


تمرین دوم درس علوم داده در مهندسی نرم افزار		 دانشکده مهندسی کامپیوتر
تاریخ: ۴۰۰/۰۹/۰۲	موعد تحویل: سه شنبه، ۴۰۰/۰۹/۱۶	
مدرس: دکتر عباس حیدرنوری	دستیاران آموزشی: مهسا مسعود (mah.masoud76@gmail.com) حامد طاهرخانی (th.hamed75@gmail.com)	

نکات مهم در مورد تحویل تکلیف
<ul style="list-style-type: none"> پاسخ‌های خود را در یک فایل فشرده با فرمت <code>HW2_fullname_studentNumber.zip</code> در صفحه CW درس ارسال کنید. هر قسمتی از کد که خطا داشته باشد، بررسی نخواهد شد و نمره‌ای هم برای آن در نظر گرفته نمی‌شود. تمامی توابعی که نیاز دارید از آن‌ها استفاده کنید یا به شما تدریس داده شده‌اند و یا با یک جستجوی ساده آن‌ها را پیدا خواهید کرد.

****در این تمرین استفاده از jupyter notebook و یا ابزارهای دیگر به منظور ارائه پاسخ در قالب ipynb الزامی می‌باشد.**

سوال اول:

دیابت diabetes مجموعه‌ای از داده‌های مربوط به پیشرفت بیماری دیابت در بدن افراد بر اساس برخی ویژگی‌های سلامتی افراد می‌باشد. این مجموعه داده‌ها یکی از دیابت‌های کتابخانه sklearn می‌باشد.

- A. این دیابت را با دستور `load_diabetes()` از این کتابخانه لود کرده، و در قالب `dataframe` آن را نمایش دهید.
*همانطور که مشاهده می‌کنید داده‌های مربوط به این دیابت نرمالایز شده‌اند.
- B. `data visualization` از مجموعه روش‌هایی است که جهت نمایش داده‌های کمی در قالب‌های مختلف گرافیکی مانند نمودارها یا نقشه‌ها، برای تحلیل هرچه بهتر داده‌های خام و شفاف‌سازی موضوع انجام می‌گیرد.
 - a. نمودار گرمایی (heat map) نشان‌دهنده میزان همبستگی (correlation) میان ستون‌های دیتافریم را نمایش دهید.
 - b. نمودار خطی (line chart) دو ستون `age`, `bmi` را در یک نمودار رسم کنید.
 - C. ستون `target` را به عنوان `target` و بقیه ستون‌ها را به عنوان `Attribute` در نظر بگیرید. با استفاده از عملیات `train_test_split` که عملیات `shuffle` را نیز انجام می‌دهد ۲۰ درصد از داده‌ها را به عنوان داده تست جدا کنید.
 - D. داده‌ها را با استفاده از الگوریتم `Linear Regression` آموزش دهید و با استفاده از مدل به دست آمده عملیات پیش‌بینی را روی داده‌های تست انجام دهید.

E. در رابطه با معیارهای ارزیابی رگرسیون مطالعه نموده و دو مورد از آن‌ها را توضیح دهید، سپس با این دو مورد نتایج پیش‌بینی شده را ارزیابی نمایید. (*استفاده از کتابخانه‌های آماده بلامانع است).

سوال دوم:

دیتاست connect-4 مربوط به بازی connect-4 است. این بازی در یک صفحه ۶*۷ بازی می‌شود که یک بازی دو نفره است. هر بازیکنی که بتواند در صفحه، ۴ نقطه خود را در یک راستا قرار دهد برنده است. بازی را می‌توانید در [این لینک](#) بازی کنید. این دیتاست ۶*۷ ستون دارد که هر کدام وضعیت یک نقطه از صفحه را نشان می‌دهد. وضعیت هر نقطه یکی از مقادیر روبرو است: {x,b,o}. مقدار 'x' نشانگر نقطه بازیکن x و 'o' هم نشانگر نقطه بازیکن o است و 'b' نیز نشان می‌دهد که این نقطه خالی است. ستون آخر نیز یکی از سه مقدار {win, loss, draw} است. صفحه بازی به شکل زیر است:

6
5
4
3
2
1
	a	b	c	d	e	f	g

A. با استفاده از کتابخانه pandas فایل connect-4.csv را load کنید.

a. ۲۰ ردیف اول دیتاست را نشان دهید و سپس با استفاده از shape تعداد سطرها و ستون‌های آن را نمایش دهید.

b. نتیجه چند درصد از بازی‌ها win بوده است؟ نتیجه چند درصد loss و چند درصد draw بوده است؟

B. با استفاده از [replace](#) مقادیر x را با ۱ و o را با -۱ و b را با صفر جایگزین کنید. ستون آخر (win) را به عنوان target جدا کنید و بقیه ستون‌ها را به عنوان Attribute در نظر بگیرید. ۲۰ درصد از داده‌ها را به عنوان داده تست در نظر بگیرید.

C. می‌خواهیم با استفاده از تکنیک kfold cross validation بهترین مدل را از بین چندین مدل انتخاب کنیم. به کمک [cross_val_score](#) (با پارامتر scoring = 'accuracy') با استفاده از داده آموزش برای هر یک از مدل‌های [SVC](#), [LinearSVC](#), [KNeighborsClassifier](#), [BernoulliNB](#), [GaussianNB](#), [DecisionTreeClassifier](#) میانگین دقت را بدست آورید و نمایش دهید. از [StratifiedKFold](#) به عنوان مدل kfold validator استفاده کنید. (خروجی cross_val_score یک آرایه از اعداد است که دقت را نشان می‌دهند).

D. با استفاده از [boxplot](#) نتایج دقت مدل‌ها را نمایش دهید.

E. بهترین مدل را انتخاب کنید و با استفاده از داده train آموزش دهید و با داده تست predict کنید. [confusion matrix](#) را برای آن نمایش دهید.

سوال سوم:

دیتاست fish مجموعه دادگان مربوط به گونه‌های مختلف ماهی‌ها می‌باشد. این دیتاست را لود کنید.

A. ستون Species را به عنوان target و مابقی ستون‌ها را به عنوان Attribute در نظر بگیرید. با استفاده از عملیات train_test_split، ۲۰ درصد از داده‌ها را به عنوان داده تست جدا کنید.

B. با استفاده از الگوریتم SVC یک دسته‌بند با کرنل موردنظر ایجاد کرده و آن را با استفاده از داده‌های train آموزش دهید.

C. برچسب دادگان تست را با مدلی که train کرده‌اید، پیش‌بینی کنید.

D. با استفاده از توابع confusion_matrix, classification_report, accuracy_score عملکرد مدل خود را ارزیابی نمایید.

E. مراحل B تا D را برای دو کرنل متفاوت انجام دهید و نتایج به دست آمده از مرحله D را مقایسه کنید. کدام کرنل در مورد این داده‌ها بهتر عمل می‌کند؟