به نام خدا

تکلیف پنجم درس مبانی داده کاوی

ترم بهار ۱۴۰۰

راهنمایی :
زبان برنامه نویسی سوالات پایتون است.
پیشنهاد می شود از محیط Jupyter notebook استفاده کنید.
پکیج‌های اصلی استفاده شده seaborn، numpy, pandas, sklearn می باشند.
دیتاست های مورد نیاز در ادامه معرفی شده اند.
روش تحویل:
الف) فایلهای مربوط به کدهای هر سوال در یک فایل با نام Qx.zip که x شماره سوال است زیپ شوند (برای نمایش خروجی دستورات، هرجا مقدور است نام دیتافریم را بزنید تا خلاصه آن را نشان دهد و در سایر حالات از دستور head استفاده کنید)، سپس کلیه این فایلهای زیپ در یک فایل واحد با نام HW5-Lastname-StudentCode.zip که Lastname نام خانوادگی و StudentCode شماره دانشجویی شما است، زیپ شده و روی سامانه تا **زمان مشخص شده** آپلود شوند.
ب) گزارش نهایی باید شامل پاسخ تمامی سوالات (سوالات تحقیقی و سوالات پیاده سازی) باشد که برای سوالات پیاده سازی شامل کد نوشته‌شده، توضیحی درمورد کد و نتیجه اجرا و تفسیر نتیجه می‌باشد.
ج) زمان و نحوه تحویل تکلیف در فایل راهنمای ترم مشخص شده است.
د) تحویل خارج سامانه و خارج ساعت مشخص شده قابل قبول نیست.

نکته ۱ : برای پاسخ به سوالات **تحقیقی** و **تفسیری**، پس از مطالعه منابع موردنیاز فقط **برداشت خود** از مسئله را توضیح دهید.

۱. شبکه عصبی

a) فایل csv دیتاست Flower را خوانده و تبدیل به دیتافریم نمایید.

b) دیتافریم موجود را از نظر داده‌های گم شده (Missing Value)، نرمالایز داده‌های عددی و همچنین encode کردن ستون‌های دسته‌ای بررسی کرده و کارهای لازم برای آماده سازی دیتافریم به منظور ایجاد مدل را انجام دهید.

c) با استفاده از نمودار scatter داده های این دیتاست را از نظر پراکندگی دسته ها نمایش دهید.

d) مقادیر همه ستون ها به جز ستون Class را در متغیر x قرار داده و ستون Class را در متغیر y قرار دهید.

e) مجموعه‌های آموزشی و تست را با نسبت ۰.۸ به ۰.۲ ایجاد کنید و توزیع دسته‌های ستون Class را در دو مجموعه نمایش دهید.

f) با استفاده از کلاس MLPClassifier مدل دسته‌بندی موردنظر خود را ایجاد نمایید.

g) داده های تست را به مدل بدهید و میزان دقت مدل را نمایش دهید.

h) با استفاده از متد plot_prediction موجود در فایل utils، نمودار مقادیر پیش‌بینی شده را نمایش دهید و نمودار را تفسیر نمایید. (راهنمایی: نمونه اجرا plot_prediction(lambda x: mlp.predict(x), X_train, y_train))

i) با استفاده از پارامترهای hidden_layer_sizes، activation، max_iter مدل خود را بهینه کنید تا بهترین دقت را بدست آورید. کدام پارامتر بیشترین تاثیر را روی دقت مدل داشته است؟

۲. خوشه بندی (Clustering)

a) در این سئوال از دیتاست Banknote استفاده می شود. این دیتاست را load کنید. می خواهیم با استفاده از الگوریتم k-mean تعداد دسته‌ها را مشخص کنیم.

b) دیتافریم موجود را از نظر داده‌های گم شده (Missing Value)، نرمالایز داده‌های عددی و همچنین اینکد کردن ستون‌های دسته‌ای بررسی کرده و کارهای لازم برای آماده سازی دیتافریم به منظور ایجاد مدل را انجام دهید.

c) ابتدا تعداد کلاستر ها را ۲ در نظر بگیرید و داده های آموزشی را به آن fit کنید و برای نمایش برچسب ها از متد predict استفاده کنید.

d) مراکز خوشه را در متغیری به نام centroids قرار دهید.

e) یک scatter plot با استفاده از داده ها ایجاد کنید طوریکه برچسب های مربوط به دسته های مختلف را با رنگ های مختلف نشان دهد. مراکز خوشه ها را با علامت ضربدر نشان دهید.

f) در مورد پارامتر algorithm در کلاس KMeans تحقیق کنید و انواع الگوریتم‌های موجود برای این فیلد را معرفی نمایید.

g) یکی از روشهای ارزیابی دقت کلاسترینگ استفاده از متد inertia_ (اینرسی) است . مقدار آن را برای کلاسترینگ فعلی نشان دهید.

h) یک حلقه بنویسید که تعداد خوشه ها را از ۱ تا ۵ افزایش دهد و هر بار k-mean را انجام دهد و مقدار inertia را بدست آورد. نتایج هر مرحله را در یک لیست اضافه کنید و در نهایت لیست را نشان دهید.

i) لیست مربوط به مقادیر اینرسی بدست آمده در قسمت قبل را روی نمودار خطی نشان دهید و آن را تفسیر کنید. در چه مرحله ای بیشترین تغییر در مقدار اینرسی دیده شده است و از نظر شما بهترین تعداد خوشه برای این دیتاست چند است؟

j) قسمتهای g تا i را این بار با استفاده از شاخص Silhouette و نمونه مطرح شده در کلاس با استفاده از sklearn انجام دهید و نمودارهای Silhouette مربوط به k=1 to 5 را نیز ترسیم نمایید.

k) مشخص کنید درهریک از قسمتهای i,j بهترین نتیجه با کدام عدد برای خوشه بندی بدست می آید و آیا این نتیجه برای هر دو روش یکسان است؟

۳. خوشه بندی سلسله مراتبی (Hierarchical Clustering)

a) ابتدا متد linkage را روی داده های iris اجرا کنید. (راهنمایی : این متد در پکیج scipy.cluster.hierarchy است.)

b) نمودار dendrogram مربوط به خوشه بندی سلسه مراتبی ایجاد شده در مرحله قبل را رسم و تفسیر نمایید.

c) همانطور که می دانید نمودار dendrogram به گونه ای است که هر چه در level بالاتری قطع شود تعداد کلاستر کمتری تولید می کند و هر چقدر level قطع پایین تر برود تعداد کلاستر ها بیشتر می شود. برای درک این موضوع از تابع fcluster استفاده کنید. ابتدا level=6 را مقدار دهی کرده و برچسب های تولید شده را که نشان دهنده تعداد کلاستر ها در این سطح است نشان دهید.

d) مقدار level را کاهش دهید و دوباره تابع fcluster را فراخوانی و برچسب های تولید شده را روی یک نمودار scatter plot نشان دهید و نمودار را تفسیر نمایید.