دانشگاه صنعتی اصفهان

دانشکده مهندسی برق و کامپیوتر

درس مبانی یادگیری ماشین

پروژه اختیاری

مهلت تحویل: ۲۲ دی ۱۴۰۲

*** در این پروژه شما مجاز به استفاده از تمامی کتابخانه‌ها هستید. این پروژه به صورت اختیاری بوده و ۱ نمره دارد .

در این پروژه می‌خواهیم داده‌های مربوط به مشتریان یک فروشگاه را به چند خوشه تقسیم کنیم. برای این کار مجموعه داده MAll_Customers در اختیار شما قرار گرفته است. در این مجموعه داده تعدادی ویژگی از افراد به دست آمده است. در این قسمت از پروژه شما باید مراحل زیر را اجرا و نتایج را گزارش کنید.

سوال ۱)

داده‌ها را با کتابخانه pandas خوانده و ۸۰ درصد را برای آموزش و ۲۰ درصد را برای آزمون در نظر بگیرید. (۳ نمره)

سوال ۲)

ستون CustomerID را از دیتافریم به دست آمده حذف کنید. (۲ نمره)

سوال ۳)

تعداد، میانگین، انحراف معیار، کوچکترین،بزرگترین، چارک اول، چارک دوم و چارک سوم داده‌های هر ستون به غیر از ستون Gender را گزارش کنید. (۱۰ نمره)

سوال ۴)

به کمک StandardScaler از کتابخانه sklearn.preprocessing داده‌ها را نرمال کنید. (به غیر از ستون جنسیت) (۵ نمره)

سوال ۵)

مقادیر Female از ستون Gender را به ۱ و سایر مقادیر را به ۰ تبدیل کنید. (۵ نمره)

سوال ۶)

تابعی بنویسید که ورودی‌های آن k و داده‌های یک دیتافریم باشد، و به کمک الگوریتم kmeans داده‌ها را به k خوشه تقسیم کند. این تابع باید موارد زیر را در خروجی مشخص کند: (۲۰ نمره)

الف) مراکز خوشه‌ها

ب) خوشه مربوط به هر نمونه داده

سوال ۷)

به کمک روش Elbow و silhouette برای مقادیر k بین ۲ تا ۱۰ تعداد خوشه مناسب را پیدا کرده و دلایل خود را ذکر کنید. (به دلیل جلوگیری از حساسیت به نقطه شروع باستی الگوریتم k_means را برای هر k چندین بار اجرا گردد و هر کدام که کمترین Loss را دارد، انتخاب شود.) (۲۰ نمره)

سوال ۸)

با استفاده از مدل به دست آمده برای گروه‌بندی داده‌ها، خوشه‌ی مربوط به هر داده را روی مجموعه داده تست پیش‌بینی کنید. این مدل هر نمونه داده جدید را به نزدیکترین مرکز خوشه به آن منتسب می‌کند. (گروه‌بندی به وسیله فراخوانی تابع predict از مدل، قابل انجام است.) (۳ نمره)

سوال ۹)

داده‌های پیش‌بینی‌شده به همراه شماره‌ی خوشه ی پیش‌بینی‌شده را در یک فایل CSV با نام predicts.csv ذخیره کنید. (۲ نمره)

سوال ۱۰)

با استفاده از روش تحلیل مؤلفه اصلی (PCA) برای کاهش بعد، ابعاد داده های آموزشی و آزمایشی را به دو ویژگی با نام‌های PCA1, PCA2 کاهش دهید. (۵ نمره)

سوال ۱۱)

k_means را با بهترین k به دست آمده روی داده‌های آموزشی کاهش بعد یافته اجرا کرده و سپس خوشه مربوط به داده‌های آزمایشی کاهش بعد یافته را با مدل گروه‌بندی به‌دست‌آمده . تعیین کنید. نتایج این خوشه‌بندی را به صورت یک ستون مجزا به فایل predicts.csv اضافه کنید. (۱۰ نمره)

سوال ۱۲)

نمودار پراکندگی [1] داده های کاهش بعد یافته را با رنگ بندی متفاوت برای هر خوشه نمایش دهید.

الف) داده های آموزشی را با دایره نمایش دهید.

ب) داده های آزمایشی را با مثلث نمایش دهید.

ج) مرکز هر خوشه را با علامت "+"نمایش دهید.

[1] Scatter plot

د) برای نمایش بهتر داده هایی که برروی هم افتاده اند از معیار آلفا)نمایانگر transparency) .در کشیدن نمودار استفاده کنید (۱۵ نمره)

نکات تکمیلی

۱. برای انجام این تکلیف استفاده از زبان پایتون الزامی است.

۲. تکالیف را در محیط jupyter notebook پیاده‌سازی کنید و فایل ipynb. را ارسال کنید.

۳. توضیح کدی که نوشته‌اید، بررسی و تحلیل نتایج آن و بیان علت نتایج و نیز مقایسه نتیجه با آنچه مورد انتظارتان بوده است، از اهمیت بالایی برخوردار است. شما می‌توانید گزارش پروژه را در همان محیط jupyter notebook بنویسید و نیازی به فایل pdf جداگانه نیست. هم‌چنین اگر برای حل سوال فرضیات خاصی مدنظر دارید حتما آن را در متن گزارش قید کنید.

۴. فرمت نامگذاری تکلیف ارسالی باید به صورت زیر باشد: HWX_Programming_LastName_StudentID که X شماره تکلیف LastName نام خانوادگی شما و StudentID شماره دانشجویی شما است.

۵. انجام این تکلیف به صورت تک نفره است. در صورت مشاهده تقلب، نمرات هم مبدا کپی و هم مقصد آن صفر لحاظ می‌شود.

۶. شما می‌توانید تا یک هفته پس از پایان مهلت تکلیف آن را در یکتا بارگذاری کنید. در این صورت به ازای هر روز تاخیر ۵ درصد از نمره تکلیف کسر می‌شود. پس از اتمام این یک هفته امکان ارسال با تاخیر وجود ندارد.

۷. در صورت وجود هر گونه ابهام و یا سوال می‌توانید سوالات خود را در گروه تلگرام بپرسید. هم چنین می‌توانید برای رفع ابهامات با دستیاران آموزشی از طریق تلگرام در تماس باشید.