دانشگاه صنعتی اصفهان

دانشکده برق و کامپیوتر

مبانی داده‌کاوی

سوالات مباحث خوشه‌بندی (تکلیف سری 5)

بهار 1403

فهرست مطالب

۱- سوالات

1- چند حالت را مثال بزنید که الگوریتم k-means در آنها بد عمل می‌کند. (حداقل سه مورد)

2- تاثیر زیادی کم و یا زیاد بودن پارامتر eps در DBSCAN، در نتیجه‌ی الگوریتم چگونه خواهد بود؟

3- درستی یا نادرستی عبارات زیر را مشخص کنید. (در صورت نادرست بود علّت آن را ذکر نمایید)

الف) برخلاف k-means، الگوریتم‌های خوشه‌بندی سلسله مراتبی نیازی به تعیین مقدار k از ابتدا ندارند.

ب) همواره می‌توان از انتروپی به عنوان معیار ارزیابی خوشه‌بندی استفاده نمود.

پ) استفاده از معیار group average به جای min در خوشه بندی سلسله مراتبی حساسیت به نویز و نیز بایاس شدن به سمت اشکال کروی را کاهش می‌دهد.

ت) الگوریتم k-means به نقاط مرکزی اوّلیه حساس است.

4- با استفاده از مجموعه داده Wholesale customers data به سوالات زیر پاسخ دهید:

1- گزارشی را از دیتاست بگیرید. (اسم ستون، تعداد فیلد غی ر null و تایپ و ...)
2- به کمک تابع pca تعداد فیچر ها را به دو فیچر کاهش دهید و آن را در یک دیتافریم مجزا ذخیره نمایید.
3- به روش ward، نمودار Dendogram داده‌ها را رسم کنید و تعداد کلاستر مناسب را بدست آورید.
4- برای مقادیر 2 تا 25 خوشه‌بندی به روش Agglomerative را انجام دهید، معیار silhouette را برای هر خوشه بدست آورده و نمودار میله ای ان را رسم کنید. نهایتاً بهترین مقدار k برای خوشه بندی بدست آورید.
5- برای بهترین k که در قسمت قبل بدست آوردید نمودارscatter را رسم کنید.
6- این بار دیتاست اولیه را براساس tsne نرمالایز کرده و آن را در دو بعد نشان دهید.
7- از الگوریتم DBScan استفاده کنید و نتیجه خوشه بندی را نشان دهید. الگوریتم را با پارامتر های مختلف انجام دهید و سعی کنید پارامترهای مناسب را تخمین بزنید. خروجی خوشه بندی را به وسیله نمودار scatter plot نشان دهید.
8- توسط الگوریتم NearestNeighbors از کتابخانه sklearn.neighbors میتوانید تخمین خود در مرحله قبل را بهبود ببخشید. از این کتابخانه استفاده کرده و پارامتر های dbscan را با آن تخمین بزنید (لزومی ندارد که تفکیک دقیقی بین کلاستر های موجود صورت بگیرد، صرفا نحوه کار با این کتابخانه مد نظر است)

5- دیتاست EastWestAirlines اطلاعات سفر 3999 مسافر را دارد. هدف پیدا کردن خوشه‌های مشابه مسافران به منظور انجام تبلیغات موثر (با دادن پیشنهادهای سفر های مناسب برای هر خوشه) می باشد.

1- دیتاست را نرمالایز نموده و سپس با روش ward و Euclidian distansce خوشه‌بندی سلسله‌مراتبی انجام دهید.(تعداد خوشه‌های بدست آمده چقدر است؟)
2- چه اتفاقی می‌افتاد اگر بدون نرمالایز کردن، قسمت (1) انجام می‌شد؟
3- ویژگی‌های مراکز خوشه را مقایسه کنید و سعی کنید به هر کدام یک لیبل بدهید.
4- برای آزمودن پایداری خوشه ها به صورت رندم 5٪ داده‌ها را حدف کنید و آنالیز را با 95% باقی‌مانده تکرار کنید. آیا نتیجه مانند قبل می‌شود؟
5- حال از الگوریتم k-means برای خوشه‌بندی استفاده کنید و مقدار k را همان تعداد خوشه‌ای که در مراحل قبل بدست آوردید قرار دهید. نتیجه را با حالت قبل مقایسه کنید.
6- برای مقادیر k از 1 تا 25 الگوریتم K- means را اجرا کنید و نمودار SSE بر حسب k را رسم کنید. بر اساس روش elbow تعیین کنید کدام k مناسب‌تر است.
7- برای مقادیر 2 تا 25 معیار silhouette را با روش K- means به دست آورید. نمودار میله ای معیار silhouette بر اساس k رسم کنید و سپس بهترین k را بدست آورید.
8- برای بهترین k که درقسمت های 4 و 5 بدست آورده‌اید خوشه بندی را انجام داده و نتیجه را به کمک scatter plot نشان دهید.

۲- نکات پاسخ‌دهی

- تمرینات به صورت مرتب و خوانا بارگذاری شود.
- برای تمرینات غیر عملی که به صورت تایپی ارسال شوند امتیاز تشویقی درنظر گرفته می شود.
- کدهای خود را حتماً در فایل PDF نیز قرار دهید.
- در سوالات توضیحی، قدرت تحلیل افراد ملاک مقایسهٔ پاسخ ها خواهد بود.
- فایل پایتون و یا Notebook برای تمرینات ضمیمه شود و همه به صورت یک فایل zip بارگذاری شوند. فایل zip را با فرمت DM4022_HW5_[StudentNumber].zip نام‌گذاری کنید.
- درصورت وجود ابهام خاص می توانید موارد را با دستیار آموزشی مطرح کنید.