به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس داده‌کاوی

تمرین خوشه‌بندی

اردیبهشت‌ماه ۱۴۰۲

*فهرست

بخش ۱- سوالات عملی

خوشه بندی یک تکنیک یادگیری بدون ناظر است که در داده کاوی برای گروه‌بندی نقاط داده مشابه بر اساس ویژگی‌های آن‌ها استفاده می‌شود. هدف این تمرین، مروری بر الگوریتم‌های خوشه‌بندی رایج می‌باشد.

شرح دادگان

در این تمرین با دو مجموعه داده متفاوت کار خواهیم کرد:

- مجموعه داده اطلاعات یک شرکت بیمه خودرو
- مجموعه داده اطلاعات وقوع زلزله در ایران از سال ۲۰۱۰ تاکنون. .

- مجموعه داده بیمه شامل ۱۰۳۰۲ سطر با ۲۷ ویژگی است. هر رکورد (ردیف) مجموعه‌ای از ویژگی‌های یک مشتری منفرد شرکت بیمه را نشان می‌دهد که با مشخصات اجتماعی-جمعیت‌شناختی و وسیله نقلیه بیمه‌شده مرتبط است. شرح این ویژگی‌ها در جدول زیر آورده شده است.

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_FLAG	Was Car in a crash? 1=YES 0=NO	None
TARGET_AMT	If car was in a crash, what was the cost	None
AGE	Age of Driver	Very young people tend to be risky. Maybe very old people also.
BLUEBOOK	Value of Vehicle	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_AGE	Vehicle Age	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_TYPE	Type of Car	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_USE	Vehicle Use	Commercial vehicles are driven more, so might increase probability of collision
CLM_FREQ	# Claims (Past 5 Years)	The more claims you filed in the past, the more you are likely to file in the future
EDUCATION	Max Education Level	Unknown effect, but in theory more educated people tend to drive more safely
HOMEKIDS	# Children at Home	Unknown effect
HOME_VAL	Home Value	In theory, home owners tend to drive more responsibly
INCOME	Income	In theory, rich people tend to get into fewer crashes
JOB	Job Category	In theory, white collar jobs tend to be safer
KIDSDRIV	# Driving Children	When teenagers drive your car, you are more likely to get into crashes
MSTATUS	Marital Status	In theory, married people drive more safely
MVR_PTS	Motor Vehicle Record Points	If you get lots of traffic tickets, you tend to get into more crashes
OLDCLAIM	Total Claims (Past 5 Years)	If your total payout over the past five years was high, this suggests future payouts will be high
PARENT1	Single Parent	Unknown effect
RED_CAR	A Red Car	Urban legend says that red cars (especially red sports cars) are more risky. Is that true?
REVOKED	License Revoked (Past 7 Years)	If your license was revoked in the past 7 years, you probably are a more risky driver.
SEX	Gender	Urban legend says that women have less crashes then men. Is that true?
TIF	Time in Force	People who have been customers for a long time are usually more safe.
TRAVTIME	Distance to Work	Long drives to work usually suggest greater risk
URBANICITY	Home/Work Area	Unknown
YOJ	Years on Job	People who stay at a job for a long time are usually more safe

- مجموعه داده زلزله شامل ۲۷۹۵ سطر و ۲۲ ویژگی می‌باشد. اطلاعات مربوط به متغیرها و ویژگی‌های موجود در مجموعه داده زلزله از اینجا قابل دسترسی می‌باشد.

سوال ۱– K-Means

بسیاری از مجموعه داده‌های دنیای واقعی شامل ترکیبی از ویژگی‌های داده عددی، ترتیبی (به عنوان مثال کوچک، متوسط، بزرگ) و اسمی (مانند فرانسه، چین، هند) هستند. در این سوال می‌خواهیم خوشه‌بندی را برای مجموعه داده بیمه خودرو اعمال کنیم. برای این کار ابتدا لازم است تا داده‌ها را با توجه به دانش پیشین خود برای تصحیح و مدیریت ویژگی‌ها پیش‌پردازش کنید (لازم است تا دلیل هر اقدام را به صورت مشروح در فایل گزارش بیان کنید)

سپس با استفاده از معیار سیلوئت، تغییرات دقت خوشه‌بندی نسبت به تعداد خوشه‌ها را در روش kmeans به‌دست آورید و نتایج را تحلیل کنید.

در آخر خوشه‌بندی به کمک روش kmeans را با تعداد خوشه مناسب بر روی مجموعه داده نرمال‌شده اجرا و نتیجه خوشه‌بندی را بر روی نمودار گزارش کنید.

سوال ۲- DBSCAN

در این سوال می‌خواهیم خوشه‌بندی را برای مجموعه داده زلزله اعمال کنیم. برای این سوال باید از روش DBSCAN استفاده کنید. برای تجزیه و تحلیل اولیه خوشه‌بندی، مجموعه ویژگی‌های خود را فقط به طول و عرض جغرافیایی (latitude, longitude) محدود می‌کنیم.

قبل از اعمال خوشه‌بندی، مقادیر بهینه epsilon و min_points را با استفاده از معیارسیلوئت به دست بیاورید و نتیجه را گزارش کنید.

- مقادیر مختلف (خیلی کم یا خیلی زیاد) اپسیلون چه تاثیری در خوشه‌بندی دارند؟
- مقادیر مختلف (خیلی کم یا خیلی زیاد) min_points چه تاثیری در خوشه‌بندی دارند؟

سپس، الگوریتم DBSCAN را با استفاده از مقادیر بهینه epsilon و min_points که قبلاً تعیین کردید، روی مجموعه داده اعمال کرده و خوشه‌های حاصل را گزارش کنید.

اگر افزودن ویژگی‌های بیشتری از دیتاست به مجموعه ویژگی‌های انتخابی منجر به بهبود خوشه‌بندی می‌شود، آن‌ها را در مجموعه ویژگی‌ها اضافه کرده و دلایل انتخاب و نتایج نهایی را گزارش کنید.

بخش ۲- سوالات تئوری

سوال ۱ – Agglomerative Hierarchical Clustering

نقاط زیر را در نظر بگیرید، با استفاده از معیار فاصله اقلیدسی Dendrogramهای الگوریتم خوشه‌بندی Agglomerative را با دو روش Single-Link و Complete-Link رسم کنید.

Point	x	y
A	0.18	0.76
B	0.02	0.27
C	0.55	0.52
D	0.88	0.53
E	0.38	0.77
F	0.35	0.05

سوال ۲– K-Modes

جدول زیر نتایج تست روانشناسی شخصیت ۱۰ فرد در ۵ دسته‌بندی می‌باشد. به کمک روش K-Modes این داده‌ها را خوشه‌بندی کنید و مراحل خوشه‌بندی و خوشه‌های نهایی را به طور کامل گزارش کنید.

- تعداد خوشه‌ها را برابر ۳ و معیار فاصله را تعداد عناصر نابرابر در نظر بگیرید.
- عناصر مرکزی اولیه را افراد ۳، ۸ و ۹ در نظر بگیرید.

	Trait	Comm	Decision	Problem-solve	Leadership
0	Extroverted	Assertive	Intuitive	Practical	Laissez-faire
1	Extroverted	Direct	Intuitive	Analytical	Authoritarian
2	Extroverted	Assertive	Emotional	Creative	Democratic
3	Open-minded	Direct	Emotional	Analytical	Democratic
4	Open-minded	Assertive	Logical	Analytical	Authoritarian
5	Open-minded	Assertive	Emotional	Practical	Democratic
6	Open-minded	Indirect	Intuitive	Creative	Authoritarian
7	Introverted	Assertive	Intuitive	Analytical	Authoritarian
8	Introverted	Indirect	Emotional	Creative	Authoritarian
9	Extroverted	Assertive	Logical	Analytical	Laissez-faire

فرض کنید نتایج آزمایش ۲۰۰۰ نفر را در اختیار داریم و آن‌ها را خوشه‌بندی کرده‌ایم. از طرفی این افراد متعلق به 4 کلاس شخصیتی کلی بوده‌اند. جدول زیر تعداد آیتم های موجود در هر خوشه به تفکیک دسته‌بندی آنها را گزارش میکند. با استفاده از این جدول، مقادیر precision, recall, f1, entropy را به‌دست بیاورید

	Class_1	Class_2	Class_3	Class_4
Cluster_1	2	0	22	540
Cluster_2	23	333	242	89
Cluster_3	36	12	700	1

ملاحظات (حتما مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان DM_CA3_StudentID تحویل داده شود.

خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد.

مهلت تحویل تمرین به هیچ عنوان تمدید نخواهد شد. مجموعا ۱۴ روز برای تمامی تمرین‌ها و پروژه‌ی درس به عنوان Grace day در نظر گرفته می‌شود و پس از پایان مجموعا ۱۴ روز، برای هر تمرینی که پس از زمان اختصاص یافته ارسال شود روزی ۱۵ درصد از نمره آن تمرین کسر خواهد شد.

- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارائه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر و به استاد نیز گزارش می‌گردد.
- در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

mailto:mohammad.na3ri@gmail.com

مهلت تحویل: ۳ خرداد ۱۴۰۲