سوال ۱: معیارهای شباهت (تئوری)

همانطور که از مطالب درس آموختید، برای خوشه‌بندی داده‌ها نیاز به معیاری برای سنجش میزان شباهت بین آن‌ها داریم. تعیین این معیار شباهت به نوع داده‌ها و هدف مسئله بستگی دارد. فاصله اقلیدسی[1] مثالی از یک معیار شباهت است که معمولا برای داده‌های عددی استفاده می‌شود و شما با آن در طول درس آشنا شدید. در این سوال به دنبال آشنایی با برخی معیارهای شباهت دیگر و تمرین آموخته‌هایمان هستیم.

قسمت الف: انتخاب معیار شباهت مناسب

در ستون سمت راست جدول ۱-۱ تعریف چند مسئله و در ستون سمت چپ چند معیار شباهت را مشاهده می‌کنید. هر مسئله را به معیار شباهت مناسبش متصل کنید و دلیل انتخاب خود را شرح دهید.

مسئله	معیار شباهت
خوشه‌بندی ستاره‌های آسمان	نقاط چگالی دردسترس[2] (الگوریتم DBSCAN را مطالعه کنید.)
خوشه‌بندی اسناد متنی	فاصله اقلیدسی
خوشه‌بندی نتایج آزمایشات پزشکی به صورت اعداد باینری و بازه‌ی عددی	شباهت جاکارد[3]
خوشه‌بندی خانه‌های برای تعیین محل احداث مراکز پستی با وجود موانع	شباهت کسینوسی[4]

جدول ۱-۱: انتخاب معیار شباهت

قسمت ب: ماتریس بی‌شباهتی[5]

ماتریس بی‌شباهتی همه اطلاعات مربوط به تفاوت داده‌ها را، بر اساس یک معیار تفاوت $d(.,.)$، در خود ذخیره می‌کند. برای n داده ماتریس بی‌شباهتی D عبارت است از:

$$D = (d_{ij});\ d_{ij} = d(i,j), 1 \le i,j \le n$$

$$0 \le d_{ij} \le 1$$

بنابراین برای شباهت خواهیم داشت:

$$s_{ij} = 1 - d_{ij}$$

[1] Euclidian distance (L2 norm)
[2] Density reachable
[3] Jaccard similarity
[4] Cosine similarity
[5] Dissimilarity matrix

اکنون جدول ۱-۲ را در نظر بگیرید. داده‌های این جدول ترکیبی از ویژگی‌های عددی، اسمی و ترتیبی دارند. ماتریس بی‌شباهتی را برای این داده‌ها بدست آورید.

شماره	ویژگی ۱ (اسمی)	ویژگی ۲ (ترتیبی)	ویژگی ۳ (عددی)
۱	A	عالی	۴۵
۲	B	متوسط	۲۲
۳	C	خوب	۶۴
۴	A	عالی	۲۸

جدول ۱-۲

نکته ۱: برای معیار شباهت می‌توانید هم از مطالبی که در درس آموخته‌اید استفاده کنید و هم روش‌های دیگری به کار ببرید اما باید بتوانید روش خودتان را در گزارش توجیه کنید.

نکته ۲: برای ویژگی ترتیبی می‌توانید ابتدا با یک نگاشت، هر مقدار را به یک عدد تبدیل کنید و سپس از فاصله اقلیدسی استفاده کنید.

نکته ۳: در نهایت فاصله هر دو داده عبارت از میانگین فاصله ویژگی‌هایشان خواهد بود.

سوال ۲: الگوریتم‌های خوشه‌بندی (تئوری)

قسمت الف: خوشه‌بندی با روش کا-میانگین[6]

با استفاده از روش خوشه‌بندی کا-میانگین ، داده های جدول ۲-۱ را به ۲ خوشه تقسیم کنید. مراکز اولیه خوشه‌ها را B و C در نظر بگیرید. برای درک بهتری از مسئله، نقاط را رسم کرده و خوشه‌ها را روی شکل مشخص کنید. همچنین مشخص کنید نقطه $x = (3, 2.5)$ در کدام دسته قرار می‌گیرد. برای فاصله از فاصله اقلیدسی استفاده کنید.

i	x_1	x_2
A	1	1
B	2	1
C	2	3
D	3	2
E	4	3
F	5	5

جدول ۲-۱: داده‌های مربوط به سوال کا-میانگین

قسمت ب: خوشه‌بندی با روش سلسله مراتبی[7]

یکبار با استفاده از پیوند واحد[8] و یکبار هم با پیوند کامل[9]، با محاسبه فاصله اقلیدسی، داده‌های زیر را خوشه‌بندی کرده و نمودار درختی آن را رسم کنید.

i	x_1	x_2
A	0.45	**0.3**
B	0.22	**0.38**
C	0.08	**0.41**
D	0.26	**0.19**
E	0.35	**0.32**

جدول ۲-۲: داده‌های مربوط به سوال پیوند واحد

[6] K-means clustering
[7] Hierarchical clustering
[8] Single-linkage
[9] Complete-linkage

سوال ۳: الگوریتم‌های خوشه‌بندی (پیاده‌سازی)

قسمت اول: الگوریتم کا-میانگین با انتخاب اولیه هوشمندانه

در الگوریتم کا-میانگین ابتدا باید به طور تصادفی تعدادی مرکز خوشه انتخاب می‌کردیم. در درس آموختید که انتخاب مراکز خوشه اولیه ممکن است تاثیر زیادی روی نتایج نهایی الگوریتم و سرعت همگرایی آن داشته باشد. در این قسمت می‌خواهیم به پیاده‌سازی الگوریتم کا-میانگین با استفاده از یکی از روش‌های هوشمندانه برای انتخاب مراکز خوشه اولیه بپردازیم.

الف) الگوریتم کا-میانگین را با معیار فاصله اقلیدسی مطابق آنچه که در درس آموختید برای دادگان iris پیاده‌سازی کنید اما با این تفاوت که به جای انتخاب مراکز خوشه در ابتدا به صورت کاملا تصادفی، از الگوریتم زیر استفاده کنید.

الگوریتم جدید انتخاب مراکز خوشه اولیه: اولین مرکز خوشه را به طور تصادفی و با احتمال یکنواخت از بین کل نقاط مجموعه داده‌ها انتخاب کنید. مراکز خوشه بعدی را باز هم به تصادف اما اینبار با احتمال متفاوتی انتخاب کنید. احتمال انتخاب هر مرکز خوشه (غیر از مرکز خوشه اول) متناسب با فاصله‌اش از نزدیکترین مرکز خوشه (نقاطی که تا آن لحظه به عنوان مرکز خوشه انتخاب شده‌اند) است. برای فاصله هم می‌توانید از همان فاصله اقلیدسی استفاده کنید.

ب) الگوریتم پیاده‌سازی‌شده را برای تعداد خوشه‌های مختلف ۱ تا ۵ خوشه اجرا کنید. برای هر تعداد خوشه الگوریتم را به اندازه کافی تکرار کنید تا همگرایی حاصل شود. از مجذور میانگین فاصله اقلیدسی نقاط با مرکز خوشه‌هایشان به عنوان تابع هزینه استفاده کنید.

نکته: کل فرآیند خوشه‌بندی را چند بار با انتخاب مرکز خوشه اولیه متفاوت تکرار کنید و بهترین حالت را بر اساس مقدار تابع هزینه در نظر بگیرید.

ج) برای هر تعداد خوشه، نمودار تابع هزینه را بر حسب تکرارهای الگوریتم، تا رسیدن به همگرایی رسم کنید.

د) نمودار مقدار همگرایی تابع هزینه بر حسب تعداد خوشه‌ها را رسم کنید و سپس بر اساس آموخته‌هایتان در درس، بگویید که چه تعداد خوشه برای این دادگان مناسب است.

تذکر۱: در این قسمت مجاز به استفاده از کتابخانه نیستید.

تذکر۲: دادگان iris را می‌توانید از کتابخانه sklearn دریافت کنید.

قسمت دوم: کا-میانه[10]

در روش کا-میانه به جای فاصله اقلیدسی، از فاصله منهتن[11] در تابع هزینه خود استفاده می‌کند. در این قسمت می‌خواهیم مقاومت این روش را با روش کا-میانگین نسبت به داده‌های تقلبی[12] بررسی کنیم.

الف) قسمت قبل را برای روش کا-میانه هم انجام دهید. (فقط به ازای ۳ خوشه)

[10] K-median
[11] Manhattan
[12] Outlier

ب) مراکز خوشه‌ها را که در قسمت‌های قبل با دو روش کا-میانگین و کا-میانه به ازای ۳ خوشه بدست آورده‌اید در نظر بگیرید. با توجه به ساختار دادگان iris و اندازه عددی ویژگی‌های هر داده، ۲۰ داده پرت به دادگان اضافه کنید.

اکنون مجددا فرآیندهای خوشه‌بندی کا-میانه و کا-میانگین را با حضور داده‌های تقلبی اجرا کنید. مقدار اختلاف مراکز خوشه را در حالتی که داده تقلبی نداشتیم با حالتی که داده تقلبی داریم، برای هر دو الگوریتم مقایسه کنید. کدام یک به داده‌های تقلبی حساس‌تر است؟

نکات تحویل

۱- مهلت تحویل بخش تئوری این تمرین ۱۵ آذرماه و بخش پیاده‌سازی ۲۱ آذرماه می‌باشد. (با توجه به اینکه بعد از تمرین پاسخ سوالات تئوری بارگذاری می‌شود امکان استفاده از گریس برای بخش تئوری را ندارید.)

۲- انجام این تمرین به صورت یک‌نفره است.

۳- برای انجام این تمرین تنها مجاز به استفاده از زبان برنامه نویسی پایتون هستید.

۴- در صورت وجود تقلب نمره تمامی افراد شرکت کننده در آن نمره صفر لحاظ می‌شود.

۵- در صورتی که از منبعی برای هر بخش استفاده می‌شود، حتماً لینک مربوط به آن در گزارش آورده شود. وجود شباهت بین منبع و پیاده‌سازی در صورت ذکر منبع بلامانع است. اما در صورت مشاهده شباهت با مطالب موجود در سایت‌های مرتبط نمره کسر می‌گردد.

۶- نتایج و تحلیل‌های شما در روند نمره‌دهی دستیاران آموزشی تأثیرگذار است.

۷- لطفا پاسخ تمرین خود را (به همراه کد/گزارش سوال کامپیوتری) به صورت زیر در صفحه درس آپلود نمایید:

HW[HW number]_[Last_name]_[Student number].zip

۸- در صورت وجود هر گونه ابهام یا مشکل می‌توانید از طریق ایمیل با طراحان تمرین در تماس باشید:

- سیدمحمدمتین آل‌محمد: sm.matin.alemohammad@gmail.com
- نیلوفر فریدنی: nilu.1380@gmail.com