به نام خدا

دانشکده مهندسی کامپیوتر
دانشگاه اصفهان

پروژه دوم درس داده‌کاوی

نام استاد :دکتر آزاده محمدی
دستیاران آموزشی : نرگس اسدی – مهسا علی‌پور

تاریخ تحویل : ۱۵ دی ۱۴۰۰

نکات

- در صورت مشاهده‌ی هر گونه تقلب یا تشخیص کپی از اینترنت، نمره صفر برای طرفین در نظر گرفته می‌شود.
- همراه با کد ارسالی یک گزارش از نتایج به دست آمده به صورت PDF ارسال نمایید.

معرفی

در این پروژه درباره‌ی نرم‌افزار رپیدماینر و همچنین الگوریتم‌های خوشه‌بندی بیشتر خواهید آموخت.

مراحل پروژه

بخش اول

۱-۱- مجموعه داده Absenteeism_at_work که همراه با این فایل ارسال شده‌است مربوط به اطلاعات غیبت افراد در سرکار می‌باشد. مفاهیم ستون‌های این مجموعه داده را می‌توانید در این لینک[۱]مشاهده کنید.
۱-۲- مجموعه داده را به نرم‌افزار رپیدماینر وارد کنید و به تحلیل و پیش‌پردازش آن بپردازید (برای انجام تحلیل از نمودارهای توزیع داده‌ها استفاده کنید و در گزارش ارسالی، تحلیل‌ها، نمودارها و دلایل انجام هر نوع پیش پردازش را شرح دهید).
سپس مجموعه داده تمیز شده را از رپیدماینر دانلود و بقیه مراحل را با برنامه نویسی انجام دهید.
۱-۳- یک الگوریتم خوشه‌بندی را با استفاده از کتابخانه‌های یادگیری ماشین بر روی این مجموعه داده اعمال کنید.

بخش دوم

۲-۱- الگوریتم k-means را پیاده‌سازی کنید. (از بسته‌های نرم‌افزاری آماده استفاده نکنید.)
۲-۲- مجموعه داده Stars که همراه این فایل ارسال شده است شامل اطلاعات حدود ۲۰۰ ستاره و ویژگی‌های آن‌ها است. کلاس و دسته‌ی صحیح هر ستاره در ستون type با اعداد ۰ تا ۵ مشخص شده است که فعلا برای فرآیند خوشه‌بندی با این ستون کاری نداریم اما هنگام ارزیابی برای ما مهم است. ۴ ستون اول، مقادیر

[۱] https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work

عددی پیوسته هستند که به ترتیب دما، درخشش، شعاع و قدر مطلق ستاره است. دو ستون بعدی، رنگ و نوع خاص ستاره است که هر دو کمیت‌هایی nominal هستند. حال می‌خواهیم این داده‌ها را بدون استفاده و توجه به ستون type خوشه‌بندی نمایید.

الگوریتم پیاده‌سازی شده را با تعداد خوشه ۶ بر روی مجموعه داده Stars اجرا کنید.

۲-۳- دقت خوشه‌بندی خود را محاسبه کنید.

نحوه محاسبه دقت خوشه‌بندی: راه‌های زیادی برای ارزیابی دقت خوشه‌بندی وجود دارد. برای این مساله، بعد از خوشه‌بندی، ببینید برچسب (type) اکثریت داده‌های موجود در یک خوشه چیست. سپس برچسب تمام داده‌های آن خوشه را همین برچسب اکثریت در نظر بگیرید. این کار را برای تمام خوشه‌ها انجام دهید. لذا تمام داده‌های خوشه‌بندی شده، یک برچسب خوشه دارد. در نهایت ببینید چند درصد از کل داده‌ها برچسب صحیح دریافت نموده‌اند.