تمرین سری سوم درس یادگیری ماشین

در این تمرین، به مباحث فصل ۵ و فصل ۶ کتاب ISLP پرداخته شده است. هدف از تنظیم و طراحی این تمرین، تسلط شما به مباحث عملی و همچنین گذری از مباحث تئوری به مباحث عملی و رایج تر است. در این تمرین، شما از زبان Python استفاده خواهید کرد. نکته مورد توجه در این تمرین، توجه شما به مفاهیم و درک نحوه حل مسئله است که از کد زدن و حتی طراحی مدل هم مهمتر است. در صورت وجود هرگونه ابهام، مشکل و یا عدم درک مسئله، در گروه درسی و یا با آیدی kodak08 در تلگرام و یا اگر به تلگرام هم دسترسی نداشتید، به ایمیل yousefisoroush1@gmail.com پیام بدهید.

سوال اول)

الف) یک مجموعه داده شبیه‌سازی شده را به صورت زیر تولید کنید:

```
rng = np.random.default_rng(1)
x = rng.normal(size=100)
y = x - 2 * x**2 + rng.normal(size=100)
```

در این مجموعه داده، مقدار n و مقدار p چقدر است؟ فرم معادله‌ای مدلی که برای تولید این داده‌ها استفاده شده است را بنویسید.

ب) نمودار پراکندگی (Scatterplot) مقادیر X را در مقابل Y رسم کنید. در مورد الگویی که مشاهده می‌کنید توضیح دهید.

ج) یک Seed تنظیم کنید و سپس خطاهای اعتبارسنجی LOOCV حاصل از برازش چهار مدل زیر را محاسبه نمایید:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \varepsilon$$

د) مراحل بخش (ج) را با استفاده از یک Seed تصادفی دیگر تکرار کرده و نتایج خود را گزارش دهید. آیا نتایج شما مشابه نتایجی است که در بخش (ج) به دست آوردید؟ دلیل آن چیست؟

ه) کدام‌یک از مدل‌های بخش (ج)، کمترین خطای LOOCV را داشت؟ آیا انتظار چنین نتیجه‌ای را داشتید؟ پاسخ خود را تشریح کنید.

و) قسمت (ج) را اینبار با روش cross validation انجام داده و خطاهای اعتبارسنجی را محاسبه کنید.

و) در مورد معناداری آماری (Statistical Significance) ضرایب تخمین‌زده شده حاصل از برازش هر یک از مدل‌های بخش (ج) با روش حداقل مربعات، توضیح دهید. آیا این نتایج با نتیجه‌گیری‌های به‌دست‌آمده بر اساس اعتبارسنجی متقابل (Cross-Validation) همخوانی دارند؟

سوال دوم)

اکنون مجموعه داده مسکن بوستون (Boston housing data set) از کتابخانه ISLP را بررسی می‌کنیم. در صورتی که موفق به دریافت دیتاست به کمک کتابخانه ISLP نشدید، میتوانید آن را از این لینک دریافت کنید.

الف) بر اساس این مجموعه داده، برآوردی برای میانگین جامعه (Population Mean) متغیر medv ارائه دهید. این برآورد را μ_hat بنامید

ب) برآوردی از خطای معیار (Standard Error) برای μ_hat ارائه دهید. این نتیجه را تفسیر کنید.

راهنمایی: می‌توانیم خطای معیار میانگین نمونه را با تقسیم انحراف معیار نمونه بر جذر (رادیکال) تعداد مشاهدات محاسبه کنیم (فرمول: σ / √n).

ج) اکنون با استفاده از روش بوت‌استرپ (Bootstrap)، خطای معیار μ_hat را برآورد کنید. این مقدار را با پاسخ خود در بخش (ب) مقایسه کنید.

ه) ر اساس برآورد بوت‌استرپ خود در بخش (ج)، یک بازه اطمینان ۹۵٪ برای میانگین medv ارائه دهید. آن را با نتایجی که با استفاده از فرمول استاندارد پایین به دست می‌آید، مقایسه کنید:

راهنمایی: می‌توانید بازه اطمینان ۹۵٪ را با فرمول تقریبی [μ_hat – 2*SE, μ_hat + 2*SE] محاسبه کنید.

و) اکنون می‌خواهیم خطای معیار μ_med را برآورد کنیم. متاسفانه فرمول ساده‌ای برای محاسبه خطای معیار میانه وجود ندارد. در عوض، با استفاده از بوت‌استرپ، خطای معیار میانه را برآورد کنید. در مورد یافته‌های خود توضیح دهید.

ز)) بر اساس این مجموعه داده، برآوردی برای صدک دهم (10th Percentile) متغیر medv در مناطق سرشماری بوستون ارائه دهید. این مقدار را μ_0.1 بنامید. **(می‌توانید از تابع np.percentile استفاده کنید.)**

ح) از روش بوت‌استرپ برای برآورد خطای معیار μ_0.1 استفاده کنید. در مورد یافته‌های خود توضیح دهید.

سوال سوم)

الف) کد زیر را به دقت بررسی و سپس اجرا کنید.

```python
import numpy as np
from sklearn.feature_selection import SelectKBest, f_classif
from sklearn.model_selection import cross_val_score
from sklearn.neighbors import KNeighborsClassifier


np.random.seed(42)
n_samples = 50
n_features = 5000
X = np.random.normal(size=(n_samples, n_features))
y = np.random.randint(0, 2, size=n_samples)


selector = SelectKBest(f_classif, k=100)
X_selected = selector.fit_transform(X, y)

# Estimate accuracy using 5-Fold Cross-Validation
knn = KNeighborsClassifier(n_neighbors=1)
scores = cross_val_score(knn, X_selected, y, cv=5)

print(f"CV Accuracy: {np.mean(scores):.2%}")
```

ب) به نظر شما، خروجی CV Accuacy این کد صحیح است؟ در مورد دلیل صحیح یا غلط بودن آن توضیح دهید.

ج) این کد را تصحیح کرده و در مورد روش خود توضیح دهید.

د) خروجی کد تصحیح شده را گزارش کنید. در مورد روش پیشگیری از این مشکل توضیح دهید.

سوال چهارم)

در این تمرین، ما تعداد درخواست‌های پذیرش دریافت شده را با استفاده از سایر متغیرهای موجود در مجموعه داده College پیش‌بینی خواهیم کرد

الف) مجموعه داده را به دو بخش مجموعه آموزش (Training set) و مجموعه آزمون (Test set) تقسیم کنید.

ب) یک مدل خطی را روی مجموعه آموزش برازش (Fit) دهید و خطای آزمون (Test Error) به دست آمده را گزارش کنید.

ج) یک مدل Ridge Regression روی مجموعه آموزش برازش دهید، به طوری که پارامتر λ با استفاده از اعتبارسنجی متقابل (Cross-Validation) انتخاب شود. خطای آزمون به دست آمده را گزارش کنید.

د) یک مدل Lasso Regression روی مجموعه آموزش برازش دهید، به طوری که پارامتر λ با اعتبارسنجی متقابل (Cross-Validation) انتخاب شود. خطای آزمون به دست آمده را به همراه تعداد ضرایب تخمین‌زده شده‌ی غیرصفر گزارش کنید.

ه) یک مدل PCR (رگرسیون مؤلفه‌های اصلی) روی مجموعه آموزش برازش دهید، که در آن M (تعداد مؤلفه‌ها) با اعتبارسنجی (Cross-Validation) متقابل انتخاب شده باشد. خطای آزمون و مقدار M انتخاب شده توسط اعتبارسنجی متقابل را گزارش کنید.

و)) یک مدل PLS (حداقل مربعات جزئی) روی مجموعه آموزش برازش دهید، که در آن M با اعتبارسنجی متقابل (Cross-Validation) انتخاب شده باشد. خطای آزمون و مقدار M انتخاب شده را گزارش کنید.

ز)در مورد نتایج به دست آمده اظهار نظر کنید. با چه دقتی می‌توانیم تعداد درخواست‌های دانشگاه دریافت شده را پیش‌بینی کنیم؟ آیا تفاوت زیادی بین خطاهای آزمون حاصل از این پنج روش وجود دارد؟ کدام مدل بهتر عمل کرده است؟ با رسم نمودار مناسب آن را شرح دهید.

سوال پنجم)

در این تمرین، ما داده‌های شبیه‌سازی شده تولید می‌کنیم و سپس از این داده‌ها برای انجام انتخاب ویژگی Forward Stepwise و Backward Stepwise استفاده خواهیم کرد.

الف) یک بردار X با ۱۰۰ نمونه بسازید که مقادیر آن از توزیع نرمال پیروی کنند. هم‌زمان، یک بردار نویز ε نیز با طول ۱۰۰ (شامل ۱۰۰ عدد تصادفی نرمال دیگر) تولید کنید.

ب) یک بردار پاسخ Y به طول n = 100 را بر اساس مدل زیر تولید کنید:

$$Y = \beta0 + \beta1X + \beta2X^2 + \beta3X^3 + \varepsilon$$

β0، β1، β2 و β3 را به صورت دلخواه قرار دهید.

ج) از روش انتخاب گام‌به‌گام رو به جلو (Forward Stepwise Selection) استفاده کنید تا مدلی شامل پیش‌بینی‌کننده‌های X, X^2, ..., X^10 را انتخاب کنید. بر اساس معیار Cp، چه مدلی به دست می‌آید؟ ضرایب مدل به دست آمده را گزارش کنید.

د) بخش (ج) را با استفاده از انتخاب گام‌به‌گام رو به عقب (Backward Stepwise Selection) تکرار کنید. پاسخ خود را با بخش ج مقایسه کنید.

ه) اکنون یک مدل لاسو (Lasso) را روی داده‌های شبیه‌سازی شده برازش دهید، و دوباره از X, X^2, ..., X^10 به عنوان پیش‌بینی‌کننده‌ها استفاده کنید. از اعتبارسنجی متقابل (Cross-Validation) برای انتخاب مقدار بهینه λ استفاده کنید. نمودارهای خطای اعتبارسنجی متقابل را به عنوان تابعی از λ رسم کنید. تخمین‌های ضرایب حاصل را گزارش کنید و نتایج به دست آمده را توضیح دهید.

و) یک بردار پاسخ Y بر اساس مدل روبرو تولید کنید: $Y = \beta0 + \beta7X^7 + \varepsilon$ و سپس انتخاب گام‌به‌گام رو به جلو و لاسو را انجام دهید. در مورد نتایج به دست آمده توضیح دهید.

سوال ششم) (امتیازی)

در این سوال، هدف ما پیش‌بینی نرخ جرم و جنایت سرانه (ستون ecrim) در مجموعه داده Boston است. برای حل این سوال، میتوانید یکی از دو روش پیشنهادی در زیر و یا ترکیبی از آن دو را انجام دهید. روش اول،روش عادی و با استفاده از پایتون و در روش دوم از ابزار Zerve میتوانید استفاده کنید. پیشنهاد می شود که با روش دوم سوال را حل کنید تا با ابزار Zerve نیز آشنایی پیدا کنید.
نکته : انجام این تسک خواسته شده به تنها یک روش کافی است (انجام هر دو روش نمره ای ندارد).

روش اول)

الف)چند مورد از روش‌های رگرسیونی که در این فصل آموخته‌اید را روی این داده‌ها پیاده‌سازی کنید. پیشنهاد می‌شود از روش‌هایی مانند:

- انتخاب بهترین زیرمجموعه (Best Subset Selection)
- رگرسیون لاسو (Lasso)
- رگرسیون ریج (Ridge Regression)
- رگرسیون مؤلفه‌های اصلی (PCR)

استفاده کنید. سپس نتایج به دست آمده (مانند خطای تست) را برای هر روش گزارش کرده و با هم مقایسه کنید.

ب) برای توجیه انتخاب خود، حتما باید از معیارهای معتبر مانند خطای اعتبارسنجی متقابل (Cross-Validation Error) یا خطای مجموعه آزمون (Test Set Error) استفاده کنید. به هیچ وجه بر اساس خطای آموزش (Training Error) تصمیم‌گیری نکنید.

ج) مدل انتخابی خود را به نحوی توضیح دهید که پاسخ سوالات زیر را در بر داشته باشد.

- آیا مدلی که در بخش (ب) انتخاب کردید، شامل تمام ویژگی‌های موجود در دیتاست است یا فقط از تعدادی از آن‌ها استفاده می‌کند؟
- اگر تمام ویژگی‌ها را دارد، چرا؟
- اگر برخی ویژگی‌ها حذف شده‌اند، دلیل آن چیست؟ (حذف متغیرهای کم‌اهمیت یا ویژگی‌های روش‌هایی مثل لاسو).

روش دوم)

الف) بارگذاری داده و آماده‌سازی :

- در محیط Canvas ابزار Zerve، یک بلوک (Block) اولیه ایجاد کنید
- داده‌های Boston را بارگذاری کنید.
- تقسیم‌بندی Train/Test را انجام دهید.
- خروجی این بلوک باید متغیرهای X_train, X_test, y_train, y_test باشد که به عنوان ورودی به تمام بلوک‌های بعدی ارسال می‌شود.

ب)مدلسازی موازی (Parallel Modeling):

به جای نوشتن کدها پشت سر هم، از قابلیت Branching در Zerve استفاده کنید و سه بلوک جداگانه که به طور موازی از بلوک ریشه تغذیه می‌شوند، بسازید:

- بلوک اول (Best Subset): الگوریتم انتخاب بهترین زیرمجموعه را اجرا کنید
- بلوک دوم (Regularization): در این بلوک دو زیر-شاخه ایجاد کنید:یکی برای Ridge (با انتخاب λ توسط CV).یکی برای Lasso (با انتخاب λ توسط CV).
- بلوک سوم (Dimension Reduction): مدل PCR را اجرا کرده و تعداد کامپوننت‌های بهینه (M) را پیدا کنید.

ج) تجمیع و مقایسه نتایج (Fusion & Comparison):

یک بلوک نهایی به نام Results Comparison ایجاد کنید که خروجی‌های (MSE و R^2) تمام بلوک‌های بالا را به عنوان ورودی دریافت کند. سپس یک دیتافریم بسازید که خطای تست تمام مدل‌ها را کنار هم نشان دهد.بهترین مدل را مشخص کنید.توضیح دهید که چگونه معماری DAG (گراف جهت‌دار) در Zerve به شما کمک کرد تا از نشت داده (که در نوت‌بوک‌های خطی رایج است) جلوگیری کنید؟

برای انجام این سوال، مشاهده ویدیوی موجود در این لینک میتواند به شما کمک کند.

نکات تحویل :

فایل پاسخ خود را که شامل یک نوتبوک و یک گزارش با موارد خواسته شده است را به صورت :

[HW3_Student Number_ First name_Last name].zip

بارگذاری کنید.