به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس سیستم‌های هوشمند

تمرین شماره 3

آذر 1401

فهرست سوالات

سوال ۱: تحلیلی

در این بخش به بررسی و حل دستی الگوریتم‌های خوشه‌بندی[1] می‌پردازیم.

الف) خوشه‌بندی به روش کا-میانگین[2]

با استفاده از روش خوشه‌بندی کا-میانگین ابتدا ماتریس فواصل(مربعات) داده‌های زیر را بدست آورید و سپس آنها را به دو خوشه تقسیم کنید. در نهایت تعیین کنید نقطه‌ی $x^* = (1,1)$ در کدام خوشه قرار می‌گیرد.

جدول ۱-۱: داده‌های 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

نکته ۱: فاصله‌ی بین دو نقطه‌ی x_i و x_j بصورت زیر بدست می‌آید:

$$D_{ij} = dist_{eucl.}(x_i, x_j)^2$$

نکته ۲: نقاط ابتدایی و انتهایی داده‌ها را به عنوان مراکز اولیه خوشه‌بندی استفاده کنید.

نکته ۳: برای دستیابی به دید بهتری از مساله، نقاط را رسم کنید و مرکز خوشه ها را مشخص کنید.

[1] Clustering
[2] K-Means

ب) خوشه‌بندی سلسله‌مراتبی

جدول ۱-۲ ماتریس فواصل اقلیدسی نقاط را نشان می‌دهد.

ب.۱) پیوند واحد

با استفاده از روش پیوند واحد[1]، داده های زیر را خوشه بندی کرده و نمودار درختی آن را رسم کنید.

ب.۲) پیوند کامل

با استفاده از روش پیوند کامل[2]، داده های زیر را خوشه بندی کرده و نمودار درختی آن را رسم کنید.

ب.۳) مقایسه

دو مقدار از ماتریس داده‌شده را طوری تغییر دهید که نتیجه‌ی استفاده از پیوند واحد و پیوند کامل مشابه یکدیگر شوند. روند فکر خود را در گزارش بیاورید.

جدول ۱-۲: مراتبی‌سلسه روش به بندی‌خوشه فواصل ماتریس

	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

[1] Single Linkage
[2] Complete Linkage

سوال 2: پیاده سازی الگوریتم خوشه بندی

قسمت اول: خوشه‌بندی کا-میانگین ساده

در این سوال به پیاده‌سازی الگوریتم خوشه‌بندی کا-میانگین می‌پردازیم. هدف این الگوریتم بدون نظارت، کمینه کردن تابع هزینه‌ای[1] است که به صورت مجموع فاصله‌ی نمونه‌های متعلق به هر خوشه تا مرکز آن خوشه (فاصله درون کلاسی) تعریف می‌شود.

با در نظر گرفتن مجموعه داده Iris سعی داریم الگوریتم خوشه‌بندی کا-میانگین ساده را با در نظر گرفتن نرم اقلیدسی پیاده‌سازی کنیم.

```
from sklearn.datasets import load_iris
data = load_iris()
```

الف.۱) تاثیر تکرار خوشه‌ها

پس از پیاده‌سازی الگوریتم، برای تعداد خوشه‌های ۵، ۱۰ و ۲۰، الگوریتم را ۲۰ بار تکرار کنید. نمودار مقدار هزینه را در حین اجرای الگوریتم رسم نمایید. آیا میتوانید تحلیل کنید که کدام تعداد خوشه مناسب تر است؟ (می‌توانید از معیار نسبت شباهت درونی[2] به شباهت بیرونی[3] استفاده نمایید).

الف.۲) تاثیر تکرار آزمایش

برای تعداد خوشه‌های ۵، ۱۰ و ۲۰، الگوریتم را به تعداد کافی تکرار کنید. Ratio را به صورت رابطه ۲-۱، تعریف می‌کنیم.

$$Ratio = \frac{Outer_{distance}}{Outer_{distance}+Inner_{distance}}$$ رابطه‌ی ۱-۲.

نمودار Ratio بر حسب تعداد دفعات تکرار الگوریتم را برای هر سه تعداد خوشه در یک نمودار رسم کنید. تحلیل کنید که کدام تعداد خوشه مناسب‌تر است؟

[1] Cost Function
[2] Intra-cluster Distance
[3] Inter-cluster Distance

حال برای کل تعداد تکرارهای انجام شده به ازای تعداد خوشه‌های ۵، ۱۰ و ۲۰، میانگین و واریانس Ratio و تابع هزینه را محاسبه کنید. تحلیل کنید کدام تعداد خوشه با توجه به میانگین‌های به دست آمده، مناسب تر است. (برای درک بهتر می‌توانید، نمودار میانگین Ratio بر حسب تعداد خوشه و یا میانگین تابع هزینه بر حسب تعداد خوشه را رسم کنید، در نهایت با توجه به نمودار نتیجه‌گیری کنید.)

قسمت دوم: خوشه بندی کا-میانگین هوشمند

در این قسمت، ناظری از بیرون اطلاعاتی در مورد قسمتی از دادگان[1] به شما می‌دهد که شما را وادار می‌کند در فرآیند خوشه بندی محتاط تر عمل کنید. به این خوشه بندی، خوشه بندی شرطی[2] گفته می‌شود. در خوشه بندی شرطی کا-میانگین تمامی مراحل مانند کا-میانگین ساده طی می‌شود با این تفاوت که در فرآیند تعلق هر داده به خوشه مربوطه نباید شرطی از شروط داده شده نقض گردد.

نگارش هر شرط داده شده به صورت سه تایی زیر می‌باشد:

$$[d_i, d_j, State]\ ; d_{i,j} \in D\ (Data\ set), State\ \in \{-1,1\}$$

شرط بالا بیان می‌کند که هر دو داده دلخواه با اندیس[3]های متفاوت درون یک خوشه هستند یا خیر. در صورتی که عبارت State برابر 1 باشد، دو داده تحت هر شرایطی باید درون یک خوشه قرار بگیرند[4] و در صورتی که برابر 1- باشد، دو داده تحت هر شرایطی باید در دو خوشه متفاوت قرار گیرند[5]. بدین صورت که یک داده به نزدیکترین خوشه‌ای تعلق می‌گیرد که هیچ کدام از دو نوع شرط ذکر شده نقض نشود.

الف) طراحی الگوریتم

الگوریتمی طراحی کنید که مرحله به مرحله مانند الگوریتم کا-میانگین ساده به سمت جواب بهینه همگرا شود.

[1] Background knowledge
[2] Constrained clustering
[3] Index
[4] Must-linked constraint
[5] Cannot-link constraint

ب) پیاده سازی الگوریتم

در پوشه پروژه فایلی بنام "Constraint.txt" وجود دارد که 60 شروط برای پیاده سازی الگوریتم را در بر دارد. به ترتیب برای سه مقدار موجود در {20,40,60}، قسمتی از شروط رو به صورت رندوم انتخاب کنید و الگوریتم را اجرا کنید.

(اندیس‌های موجود در فایل قرار داده شده متناظر با اندیس‌های دادگان Iris خوانده شده توسط کتابخانه Scikit-learn می‌باشند)

ج) ارزیابی الگوریتم

1. با توجه به معیارهای ارزیابی در درس، کیفیت خوشه بندی خود را برای سه حالت ذکر شده با رسم نمودار مقایسه و تحلیل کنید.
2. به طور خاص تر، برای تعداد 3 خوشه، به کمک برچسب موجود برای دادگان، دقت را برای دو حالت خوشه بند کا-میانگین و کا-میانگین هوشمند مقایسه کنید. آیا اطلاعات زمینه‌ای داده شده در حالت کا-میانگین هوشمند باعث افزایش دقت شده است ؟

نکات تحویل:

- مهلت تحویل این تمرین 17 آذر میباشد.
- انجام این تمرین به صورت یک نفره است.
- برای انجام این تمرین تنها مجاز به استفاده از زبان برنامه نویسی پایتون هستید.
- در صورت وجود تقلب نمره تمامی افراد شرکت کننده در آن -۱۰۰ لحاظ میشود.
- 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

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

مسئول تمرین سوال۱: شیوا شاکری (shiva.shakeri@ut.ac.ir)

مسئول تمرین سوال۲ قسمت اول: عاطفه ملاباقر (ati.mollabagher@ut.ac.ir)

مسئول تمرین سوال۲ قسمت دوم: شایان واصف (sh.vassef@ut.ac.ir)