تمرینات کلاسی هفته ششم - سری اول

==========

10 سوال سرعتی

سؤال ۱ — قاعده‌ی بیز
فرض کنید P(girl) = 0.6 و P(boy) = 0.4. در نقطه‌ی x، چگالی‌ها این‌طور اندازه‌گیری شده‌اند:
p(x|boy) = 0.2 و p(x|girl) = 0.5
مدل بیز در این نقطه چه کلاسی را پیش‌بینی می‌کند؟

سؤال ۲ — رگرسیون خطی
یک مدل معادله‌ی y_hat = 10 + 2x را یاد گرفته. برای x=4، مدل چه چیزی پیش‌بینی می‌کند؟ اگر مقدار واقعی y برابر ۲۱ باشد، باقی‌مانده (residual) چقدر است؟

سؤال ۳ — یک KNN خیلی ساده
با numpy دو گروه داده بسازید: گروه اول حول عدد ۲ و گروه دوم حول عدد ۸ (هرکدام ۵۰ نمونه، انحراف‌معیار ۱). یک KNeighborsClassifier با k=3 آموزش دهید و دقتش را روی همان داده چاپ کنید.

سؤال ۴ — مقایسه‌ی چگالی با scipy
با scipy.stats.norm، چگالی x=5 را برای دو توزیع N(4,1) و N(7,1) حساب کنید و چاپ کنید کدام‌یک بزرگ‌تر است.نکته: چون x=5 به میانگین گروه دختر (۴) نزدیک‌تر است تا میانگین گروه پسر (۷)، انتظار داریم چگالی دختر بزرگ‌تر باشد.

سؤال ۵ — برازش رگرسیون خطی
داده‌ی مصنوعی حول رابطه‌ی y = 3x + 5 (با کمی نویز) بسازید و یک LinearRegression رویش برازش دهید. شیب و عرض از مبدأ یادگرفته‌شده را چاپ کنید.

سؤال ۶ — یک رگرسیون لجستیک ساده
یک داده‌ی دوکلاسه (شبیه سؤال ۳) بسازید و یک LogisticRegression رویش آموزش دهید. دقت مدل را چاپ کنید.

سؤال ۷ — پیاده‌سازی تابع سیگموید
تابعی به نام sigmoid بنویسید که مقدار 1 / (1 + e^-z) را برمی‌گرداند و آن را برای z = -5, 0, 5 تست کنید.

سؤال ۸ — امتحان چند مقدار k در KNN
با داده‌ی سؤال ۳، سه مدل KNN با k=1، k=5، و k=20 بسازید و دقت هرکدام را چاپ کنید.

سؤال ۹ — مرور آمار توصیفی با pandas/numpy
یک آرایه از ۱۰ عدد دلخواه به‌عنوان «سطح هورمون» بسازید و میانگین و انحراف‌معیارش را چاپ کنید.

سؤال ۱۰ — ماتریس درهم‌ریختگی (Confusion Matrix)
روی مدل رگرسیون لجستیک سؤال ۶، ماتریس درهم‌ریختگی را چاپ کنید و بگویید مدل چند مورد را درست و چند مورد را اشتباه پیش‌بینی کرده.

==========

راند ۱ — فکر کردن به سبک بیز

۱.۱ — بحث کنید (نیازی به محاسبه نیست). یک هواشناس می گوید: «طبق آمار سال های قبل، در فصل باران ۸۰٪ روزها تا عصر باران می بارد.» این باور اولیه اوست، قبل از این که به بیرون نگاه کند.

بعد سرش را بلند می کند و ابرهای سیاه و سنگین می بیند. حالا تخمینش از «احتمال بارش امروز» به ۹۵٪ می رسد.
به آن باور اولیه ۸۰٪ اصطلاحا چه می گویند؟
به این باور به روزشده ۹۵٪ (بعد از دیدن شواهد) اصطلاحا چه می گویند؟
در یک جمله، با کلمات خودتان توضیح دهید چه چیزی عدد اول را به عدد دوم تبدیل می کند.

۱.۲ — درست یا غلط (دلیل بیاورید). «احتمال این که یک دختربچه به طور تصادفی انتخاب شده، سطح تاو-هورمونش دقیقا ۴.۰۰۰۰۰ باشد، ۷۰٪ است.»

۱.۳ — فرض کنید در سطح هورمونی x=5، چگالی های شرطی زیر را (ساده شده، فقط برای این تمرین) اندازه گیری کرده ایم:
p(x=5 | کلاس)
دختر: ۰.۲۸
پسر: ۰.۱۵
احتمال های پیشین را هم به یاد بیاورید:
P(دختر) = 0.7 ، P(پسر) = 0.3.
مقادیر وزن دار
P(دختر) * p(x | دختر) و P(پسر) * p(x | پسر)
را حساب کنید.
قاعده بهینه بیز در x=5 کدام کلاس را پیش بینی می کند؟ چرا؟

==========

راند ۲ — آشنایی با KNN

۲.۱ — در دو یا سه جمله، به هم گروهی هایتان (به زبان ساده، بدون فرمول) توضیح دهید KNN چطور کلاس یک نقطه جدید را تعیین می کند.
۲.۲ — دستی حساب کنید. این هم یک مجموعه آموزشی کوچک از نوزادان و سطح تاو-هورمون شان:

سطح هورمون	برچسب
۳.۸	دختر
۴.۲	دختر
۵.۰	دختر
۵.۹	پسر
۶.۵	پسر
۷.۰	پسر
۷.۸	پسر

یک نوزاد جدید x=6.2 اندازه گیری شده. با k=3:
سه نقطه آموزشی نزدیک تر را لیست کنید (فاصله ها را نشان دهید).
رای اکثریت چیست؟ KNN چه چیزی پیش بینی می کند؟

۲.۳ — ارتباط را پیدا کنید. در نوت بوک گفته شده بود: «اگر k_دختر تا از k همسایه نزدیک شما دختر باشند، آنگاه k_دختر / k تقریبا برابر است با P(دختر | x).» با کلمات خودتان توضیح دهید چرا تراکم بیشتر نقاط دختر نزدیک x به طور طبیعی باعث می شود دختر بیشتری بین نزدیک ترین همسایه ها ظاهر شود. (راهنمایی: به این فکر کنید که «چگالی بالا» و «پیشین بالا» واقعا چه تاثیری روی تعداد نقاطی که در آن نزدیکی قرار می گیرند دارند.)

۲.۴ — جای خالی را پر کنید (موازنه بایاس-واریانس). جدول را با هم تکمیل کنید:
مقدار k | رفتار مدل | بیش برازش یا کم برازش؟
k = 1 | _____ | _____
k ی متوسط (مثلا ۱۵) | _____ | _____
k = n (همه نقاط آموزشی) | همیشه _____ را پیش بینی می کند | _____

==========

راند ۳ –

یک نوت بوک/اسکریپت تازه باز کنید. این ایمپورت ها را اجرا کنید:

```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
np.random.seed(42)
```

۳.۱ – دو توزیع زیر را تعریف کنید:

```
girl_dist = stats.norm(4.0, 1.2)
boy_dist = stats.norm(7.0, 1.5)
```

هر دو تابع چگالی احتمال را روی

```
x = np.linspace(0, 12, 1000)
```

رسم کنید، درست مثل کلاس.

۳.۲ – حالا فرض کن میانگین پسرها به‌جای ۷، برابر ۶ باشه (یک توزیع جدید بساز، توزیع قبلی رو دستکاری نکن). دوباره هر دو منحنی رو رسم کن.
بین این دو حالت (میانگین ۷ در مقابل میانگین ۶)، در کدوم حالت دو منحنی هم‌پوشانی بیشتری دارن؟ با توجه به این هم‌پوشانی، انتظار داری خطای بهینه‌ی بیز (Bayes error) در کدوم حالت بیشتر باشه؟ توضیح بده چرا.

۳.۳ – مرز تصمیم را به صورت عددی پیدا کنید. با استفاده از توزیع های اصلی (میانگین پسر = ۷)، کدی بنویسید که:

۱. g = girl_dist.pdf(x) * P_GIRL و b = boy_dist.pdf(x) * P_BOY

رو حساب کنه (با `P_GIRL=0.7`, `P_BOY=0.3`)

نقطه‌ای که `g` برای اولین بار از `b` کمتر می‌شه رو پیدا کنه و مقدارش رو چاپ کنه

۳.۴ – `n = 5000` نوزاد بساز (۷۰٪ دختر و 30 % پسر، آرایه‌ی `X` (ستونی) و `y`، عدد 0 یعنی دختر, ۱ یعنی پسر). داده را به آموزش/آزمون تقسیم کنید (۷۰/۳۰)، و یادتان باشد stratify=y را بگذارید.

یک KNeighborsClassifier با n_neighbors=15 آموزش دهید.
دقت روی داده آزمون را چاپ کنید.
آیا انتظار دارید این دقت کامل (۱.۰) باشد؟ چرا؟

۳.۵ — کد زیر خطا می دهد. دو باگ را پیدا کرده و درستش کنید.

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
X = girl_dist.rvs(3500) # آرایه یک بعدی از سطوح هورمونی
y = [0]*3500
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
knn = KNeighborsClassifier(n_neighbors=15)
knn.fit(X_train, y_train)
```

(راهنمایی: دو ایراد جدا از هم اینجا هست — یکی درباره شکلی که scikit-learn انتظار دارد، و یکی درباره این که آیا اصلا این مجموعه داده چیز مفیدی برای یاد دادن به یک طبقه بند دارد یا نه.)

==========

راند ۴ — مبانی رگرسیون خطی

۴.۱ — پنج دانشجو ساعت مطالعه هفتگی و نمره امتحانشان را در جدول زیر گزارش کرده اند:

ساعت مطالعه (x)	نمره (y)
2	65
4	70
6	75
8	85

10	90

با استفاده از فرمول های کلاس:

w1 = sum((xi - mean(x)) * (yi - mean(y))) / sum((xi - mean(x))^2)

w0 = mean(y) - w1 * mean(x)

mean(x) و mean(y) را حساب کنید.

w1 و w0 را حساب کنید.

معادله برازش شده نهایی y_hat = w0 + w1 * x را بنویسید.

۴.۲ –توضیح دهید w1 و w0 در بافت این مسئله (ساعت مطالعه و نمره امتحان) به چه معنایند.

۴.۳ – با استفاده از معادله خودتان در ۴.۱، نمره دانشجویی را که ۶ ساعت مطالعه کرده پیش بینی کنید. این پیش بینی را با نمره واقعی گزارش شده برای ۶ ساعت در جدول بالا مقایسه کنید. باقی مانده (residual) چقدر است؟ آیا مدل برای این دانشجو بیش از حد یا کمتر از حد پیش بینی می کند؟

۴.۴ – برای یک مجموعه داده دیگر، SSres = 50 و SStot = 500 است. R^2 را حساب کنید.
به زبان ساده، این مقدار چه چیزی درباره‌ی کیفیت مدل می گوید؟

۴.۵ – چرا در تابع هزینه (MSE) به جای این که باقی مانده های خام را مستقیم با هم جمع کنیم، آنها را به توان دو می رسانیم؟ (راهنمایی: فکر کنید چه اتفاقی می افتد اگر بعضی باقی مانده ها مثبت و بعضی منفی باشند.)

==========

راند ۵ – برازش یک خط

۵.۱ –دکتر لیند، دانه‌های قهوه را برشته می کند و می خواهد درصد کاهش وزن حین برشته شدن را از روی زمان برشته شدن (دقیقه) پیش بینی کند. داده مصنوعی زیر را بسازید:

```
import numpy as np
```

```
np.random.seed(0)
n = 100
roast_time = np.random.uniform(8, 18, n)
weight_loss = 2.0 + 0.9 * roast_time + np.random.normal(0, 1.5, n)
```

نمودار پراکندگی roast_time در مقابل weight_loss را رسم کنید.
یک مدل LinearRegression برازش دهید (یادتان باشد: roast_time باید به شکل ستونی درآید).
عرض از مبدا و شیب یادگرفته شده، مقدار MSE، و مقدار R^2 را چاپ کنید.

۵.۲ — داده‌ی خام رو به‌صورت پراکندگی رسم کنید، خط برازش شده را رویش بکشید.

۵.۳ — ۵.۱ را دوباره انجام دهید، اما این بار نویز را به

```
np.random.normal(0, 4.0, n)
```

تغییر دهید (داده پرنویزتر).
R^2 نسبت به ۵.۱ چطور تغییر می کند؟

۵.۴ — از جدول ۴.۱ آرایه های X و y بسازید، یک LinearRegression رویشان برازش دهید، آیا شیب و عرض از مبدأیی که sklearn پیدا می‌کنه، با چیزی که در ۴.۱ دستی حساب کردید یکیه؟

==========

راند ۶ — رگرسیون لجستیک و تابع سیگموید

۶.۱ — یک خط w0 + w1 * x می تواند هر عدد حقیقی ای، از منفی بینهایت تا مثبت بینهایت بدهد. اگر بخواهیم یک احتمال پیش بینی کنیم، این چرا مشکل ساز است؟ رگرسیون لجستیک برای حل این مشکل چه کار می کند؟

۶.۲ — سیگموید

```
sigma(z) = 1 / (1 + e^-z)
```

را به یاد بیاورید:
sigma(0) دقیقا چقدر است؟

sigma(-2) به ۰ نزدیک تر است یا به ۱؟ با آستانه معمول ۰.۵، این چه کلاسی را پیش بینی می کند؟
sigma(3) به ۰ نزدیک تر است یا به ۱؟ این چه کلاسی را پیش بینی می کند؟

۶.۳ — یک مدل رگرسیون لجستیک روی داده تاو-هورمون برازش داده شده و w1 = 1.2 ، w0 = -7.26 به دست آمده.
مقدار x ای را حل کنید که در آن w0 + w1 * x = 0 (این همان مرز تصمیم است، جایی که مدل دقیقا ۵۰/۵۰ پیش بینی می کند).
این مقدار را با مرز بهینه بیز که در راند ۳.۲ به صورت عددی حساب کردید مقایسه کنید. چه چیزی متوجه می شوید؟

۶.۴ — درست یا غلط (هرکدام را توجیه کنید).
«رگرسیون لجستیک، درست مثل رگرسیون خطی، یک راه حل بسته (closed-form) دارد.»

==========

راند ۷ — امتیازی - طبقه بندی با رگرسیون لجستیک

۷.۱ — با همان سبک مجموعه داده راند ۳.۳ (با n = 2000 دوباره بسازید، stratify=y را در تقسیم بندی فراموش نکنید)، یک LogisticRegression آموزش دهید.
w1 ، w0 ، و مرز تصمیم به دست آمده x* = -w0 / w1 را چاپ کنید.
دقت روی داده آزمون را چاپ کنید.

۷.۲ — این کد رو اجرا کن تا منحنی پسینِ واقعی ساخته بشه:

```
grid = np.linspace(0, 12, 500).reshape(-1, 1)
num = P_BOY * boy_dist.pdf(grid.ravel())
true_post = num / (num + P_GIRL * girl_dist.pdf(grid.ravel()))
```

true_post را در کنار خروجی

```
predict_proba(grid)[:, 1]
```

مدل رگرسیون لجستیک، روی همان محورها رسم کنید.
این دو منحنی چقدر به هم شبیه اند؟

۷.۳ — ماتریس درهم ریختگی (confusion matrix) و classification_report را برای داده آزمون تان حساب و چاپ کنید.
کدام کلاس (دختر یا پسر) recall بالاتری دارد؟ می توانید دلیلش را با توجه به نامتوازن بودن کلاس ها (۷۰/۳۰) توضیح دهید؟
۷.۴ — دقت روی داده آزمون مدل KNN تان (راند ۳.۳، با k=15) را با مدل رگرسیون لجستیک (۷.۱) روی همین نوع داده مقایسه کنید.
کدام یک بهتر عمل کرد، یا نتایج نزدیک بودند؟

==========

راند ۸ — سبد ترکیبی

۸.۱ — برای هر سناریوی زیر، تصمیم بگیرید کدام ابزار بهترین گزینه است: رگرسیون خطی، رگرسیون لجستیک، KNN
الف) پیش بینی قیمت فروش یک خانه از روی متراژ و تعداد اتاق خواب.
ب) پیش بینی این که آیا یک تراکنش ورودی تقلبی است یا نه، در حالی که یک تخمین احتمالی هم می خواهیم.
ج) طبقه بندی یک میوه جدید به عنوان رسیده یا نارس، با مقایسه رنگ و سفتی اش با میوه های مشابه از قبل برچسب خورده.
د) تخمین فروش بستنی ماه آینده از روی میانگین دمای روزانه.

۸.۲ — یک مدل با واریانس بالا با داده‌ی آموزشی چیکار می‌کنه، و روی داده‌ی جدید چطور عمل می‌کنه؟

یک مدل با بایاس بالا چه مشکلی داره، و روی داده‌ی آموزشی و جدید چطور عمل می‌کنه؟

۸.۳ — درست یا غلط
الف) «KNN در زمان آموزش، یک معادله ریاضی ثابت یاد می گیرد، شبیه رگرسیون خطی.»
ب) «R^2 برابر با صفر همیشه یعنی مدل کاملا بی فایده است و هیچ اطلاعاتی نمی دهد.»

==========

راند ۹ — امتیازی _ دستگاه جداسازی قهوه دکتر لیند

دکتر لیند حالا می خواهد دانه‌های قهوه را به صورت خودکار به دو دسته برشته شده یا سبز (برشته نشده) با استفاده از یک سنجه پیوسته به نام شاخص آروما دسته بندی کند. طبق بچ های قبلی:
۵۵٪ دانه‌ها برشته اند -> P(برشته) = 0.55
۴۵٪ دانه‌ها سبزند -> P(سبز) = 0.45
دانه‌های برشته: شاخص آروما دارای توزیع نرمال با میانگین ۸ و انحراف معیار ۱.۰
دانه‌های سبز: شاخص آروما دارای توزیع نرمال با میانگین ۵ و انحراف معیار ۱.۳
این را به صورت یک خط لوله (pipeline) کامل و کوچک جلو ببرید.

۹.۱ — با استفاده از scipy.stats.norm، هر دو توزیع را تعریف کنید و دو چگالی شرطی کلاس را روی یک بازه معقول (مثلا np.linspace(0, 14, 1000)) رسم کنید.

۹.۲ — آستانه بهینه بیز را پیدا کنید. با همان روش عددی راند ۳.۲، مقدار شاخص آروما ای را پیدا کنید که در آن دو منحنی پسین وزن دار همدیگر را قطع می کنند. این بهترین آستانه ممکن برای جداسازی دانه‌هاست — هیچ الگوریتمی نمی تواند از آن بهتر عمل کند.

۹.۳ — مجموعه داده را بسازید و تقسیم کنید. n = 4000 دانه را با نسبت های متناظر با پیشین های بالا شبیه سازی کنید، X (به شکل ستونی) و y (۰ = سبز، ۱ = برشته، یا برعکس — فقط سازگار باشید) بسازید، و به آموزش/آزمون (۷۵/۲۵، لایه بندی شده) تقسیم کنید.

۹.۴ — دو طبقه بند آموزش دهید.
سه مدل KNN با k=25 ، k=1 ، و k=500 آموزش دهید. دقت آزمون هر کدام را گزارش کنید و الگویی که می بینید را با توجه به بحث بایاس-واریانس راند ۸.۲ به طور خلاصه توضیح دهید.
یک مدل رگرسیون لجستیک آموزش دهید. w1 ، w0 ، مرز تصمیم به دست آمده، و دقت آزمون را گزارش کنید.

۹.۵ — ارزیابی کنید و پیشنهاد بدهید.

ماتریس درهم ریختگی و گزارش طبقه بندی را برای بهترین مدل KNN تان و برای مدل رگرسیون لجستیک چاپ کنید.
مرز تصمیم هر دو مدل را با آستانه واقعی بهینه بیز از ۹.۲ مقایسه کنید. کدام یک نزدیک تر بود؟

==========

راند ۱۰ - امتیازی - کار با csv واقعی

۱۰.۱ - ساخت دیتاست (یک بار، برای همه سوالات این راند)
این کد را اجرا کنید تا یک فایل csv شبیه داده واقعی بیمارستان بسازد (سطح تاو-هورمون و چند ویژگی دیگر برای نوزادان):

```
import numpy as np
 import pandas as pd
 from scipy import stats

np.random.seed(7)
 n = 800
 P_GIRL = 0.65

is_girl = np.random.rand(n) < P_GIRL
 hormone = np.where(is_girl, stats.norm(4.0, 1.2).rvs(n), stats.norm(7.0, 1.5).rvs(n))
 birth_week = np.where(is_girl, np.random.normal(38, 1.5, n), np.random.normal(39, 1.3, n))
 weight_kg = np.where(is_girl, np.random.normal(3.1, 0.4, n), np.random.normal(3.4, 0.4, n))

df = pd.DataFrame({
 "hormone_level": hormone,
 "birth_week": birth_week,
```

```
"weight_kg": weight_kg,
"label": np.where(is_girl, "girl", "boy")
})
df.to_csv("newborns.csv", index=False)
print(df.head())
```

۱۰.۲ - آشنایی با داده (pandas)

فایل newborns.csv را با pd.read_csv بخوانید.

با df.describe() آمار توصیفی هر ستون عددی را ببینید.

با df.groupby("label")["hormone_level"].agg(["mean","std"]) میانگین و انحراف‌معیار سطح هورمون را به تفکیک کلاس چاپ کنید.

سوال: آیا میانگین و انحراف‌معیاری که از داده به‌دست می‌آید با پارامترهای واقعی توزیع‌هایی که در راند ۱۰.۱ استفاده کردیم (۴و۱.۲ برای دختر، ۷و۱.۵ برای پسر) نزدیک است؟ چرا کاملاً یکسان نیست؟

۱۰.۳ - طبقه‌بندی با KNN روی چند ویژگی (pandas + sklearn)

از df، دو ستون hormone_level و weight_kg را به‌عنوان ویژگی (X) و ستون label را به‌عنوان برچسب (y، پس از تبدیل به عدد) استفاده کنید. یک KNeighborsClassifier با k=10 آموزش دهید و دقت را روی داده آزمون چاپ کنید. سپس این کار را فقط با ستون hormone_level (تک‌ویژگی) هم انجام دهید و دو دقت را مقایسه کنید.

۱۰.۴ - رگرسیون خطی روی csv واقعی

با همان df، فرض کنید می‌خواهیم weight_kg را از روی birth_week پیش‌بینی کنیم (بدون توجه به جنسیت).

یک LinearRegression بین birth_week (ویژگی) و weight_kg (هدف) برازش دهید، شیب و عرض از مبدأ و R^2 را چاپ کنید.

سپس همین کار را جدا برای گروه دختر و گروه پسر (با df[df.label=="girl"] و df[df.label=="boy"]) انجام دهید.

سوال: آیا رابطه بین هفته تولد و وزن برای دو گروه یکسان است یا متفاوت؟ این چه معنایی برای مدل‌سازی دارد (آیا بهتر است یک مدل مشترک بسازیم یا مدل جدا برای هر گروه)؟

۱۰.۵ - رگرسیون لجستیک با چند ویژگی از csv

با ستون‌های birth_week، hormone_level و weight_kg (هر سه به‌عنوان ویژگی) یک LogisticRegression برای پیش‌بینی label_num آموزش دهید. ضرایب هر ویژگی (model.coef_) را چاپ کنید و بگویید کدام ویژگی وزن (اهمیت) بیشتری در تصمیم مدل دارد. توجه کنید قبل از این کار حتماً باید ویژگی‌ها را استاندارد کنید تا مقایسه ضرایب معنادار باشد.