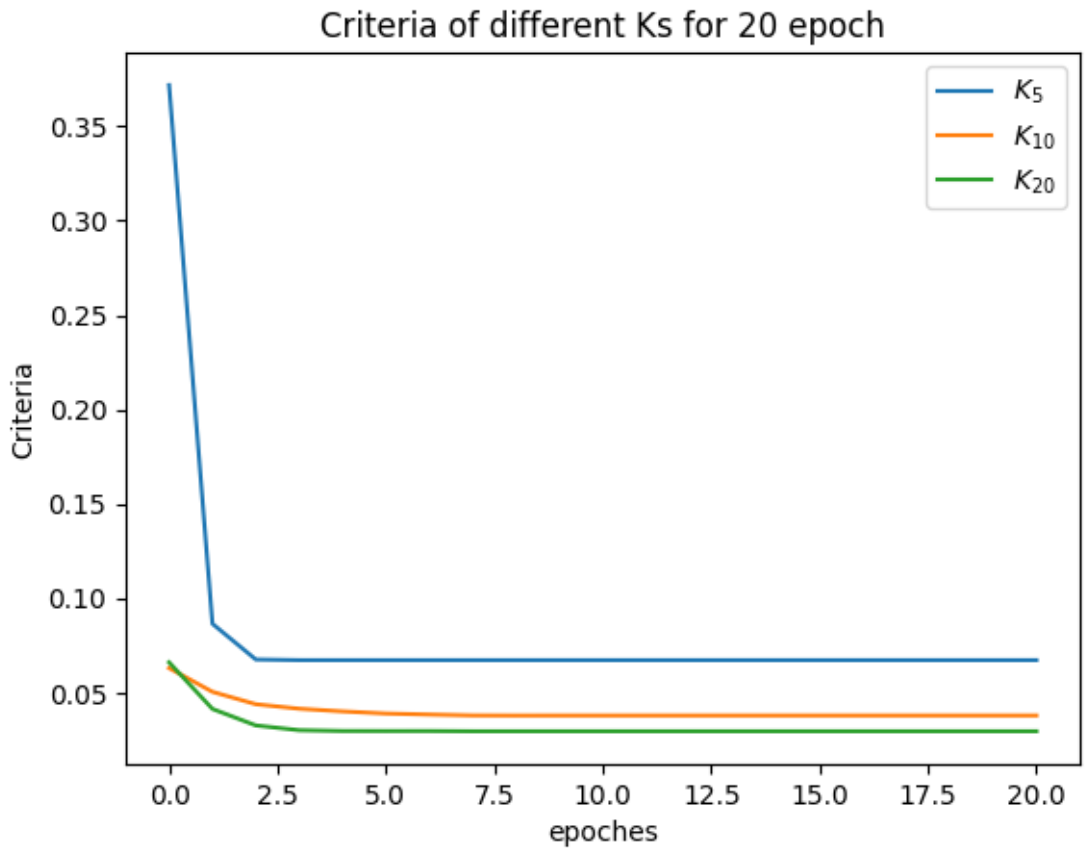

Figure 5 نمودار معیار نسبت شباهت درونی به شباهت بیرونی به ازای **K** مختلف

با توجه به معیار های داده شده و تابع هزینه مدل، حالت K = 20 توانست تابع هزینه را کاهش زیادی بدهد. که این به علت تعداد زیاد خوشه‌های آن است. که فاصله از هر مرکز کم شده است. اما به همین علت نیز در حضور یک دیتاست آزمون؛ می‌بینیم که مدل overfit می‌شود.

الف.2)

با تعریف Ratio به فرم گفته شده؛ نمودار Ratio بر حسب تعداد تکرار در تعداد خوشه های مختلف بصورت زیر درمی‌آید

این معیار در واقع نشان می‌دهد برای هر خوشه بندی که روی داده‌ها لحاظ می‌شود. چقدر داده های یک خوشه نزدیک به هم و چقدر از داده های بقیه خوشه ها دور هستند. فاصله هر چقدر بیشتر شود. این مقدار به یک نزدیک تر است. و مدل بهتر است. مدل K = 20 با اختلاف کمی از بقیه بهتر است.

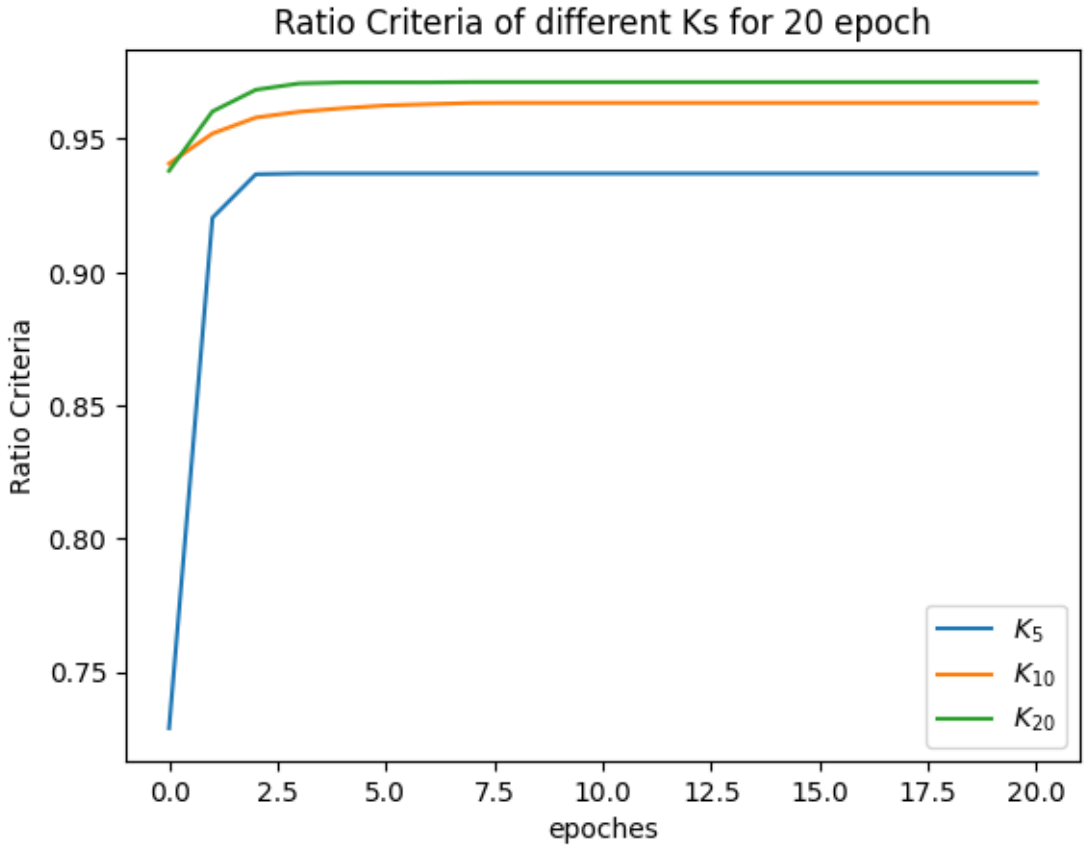

Figure 6 نمودار **Ratio** برای **K** های مختلف

Table 1 جدول میانگین و واریانس تابع هزینه و **Ratio**

K	5	10	20
Mean(cost)	90.44	72.91	64.73
Variance (cost)	1613.64	605.74	372.45
Mean(Ratio)	0.926	0.961	0.969
Variance (Ratio)	196e-5	3e-5	5e-5

در این حالت نیز K = 20 از همه بهتر عمل کرده است. و میانگین و واریانس کمتری دارد.

قسمت دوم: خوشه‌بندی کا-میانگین هوشمند

طراحی الگوریتم

الگوریتم مورد نظر از همه نظر همانند الگوریتم معمولی کا-میانگین است. اما در قسمت assign داده ها به مرکز خوشه ها تفاوت دارد. برای اساین ابتدا در یک حلقه همه داده هارا به نزدیک مرکز خوشه اساین می‌کنیم. به شرطی که در لیست قیود در جایگاه Index2 نباشد.(فرقی ندارد میتوان Index1 گذاشت و همه 1و2 ها با هم جابجا شوند)

اما اگر در لیست قیود بود بررسی می‌کنیم. که در چند قید است. اگر در یک قید بود. صرفا بررسی می‌کنیم که اگر قید 1 بود آنگاه به مرکز خوشه جفت خود اساین می‌شود. اما اگر قید 1- بود. به نزدیکترین مرکز خوشه‌ای که مرکز نزدیکترین خوشه جفت خودش نباشد اساین می‌شود.

حال اگر در دو قید بود. اگر قید اولی 1 و دومی 1- بود. جفت اول به مرکز نزدیکترین خوشه به دیتاپوینت اساین می‌شود و خود دیتاپوینت به نزدیکترین مرکز خوشه‌ای که مرکز نزدیکترین خوشه جفت خودش نباشد اساین می‌شود.

اگر هر دو قید 1- بود نیز به نزدیکترین مرکز خوشه‌ای که مرکز نزدیکترین خوشه جفت اول و دوم خودش نباشد اساین می‌شود.

به این ترتیب قید ها اعمال می‌شوند. و بقیه مراحل(اندازه گیری فاصله ها از مرکز و حساب کردن مرکز خوشه‌های جدید مثل روش قبلی است.)

ج) ارزیابی الگوریتم

.1

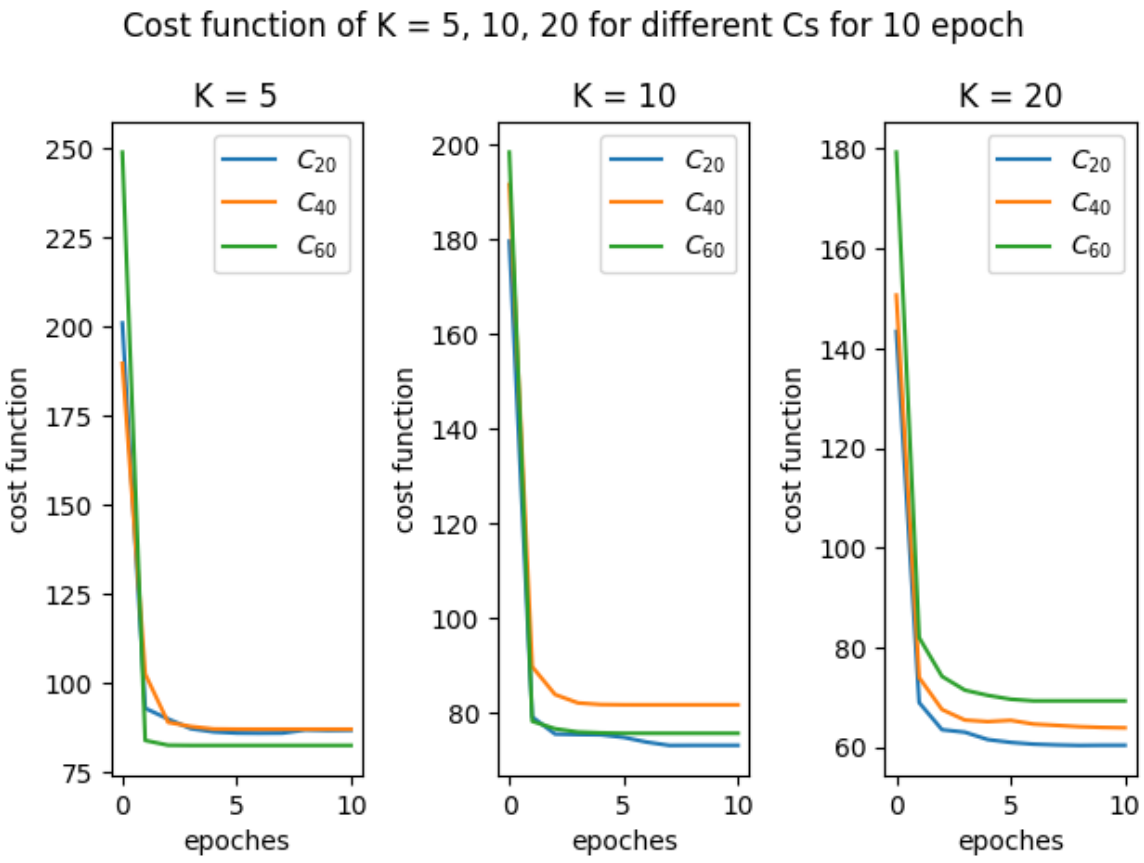

Figure 7 نمودارتابع هزینه به ازای تعداد قید متفاوت

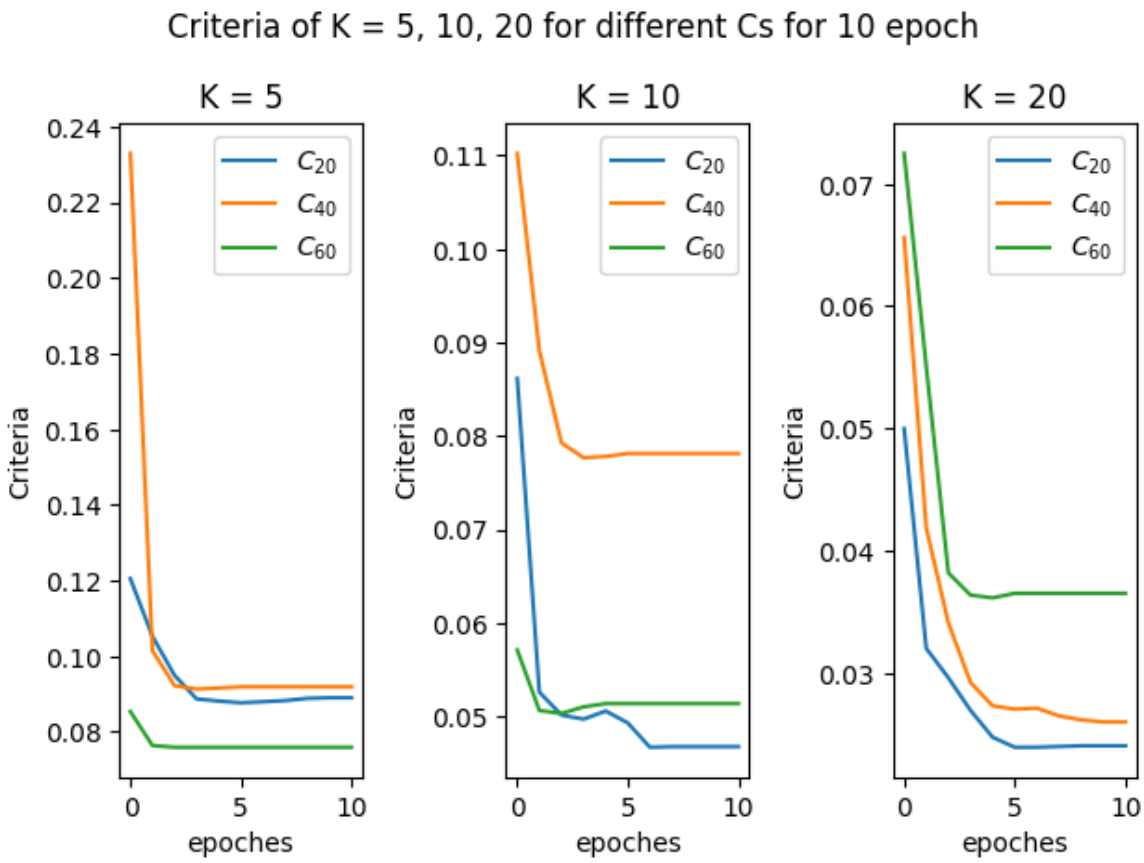

Figure 8 نمودار معیار نسبت شباهت درونی به شباهت بیرونی به ازای تعداد قید متفاوت

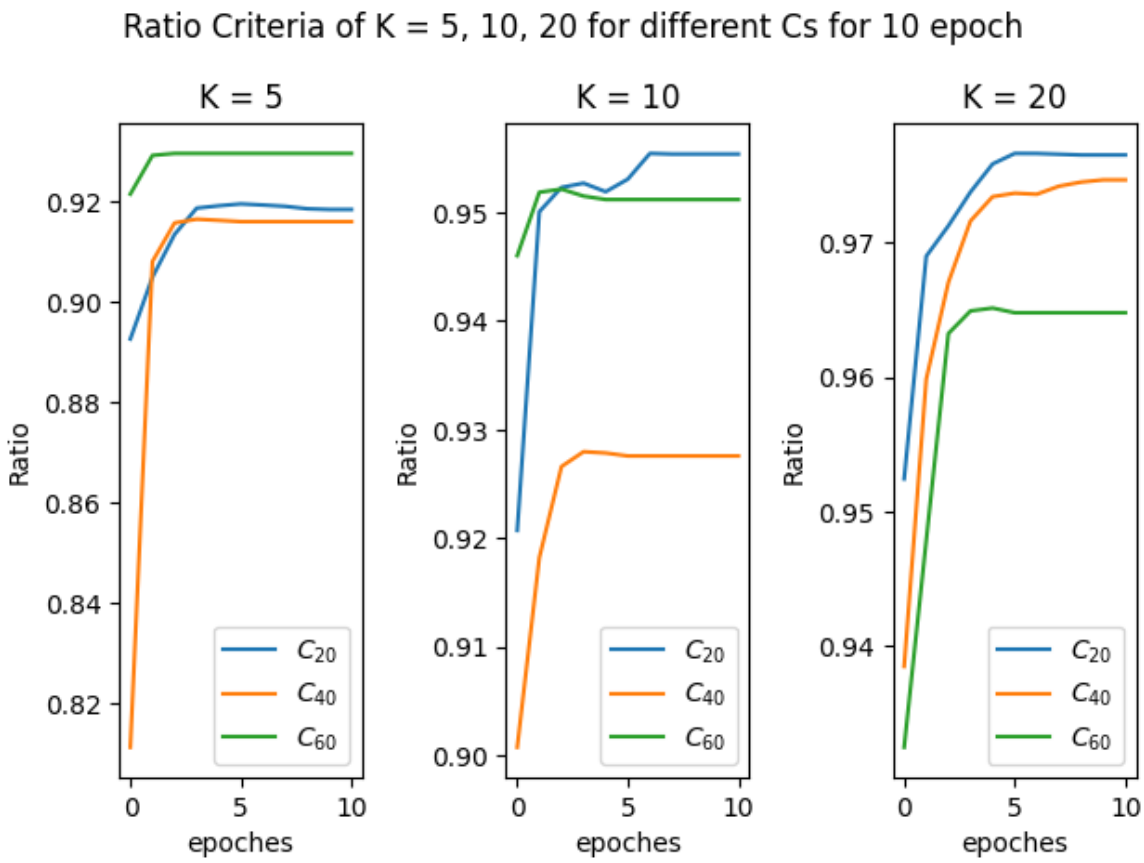

Figure 9 نمودار معیار **Ratio** به ازای تعداد قید مختلف

هماطور که می‌بینیم افزایش قید ها باعث افزایش تابع هزینه می‌شود. و همه مقادیر از تابع هزینه در حالت کا-میانگین ساده بیشتر است. این به این علت است که قیود برای حالت $K = 3$ نوشته شده‌اند. ولی در اینجا K بیشتر است. این مورد باعث می‌شود که در خوشه ها بی‌نظمی به وجود آید و داده ها پراکنده شوند. مثلا در $K = 5$ تعداد 20 قید توانست کمی داده ها را متمرکز تر کند ولی با افزایش تعداد خوشه به 20 عدد، اطلاعات موجود برای 3 خوشه کمکی به خوشه بندی نکرد و تعداد کمی خوشه از بقیه بزرگتر شد

2. برای حالت 3 خوشه برای قسمت کا-میانگین دقت برابر 88.67٪ و کا-میانگین هوشمند 87.33٪ است. اطلاعات زمینه ای باعث افزایش چشمگیر دقت نشده است. و دقت در همان حد 88٪ مانده است. که ممکن است بدلیل حالات خاص موجود در ترتیب اجرای قیود باشد. یعنی داده هایی در قیود باشند که مدل نیز خودش درست تشخیص می‌دهد.

پیوست

توضیح کد سوال یک:

برای سوال یک ابتدا نقاط را در یک آرایه ریخته و آن‌ها را رسم می‌کنیم. سپس در یک ماتریس 6*6 فاصله هر نقطه از این آرایه را به کمک دستور np.linalg.norm() حساب می‌کنیم.

توضیح کد سوال دو:

K-Means

ابتدا کتابخانه ها را import کرده و دیتاست را لود می‌کنیم. سپس توابع مورد نظر را جداگانه ساخته و در مدل نهایی ترکیب می‌کنیم. ابتدا در تابع get_random_initials() به تعداد خوشه ها مرکز خوشه رندوم تولید می‌کنیم. سپس در تابع get_distances() فاصله همه نقاط را از همه مرکز خوشه ها را به کمک تابع np.linalg.norm() حساب می‌کنیم. سپس در تابع get_cluster_assign() برای فاصله هایی که در تابع get_distances() حساب شد آرگومان کمترین فاصله که همان اندیس خوشه است را برمی‌گردانیم. در تابع make_new_K() نیز با توجه به اندیس های داده شده به نقاط، مرکز خوشه های جدید را حساب می‌کنیم. توابع cost_function() تابه هزینه تعریف شده در سوال را حساب می‌کند. و criteria() نیز معیار اندازه بیرونی و درونی تعریف شده در درس را حساب میکند.

حالا با همه این توابع تابع K_means() را تعریف می‌کنیم. که ابتدا یک K مرکز رندوم انتخاب نموده و در یک حلقه، فاصله نقاط از مرکز خوشه ها را حساب می‌کنیم. سپس نزدیک ترین مرکز خوشه به هر نقطه را در لیست cluster_assign می‌ریزد. و بر اساس آن تابع هزینه را حساب کرده و مراکز خوشه جدید را می‌سازد. این حرکت را به اندازه آرگومان داده‌شده epoch انجام می‌دهد. و تابع هزینه، لیست مراکز اساین شده و معیار اندازه دورنی و بیرونی را برمی‌گرداند.

حال پس از تعریف مدل، آن را مطابق خواسته های سوال اجرا کرده و نمودار های خواسته شده را رسم می‌کنیم.

Smart K-Means

مدل کا-میانگین هوشمند همانند مدل کا-میانگین است. با این تفاوت که تابع get_cluster_assign_smart() در آن باید قیود را نیز اعمال کند. روش اعمال قیود در بخش 2.الف توضیح داده شد. با این حساب مدل smart k-means نیز ساخته شد. حالا فایل قیود را به کمک pd.read_csv()

خوانده و آن را به دیتافریم[1] تبدیل می‌کنیم. و برای حالات خواسته شده مدل را اجرا می‌کنیم. در نهایت نیز برای حالت K = 3 نیز اجرا می‌کنیم. برای محاسبه دقت مدل از ستون target دیتاست لیبل هر دیتاپوینت را استخراج کرده و خروجی مدل را نیز دریافت می‌کنیم. حال از آنجایی که لیبل ها به ترتیب هستند.(یعنی 50 گل اول نوع 0 و 50 گل بعدی نوع 1 و 50 گل آخر نوع 2) فرض کنیم که مدل حدودا درست جواب می‌دهد. حال برای نظیر کردن شماره لیبل target به خروجی مدل، مد 20 داده اول را پیدا کرده و آن را به نوع گل 0 نظیر نظیر می‌کنیم. سپس مد 20 داده وسط را پیدا کرده و به گل نوع 1 نظیر می‌کنیم. و مد 20 داده آخر را پیدا کرده و به گل نوع 2 نظیر می‌کنیم. در این حالت شماره لیبل لیست target و خروجی مدل برابر می‌شود و به راحتی می‌توان تعداد پیش‌بینی های درست را اندازه گرفت

[1] DataFrame