تمرین سری چهارم

مهلت تحویل: پایان روز ۱۶ آذر ۱۴۰۳

طراح تمرین: تکین جزایری ـ مهدی شفیعی

نکات ضروری

۱. لازم است کد تمرین به همراه گزارش تحویل داده شود که شامل پاسخ سوالات نظری، توضیحات و تحلیل کد و تصاویر و نمودارهای لازم می باشد.

۲. کد تمرین باید در زبان پایتون، به صورت notebook و شامل نتایج اجرا باشد.

۳. فایل گزارش باید به فرمت pdf باشد. حجم گزارش به هیچ عنوان معیار نمره دهی نیست، لذا لطفا در حد نیاز توضیح دهید.

۴. استفاده از کتابخانه های از پیش آماده مجاز است، مگر اینکه نقیض آن گفته شود. در گزارش کتابخانه ها ذکر شوند.

۵. نمره هر تمرین از ۱۰۰ است (مگر آنکه نمره امتیازی در تمرین مشخص شده باشد). تا سه روز بعد از مهلت اعلام شده برای تمرین امکان بارگذاری وجود دارد اما به ازای هر روز ۱۰ درصد از نمره تمرین کسر خواهد شد.

۶. مشاهده شباهت نامتعارف در گزارش و کد به منزله تقلب می باشد و طرفین مشمول کسر نمره خواهند شد.

۷. برای این تمرین، می‌توانید گزارش و پاسخ‌هایتان را به زبان‌های فارسی و یا انگلیسی تحویل دهید.

مسئله ۱.

مجموعه‌داده زیر شامل ۱۲ نمونه است که هر کدام دارای دو ویژگی و یک برچسب هستند.

(آ) با استفاده از این مجموعه‌داده و دسته‌بند بیز ساده[1]، برچسب داده $(X_۱, X_۲) = (۰, ۷)$ را تخمین بزنید. مراحل و محاسبات انجام شده را تشریح کنید.

No.	X_1	X_2	Y	No.	X_1	X_2	Y
1	0.6	6.1	1	7	-1.3	7.6	2
2	-1.8	5.5	1	8	1.3	7.9	3
3	-0.8	6.1	1	9	1.1	4.9	3
4	0	6.3	1	10	1.1	6.3	3
5	-0.3	7.8	2	11	2.9	7.1	3
6	-0.8	7.1	2	12	1.1	6.3	3

[۱۰ نمره]

(ب) فرض کنید برخلاف قسمت (الف)، هزینه متفاوتی برای دسته‌بندی‌های اشتباه[2] در نظر بگیریم. ماتریس هزینه[3] را که در آن درایه i, j نشان‌دهنده هزینه تشخیص کلاس iام برای نمونه‌های کلاس jام است، به صورت در نظر بگیرید. این ماتریس بیانگر چه تغییر رویکردی در حل این مسئله دسته‌بندی است؟
با اعمال این ماتریس هزینه، برچسب داده $(X_۱, X_۲) = (۰, ۷)$ را مجدداً تخمین بزنید. مراحل و محاسبات انجام شده را تشریح کنید.

$$L = \begin{bmatrix} 0 & 1 & 3 \\ 1 & 0 & 3 \\ 1 & 1 & 0 \end{bmatrix}$$

[۱۰ نمره]

مسئله ۲.

هدف این مسئله، پیاده سازی الگوریتم‌های خوشه بندی بدون استفاده از کتابخانه‌هایی است که آن‌ها را به صورت آماده در اختیار می‌گذارند. سه مجموعه‌داده xclara، engytime و target در فضای دوبعدی به پیوست این تمرین در اختیار شما قرار داده شده‌است.[4] نمودار نقطه‌ای[5] آن‌ها را بدون استفاده از برچسب‌های موجود رسم کنید. (در تمامی قسمت‌های این مسئله، نتایج نهایی، نمودارهای رسم‌شده و تحلیل‌ها را در گزارش بیاورید.)

[1] Naive Bayes Classifier
[2] Misclassifications
[3] Loss Matrix

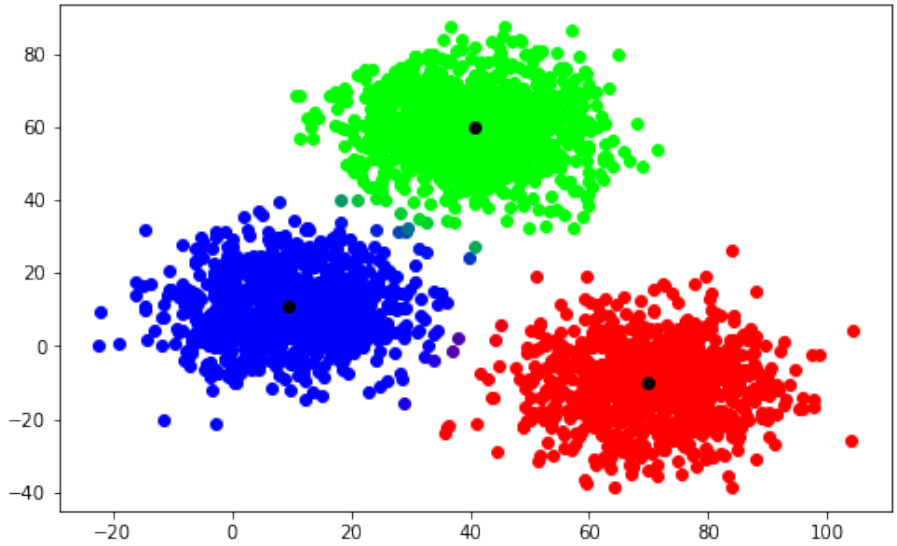
[4] منبع مجموعه‌داده‌ها:
https://github.com/milaan9/Clustering-Datasets

[5] Scatter Plot

(آ) الگوریتم k-mean را بدون استفاده مستقیم از کتابخانه‌ها پیاده‌سازی کنید. سپس برای هر مجموعه‌داده به ازای k از ۱ تا ۱۰، این الگوریتم را اجرا کنید و نتیجه را به صورت یک نمودار نقطه‌ای ترسیم کنید که در آن، داده‌های هر کلاس با رنگ‌های متفاوت از یکدیگر جدا شده باشند. در نهایت، برای هر مجموعه‌داده نمودار مقدار نهایی تابع هزینه (J) نسبت به k ترسیم کنید و بر اساس آن مناسب‌ترین مقدار برای k را تعیین کنید.

[۱۵ نمره]

(ب) الگوریتم EM برای مدل ترکیبی گاوسی[۶](GMM) را بدون استفاده مستقیم از کتابخانه‌ها پیاده‌سازی کنید و آن را روی مجموعه‌داده‌های xclara (با سه خوشه)، engytime (با دو خوشه) و target (با دو و سه خوشه) اجرا کنید. پس از هر یک از اجراها، نمودار نقطه‌ای داده‌ها را ترسیم کنید به طوری که هر خوشه دارای یک رنگ خاص باشد و هر داده بسته به احتمال تعلق به هر خوشه، ترکیبی از رنگ خوشه‌ها را داشته باشد. به طور مثال، نتیجه می‌تواند مشابه تصویر زیر باشد:



[۲۰ نمره]

(ج) (امتیازی) علاوه بر روش‌های خوشه‌بندی k-mean، GMM و سلسله‌مراتبی که با آن‌ها آشنا شده‌اید، الگوریتم مشهور دیگری به نام DBSCAN[۷] برای خوشه‌بندی در سال ۱۹۹۶ ارائه شده است که ایده آن متفاوت از روش‌های ذکرشده است. این روش یک الگوریتم مبتنی بر چگالی[۸] است؛ یعنی نقاط نزدیک به هم را در یک خوشه قرار می‌دهد و نقاطی را نیز که نزدیک به نقاط دیگر نیستند و در منطقه کم‌چگالی قرار گرفته‌اند، داده پرت در نظر می‌گیرد.
این الگوریتم را بدون استفاده مستقیم از کتابخانه‌ها پیاده‌سازی کنید و با تغییر پارامترهای مدل، تلاش کنید به نتایج مناسبی روی هر یک از سه مجموعه‌داده برسید. برای هر اجرا، نمودار نقطه‌ای نمونه‌ها را ترسیم کنید.

[۲۰ نمره]

مسئله ۳.

در این تمرین شما باید روی مجموعه‌داده‌ای کار کنید که از طریق فایل Loan_Data.csv به شما داده شده‌است و شامل اطلاعاتی درباره متقاضیان بیمه است. هدف نهایی این است که پیش‌بینی کنید آیا بیمه به یک متقاضی تعلق می‌گیرد یا خیر. این تمرین شامل مراحل آشنایی داده و تحلیل کاوشگرانه داده[۹](EDA)، پیش‌پردازش داده و پیاده‌سازی یک مدل رگرسیون لجیستیک[۱۰] است.

[6] Gaussian Mixture Models
[7] Density-Based Spatial Clustering of Applications with Noise
[8] Density-Based
[9] Exploratory Data Analysis
[10] Logistic Regression

(آ) وجود مقادیر گمشده را بررسی کنید. سطر های تکراری را در صورت وجود حذف کرده و ۱۰ سطر اول آن را چاپ کنید.

[۵ نمره]

(ب) تحلیل کاوشگرانه داده:

(i) تعداد مقادیر یکتا در ستون property_area را محاسبه کرده و نمایش دهید.

(ii) ماتریس همبستگی [۱۱] را رسم کنید. (می‌توانید از نقشه گرمایی[۱۲] استفاده کنید.)

(iii) برای بررسی رابطه بین ویژگی‌های Gender و Loan_Status نمودار میله‌ای رسم کنید.

(iv) یک نمودار دایره‌ای برای ستون Education رسم کنید.

(v) هیستوگرام ستون ApplicantIncome را با ده bin رسم کنید.

[۱۰ نمره]

(ج) پیش‌پردازش: تمام پیش‌پردازش‌های لازم (حذف مقادیر null، نرمالسازی، تبدیل داده‌های کیفی و ...) را انجام دهید و داده‌ها را به ۲۰ درصد برای آموزش و ۸۰ درصد برای تست تقسیم کنید.

[۵ نمره]

(د) با استفاده از کتابخانه‌های آماده، یک مدل رگرسیون لجیستیک آموزش دهید و ماتریس درهم‌ریختگی[۱۳] را محاسبه کنید.

[۱۰ نمره]

(ه) مدل رگرسیون لجیستیک را با استفاده از قالب زیر بپیاده‌سازی کنید. در کد زیر باید توابع loss، loss_derivate، predict و حلقه for در تابع fit را کامل کنید. سپس precision و recall و f1-score را بر روی داده‌های تست محاسبه کنید.

```
import numpy as np
class GDLogisticRegression:
    def __init__(self, n_features, max_iter=50000, lr=0.0001, tol=1e
       -6, momentum=0.9):
        self.N = n_features
        self.beta = np.zeros((self.N+1,))
        self.max_iter = max_iter
        self.lr = lr
        self.tol = tol
        self.momentum = momentum

    def loss(self, X, y):
        pass

    def loss_derivative(self, X, y):
        pass
```

[11] Correlation Matrix
[12] Heatmap
[13] Confusion Matrix

```python
    def predict(self, X_test, threshold=0.5):
        pass

    def fit(self, X_train, y_train):
        X_train_new = np.concatenate((X_train, np.ones((X_train.shape
            [0], 1))), axis=1)
        last_loss = 0
        momentum = 0

        for _ in range(self.max_iter):
            pass
```

[۱۵ نمره]

[بارم کل: ۲۰ + ۱۰۰ نمره]

با آرزوی موفقیت