

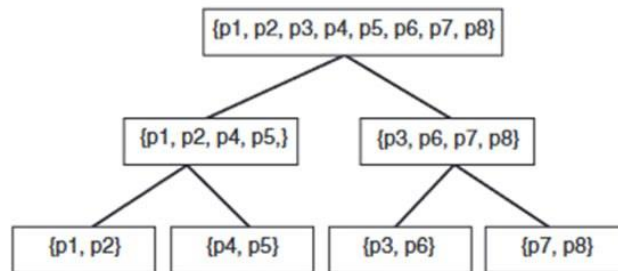
تمرین سوم درس مبانی داده کاوی

(بهار ۴۰۲)

مهلت تحویل تمرین: ۸ اردیبهشت ماه

سوالات تئوری

سوال ۱- برای ۸ مورد داده های $p1$ تا $p8$ و نیز خوشه بندی سلسله مراتبی شکل زیر مقدار **F-measure** سلسله مراتبی (**hierarchical F-measure**) را محاسبه نمایید. کلاس A شامل $p1$ ، $p2$ و $p3$ است و کلاس B شامل $p4$ تا $p8$ می باشد.



سوال ۲- یک مجموعه داده حاوی ۱۰۰ رکورد به شما داده می شود و از شما خواسته شده که داده ها را خوشه بندی کنید. شما از **K-means** برای خوشه بندی داده ها استفاده می کنید، اما به ازای تمام k های ممکن ($k = \{1, 2, \dots, 100\}$) الگوریتم **K-means** تنها یک خوشه غیر خالی را بر می گرداند. سپس یک نسخه افزایشی از **K-means** را اعمال کنید، اما دقیقاً همان نتیجه قبلی را به دست می آورید. چه گونه چنین اتفاقی ممکن است؟ چگونه یک خوشه بندی **single link** یا **DBSCAN** می تواند چنین مشکلی را رفع کند؟

سوال ۳- با یک مثال نشان دهید که انتخاب مرکزهای تصادفی ممکن است در **K-means** باعث عملکرد ضعیف این الگوریتم خوشه بندی شود. اجرای چندگانه می تواند این مشکل را برطرف کند، اما گاهی ممکن است این کار نیز به خوبی عمل نکند. این مورد را نیز با یک مثال نشان دهید.

سوال ۴- هر دو الگوریتم **K-means** و **K-medoids** می توانند خوشه بندی موثری را انجام دهند. (تمرین ۱۰، ۶ سوال ۶ فصل دهم) کتاب آقای هان

- قدرت و ضعف **K-means** را در مقایسه با **K-medoids** نشان دهید.
- قدرت و ضعف این طرح ها را در مقایسه با یک طرح خوشه بندی سلسله مراتبی (مانند **AGNES**) نشان دهید.

سوالات عملی

سوال ۵- دیتاست حملات قلبی بیماران (**heart_diagnose.csv**) یک بیمارستان در اختیار شما قرار گرفته است. در این دیتاست کوشیده شده است که احتمال حمله قلبی براساس سایر فیلدها بیمار پیش بینی شود.

- ۱- گزارشی را از دیتاست بگیرید. (شامل اسم ستون، تعداد فیلد غیر **null** و تایپ و ...)
- ۲- فیلدهای دسته ای را به روش **one-hot** به فیلدهایی عددی تبدیل کنید.
- ۳- به کمک تابع **pca** از کتابخانه **sklearn** تعداد فیچر ها را به دو فیچر کاهش دهید، آن را در یک دیتافریم مجزا ذخیره و در ادامه کار استفاده کنید.
- ۴- برای مقادیر k از ۱ تا ۲۰ الگوریتم **K-means** را اجرا کنید و نمودار **SSE** بر حسب k را رسم کنید. بر اساس روش **elbow** بگویید که کدام k مناسب تر است.
- ۵- برای مقادیر ۲ تا ۲۰ معیار **silhouette** را با روش **K-means** به دست آورید. نمودار میله ای معیار **silhouette** بر اساس k رسم کنید. همچنین بهترین k در این مرحله را بدست آورید.
- ۶- برای بهترین k که در قسمت های ۴ و ۵ بدست آورده اید خوشه بندی را انجام داده و نتیجه را به کمک **scatter plot** نشان دهید.
- ۷- برای هریک از خوشه های بدست آمده با بهترین مقدار k در قسمت ۶، تابع **describe** را فراخوانی و آن را توصیف کنید.

سوال ۶- با استفاده از مجموعه داده **BankNote** به سوالات زیر پاسخ دهید:

- ۱- مشابه سوال قبل قسمت های ۱ و ۳ را برای این دیتاست انجام دهید.
- ۲- به روش ward، نمودار Dendrogram داده ها را رسم کنید و تعداد کلاستر مناسب را بدست آورید. (برای تمرین میتوانید درمورد دو روش دیگر single, complete نیز مطالعه نموده، آن را اعمال و نتایج آن را مقایسه کنید)
- ۳- برای مقادیر ۲ تا ۲۰ خوشه بندی به روش Agglomerative را انجام دهید، معیار silhouette را برای هر خوشه بدست آورده و نمودار میله ای آن را رسم کنید. در نهایت بهترین مقدار k برای خوشه بندی بدست آورید.
- ۴- برای بهترین k که در قسمت قبل بدست آوردید نمودار scatter را رسم کنید.
- ۵- هریک از خوشه های بدست آمده را با تابع describe توصیف کرده و با نتایج قسمت قبل مقایسه کنید. (صرفا نتیجه بدست آمده در تعداد کلاستر مورد نیاز و نمودار های scatter). (همچنین برای تمرین میتوانید قسمت های ۳ و ۴ را بر مبنای معیار elbow انجام دهید و نتایج را مقایسه کنید).
- ۶- این بار از الگوریتم KElbowVisualizer استفاده کرده (واقع در کتابخانه yellowbrick.cluster) و نمودار معیار elbow بر حسب k را رسم کنید. مقدار بهترین k را بدست آورید. نتیجه خوشه بندی را به صورت scatter plot نشان دهید (راهنمایی: معیار elbow را براساس مدلی که از مدل Agglomerative نشان دهید. برای تمرین میتوانید همین معیار را برای الگوریتم K-means انجام دهید و خروجی نمودار های ElbowVisualizer را مقایسه کنید).
- ۷- یکی از روش های ارزیابی دقت کلاسترینگ ارزیابی به روش inertia_ است. برای تعداد کلاستر بین ۱ تا ۵ (به روش K-means) این معیار را محاسبه و سپس آن را روی یک نمودار scatter نشان دهید.
- ۸- این بار دیتاست اولیه را براساس tsne نرمالایز کرده و آن را در دو بعد نشان دهید.
- ۹- از الگوریتم DBScan استفاده کنید و نتیجه خوشه بندی را نشان دهید. می توانید الگوریتم را با پارامتر های مختلف انجام دهید (مثال: $eps=1, min_samples=4$ گرچه که نتیجه خیلی خوبی را به همراه ندارد). تاثیر مقادیر مختلف eps را بررسی کنید و درباره آن توضیح دهید. خروجی خوشه بندی (لیبل بندی داده) را به وسیله نمودار scatter plot نشان دهید.
- ۱۰- 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 بنزید (لزومی ندارد که تفکیک دقیقی بین کلاستر های موجود صورت بگیرد، صرفا نحوه کار با این کتابخانه مد نظر است).

نحوه تحویل: سوالات تئوری را به صورت تایپ شده و در قالب یک فایل PDF تحویل دهید. به علاوه هر یک از سوالات عملی را در قالب یک فایل ipynb به همراه نتایج قرار داده و فایل را به صورت Qn نام گذاری نمایید که n شماره سوال مربوطه می باشد. در انتها فایل های پایتون را به همراه فایل PDF تماما در قالب یک فایل zip نامگذاری شده به صورت NAME_STUDENTID در سامانه درس بارگذاری کنید. برای سوالات عملی توضیحات خود را به صورت Markdown در فایل پایتون بنویسید.

"If people valued life and love more than gold and power, the world would be a better place." - Therin Oakensfield