به نام خدا

تکلیف دوم درس داده کاوی

ترم دوم ۹۸-۹۷

راهنمایی :

زبان برنامه نویسی سئوالات پایتون است.

پیشنهاد می شود از محیط jupyter notebook استفاده کنید.

پکیج های اصلی استفاده شده numpy, pandas, sklearn می باشند.

سایر کتابخانه ها مورد نظر در هر سئوال اشاره شده است.

دیتاست های مورد نیاز در ادامه معرفی شده اند.

روش تحویل:

الف) فایلهای مربوط به کدهای هر سوال در یک فایل با نام Bx.zip که x شماره سوال است زیپ شوند، سپس کلیه این فایلهای زیپ در یک فایل واحد با نام HW2-Lastname.zip که Lastname فامیل شماست، زیپ شده و روی سامانه تا **ساعت ۷،۳۰ صبح سه شنبه ۴ تیر ۹۸ (روز قبل از امتحان)** اپلود شوند.

ب) تحویل بصورت حضوری بعدازظهر روز امتحان داده کاوی خواهد بود. برای هر سئوال کد نوشته شده و نتیجه اجرا را در فایل نهایی وارد کنید. فایل نهایی باید به صورت pdf باشد.

۱. KNN

دیتاستی که در این سئوال استفاده میشود مربوط به Breast Cancer می باشد. پکیج sklearn.datasets را ایمپورت کرده و داده های مربوط به این دیتاست را load کنید. برای راهنمایی بیشتر می توانید به آدرس زیر مراجعه فرمایید:
https://scikit-learn.org/stable/datasets/index.html

۱،۱. متد DESCR را برای این دیتاست فراخوانی و نتیجه را نمایش دهید.

۱،۲. با استفاده از متد های keys()، feature_names، data اطلاعات بدست آمده در مورد داده ها را نشان دهید.

۱،۳. داده ها را به فرمت Pandas DataFrame تبدیل کنید.

۱،۴. متد describe() را در مورد داده هایی که به صورت دیتافریم تبدیل شده اند اجرا نمایید.

۱،۵. یک ستون (فیلد) به نام target ایجاد نموده و کلید target را در آن قرار دهید

۱،۶. دستور value_counts() و target_names را در مورد فیلد target اجرا کنید. نتیجه اجرا چه اطلاعاتی در بردارد؟

۱،۷. برای تقسیم داده ها به مجموعه تست و آموزش ، تابع train_test_split را مقداردهی و اجرا نمایید.

۱،۸. ابعاد مجموعه های X_train ، X_test، y_train و y_test را نشان دهید.

۱,۹. دسته بند KNeighborsClassifier را با مقدار ۶ روی داده های آموزشی اجرا نموده (مدل را آموزش دهید) و دقت دسته بندی را روی داده های تست با تابع score نشان دهید.

۱,۱۰. مقدار هدف را برای مجموعه X_test با استفاده از اتابع predict بدست آورید.

۱,۱۱. به زبان ساده عملکرد predict را توضیح دهید.

۱,۱۲. از پکیج preprocessing تابع MinMaxScaler را ایمپورت کرده و با استفاده از آن داده های X_train و X_test را نرمال سازی کنید.

۱,۱۳.بار دیگر مدل را با استفاده از داده های آموزشی نرمال سازی شده ، آموزش دهید.

۱,۱۴.دقت مدل را روی داده های آموزشی و روی داده های تست با استفاده از تابع score بدست آورید.

۱,۱۵.برای بررسی اثر تعداد همسایه ها ، یک آرایه به نام train_accuracy و یک آرایه به نام test_accuracy ایجاد نموده ، سپس در یک حلقه for مقدار همسایگی را از ۱ تا ۱۰ افزایش داده و هر بار دقت مدل را روی داده های آموزشی و تست در ایندکس مورد نظر از آرایه های مربوطه ذخیره کنید. (دقت مدل روی داده های آموزشی در آرایه train_accuracy و دقت مدل روی داده های تست در آرایه test_accuracy ذخیره شود.)

۱,۱۶.با استفاده از کتابخانه matplotlib.pyplot روند تغییرات دقت بدست آمده روی داده های آموزشی و تست را که در قسمت قبل در آرایه های مورد نظر ذخیره نمودید به صورت نمودار نشان داده و جزئیات نمودار را مشخص کنید.

۱,۱۷.تقسیر خود را از نمودار بنویسید.

۲. **Decision Tree**

۲,۱. فایل csv دیتاست Vehicle را از آدرس https://www.openml.org/d/54 دانلود کنید.

۲,۲. فایل csv را بخوانید و در یک متغیر قرار دهید و از آن head بگیرید.

۲,۳. مقادیر موجود در فیلد هدف (Class) را با تابع unique نشان دهید.

۲,۴. همبستگی متغیر ها را نسبت به یکدیگر محاسبه و نمودار heatmap آن را رسم کنید. (راهنمایی : از کتابخانه matplotlib و seaborn استفاده کنید.)

۲,۵. مقادیر همه ستون ها به جز ستون class را در متغیر x قرار داده و ستون class را در متغیر y قرار دهید.

۲,۶. با استفاده از تابع train_test_split و انتخاب مقدار test_size=0.2 مجموعه های آموزشی و تست را ایجاد کنید.

۲,۷. با استفاده از متد decisionTreeClassifier از پکیج sklearn.tree داده ها را دسته بندی کنید . (راهنمایی : مقدار پارامترهای ورودی را به صورت زیر قرار دهید: 'max_depth = 5, max_features=4, criterion='entropy)

۲,۸. یک دیکشنری به نام params از سه پارامتر زیر تشکیل دهید. به آدرس زیر مراجعه کنید و سه پارامتر max_depth و max_features و min_sample_leaf را انتخاب کنید. توضیح دهید که هر یک از این پارامتر ها چه چیزی را کنترل می کند. می خواهیم مجموعه ای از مقادیر را به این سه پارامتر انتساب دهیم و مقدار بهینه را برای هر یک از پارامتر ها پیدا کنیم. لازم است که برای هر یک از سه پارامتر مجموعه ای مقادیر را معرفی کنید. برای پارامتر اول یک

لیست شامل None و ۳ ایجاد کنید ، و برای پارامتر های دوم و سوم از از تابع تولید کننده اعداد تصادفی randint که اعداد بین ۱ تا ۹ را ایجاد می کند استفاده کنید. خروجی این سئوال کد مربوط به ساخت دیکشنری است .

https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

۲,۹. مثل قسمت 2.7 ابتدا با استفاده از متد decisionTreeClassifier یک مدل ایجاد کنید و در متغیر tree قرار دهید. سپس از متد RandomizedSearchCV برای پیدا کردن مقدار بهینه هر یک از سه پارامتر تعیین شده در قسمت ۲,۸ استفاده کنید. مدل tree و دیکشنری params را با مقدار cv=5 به تابع دهید. و نتایج اجرا را در متغیری به نام tree_cv قرار دهید. داده های آموزشی را به tree_cv ، fit کنید.

۲,۱۰.بهترین مقدار پارامتر ها و بهترین مقدار دقت درخت بدست آمده را نشان دهید.

۲,۱۱.داده های تست را به مدل بدهید و میزان دقت را نمایش دهید.

۲,۱۲.افزایش مقدار cv چه تاثیری روی دقت خواهد داشت.

۲,۱۳.توضیح دهید متد feature_importances_ نشان دهنده چیست و مقدار آن را برای classifier بدست آورید.

۲,۱۴.خروجی تابع export_graphviz را بر روی classifier ی که با بهترین پارامتر های بدست آمده خواهید ساخت بدست آورید و با عنوان dot_data ذخیره کنید. (راهنمایی: graphviz را آدرس زیر دانلود کنید. https://graphviz.gitlab.io/download/) .

۲,۱۵.کتابخانه pydotplus را نصب کنید و با استفاده از آن فایل dot_data را به گراف تبدیل کنید و آن را نمایش دهید.

۲,۱۶.فایل های png و pdf گراف را ایجاد کنید.

۳. Clustering

۳,۱. در این سئوال از دیتاست iris استفاده می شود. این دیتاست را load کنید. می خواهیم با استفاده از الگوریتم k-mean تعداد گونه ها را مشخص کنیم.

۳,۲. ابتدا تعداد کلاستر ها را ۳ در نظر بگیرید و داده های آموزشی را به آن fit کنید و برای نمایش برچسب ها از متد predict استفاده کنید.

۳,۳. مراکز خوشه را در متغیری به نام centroids قرار دهید.

۳,۴. یک scatter plot با استفاده از داده های اول و سوم 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.

۳,۵. یکی از روشهای ارزیابی دقت کلاسترینگ استفاده از متد inertia_ (اینرسی) است . مقدار آن را برای کلاسترینگ فعلی نشان دهید.

۳,۶. یک حلقه for 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.

۳,۷. 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؟

۴. Hierarchy clustering

۴,۱. ابتدا متد linkage را روی داده های iris اجرا کنید. (راهنمایی : این متد در پکیج scipy.cluster.hierarchy است. در مرحله اول متد را برابر با complete قرار دهید)

۴,۲. نمودار dendrogram مربوط به خوشه بندی سلسه مراتبی ایجاد شده در مرحله قبل را رسم کنید.

۴,۳. 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 تجربه 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.

۴,۴. مقدار level را کاهش دهید و دوباره تابع fcluster را فراخوانی و برچسب های تولید شده را روی یک نمودار scatter plot نشان دهید.

۵. Regression

۵,۱. دیتاست boston را load کرده و داده های مربوط به feature های آن را به صورت دیتافریم تبدیل نمایید.

۵,۲. به انتهای دیتافریم یک ستون به نام Price اضافه کرده و مقدار target را در این ستون قرار دهید و دیتاست جدید را نشان دهید.

۵,۳. در این سئوال هدف این است که مقدار قیمت خانه را بر اساس پارامتر های ورودی پیش بینی کنیم و تاثیر افزایش تعداد پارامتر ها را بر روی دقت تخمین بدست آمده ارزیابی کنیم. لذا در مرحله اول یک regression بر اساس تنها دو پارامتر اول دیتاست ایجاد کنید: مجموعه x را با دو پارامتر اول دیتاست و مجموعه y را با پارامتر target ایجاد کنید.

۵,۴. با استفاده از تابع train_test_split و انتخاب مقدار test_size=0.3 مجموعه های آموزشی و تست را ایجاد کنید.

۵,۵. سه مرحله زیر را انجام دهید.

۵,۵,۱. ساختن مدل با تابع LinearRegression()

۵,۵,۲. fit کردن داده های آموزشی به مدل

۵,۵,۳. predict بر اساس داده های تست

۵,۶. مقدار MSE را با استفاده از تابع mean_squerd_error از کتابخانه metrics بدست آورید.

۵,۷. همه مراحل ذکر شده در قسمت های ۳ تا ۶ این سئوال را اینبار با ۱۳ feature تکرار کنید.

۵،۸. مقدار MSE جدید چه تغییری می کند. ارزیابی شما از اثر افزایش تعداد پارامتر ها در دقت مدل چیست؟

۵،۹. در این قسمت برای ارزیابی مدل از روش k-Fold Cross Validation استفاده خواهیم کرد. بدین منظور از متد cross_val_score استفاده کنید . مقدار cv را ۵ قرار دهید . (۵ بار مدل را آموزش داده و هر بار با داده تست جدید آن را ارزیابی خواهید کرد.) مقادیر مربوط به score های اجرا های مختلف را نشان داده و از آن میانگین بگیرید.

۶. Confusion Matrix و ROC

۶،۱. در این سئوال از دیتاست breast cancer استفاده می شود. این دیتاست را load کنید.

۶،۲. با استفاده از تابع train_test_split و انتخاب مقدار test_size=0.2 مجموعه های آموزشی و تست را ایجاد کنید. با استفاده از knn و تعداد همسایه های ۸ مدل را ایجاد کرده و داده های آموزشی را به مدل fit کنید و سپس تابع predict را برای آن فراخوانی کنید و نتیجه را در y_pred ذخیره کنید.

۶،۳. متد های confusion_matrix و classification_report را از ساب پکیج metrics ایمپورت کنید.

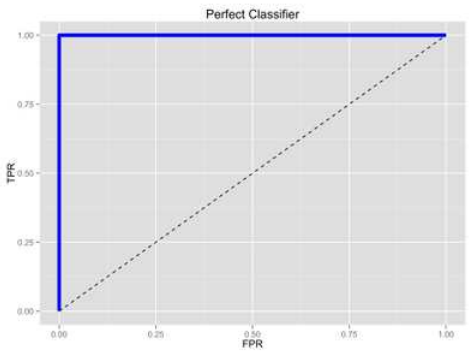
۶،۴. متد confusion_matrix را با داده های y_test و y_pred و مقدار برچسب ها مقداردهی کنید و از خروجی print بگیرید و نتیجه را تفسیر کنید. هر کدام از ۴ عدد نشان داده شده در خروجی نشان دهنده چیست؟

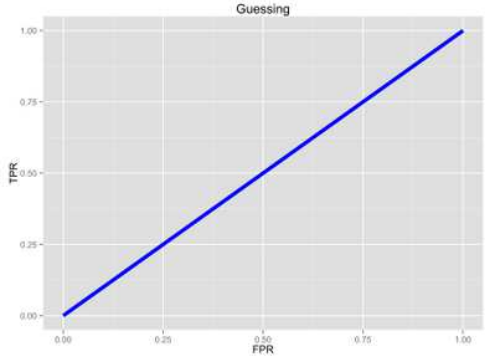
۶،۵. متد classification_report را با داده های y_test و y_pred مقداردهی کنید و از خروجی print بگیرید و نتیجه را تفسیر کنید.

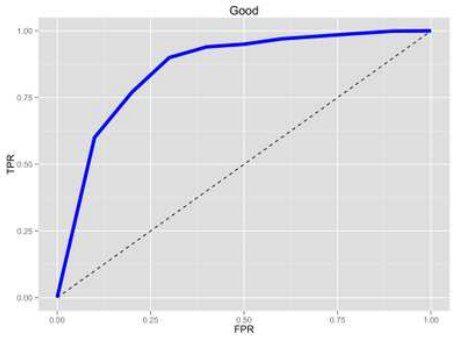
۶،۶. نتیجه حاصل از confusion_matrix را نرمال سازی کنید . (راهنمایی برای نرمال سازی از متد normalize از ساب پکیج preprocessing استفاده کنید. و مقدار norm را برابر l1 قرار دهید.)

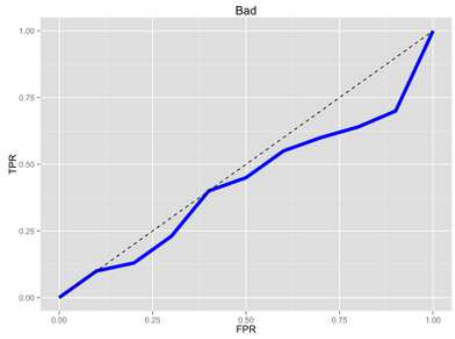
۶،۷. نتیجه اجرای مرحله قبل را به صورت یک دیتافریم درآورید که سطر ها و ستون های آن با نام مقادیر target که همان عبارت های malignant و benign هستند مقدار دهی شده باشد. (راهنمایی: با مقدار دهی پارامتر های columns و index در متد dataframe).

۶،۸. همانطور که می دانید منحنی ROC برای ارزیابی روش های دسته بندی باینری کاربرد دارد. تفاوت وضعیت های مختلف نشان داده شده در شکلهای زیر را توضیح دهید:









۶,۹. ابتدا با استفاده از متد predict_proba احتمال انتساب هر یک از مقادیر داده های آموزشی x_test را به کلاسهای هدف بدست آورید و در متغیری به نام y_pred_prob ذخیره کنید و ان را نشان دهید

۶,۱۰.با استفاده از roc_curve و با تنظیم ورودی های این متد مقادیر fpr , tpr و threshold را بدست آورید.

۶,۱۱.از داده های fpr , tpr یک plot رسم کنید و نتیجه را تفسیر کنید. مدل شما چقدر خوب عمل کرده است؟

۶,۱۲.در نهایت خروجی متد roc_auc_score را برای این مدل فراخوانی کنید و نتیجه را نشان دهید. تفاوت AUC و ROC چیست؟

۷. Association Rules

۷,۱. دیتاست مربوط به این سئوال را می توانید از لینک زیر دریافت کنید. (راهنمایی : پیشنهاد می شود دیتاست را ابتدا با فرمت csv ذخیره و سپس در محیط notebook وارد کنید.)

http://archive.ics.uci.edu/ml/datasets/Online+Retail

۷,۲. برای این کار روی این سئوال نیاز به نصب پکیج mlxtend وجود دارد. (برای نصب این پکیج از دستور زیر استفاده کنید: `conda install -c conda-forge mlxtend`

۷,۳. توابع apriori و association_rules را از این پکیج ایمپورت کنید.

۷,۴. از فیلد Description ، فاصله های موجود (بلانک) ها را حذف کنید. (راهنمایی : استفاده از متد (strip()

۷,۵. رکوردهایی که InvoiceNO آنها خالی است را حذف کنید. سپس نوع داده ای این فیلد را به str تبدیل کنید. (با استفاده از astype

۷,۶. InvoiceNO هایی که دارای حرف C هستند را حذف کنید.

۷,۷. دستور زیر را روی داده ها اجرا کنید. توضیح دهید این دستور دقیقا چه می کند؟

```
basket = (df[df['Country'] =="France"]
          .groupby(['InvoiceNo', 'Description'])['Quantity']
          .sum().unstack().reset_index().fillna(0)
```

```
        .set_index('InvoiceNo'))
```

۷,۸. یک تابع بنویسید که مقادیر بیشتر از صفر را به یک و سایر مقادیر را به صفر تبدیل کند. سپس این تابع را روی کل داده های basket اعمال کنید. (راهنمایی : با استفاده از applymap)

۷,۹. ستون POSTAGE را از مجموعه داده های حاصل از مرحله قبل حذف کنید. در این تحلیل نیازی به این ستون نیست.

۷,۱۰. frequent item sets ها را با حداقل support برابر ۷٪ بدست آورید. (راهنمایی : با استفاده از تابع apriori)

۷,۱۱.قوانین وابستگی را تولید کنید. (راهنمایی metric را برابر با lift قرار بدهید.)

۷,۱۲.آن دسته از قوانینی که مقدار lift آنها بیشتر از ۶ و مقدار confidence آنها بیشتر از ۰,۸ است را فیلتر کنید.

۷,۱۳.چند مورد از نتایج بدست آمده را تفسیر کنید.