به نام خدا

تکلیف دوم درس داده کاوی

ترم دوم ۹۹-۹۸

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 **زمان مشخص شده** اپلود شوند.

ب) گزارش نهایی باید شامل پاسخ تمامی سوالات باشد که شامل کد نوشته‌شده، توضیحی درمورد کد و نتیجه اجرا و تفسیر نتیجه می‌باشد.

ج) نحوه تحویل در کانال مشخص خواهد شد.

توضیحات:

شماره ۱ : برای تمامی سوالات مطرح شده در این تمرین در صورت موجود بودن ستون **عددی** در دیتاست، به تشخیص خود این مقادیر را **نرمالیز** کرده و سپس به مدل خود آموزش دهید.

شماره ۲ : برای نمایش گرافیکی گراف های سوال دو از دو کتابخانه pydotplus و graphviz استفاده نمایید. برای دانلود graphviz از لینک زیر استفاده نمایید: (https://graphviz.gitlab.io/download/)

شماره ۳ : برای لود کردن دیتاست iris میتوانید از دیتاست موجود در کتابخانه sklearn استفاده نمایید.

شماره ۴ : برای لود کردن دیتاست boston میتوانید از دیتاست موجود در کتابخانه sklearn با نام boston_dataset استفاده نمایید.

شماره ۵ : برای حل سوال ۷ از کتابخانه mlxtend استفاده نمایید.

۱. Regression Decision Tree

۱/۱. فایل csv دیتاست Housing را بخوانید و در یک متغیر قرار دهید و از آن head بگیرید.

۱/۲. مقادیر ستون LSTAT را در متغیر x قرار داده و ستون MEDV را در متغیر y قرار دهید.

۱/۳. با استفاده از تابع train_test_split و انتخاب مقدار test_size=0.2 مجموعه های آموزشی و تست را ایجاد کنید.

۱/۴. با استفاده از کلاس DecisionTreeRegressor از پکیج sklearn.tree مدل پیش‌بینی کننده را ایجاد کنید. (راهنمایی : مقدار پارامترهای ورودی را به صورت زیر قرار دهید: max_depth = 3, criterion='mse')

۱/۵. داده های تست را به مدل بدهید و میزان دقت را نمایش دهید.

۱/۶. افزایش یا کاهش مقدار max_depth چه تاثیری روی دقت خواهد داشت. بهترین مقدار برای عمق درخت این مسئله چه عددی میباشد.

۲. Classification Decision Tree

۲/۱. فایل csv دیتاست Vehicle را بخوانید و در یک متغیر قرار دهید و از آن head بگیرید.

۲/۲. نام ستون های این دیتاست را به ترتیب مطابق جدول زیر نامگذاری کنید.

1. Number of times pregnant.
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
3. Diastolic blood pressure (mm Hg).
4. Triceps skinfold thickness (mm).
5. 2-Hour serum insulin (mu U/ml).
6. Body mass index (weight in kg/(height in m)^2).
7. Diabetes pedigree function.
8. Age (years).
9. Class variable (0 or 1).

۲/۳. همبستگی متغیر ها را نسبت به یکدیگر محاسبه و نمودار heatmap آن را رسم کنید. (راهنمایی: از کتابخانه matplotlib و seaborn استفاده کنید.)

۲/۴. مقادیر همه ستون ها به جز ستون class را در متغیر x قرار داده و ستون class را در متغیر y قرار دهید.

۲/۵. با استفاده از تابع train_test_split و انتخاب مقدار test_size=0.2 مجموعه های آموزشی و تست را ایجاد کنید.

۲/۶. 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')

۲/۷. داده های تست را به مدل بدهید و میزان دقت را نمایش دهید.

۲/۸. بهترین مقدار پارامتر max_depth را با استفاده از مقادیر مختلف ۳ تا ۹ بررسی کرده و بهترین مقدار دقت درخت بدست آمده را نشان دهید.

۲/۹. توضیح دهید متد feature_importances_ نشان دهنده چیست و مقدار آن را برای classifier بدست آورید.

۲/۱۰. خروجی تابع export_graphviz را بر روی classifier ی که با بهترین پارامتر های بدست آمده خواهید ساخت بدست آورده و ذخیره کنید.

۲/۱۱. کتابخانه pydotplus را نصب کنید و با استفاده از آن فایل dot_data را به گراف تبدیل کنید و آن را نمایش دهید.

۳. Clustering

۳/۱. 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.

۳/۲. 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.

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

۳/۴. 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.

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 از کتابخانه sklearn لود کرده و مقادیر موجود در دیکشنری این دیتاست را بررسی نمایید.

۵/۲. داده های مربوط به feature های آن را به صورت دیتافریم تبدیل نمایید.

۵/۳. به انتهای دیتافریم یک ستون به نام Price اضافه کرده و مقدار target این دیتاست را در این ستون قرار دهید و دیتاست جدید را ذخیره نمایید.

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. سپس یک مدل از نوع LinearRegression() ساخته و داده های آموزشی را به مدل fit کنید. سپس داده های تست را با استفاده از متد predict مدل پیش بینی کنید.

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

۵/۷. 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 از کتابخانه sklearn 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 مدل بردار ماشین پشتیبان (SVM) و کرنل Linear 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.

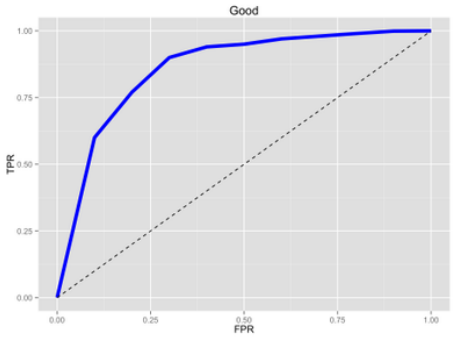
۶/۴. متد confusion_matrix را با داده های y_test و y_pred و مقدار برچسب ها مقداردهی کنید و از خروجی print بگیرید و نتیجه را تفسیر کنید. هر کدام از ۴ عدد نشان داده شده در خروجی نشان دهنده چیست؟

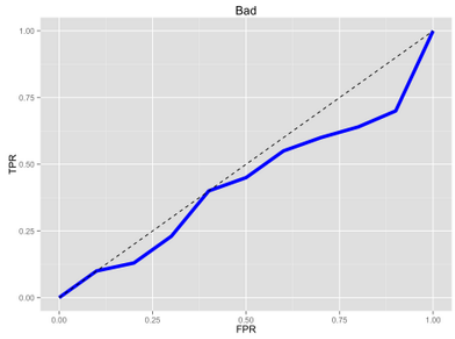
۶/۵. متد classification_report را با داده های y_test و y_pred مقداردهی کنید و از خروجی print بگیرید و نتیجه را تفسیر کنید.

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؟

۷. Association Rules

۷/۱. دیتاست مربوط به این سئوال را می توانید از لینک زیر دریافت کنید.

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

۷/۲. توابع apriori و association_rules را از پکیج mlxtend ایمپورت کنید.

۷/۳. از فیلد Description ، فاصله های موجود (بلانک) ها را حذف کنید. (راهنمایی : استفاده از متد (strip()

۷/۴. رکوردهایی که InvoiceNO آنها خالی است را حذف کنید. سپس نوع داده ای این فیلد را به str تبدیل کنید. (با استفاده از astype

۷/۵. InvoiceNO هایی که دارای حرف C هستند را حذف کنید.

۷/۶. 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'))
```

۷/۷. 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)

۷/۸. ستون POSTAGE را از مجموعه داده های حاصل از مرحله قبل حذف کنید. در این تحلیل نیازی به این ستون نیست.

۷٫۹. frequent item sets ها را با حداقل support برابر ۷٪ بدست آورید. (راهنمایی : با استفاده از تابع apriori)

۷٫۱۰. قوانین وابستگی را تولید کنید. (راهنمایی metric را برابر با lift قرار بدهید.)

۷٫۱۱. آن دسته از قوانینی که مقدار lift آنها بیشتر از ۶ و مقدار confidence آنها بیشتر از ۰.۸ است را فیلتر کنید.

۷٫۱۲. یک مورد از نتایج بدست آمده را تفسیر کنید.

۸. Naive Bayse

۸٫۱. یکی از کاربردهای مدل نایوبیز، دسته بندی نمودن متون برچسب دار میباشد. در این مسئله میخواهیم یک مدل پیش‌بینی برچسب متن طراحی کنیم تا براساس متن ورودی، دسته متن را پیش‌بینی کند.

۸٫۲. دیتاست 20NewsGroups را از کتابخانه sklearn لود کرده و مقادیر موجود در فیلد Target این دیتاست را نمایش دهید.

۸٫۳. یک آرایه ساخته و مقادیر دسته‌های مختلف زیر را در این آرایه وارد نمایید. سپس براساس این مقادیر، داده‌های آموزشی و داده های تست موردنظر را دریافت نمایید. (راهنمایی: این کار را توسط متد fetch_20newsgroups انجام دهید.)

```
targets = ['talk.religion.misc', 'soc.religion.christian','sci.space',
'comp.graphics']
```

۸٫۴. حال به ازای داده‌های دریافت شده نیاز است تا متون هر رکورد تبدیل به یک بردار عددی شود تا قابل استفاده برای مدل یادگیری ماشین شوند. برای این داده‌ها از روش TF-IDF استفاده نمایید. (راهنمایی: برای پیاده سازی این روش میتوانید از کلاس TfidfVectorizer کتابخانه sklearn استفاده نمایید.)

۸٫۵. تفاوت روش TF و روش TF-IDF را توضیح داده و مزیت روش TF-IDF را بیان کنید.

۸٫۶. سپس یک مدل از نوع MultinomialNB() ساخته و داده های آموزشی را به مدل fit کنید. سپس داده های تست را با استفاده از متد predict مدل پیش بینی کنید.

۸٫۷. با استفاده از متد confusion_matrix و داده های y_test و y_ میزان دقت مدل را بررسی نموده و خروجی ماتریس را نشان دهید.

۸٫۸. با استفاده از مدل ساخته شده و متد predict، برچسب جمله زیر را پیش‌بینی کنید.

```
she is also campaigning to remove Christmas programs.
```