به نام خدا

تکلیف سوم درس مبانی داده کاوی

ترم دوم ۱۴۰۰-۱۴۰۱

راهنمایی :
زبان برنامه نویسی سوالات پایتون است.
پیشنهاد می شود از محیط Jupyter notebook استفاده کنید.
پکیج های اصلی مورد نیاز شامل pandas،numpy می باشند.
مجموعه داده های مورد نیاز در ادامه معرفی شده اند.
روش تحویل :
a) فایلهای مربوط به کدهای هر سوال در یک فایل با نام Qx.zip که x شماره سوال است زیپ شوند، سپس کلیه این فایلهای زیپ در یک فایل واحد با نام HW3-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 تشریحی و سوالات پیاده سازی) باشد که برای سوالات پیاده سازی شامل کد نوشته‌شده، توضیحی در مورد کد و نتیجه اجرا و تفسیر نتیجه می‌باشد (گزارش سوالات پیاده سازی را میتوانید در همان محیط Jupyter notebook بنویسید).
ج) زمان و نحوه تحویل تکلیف روی سامانه و در فایل راهنمای ترم مشخص شده است.
د) تحویل خارج سامانه و خارج ساعت مشخص شده قابل قبول نیست.

۱. **شبکه عصبی** ( زمان تقریبی: ۱ ساعت)

a. داده های titanic را خوانده و پیش پردازش های لازم مانند حذف missing , ایجاد فیلد familysize و غیره را طبق تکلیف دوم انجام دهید.
b. داده ها را به ۷۰ درصد آموزشی و ۳۰ درصد تست تقسیم کنید. (train_test_split)
c. با استفاده از MLPClassifer و پارامترهای پیش فرض آن و بدون استانداردسازی داده مدلسازی انجام دهید و دقت مدل را روی داده تست گزارش دهید.
d. در صورت استاندارد سازی داده مدل MLPClassifer چقدر باعث افزایش دقت تست می شود؟
e. با استفاده از GridSearchCV و انجام HyperParameters Tuning بهترین مقدار پارامترهای زیر را از میان مقادیر زیر بدست آورید و گزارش دهید که چقدر به بهبود دقت مدل کمک شده است.

الگوریتم بهینه سازی: Adam, SGD
نرخ یادگیری : ۱e-4, 1e-3 , 5-e-2
تعداد لایه و نورون ها : (بین یک تا سه لایه پنهان و هر لایه بین ۱۰۰ تا ۱۰۰۰ نرون)
تابع فعالسازی : خطی , tanh و relu

f. برای مدل بدست آمده بخش e ماتریس Confusion رسم کنید و مقادیر Precision Recall F1-Score را برای هر کلاس جداگانه بدست آورید.

۲. **Feature Selection** (زمان تقریبی: ۰.۵ ساعت)

a. اطلاعات دانش آموزان را از فایل student خوانده و ستون های ۱G و ۲G را حذف کرده و ۳G را به عنوان هدف در نظر بگیرید

b. یکی از روشهای feature selection استفاده از مدل های دیگر از جمله OLS است. با استفاده از Recursive feature elimination with cross-validation یا RFECV موارد زیر را انجام دهید:

- با استفاده از امتیاز دهی neg_mean_squared_log_error و مدل OSL تعداد خصوصیات مهم داده را مشخص کنید.
- نمودار scoring را بر حسب تعداد متغیر ها رسم کنید.
- با استفاده از ranking_ اهمیت هر ستون را گزارش دهید.

---

۳. **Naïve Bayes** (زمان تقریبی:۱ ساعت)

اطلاعات افرادی که به بیماری Covid-19 مبتلا شده اند در جدول زیر وجود دارد. با استفاده از مدل بیزین احتمال سالم بودن کودک ۶ ساله که علائم تب و خستگی دارد ولی درد و سرفه خشک ندارد چقدر است؟ (برای این سوال نمی توانید از classifer آماده استفاده کنید).

| شدت بیماری | سن | درد | سرفه خشک | خستگی | تب |
|---|---|---|---|---|---|
| متوسط | ۰-۹ | خیر | خیر | خیر | خیر |
| متوسط | ۱۰-۱۹ | خیر | بله | بله | خیر |
| خفیف | ۱۰-۱۹ | بله | خیر | بله | بله |
| متوسط | ۰-۹ | بله | بله | خیر | خیر |
| شدید | ۱۰-۱۹ | خیر | خیر | بله | بله |
| شدید | ۰-۹ | بله | خیر | خیر | خیر |
| سالم | ۱۰-۱۹ | بله | بله | بله | بله |
| سالم | ۱۰-۱۹ | خیر | بله | بله | خیر |
| شدید | ۱۰-۱۹ | بله | بله | بله | بله |
| سالم | ۰-۹ | بله | بله | بله | بله |
| سالم | ۰-۹ | خیر | خیر | خیر | خیر |
| متوسط | ۰-۹ | بله | خیر | بله | خیر |
| سالم | ۰-۹ | بله | بله | بله | بله |
| شدید | ۰-۹ | خیر | بله | بله | بله |
| خفیف | ۰-۹ | خیر | خیر | خیر | خیر |
| سالم | ۱۰-۱۹ | خیر | بله | بله | بله |
| شدید | ۰-۹ | بله | خیر | خیر | خیر |
| متوسط | ۰-۹ | بله | بله | بله | بله |
| شدید | ۰-۹ | خیر | بله | بله | بله |
| خفیف | ۰-۹ | بله | بله | بله | خیر |
| سالم | ۰-۹ | خیر | بله | بله | بله |
| خفیف | ۰-۹ | خیر | خیر | خیر | خیر |

| | | | | | |
|---|---|---|---|---|---|
| خفیف | ۱۰-۱۹ | خیر | بله | بله | خیر |
| سالم | ۰-۹ | خیر | بله | بله | خیر |

---

۴. مجموعه داده forestfires که شامل اطلاعات مربوط به آتش سوزیهای مناطق شمالی پرتقال است در اختیار شما قرار گرفته است. هدف پیش بینی مساحت ناحیه آتش گرفته است. اطلاعات بیشتر در مورد این مجموعه داده در فایل forestfiresReadMe.txt ضمیمه شده است. با توجه به این مجموعه داده به سوالات زیر پاسخ دهید: (زمان تقریبی: ۲.۵ ساعت)

- **Preprocessing**
  a. پس از ذخیره این مجموعه داده در دیتا فریم, به کمک تابع get_dummies دو ستون month و day را تبدیل کنید. و پس از تبدیل, این دو ستون را از دیتا فریم حذف کنید.
  b. داده را از نظر وجود missing value بررسی کنید و در صورت موجود بودن با ۰ جایگزین کنید.
  c. داده را از نظر وجود سطر تکراری (duplicate) بررسی کنید . در صورت وجود داشتن آنها را نشان دهید. و در آخر آنها را حذف کنید.
  d. هیستوگرام متغیر هدف مجموعه داده (area) را رسم کرده و از نظر داشتن کجی بررسی کنید و مقدار عددی آن را نیز نشان دهید.
  e. در صورت وجود کجی روی area به کمک متد sqrt کجی را برطرف کنید.
  f. همبستگی بین ویژگی های داده با متغیر هدف (area) را به کمک heatmap نشان داده و تفسیر کنید.

- **Feature selection and Linear Regression**
  g. همه ی ستونهای دیتا فریم به جز متغیر مورد پیش بینی (area) را در X و area را در Y ذخیره کنید.
  h. به کمک کتابخانه sklearn ۸۰ درصد از داده ها را برای اموزش و ۲۰ درصد را برای تست جدا کنید.
  i. به کمک Dummyregressor از کتابخانه sklearn یک baseline model با روش میانگین بسازید. آنرا روی مجموعه آموزشی, آموزش داده و سپس به کمک این مدل پیش بینی را روی داده تست انجام دهید. میانگین مربعات خطا بین مقدار واقعی و مقدار پیش بینی شده را به دست آورید. (میتوانید از توابع آماده sklearn استفاده کنید).
  j. این بار پیش بینی را به کمک مدل LinearRegression (از کتابخانه sklearn) انجام داده و مجددا MSE را برای این مدل به دست آورید.
  k. با مقایسه MSE مدل baseline و مدل LinearRegression بگویید که آیا رگرسیون خطی خوب عمل کرده است؟
  l. به کمک OLS از کتابخانه, statsmodels pvalue هر یک از ویژگی ها را به دست آورید و نتایج را تحلیل کنید. آیا بین این نتایج و نمودار heatmap قسمت قبلی رابطه ای وجود دارد؟ توضیح دهید.
  m. به کمک OLS و نیز روش backward elimination, عملیات feature selection را روی X انجام داده و نهایتا فیچرهای انتخاب شده را در X۲ نگه دارید. (*** برای قسمت های n,o,p,q از X۲ استفاده کنید).
  n. مجددا قسمت های h, I, j, k را اجرا کنید (به جای X از X۲ استفاده کنید) نتایج را مقایسه و تفسیر کنید.

o. به کمک Ridge از کتابخانه sklearn بار دیگر رگرسیون را انجام دهید (با alpha برابر با ۱) و از نظر معیار MSE آنرا با linear regression مقایسه کنید. و به طور کلی کاربرد آن و نیز علت عملکرد بهتر آن را بیان کنید.

p. به کمک ElasticNet از کتابخانه sklearn بار دیگر رگرسیون را انجام دهید (با alpha برابر با ۱) و از نظر معیار MSE آن را با Ridge مقایسه کنید. و به طور کلی کاربرد آن و نیز علت عملکرد بهتر آن را بیان کنید.

q. قسمت p را مجددا تکرار کنید اما این بار برای مقایسه پیش بینی با مقدار واقعی از تابع هزینه mean absolute error (میتوانید از تابع آماده استفاده کنید) استفاده کرده و علت تفاوت زیاد مقدار نهایی را بیان کنید.

- **Model Selection**

برای این قسمت از X (همه ی فیچرها و نه فقط فیچر های منتخب) استفاده کنید.

r. در این قسمت می خواهیم به کمک روش k fold cross validation بهترین مقدار alpha را برای مدل قسمت p به دست آوریم. بدین منظور از توابع KFold و cross_val_score کتابخانه sklearn استفاده کنید. تعداد فولدها را برابر با ۵ و Random state را برابر با ۱ و متد scoring را neg_mean_squared_error در نظر بگیرید.
مقدار alpha را از ۱ تا ۲۰۰ یک واحد یک واحد افزایش دهید. و به ازای هر مقدار alpha یک بار مدل elasticNet را با متد 5FoldCrossValidation اجرا کرده و نهایتا به ازای هر مقدار alpha, میانگین خطای MSE را پس از ۵ فولد اجرا ذخیره کنید.

s. نمودار میانگین MSE مدل را پس از هر اجرا بر حسب مقدار alpha رسم کنید. علت افزایش مقدار میانگین MSE پس از نقطه min را بیان کنید. alpha بهینه برای elasticNet کدام است؟ و منجر به چه مقدار MSE میشود؟

---

۵. **Poisson Regression** (زمان تقریبی ۰.۵ ساعت)

مجموعه داده competition_awards_data شامل دو ستون math score و award است. هدف از این سوال پیش بینی تعداد award بر اساس math Score است.

a. پس از ذخیره این مجموعه داده در یک دیتا فریم, وجود missing value ها را در آن بررسی کنید و در صورت موجود بودن با مقدار ۰ جایگزین کنید.

b. نمودار پراکندگی award بر حسب math score را رسم و تفسیر کنید.

c. ستون Math Score را در X و ستون award را در y ذخیره کرده و مجددا به کمک کتابخانه sklearn, ۲۰ درصد از داده را برای تست و ۸۰ درصد را برای آموزش جدا کنید.

d. مدل PoissonRegressor از کتابخانه sklearn, را روی داده های آموزشی آموزش دهید. سپس به کمک آن روی داده های تست عمل پیش بینی را انجام دهید.

e. عملکرد مدل خود را روی مجموعه تست با معیار $R^2$ بررسی و تفسیر کنید.

f. روی یک نمودار پراکندگی مقدار واقعی award روی مجموعه تست و نیز پراکندگی مقدار پیش بینی شده award را نمایش داده و نتایج را تفسیر کنید.
g. آیا میتوانیم از LogisticRegression برای این مسئله استفاده کنیم؟ توضیح دهید.

---

۶. **KNN** (زمان تقریبی ۱.۵ ساعت)

a. مجموعه داده Iris در اختیار شما قرار گرفته است. هدف از این سوال پیاده سازی الگوریتم KNN روی این مجموعه داده است. تابعی بنویسید که با دریافت داده ها و نیز مقدار k, پس از نرمال سازی داده ها (میتوانید برای نرمال سازی از standardScaler از کتابخانه sklearn استفاده کنید), کلاسیفیکیشن به روش KNN را انجام دهد. (برای این کار نمیتوانید از تابع آماده KNeighborsClassifier در پایتون استفاده کنید.) همچنین برای جداسازی داده آموزشی از داده تست از روش Leave One Out, که هر بار یکی از رکوردهای داده را برای تست و مابقی را برای آموزش استفاده میکند, بهره بگیرید. (برای اینکار میتوانید از تابع LeaveOneOut در کتابخانه sklearn استفاده کنید).

b. تابعی که در قسمت قبلی پیاده سازی کردید را به ازای مقادیر k بین ۱ تا ۵۰ اجرا کرده, در هر مرحله نرخ خطای کلاس بندی را رسم کنید. نتایج نمودار را تحلیل کنید. کمترین نرخ خطا مربوط به چه مقدار k است و میزان آن چقدر است؟
c. این بار برای پیاده سازی KNN از روش weighted voting استفاده کنید و مراحل a و b را روی Iris اجرا کرده, آن را با روش unweighted مقایسه کرده و نتایج خود را تحلیل کنید.

---

۷. برای پیاده سازی قسمت های زیر میتوانید از توابع آماده استفاده کنید. همچنین در این سوال برای جداسازی داده آموزشی از تست از تابع train_test_split استفاده کرده و ۲۰ درصد از داده ها را برای تست و مابقی را برای آموزش استفاده کنید. در این سوال نیز مشابه سوال قبلی قبل از اجرای الگوریتم داده ها را نرمال کنید. (زمان تقریبی ۱.۵ ساعت)

a. بر روی مجموعه داده iris, الگوریتم KNN را با استفاده از KDTree برای مقادیر k از ۱ تا ۳۰ اجرا کنید, نمودار نرخ خطای کلاس بندی را بر اساس مقدار k رسم کنید و آن را تحلیل کنید. برای k بهینه, confusion matrix را محاسبه و تحلیل کنید.
b. قسمت قبل را این بار به کمک BallTree تکرار کنید. و پس از تحلیل نتایج این دو روش را مقایسه کنید.
c. قسمت های a و b را این بار روی مجموعه داده pop_failures که در اختیارتان قرار گرفته است اجرا کنید و نتایج را تحلیل کنید.

d. قسمت های a و b را این بار روی مجموعه داده banknote_authentications که در اختیارتان قرار گرفته است اجرا کنید و نتایج را تحلیل کنید.
e. قسمت های a و b را این بار روی مجموعه داده credit cards که در اختیارتان قرار گرفته است اجرا کنید و نتایج را تحلیل کنید.
f. با توجه به نتایج قسمتهای قبلی تحلیل کنید که چه زمان بهتر است از KDTree استفاده کنیم و چه زمان از BallTree؟ (برای تحلیل, پیچیدگی زمانی بر حسب تعداد نمونه ها و نیز تعداد ویژگی ها را در نظر بگیرید).

---

۸. **SVM** (زمان تقریبی ۰.۵ ساعت)

روی مجموعه داده Iris, پس از جداسازی ۲۰ درصد از داده ها برای تست و مابقی برای آموزش, الگوریتم SVM خطی را اجرا کنید. پس از آموزش مدل روی مجموعه آموزشی, به کمک مدل آموزش دیده روی مجموعه داده تست پیش بینی انجام دهید. با تحلیل confusion matrix نتیجه را بررسی کنید.

a. اینبار میخواهیم Kernel SVM را روی Iris اجرا کنیم. برای اینکار از متد polynomial استفاده کنید. مقادیر پارامتر چند جمله ای را از ۱ تا ۱۰ تغییر دهید و نمودار نرخ خطای کلاس بندی بر حسب درجه چندجمله ای را رسم کنید. و نتایج نمودار را تحلیل کنید.
b. با توجه به نتایج قسمت قبلی, Kernel SVM با متد polynomial و درجه بهینه را پیاده سازی کنید و نتایج را بر اساس confusion matrix تحلیل کرده و با قسمت a مقایسه کنید.

---

۹ .**کاهش ابعاد** (زمان تقریبی ۱ ساعت)

برای ۳۳ بیمار مختلف مبتلا به سرطان, اطلاعات بیان ژن آنها در اختیار است. نوع سرطان به دو دسته acute lymphoblastic leukemia (ALL) و acute myeloid leukemia (AML) تقسیم می شود.

در فایل actual.csv کد بیمار و نوع سرطان بعنوان target وجود دارد.

در فایل data_set_ALL_AML_train.csv داده های بیان ژن به تفکیک هر ژن و هر بیمار مشخص شده است که به آن فایل train می گوییم.

در فایل data_set_ALL_AML_train.csv داده های تست قرار دارد که به آن test می گوییم.

a. نیازی به ستونهای Gene Description , call و Gene Accession Number نیست آنها را در Train و Test حذف کنید.

b. داده های train و test را transpose کنید تا ستون ها مقادیر ژنها و سطر ها بیماران باشد و سپس آنها را به داده target متصل کنید..
c. با استفاده از tSNE در دو بعد بیماران را با رنگ نوع سرطان مشخص کنید. (train)
d. داده ها را استاندارد ساخته و MLPClassifer دقت مدل را روی داده test گزارش دهید.
e. با استفاده از PCA ابعاد داده را به سه بعد کاهش دهید و مجددا مدل MLPClassfier را روی داده اجرا کرده و دقت بدست آمده را گزارش دهید.

---

۱۰. **AdaBoost** (زمان تقریبی: ۰.۵ ساعت)

a. داده های تایتانیک را همانند تمرین شبکه عصبی (سوال ۱) خوانده و پیش پردازش های لازم را انجام دهید. ستون familysize را اضافه کنید. و داده ها را به نسبت ۰.۳ به train و test تقسیم کنید
b. مدل AdaBoostClassifer را توسط تخمین زننده DecisionTreeClassifier با حداکثر عمق ۱ بجای weak estimator پیش فرض بر روی داده های فوق آموزش دهید و دقت مدل و ماتریس confusion را گزارش دهید.

---

۱۱. **XGBoost** (زمان تقریبی: ۰.۵ ساعت)

داده های تایتانیک را به کمک XGBoost دسته بندی کنید. (پیش پردازش همانند قبل) و با استفاده از GridSearchCV بهترین پارامترهای learning_rate و maxdepth و alpha را مشخص کنید.

---

۱۲. **Stacking Ensemble** (زمان تقریبی: ۰.۵ ساعت)

به کمک مجموعه داده pop_failures, به سوالات زیر پاسخ دهید:

a. پس از ذخیره برچسب داده ها در متغیر y و ویژگی های آنها در متغیر X, به ترتیب سه کلاسیفایر Bayesian, KNN و SVM را به روش KFold با تعداد فولد برابر با ۱۰ و رندوم استیت برابر با ۱ روی آن اجرا کنید. متد scoring را accuracy در نظر گرفته و از طریق نمودار میله ای دقت این سه کلاسیفایر را روی مجموعه داده نشان داده و با هم مقایسه کنید.
b. در این قسمت هدف پیاده سازی روش stacking است. بدین منظور پس از تفکیک مجموعه داده به ۸۰ درصد آموزش و ۲۰ درصد تست, برای base model ها از ۳ کلاسیفایر قسمت a و برای meta model از LogisticRegression استفاده کنید و تعداد فولد ها را نیز برابر با ۱۰ در نظر بگیرید. پس از آموزش مدل روی داده آموزشی, به کمک مدل آموزش دیده روی داده تست پیش بینی انجام دهید و دقت آنرا به دست آورده و با قسمت a مقایسه کنید. هم چنین نرخ خطا را با قسمت c سوال ۷ مقایسه کرده و نتایج را تحلیل کنید.

---