دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

گزارش تمرین شماره 2

درس سیستم های هوشمند

پاییز 1401

امیرحسین بیرژندی

...

810198367

...

سوال 1 – تحلیلی

الف) خوشه بندی به روش کا-میانگین

i	x_1	x_2
A	**0**	**0**
B	**0**	**1**
C	**-1**	**2**
D	**2**	**0**
E	**3**	**0**
F	**4**	**-1**

در اینجا با استفاده فاصله اقلیدوسی مسئله را حل می‌کنیم. ابتدا با توجه به اینکه قرار است داده ها را به دو خوشه تقسیم کنیم دو نقطه را به صورت رندوم به عنوان نقاط اولیه دو گروه انتخاب می‌کنیم. در اینجا دو نقطه رندوم دو نقطه B و F هستند.

Iteration 1

-		1(0,0)	2(4,-1)	
-	Point	Distance 1	Distance 2	Cluster
A	(0,0)	0	4.123	1
B	(0,1)	1	4.472	1
C	(-1,2)	2.236	5.831	1
D	(2,0)	2	2.236	1
E	(3,0)	3	1.414	2
F	(4,-1)	4.12	0	2

Iteration 2

-		1(0.25,0.75)	2(3.5,0.5)	
-	Point	Distance 1	Distance 2	Cluster
A	(0,0)	0.791	3.536	1
B	(0,1)	0.356	3.536	1
C	(-1,2)	1.768	4.7434	1
D	(2,0)	1.904	1.581	2
E	(3,0)	2.850	0.707	2
F	(4,-1)	4.138	1.581	2

Iteration 3

-		1(-0.333,1)	2(3,0.333)	
-	Point	Distance 1	Distance 2	Cluster
A	(0,0)	1.054	3.018	1
B	(0,1)	0.333	3.073	1
C	(-1,2)	1.202	4.333	1
D	(2,0)	2.538	1.053	2
E	(3,0)	3.480	0.333	2
F	(4, -1)	4.772	1.666	2

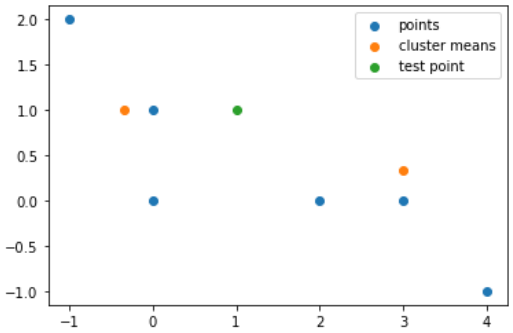
از آنجایی که در دو ایتریشن متوالی cluster های نقاط یکسان باقی ماندند در نتیجه خوشه ها کامل شده‌اند و خوشه بندی تمام شده است.

حال نقطه (1و1) را بررسی می‌کنیم و فاصله آن را از دو مرکز خوشه‌ها محاسبه می‌کنیم.

$$D_{*1} = dist_{eucl}(x^*, x_1)^2 = 1.333$$

$$D_{*2} = dist_{eucl}(x^*, x_2)^2 = 2.108$$

در نتیجه این نقطه را جزء خوشه 1 حساب می‌کنیم.



ب) خوشه بندی سلسله مراتبی

ب.1) پیوند واحد

	A	B	C	D	E	F
A	0	0.12	0.51	0.84	0.28	0.34
B	0.12	0	0.25	0.16	0.77	0.61
C	0.51	0.25	0	0.14	0.7	0.93
D	0.84	0.16	0.14	0	0.45	0.2
E	0.28	0.77	0.7	0.45	0	0.67
F	0.34	0.61	0.93	0.2	0.67	0

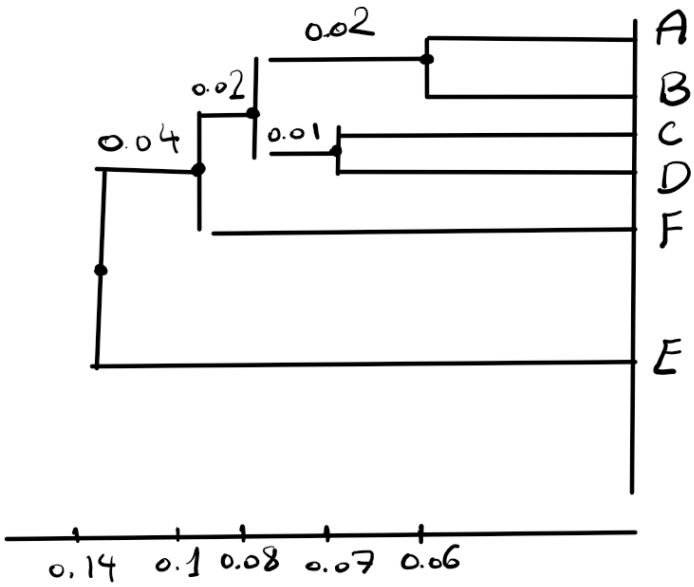
	(A, B)	C	D	E	F
(A, B)	0	0.25	0.16	0.28	0.34
C	0.25	0	0.14	0.7	0.93
D	0.16	0.14	0	0.45	0.2
E	0.28	0.7	0.45	0	0.67
F	0.34	0.93	0.2	0.67	0

	(A, B)	(C, D)	E	F
(A, B)	0	0.16	0.28	0.34
(C, D)	0.16	0	0.45	0.2
E	0.28	0.45	0	0.67
F	0.34	0.2	0.67	0

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

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

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



ب.2) پیوند کامل

	A	B	C	D	E	F
A	0	0.12	0.51	0.84	0.28	0.34
B	0.12	0	0.25	0.16	0.77	0.61
C	0.51	0.25	0	0.14	0.7	0.93
D	0.84	0.16	0.14	0	0.45	0.2
E	0.28	0.77	0.7	0.45	0	0.67
F	0.34	0.61	0.93	0.2	0.67	0

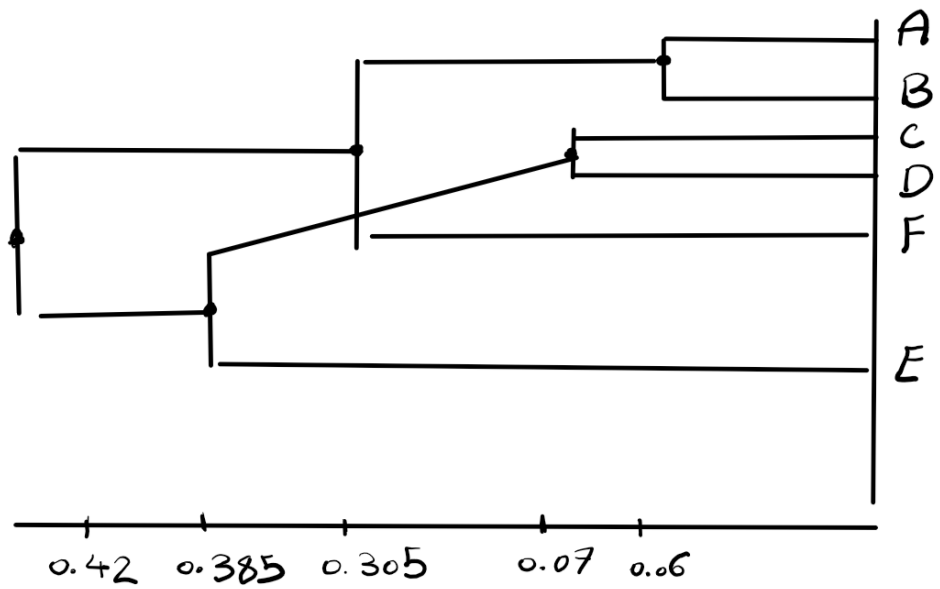
	(A, B)	C	D	E	F
(A, B)	0	0.51	0.84	0.77	0.61
C	0.51	0	0.14	0.7	0.93
D	0.84	0.14	0	0.45	0.2
E	0.77	0.7	0.45	0	0.67
F	0.61	0.93	0.2	0.67	0

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

	(A, B, F)	(C, D)	E
(A, B, F)	0	0.84	0.77
(C, D)	0.84	0	0.7
E	0.77	0.7	0

	(A, B , F)	(C, D, E)
(A, B , F)	0	0.84
(C, D, E)	0.84	0

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



ب.3) مقایسه

برای اینکه دو الگوریتم به یک نتیجه مشابه برسند کافی است

1- عدد 0.84 که فاصله A و D است را به مقداری کوچکتر از 0.61 برسانیم و در عین حال از 0.14 بزرگتر تبدیل کنیم.

2- عدد 0.93 که فاصله F و C است را به مقداری بین 0.61 و 0.67 برسانیم.

برای رسیدن به این پاسخ کافی است محل های تفاوت در جدول های دو راه حل را بررسی کنیم. اگر دقت کنیم در جدول شماره 3 هر دو روش محل کمترین فاصله متفاوت است اگر مقادیر را به گونه‌ای تغییر دهیم که محل هایلایت ها در دو جدول یکی شود دو الگوریتم یک پاسخ خواهند داشت. در نتیجه با انتخاب مقداری بین 0.61 و 0.14 این اتفاق خواهد افتاد. دلیل بزرگتر بودن از 0.14 این است که در مراحل اول و دوم تفاوتی ایجاد نشود.

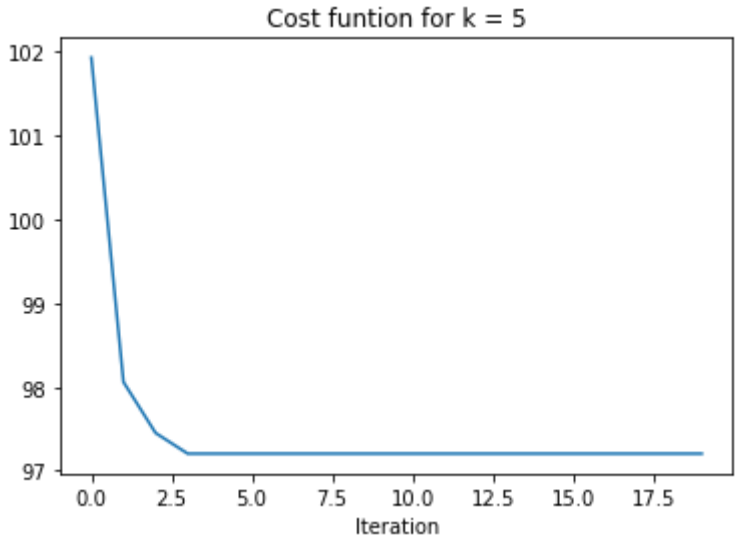
برای تغییر دوم نیز از این منطق استفاده شد که پس از تغییر 0.84 صدم به مقدار گفته شده جدول شماره 4 به شکل زیر در می آید.

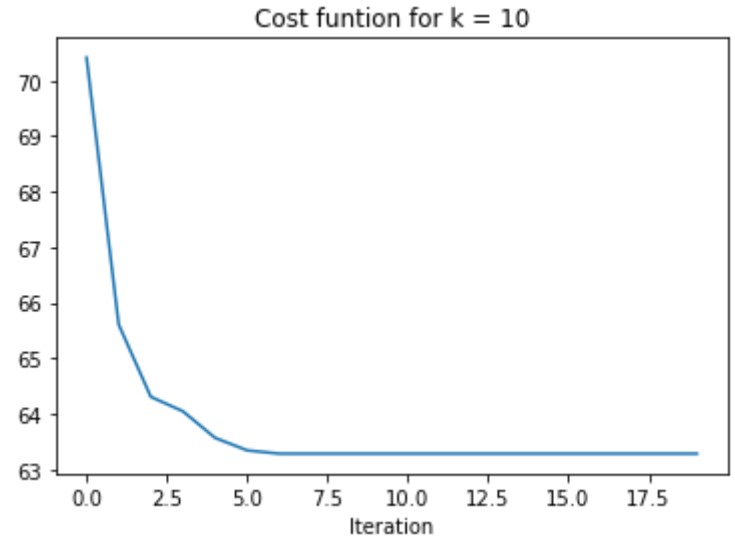
	(A, B, C, D)	**E**	**F**
(A, B, C, D)	**0**	**0.77**	**0.93**
E	**0.77**	**0**	**0.67**
F	**0.93**	**0.67**	**0**

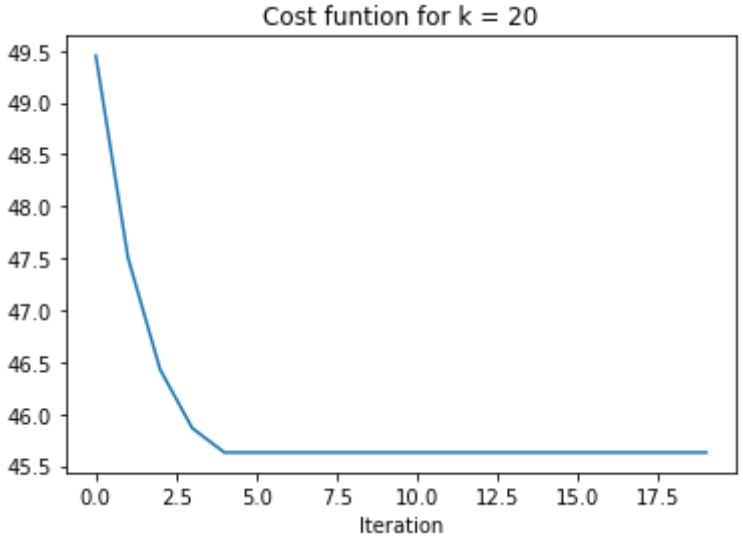
حال برای اینکه نتایج مشابه داشته باشیم باید که محل هایلایت همانند محل هایلایت در جدول 4 روش پیوند واحد باشد در نتیجه 0.93 را باید به عددی بین 0.61 و 0.67 تبدیل کنیم تا

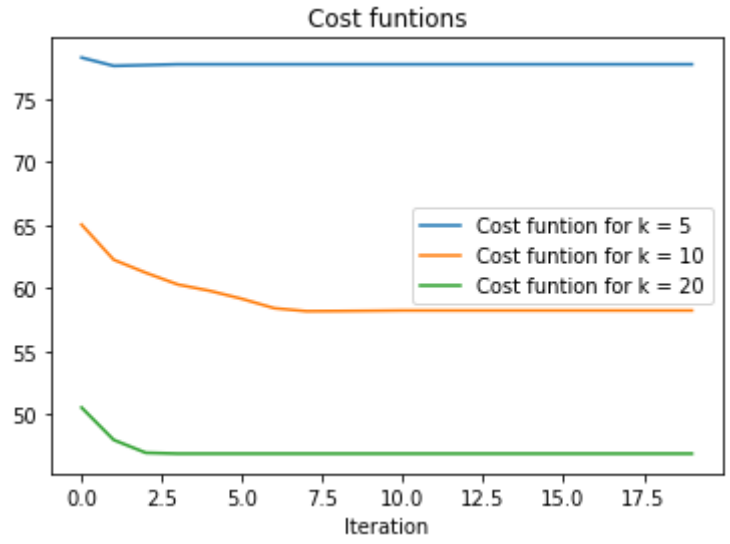
سوال 2: پیاده‌سازی الگوریتم خوشه بندی

قسمت اول: خوشه بندی کا-میانگین ساده



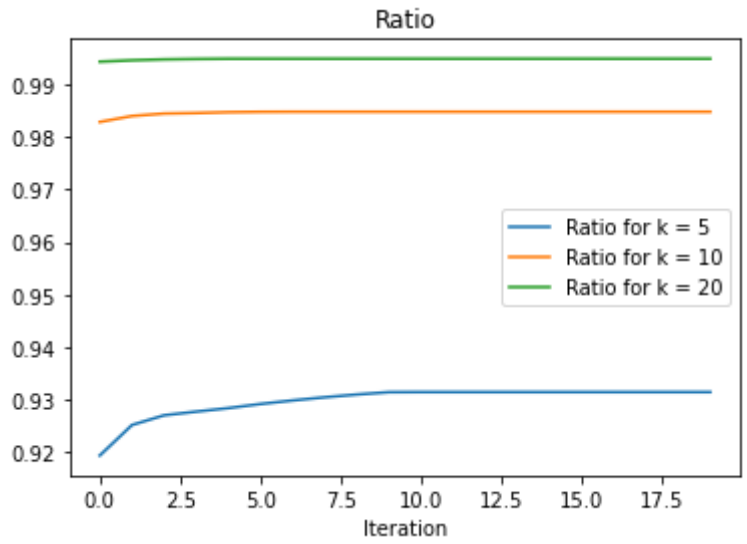






با توجه به نتایج بدست آمده در نمودار های بالا مشاهده می‌کنیم که مقدار هزینه‌ای که در تعداد خوشه بالاتر داریم کمتر است زیرا هر چه تعداد خوشه ها بیشتر باشد نقاط مختلف به مرکز خوشه خودشان نزدیکتر است. این موضوع الزاماً نشان دهنده عملکرد بهتر $k=20$ نیست.

الف.2) تاثیر تکرار آزمایش



ابتدا تعریف Outer distance و Inner distance را مطرح می‌کنیم.

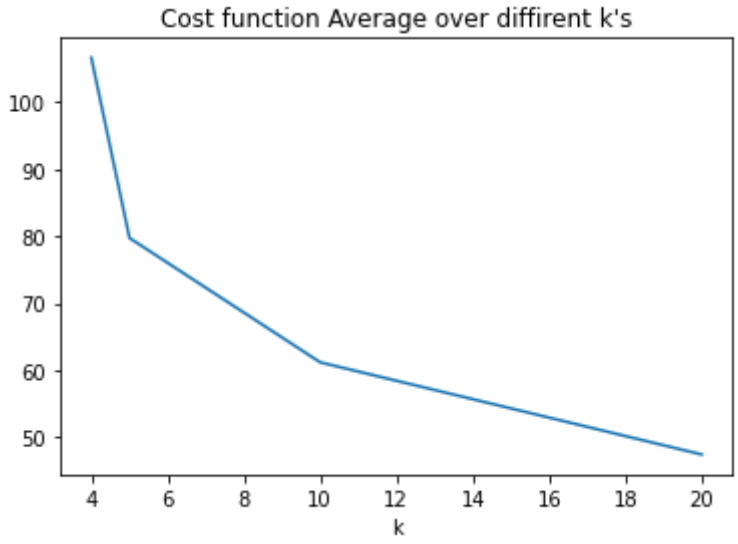
Outer distance: مجموع فواصل هر نقطه از دیگر نقاط خوشه‌های دیگر را گوییم.

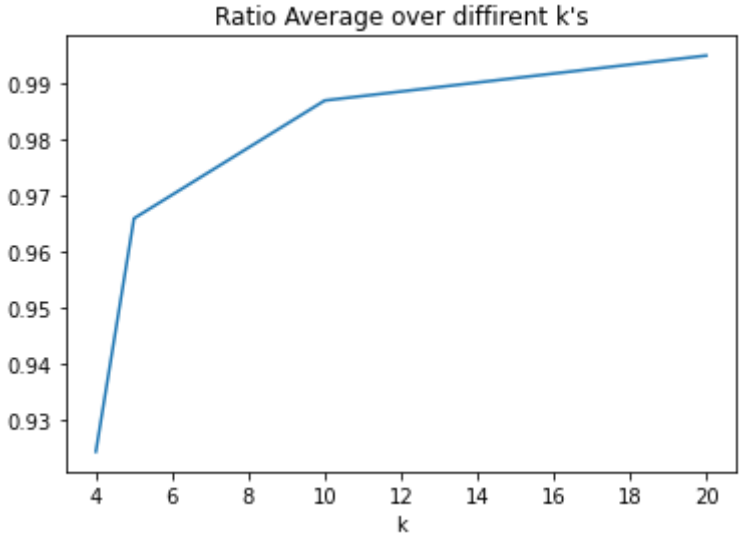
Inner distance: مجموع فواصل هر نقطه از دیگر نقاط هم خوشه‌ای را گوییم.

حال دلیل اینکه ratio را به صورت $\frac{\text{Outer distance}}{\text{Outer distance+Inner distance}}$ تعریف می کنیم این است که هدف ما دور کردن نقاط از خوشه های دیگر است و همچنین نزدیک کردن آن ها به مرکز خوشه خودی است. توقع داریم که این کسر به عدد یک همگرا شود. همانطور که مشاهده می‌کنیم در k=20 این مقدار به یک نزدیک‌تر است و عملکرد بهتری دارد و هر چه تعداد خوشه کمتر می‌شود Ratio نیز از یک کمتر می‌شود.

حال این آزمایش را چندین بار تکرار می‌کنیم و میانگین و واریانس را برای Ratio و Cost function آخر هر بار تکرار الگوریتم را محاسبه می‌کنیم. که نتایج به صورت زیر است:

```
Ratio mean for k = 5:  0.9651243472459538
Ratio variance for k = 5:  3.955623524489174e-06
Cost mean for k = 5:  80.91934393595828
Cost variance for k = 5:  21.474443101038112
-----------------------------------------------
Ratio mean for k = 10:  0.9869533479877255
Ratio variance for k = 10:  3.075017867870404e-07
Cost mean for k = 10:  60.146812860126985
Cost variance for k = 10:  3.2334535792105337
-----------------------------------------------
Ratio mean for k = 20:  0.9948334174864017
Ratio variance for k = 20:  3.8146325365078675e-08
Cost mean for k = 20:  46.96271056372344
Cost variance for k = 20:  2.326501446029138
```





با توجه به نمودار های بالا با اینکه در خوشه های با تعداد بیشتر میانگین هزینه کمتری داریم اما قرار نیست تعداد خوشه‌ها را تا جایی که می‌توانیم زیاد کنیم زیرا از هدف واقعی clustering دور می‌شویم. در نتیجه باید trade-off تعداد خوشه و هزینه را در نظر بگیریم. این موضوع را از نقطه زانو نمودار می‌توانیم به بهترین عملکرد در این سه مقدار پیدا کنیم که k=5 بهترین عملکرد را دارد.

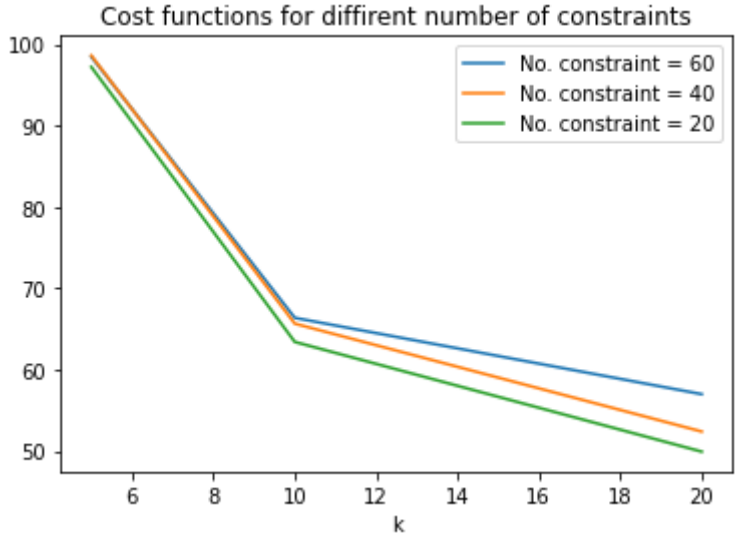
قسمت دوم: خوشه بندی کا-میانگین هوشمند

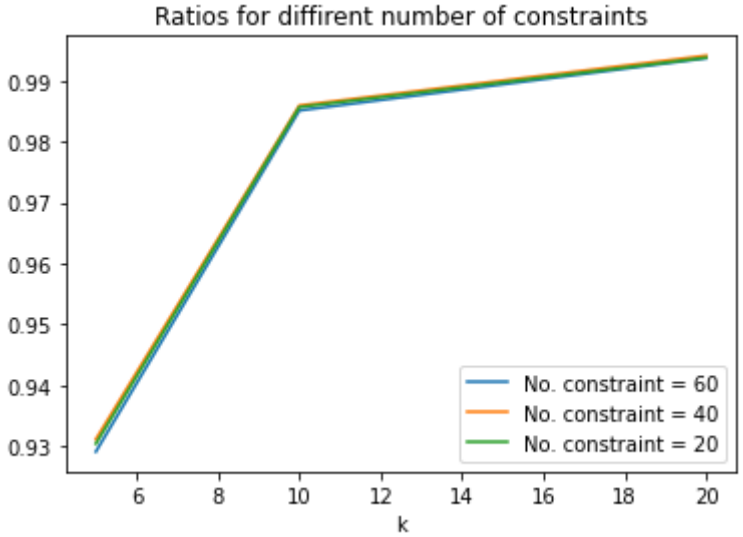
الف) طراحی الگوریتم

الگوریتم این روش بسیار مشابه روش کا-میانگین ساده است فقط تنها تفاوت آن این است که پس از هر ایتریشن که عمل clustering صورت پذیرفت شروط را چک می‌کنیم و پس از ارزیابی شروط دوباره همین روال تکرار می‌شود.

حال برای قسمت شروط به این گونه عمل می‌کنیم که اگر عدد 1 را در ستون سوم دریافت کردیم، چک می‌کنیم بین دو نقطه داده شده که باید در یک گروه باشند کدام داده به مرکز آن خوشه نزدیک‌تر است و با پیدا کردن آن، داده دیگری را به آن خوشه منتقل می‌کنیم. (البته همه این کار ها در صورتی است که دو داده در دو خوشه متفاوت قرار گرفته بودند)

اگر عدد -1 را در ستون سوم دریافت کردیم، چک می‌کنیم کدام داده به مرکز خوشه‌ای که آن دو داده نباید با هم در آن قرار بگیرند نزدیک‌تر است؛ داده نزدیک تر را نگه داشته و داده دورتر را به خوشه دیگری که پس از آن خوشه مذکور به آن نزدیک تر است را منتقل می‌کنیم.





برای رسم دو نمودار بالا به ازای تعداد خوشه های k={5,10,20} و تعداد شرط 20 و 40 و 60 استفاده کردیم. این نمودار ها Cost و ratio را به ازای k های مختلف نشان می‌دهد.

مشاهده می‌کنیم به ازای تعداد شرط کمتر عملکرد بهتری داریم و هزینه کمتری نیز داده‌ایم. این بدین معناست بعضی از شروط با اینکه متعلق به یک گروه دیگر بوده‌اند اما با توجه به معیار فاصله هزینه بیشتری داشته اند. به عبارتی الزاما حضور در یک کلاستر نشان دهنده کمترین فاصله نسبت به مرکز خوشه مذکور ندارد و می تواند به مرکز خوشه دیگری نزدیک تر باشد و در عین حال در یک کلاستر دیگر باشد.

سوال 2.

در این قسمت با توجه به اینکه لیبل های داده های iris را داریم به بررسی دقت خوشه‌بندی خود می پردازیم. البته باید توجه کنیم که شماره خوشه ها متناظر با لیبل های اصلی آن ها نیست. از این حیث این محاسبات را با جدا کردن 50 تا 50 تا داده‌ها انجام می‌دهیم و پرتکرار ترین لیبل را انتخاب می‌کنیم.

$$60\ constraint\ accuracy = \frac{141}{150} = 0.94$$

$$40\ constraint\ accuracy = \frac{137}{150} = 0.9133$$

$$20\ constraint\ accuracy = \frac{137}{150} = 0.9133$$

$$simple\ accuracy = \frac{118}{150} = 0.7866$$