به نام خدا

هیدروانفورماتیک | نیم‌سال اول سال تحصیلی ۰۳-۱۴۰۲

تمرین پایان ترم | تاریخ ارسال: ۱۴۰۲/۱۰/۱۹ | موعد تحویل پاسخ: ۱۴۰۲/۱۱/۱۹

داده‌های مربوط به اندازه‌گیری متغیرهای مختلف کیفیت آب در نمونه‌های برداشت‌شده از تعدادی ایستگاه هیدرومتری در فایل‌های پیوست با پسوند csv. ارائه شده‌اند. فایل داده‌هایی که توسط هر یک از دانشجویان باید مورد استفاده قرار گیرد با شماره‌ی دانشجویی وی مشخص شده است. ستون متغیر رسته‌ای class در گام‌های (الف) تا (ه) تمرین حاضر کاربردی ندارد و نباید در این گام‌ها اثرگذار باشد.

گام‌های این تمرین به‌شرح زیر هستند:

الف) با استفاده از روش استانداردسازی، داده‌ها را پیش‌پردازش کنید. (۱ امتیاز)

ب) خوشه‌بندی سلسله‌مراتبی را روی داده‌های استانداردشده اجرا کنید. (۴ امتیاز)

در اجرای خوشه‌بندی سلسله‌مراتبی:

۱. از ستون آخر داده‌ها که مربوط به متغیر دبی جریان است در خوشه‌بندی استفاده نکنید. (دبی جریان جزء مؤلفه‌های بردار ویژگی‌های مورد استفاده در خوشه‌بندی نباشد.)

۲. برای سنجش فواصل نقاط از رابطه‌ی فاصله‌ی اقلیدسی استفاده کنید.

۳. از بین روش‌های اجرای خوشه‌بندی سلسله‌مراتبی single linkage، complete linkage، average linkage و Ward، حداقل یک روش را برای اجرای خوشه‌بندی انتخاب کنید. استفاده از بیش از یک روش در صورت ارائه و تحلیل مناسب نتایج، دارای امتیاز اضافه خواهد بود.

۴. خوشه‌بندی‌ها را به‌ازای تعداد خوشه‌ها از ۲ تا ۵ اجرا کنید. اجرای خوشه‌بندی با تعداد خوشه‌های بیشتر، در صورت ارائه و بررسی نتایج، دارای امتیاز اضافه خواهد بود.

۵. نتیجه‌ی اصلی که در مورد هر خوشه‌بندی باید ارائه شود، چگونگی اختصاص نقطه‌داده‌ها یا ایستگاه‌ها به خوشه‌ها است. (یعنی مشخص شود در هر خوشه‌بندی کدام ایستگاه‌ها به‌عضویت کدام خوشه‌ها درآمده‌اند.)

ج) خوشه‌بندی افرازی را با استفاده از الگوریتم K-means روی داده‌های استانداردشده اجرا کنید. (۴ امتیاز)

در اجرای خوشه‌بندی افرازی:

۱. از ستون آخر داده‌ها که مربوط به متغیر دبی جریان است در خوشه‌بندی استفاده نکنید. (دبی جریان جزء مؤلفه‌های بردار ویژگی‌های مورد استفاده در خوشه‌بندی نباشد.)

۲. برای سنجش فواصل نقاط از رابطه‌ی فاصله‌ی اقلیدسی استفاده کنید.

۳. تعیین مقدار آرگومان‌های تابع اجرای الگوریتم مانند چگونگی تعیین مراکز اولیه و تعداد تکرار محاسبات الگوریتم و توجیه دلیل انتخاب آن‌ها بر عهده‌ی شما است.

۴. خوشه‌بندی‌ها را به‌ازای تعداد خوشه‌ها از ۲ تا ۵ اجرا کنید. اجرای خوشه‌بندی با تعداد خوشه‌های بیشتر، در صورت ارائه و بررسی نتایج، دارای امتیاز اضافه خواهد بود.

۵. نتیجه‌ی اصلی که در مورد هر خوشه‌بندی باید ارائه شود، چگونگی اختصاص نقطه‌داده‌ها یا ایستگاه‌ها به خوشه‌ها است. (یعنی مشخص شود در هر خوشه‌بندی کدام ایستگاه‌ها به‌عضویت کدام خوشه‌ها درآمده‌اند.)

د) شاخص صحت خوشه‌ی میانگین silhouette witdth را برای تمام خوشه‌بندی‌های اجراشده محاسبه کنید و بر اساس آن کیفیت خوشه‌بندی‌های اجراشده را با هم مقایسه کنید. محاسبه‌ی سایر شاخص‌های صحت خوشه و ارائه و تحلیل نتایج مربوط به آن‌ها دارای امتیاز اضافه خواهد بود. اجرای این بند اجباری نیست و اجرای آن دارای امتیاز اضافه خواهد بود.

ه) دو مقدار مفقود (missing value) موجود در داده‌های دبی جریان را با استفاده از الگوریتم kNN محاسبه کنید. (۴ امتیاز)

در اجرای الگوریتم kNN:

۱. روشن است که دبی جریان متغیر هدف است و داده‌های مربوط به آن باید مورد استفاده قرار گیرند. در مورد این متغیر از داده‌های استانداردشده استفاده نکنید. چنانچه از داده‌های استانداردشده استفاده

کردید، لازم است پس از محاسبه‌ی مقادیر متغیر هدف با استفاده از الگوریتم kNN، تبدیل استانداردسازی را به‌صورت معکوس اجرا کنید تا مقادیر دبی جریان حاصل شود.

۲. در مورد داده‌های مربوط به متغیرها یا ویژگی‌های غیر از دبی جریان از داده‌های استانداردشده استفاده کنید.

۳. مقدار k را از ۱ تا ۵ تغییر دهید. اجرای الگوریتم به‌ازای سایر مقادیر k و ارائه و تحلیل نتایج دارای امتیاز اضافه خواهد بود.

۴. نتیجه‌ی این بخش باید شامل مقادیر برآوردشده برای دو داده‌ی مفقود دبی به‌ازای هر یک از مقادیر k همسایگی باشد.

و) فرض کنید مقدار متغیر رسته‌ای class را برای سه ایستگاه ۰۲۰-۲۱، ۲۴۳-۲۱ و ۳۰۷-۲۱ نامشخص است. مقدار (برچسب) این متغیر را برای این سه ایستگاه با استفاده از الگوریتم kNN مشخص کنید. (۴ امتیاز)

در اجرای الگوریتم kNN:

۱. روشن است که متغیر class متغیر هدف است یا همان برچسب است.

۲. در این بخش از داده‌های دبی جریان استفاده نکنید. (این متغیر را از مجموعه‌داده کنار بگذارید.)

۳. در مورد داده‌های مربوط به متغیرهای پیش‌بینی‌کننده از داده‌های استانداردشده استفاده کنید.

۴. مقدار k را از ۱ تا ۵ تغییر دهید. اجرای الگوریتم به‌ازای سایر مقادیر k و ارائه و تحلیل نتایج دارای امتیاز اضافه خواهد بود.

۵. نتیجه‌ی این بخش باید شامل طبقه یا برچسب class برای هر یک از سه ایستگاه ۰۲۰-۲۱، ۲۴۳-۲۱ و ۳۰۷-۲۱ به‌ازای هر یک از مقادیر k همسایگی باشد.

ز) با استفاده از روش انتخاب ویژگی Exhaustive دو ویژگی را که ترکیب آن‌ها بهترین ترکیب دوتایی برای اجرای طبقه‌بندی داده‌ها (مشخص کردن برچسب class) با استفاده از الگوریتم kNN با تعداد ۳ نزدیک‌ترین

همسایه است را مشخص کنید. برای تعیین بهترین ترکیب رویکرد مدل‌مبنا (model-based) را به‌کار بگیرید و تشخیص صحیح برچسب تمام ایستگاه‌ها را مدنظر قرار دهید و نه فقط سه ایستگاه مطرح‌شده در بند پیشین را. در این بخش نیز داده‌های دبی جریان را لحاظ نکنید. (۳ امتیاز)

نکات مربوط به نحوه‌ی ارائه‌ی پاسخ تمرین

- پاسخ تمرین شامل فایل (یا فایل‌های) برنامه‌ی (یا برنامه‌های) نوشته‌شده به زبان Python (۱۲ نمره شامل ۱۰ نمره برای صحت کدها و نتایج و ۲ نمره برای کیفیت کدنویسی)، یک گزارش متنی (فایل با قالب doc. یا docx.) حداکثر در ۱۰ صفحه (۴ نمره شامل ۲ نمره برای صحت نتایج و ۲ نمره برای کیفیت گزارش) و یک فایل ارائه (با قالب ppt. یا pptx.) (۴ نمره شامل ۲ نمره برای صحت نتایج و ۲ نمره برای کیفیت ارائه) است. شایان ذکر است که هیچ بخشی به‌تنهایی دارای امتیاز نیست. زمان برگزاری جلسه‌ی ارائه نیز متعاقباً تعیین و اعلام خواهد شد.
- گزارش متنی، غیر از صفحه‌ی عناوین و مشخصات باید حداقل شامل سه بخش بیان مسأله، مواد و روش‌ها (شامل معرفی داده‌ها و روش حل مسأله) و نتایج و بحث (شامل ارائه‌ی نتایج و بحث و تحلیل روی آن‌ها) باشد.
- سعی کنید برای ارائه‌ی نتایج و تحلیل آن‌ها حتی‌الامکان از نمودارها و جداول استفاده کنید.
- برای ترسیم نمودارها از زبان و محیط نرم‌افزاری Python استفاده کنید.
- تعداد اسلایدها و نحوه‌ی ارائه را برای زمان ۱۲ تا ۱۵ دقیقه تنظیم کنید.
- مجموعه‌ی فایل‌های مذکور را در قالب یک پرونده‌ی فشرده با قالب zip. یا rar. با شماره‌ی دانشجویی خود نام‌گذاری نمایید و آن را از طریق سامانه‌ی مدیریت یادگیری الکترونیک دانشگاه (https://lms2.sbu.ac.ir/) ارسال کنید.

نکته‌ی مربوط به بخش‌های دارای امتیاز اضافه

امتیازات مربوط به بخش‌هایی که دارای امتیاز اضافه هستند فقط می‌توانند کمبود امتیازات در سایر بخش‌های تمرین را جبران کنند و در صورتی که مجموع امتیازات به سقف امتیاز این تمرین برسد، امتیازات اضافه به نمره‌ی سایر مؤلفه‌های ارزیابی درس مانند آزمون کتبی یا پروژه‌های دیگر انتقال نخواهند یافت.

سربلند و پیروز باشید!

آهنی