به نام خدا

تمرین ششم داده کاوی

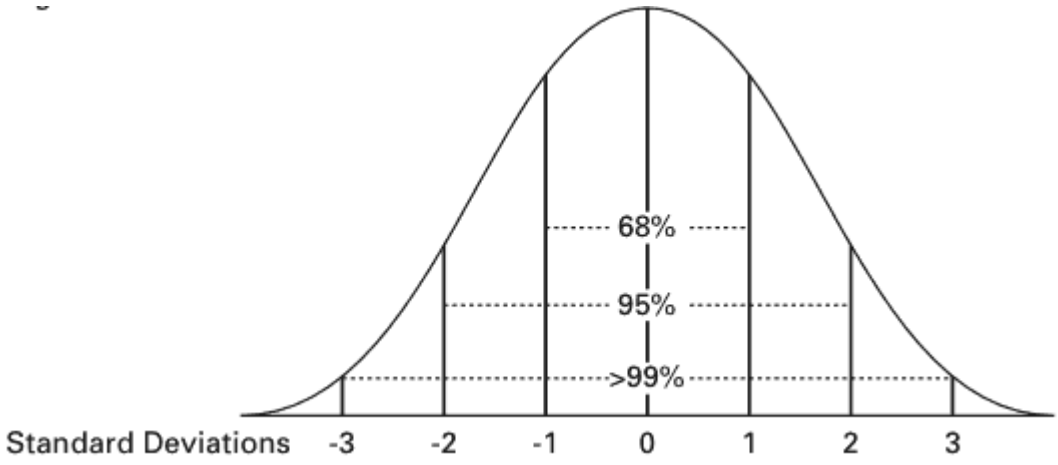
مهلت تحویل: ۲۰ تیر

هدف از این تمرین آشنایی با روش‌ها و الگوریتم‌های تحلیل داده‌ها در حوزه‌ی Unsupervised Learning است. ملاک اصلی امتیازدهی به این تمرین گزارش و پاسخ به سوالات هر بخش به صورت دقیق و کوتاه است. هرچند پیاده‌سازی هر بخش الزامیست و در صورت عدم پیاده‌سازی نمره‌ای به گزارش یا پاسخ سوالات تعلق نمی‌گیرد.

بخش اول: تشخیص داده‌های پرت(۵۰ نمره)

یافتن داده‌هایی که از توزیع کلی مجموعه داده‌ها پیروی نمی‌کنند، یکی از مباحث حوزه‌ی یادگیری بدون نظارت است. حذف کردن داده‌های پرت عموما موجب بهبود دقت و عملکرد الگوریتم‌ها می‌شود و می‌توان به آن به عنوان یکی از ابزارهای کلیدی پیش‌پردازش نیز نگاه کرد. هدف این بخش آشنایی و پیاده‌سازی یکی از شهودی‌ترین روش‌های یافتن داده‌های پرت یعنی روش مبتنی بر robust estimator of covariance است.

در این روش برای داده‌ها توزیع مشخصی(عموما توزیع نرمال گاوسی) در نظر گرفته می‌شود و بنابراین داده‌ی پرت به صورت داده‌ای که با احتمال کمتری از توزیع پیروی می‌کند شناخته می‌شود.



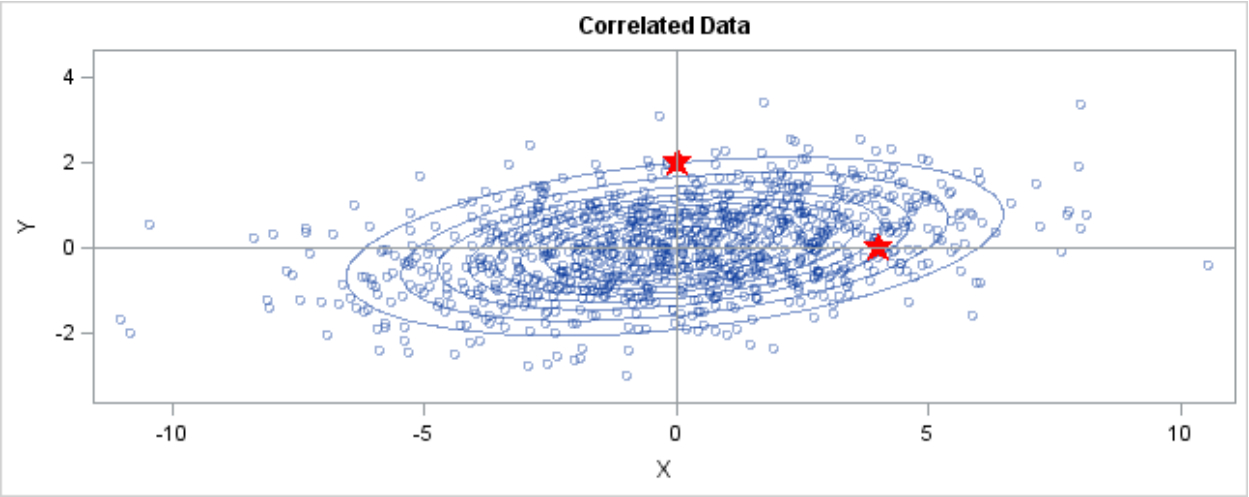
مطابق شکل اگر توزیع داده‌ها به صورت گاوسی باشد، احتمال حضور داده در هر ناحیه با توجه به مقدار میانگین و واریانس مشخص می‌شود. بنابراین برای تشخصی پرت بودن یا نبودن داده‌ها نیاز به مقایسه‌ی آن با توزیع کلی داده‌ها هستیم. برای درک بهتر این روش به سوالات زیر پاسخ دهید:

۱) فاصله‌ی ماهالانوبیس[1] چیست و چه تفاوتی با فاصله‌ی اقلیدسی دارد؟

[1] Mahalanobis Distance

۲) برای محاسبه‌ی این فاصله در یک فضای n بعدی، نیازمند محاسبه‌ی چه پارامترهایی هستیم؟

۳) با توجه به روش گفته شده و فاصله‌ی ماهالانوبیس توضیح دهید کدام یک از نقاط زیر کاندیدای مناسب‌تری برای معرفی به عنوان داده‌ی پرت هستند؟ مرکز داده‌ها (0,0) و دو نقطه‌ی علامت‌گذاری شده، (4,0) و (0, 2) هستند:



۴) مقصود از Maximum Likelihood چیست؟

۵) دلیل استفاده از Robust Estimator به جای Maximum Likelihood معمولی برای یافتن داده‌های پرت چیست؟

۶) برای پیاده‌سازی الگوریتم توصیه می‌شود از بسته‌ی covariance.EllipticEnvelope در Scikit استفاده نمایید. هرچند می‌توانید از پیاده‌سازی با روش‌های مشابه به زبان‌های دیگر استفاده نمایید.

۷) با اجرای الگوریتم بر مجموعه داده و به دست آوردن برچسب هر کدام از داده‌ها(۰ به معنای داده‌ی غیرپرت و ۱ به معنای پرت، دقت الگوریتم خود را محاسبه و گزارش کنید.)

مجموعه داده‌ی مورد استفاده را از آدرس

http://odds.cs.stonybrook.edu/cardiotocogrpahy-dataset دانلود کنید.

بخش دوم: خوشه‌بندی داده‌ها و ارزیابی خوشه‌ها(۵۰ نمره)

در این بخش هدف اجرای الگوریتم‌های خوشه‌بندی و ارزیابی نتایج به دست‌آمده بر مجموعه‌ی داده‌هاست.

الف) برای اجرای روش خوشه‌بندی k-means مهمترین مسئله تعیین K مناسب است. برای این کار روش‌های گوناگونی وجود دارد. در این قسمت شما هر کدام از دو روش زیر را به کار می‌گیرید و k مناسب را انتخاب می‌کنید

و سپس مقادیر به دست آمده از دو روش را با هم مقایسه می‌کنید. بدین منظور مقدار k را از ۲ تا ۲۰ تغییر دهید و در هر مرحله مقدار معیارهای زیر را محاسبه کنید. مقادیر به دست آمده و بیشترین مقدار هر معیار و k متناظر با آن را گزارش کنید. همچنین نمودار مقدار هر معیار بر حسب k را نیز رسم نمایید. مجموعه داده‌ی استفاده شده برای بخش الف http://archive.ics.uci.edu/ml/datasets/Water+Treatment+Plant است.

۱) V-measure برای مجموعه‌ای از داده‌ها و خوشه‌های نسبت داده شده به آن‌ها چگونه محاسبه می‌شود و چه فاکتورهایی را در نظر دارد؟ با استفاده از این معیار k بهینه را پیدا کنید.
۲) ضریب Silhouette چگونه محاسبه می‌شود؟ با استفاده از این معیار k بهینه را پیدا کنید و نمودار آن را با نمودار سوال قبل مقایسه کنید.

ب) برای اجتناب از تعیین k به عنوان Hyper parameter در الگوریتم K-means می‌خواهیم از الگوریتم Agglomerative Clustering استفاده نماییم. این الگوریتم را روی مجموعه داده‌ی http://archive.ics.uci.edu/ml/datasets/HTRU2 اجرا نمایید. (برای اجرای الگوریتم ویژگی آخر که کلاس مقصد است را از ورودی الگوریتم حذف نمایید.)

۱) مقصود از استراتژی‌های Linkage در این الگوریتم چیست؟ دو استراتژی Ward و Complete را توضیح دهید و با هر کدام از این استراتژی‌ها یک بار الگوریتم را اجرا کنید.
۲) مقصود از مجموعه داده‌های دارای Ground truth برای خوشه‌بندی چیست؟
۳) معیار Normalized Mutual Information چیست؟ شباهت و تفاوت آن با Correlation چیست؟
۴) برای ارزیابی روش پیاده‌سازی شده NMI را به ازای دو استراتژی Linkage محاسبه کنید و با هم مقایسه کنید.

بخش سوم: Association Rule Mining (۵۰ امتیاز)

برای اجرای این بخش پیشنهاد می‌شود که از ابزار RapidMiner استفاده نمایید. در غیر این صورت پیاده‌سازی کلیه‌ی بخش‌ها به عهده‌ی خودتان است و هیچ امتیاز اضافه‌ای نخواهد داشت.

سناریو: شهردار یک شهر در حال توسعه قصد دارد از گروه‌ها و اصناف مختلف شهر برای همکاری در تخصیص منابع استفاده نماید. شهردار می‌خواهد با استفاده از مجموعه داده‌ی **communities** **(به همراه فایل تمرین**

قرار داده شده است) متوجه شود که بین کدام گروه‌ها به طور پیش‌فرض ارتباط و همکاری بیشتری وجود دارد و بر این اساس گروه‌های همکاری را تشکیل دهد. در واقع هدف از تحلیل این مجموعه داده یافتن گروه‌های مختلف با بیشترین سطح طبیعی ارتباط و همکاریست. بدین منظور از روش Association Rule Mining استفاده شده است.

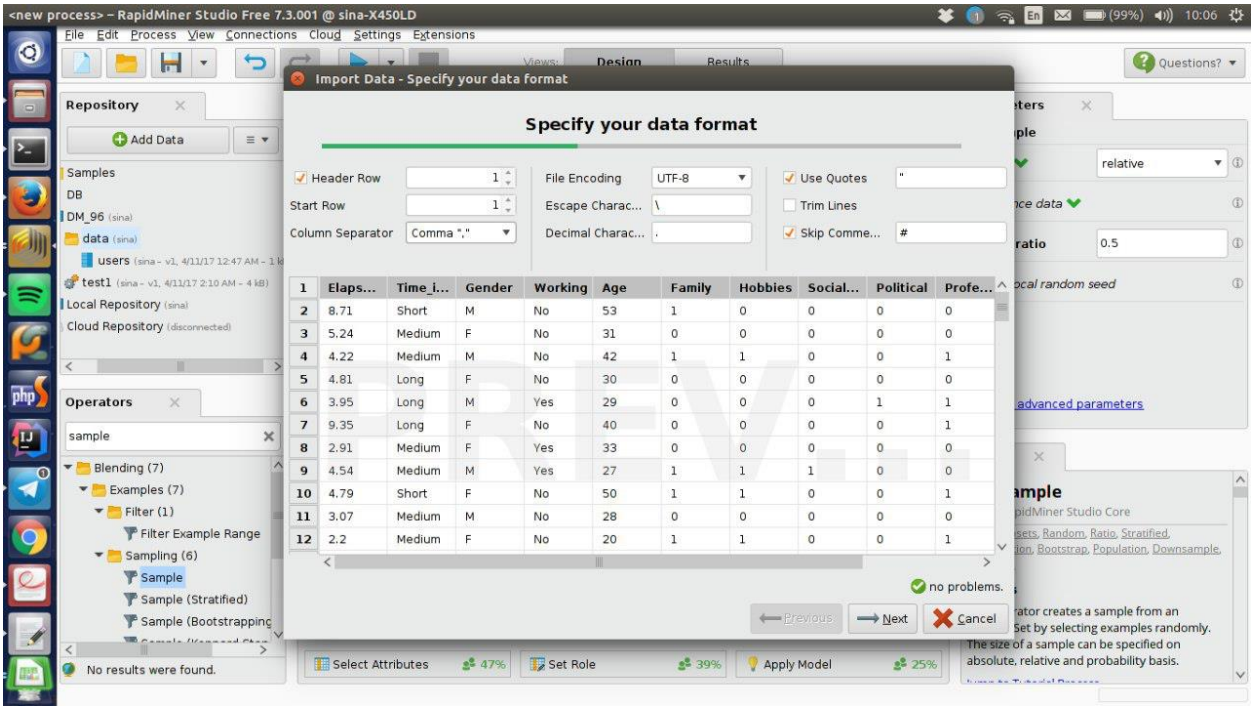
الف) توضیح مختصری در مورد هدف Association Rule Mining ارائه دهید.

ب) منظور از یک Rule که به صورت A → B نمایش داده می‌شود، چیست؟

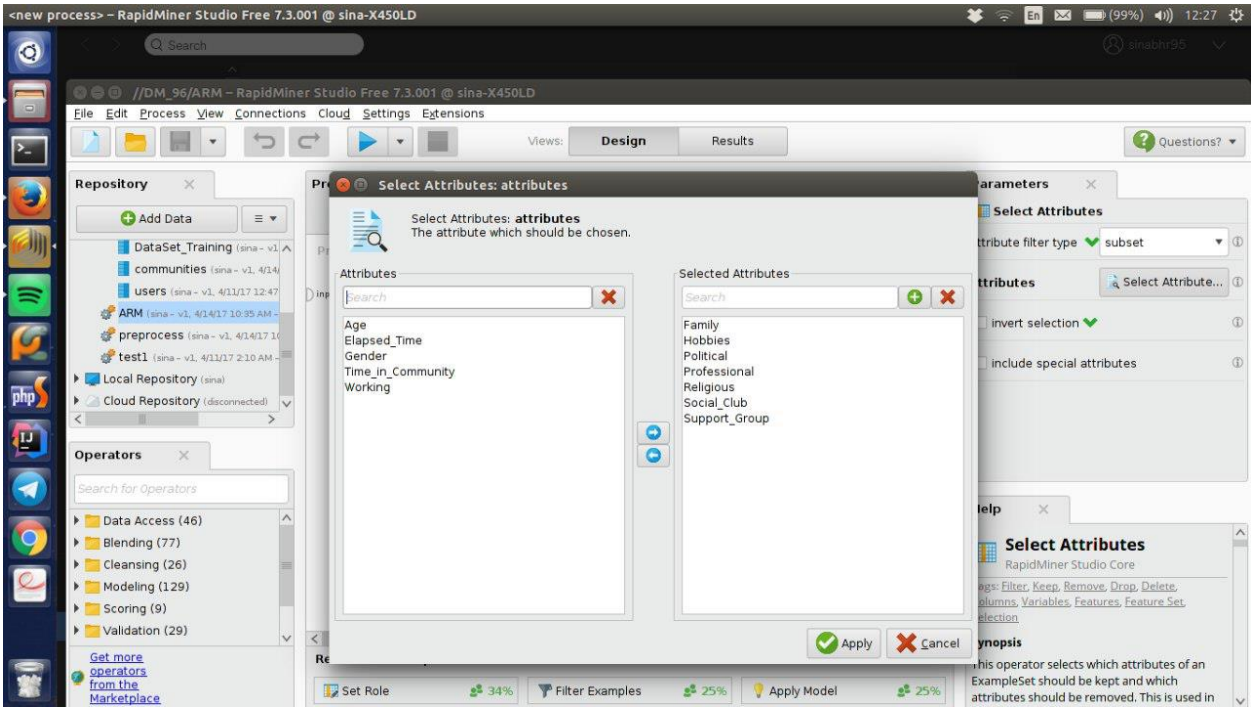
ج) منظور از Support و Confidence یک Rule چیست؟

حال برای یافتن پاسخ مناسبی برای سناریوی بالا مراحل زیر را در یک فرآیند جدید در RapidMiner اجرا کنید.

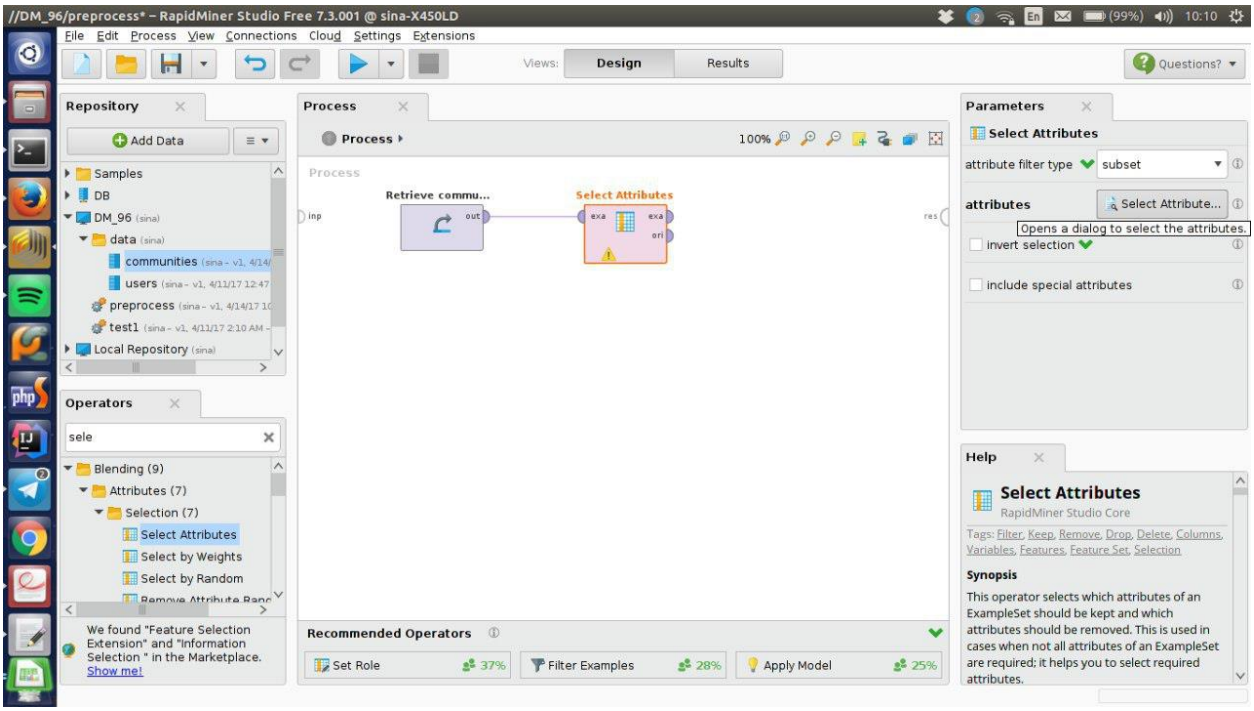
۱) ابتدا داده‌ها را مطابق شکل به بخش داده‌ی Repository خود اضافه کنید.

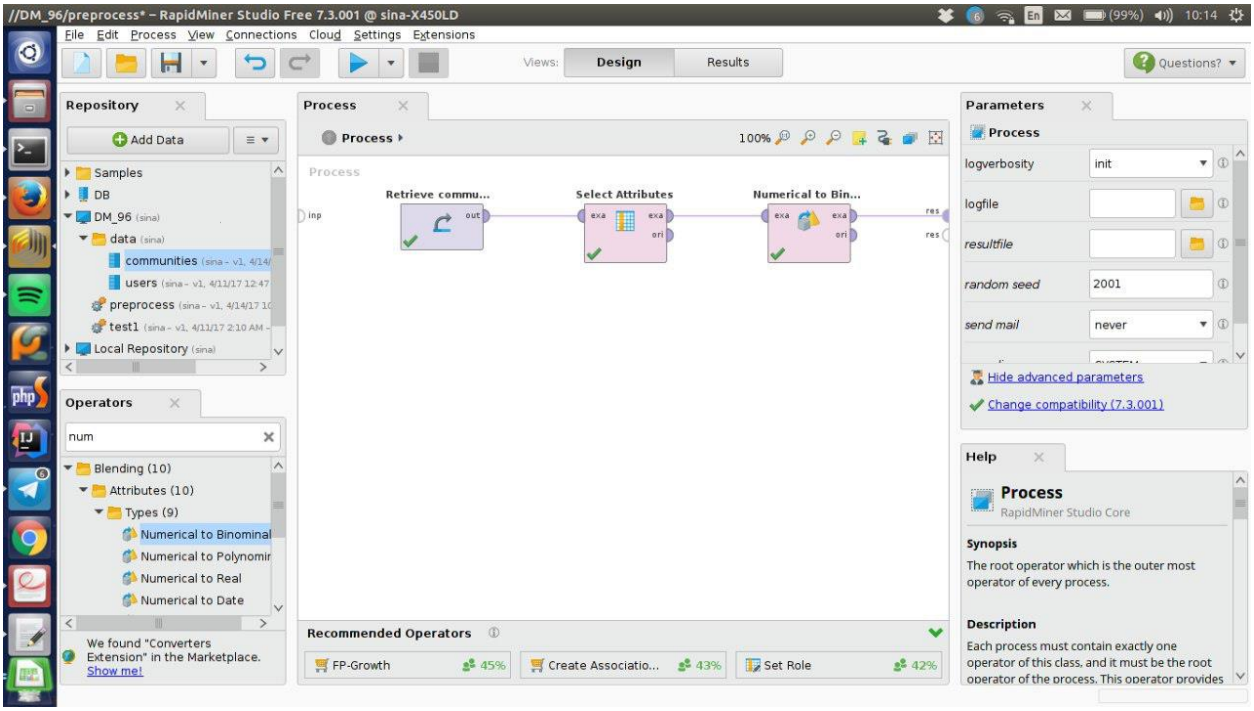


۲) با مشاهده‌ی داده‌ها متوجه می‌شوید که هیچ‌گونه مقدار از دست رفته‌ای وجود ندارد و نیازی به پیش‌پردازش‌هایی از این نوع نیست. اما قصد داریم تنها از برخی Featureها برای تحلیل خود استفاده کنیم. از این رو مطابق شکل یک اپراتور Select Attribute را به فرآیند خود اضافه کنید و فیلترتایپ subset را انتخاب کنید.



۳) برای عملکرد بهتر فرآیند ویژگی‌های دومقداری بهتر است به فرم Binominal در آیند. ابتدا توضیح دهید فرق نوع Binomial و Binominal چیست؟ سپس مطابق شکل عملگر Numerical to Binominal را انتخاب کرده و از آن برای تبدیل داده‌ها استفاده کنید.





۴) اپراتور FP-Growth را انتخاب کرده و مطابق شکل به فرآیند اضافه کنید. توضیح دهید که این اپراتور چه عملی را انجام می‌دهد.

۵) اپراتور Create Association Rules را مطابق شکل به فرآیند اضافه کنید و به ازای min confidenceهای 0.3 و 0.5 و 0.7 نتایج را بررسی کرده و گزارش کنید و بیان کنید در هر حالت کدام یک از گروه‌ها ارتباط بیشتری را با هم نشان می‌دهند.

این فرآیند را ذخیره کنید و همراه با نتایج در یک پوشه برای بخش سوم تمرین قرار دهید.

بخش چهارم: Graph Mining(۴۰ امتیاز)

در این بخش قصد یافتن خوشه‌ها یا اجتماع[2]های یک گراف را داریم. به طور عمومی یک خوشه یا اجتماع بر این اساس تعریف می‌شوند که گره‌های درون یک خوشه یا اجتماع تعداد ارتباط و یال بیشتری در داخل اجتماع خود نسبت به بیرون خود داشته باشند. برای این بخش یکی از معروفترین دیتاست‌های گرافی در نظر گرفته شده است که باید یک الگوریتم یافتن خوشه‌ها یا اجتماعات گراف را روی آن اجرا کنید.

[2] Community

۱) مجموعه داده‌ی Zachary's karate club که به فرمت GML است را دریافت کنید. این مجموعه داده ارتباطات دوستی ۳۴ نفر را در یک باشگاه کاراته نشان می‌دهد. فایل‌های GML به راحتی در اکثر زبان‌های برنامه نویسی از جمله R و Python قابل خواندن هستند.

۲) پس از خواندن فایل با یکی از ابزارهای پردازش گراف، این گراف را رسم نمایید(گره‌ها و یال‌ها). برای این کار می‌توانید از igraph package، Gephi یا هر ابزار دیگری که می‌شناسید استفاده کنید. این ابزارها امکان بارگذاری، تحلیل و نمایش گراف را به شما می‌دهند.

۳) حال یکی از الگوریتم‌های یافتن خوشه‌ها یا اجتماعات گراف را روی این مجموعه داده اجرا کنید. الگوریتم‌های Girvan-Newman و Louvain از مشهورترین این الگوریتم‌ها هستند و اکثر این بسته‌ها این الگوریتم‌ها را پیاده سازی کرده‌اند. از هر الگوریتمی که استفاده می‌کنید، شرح مختصری از نحوه‌ی عملکرد الگوریتم بدهید.

۴) پس از یافتن اجتماعات گراف، گراف را دوباره رسم کنید و گره‌های درون یک اجتماع را به یک رنگ متمایز با سایر گره‌ها در آورید.

۵) ماژولاریتی[3] یکی از مهمترین معیارهای سنجش کیفیت خوشه‌بندی است. این معیار را تعریف کنید و مقدار ماژولاریتی را برای اجتماع‌های یافته شده، به دست آورید.

تمام کدها و تصاویر را در پوشه‌ای برای بخش چهارم ذخیره کنید.

نکات

- گزارش تمرین را به صورت report6_stdNum.pdf نامگذاری کنید.
- پوشه‌ی اصلی شامل پیاده‌سازی، گزارش و پاسخ به سوالات را در یک پوشه به نام DM6_stdnum قرار داده و به صورت فشرده شده بارگذاری کنید.
- تمرین خود را قبل از زمان مشخص شده در مودل آپلود کنید.
- ۱۰۰ نمره از این تمرین به صورت اجباری و مابقی آن به عنوان نمره‌ی امتیازی درنظر گرفته می‌شود.

توجه کنید که اجرای بخش سوم الزامی است.

[3] Modularity