موضوع هفته:

آمار

تمرین ۱، ایجاد داده و نمونه‌برداری کامل و ناقص

۱- با استفاده از NumPy یک دیتاست با ۵۰۰ نمونه بسازید.
داده‌ها باید از توزیع نرمال با میانگین 50 و انحراف معیار 10 ساخته شوند.

۲- دو نوع نمونه‌برداری انجام دهید:

- **نمونه‌برداری کامل** :یعنی کل ۵۰۰ داده
- **نمونه‌برداری ناقص** :انتخاب تصادفی فقط ۳۰ داده

۳- برای هر دو حالت، موارد زیر را محاسبه و مقایسه کنید:

- میانگین
- واریانس
- انحراف معیار

۴- توضیح دهید که چرا نمونه‌برداری کوچک باعث نوسان بیشتر در میانگین و واریانس می‌شود.

تمرین ۲، اثر واریانس بر شکل تابع احتمال گوسی

تابع گوسی زیر را در نظر بگیرید:

$$f(x) = e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

۱- برای متغیر **X**، از 10- تا 10+ به اندازه ۱۰۰۰ نمونه بسازید

۲- برای سه حالت زیر مقدار گوسی را رسم کنید:

- $\sigma = 0.5$
- $\sigma = 1$
- $\sigma = 3$

۳- اثر انحراف معیار را بررسی کنید

تمرین ۳، هیستوگرام، binsو باکس‌پلات + مقایسه توزیع نرمال و یکنواخت

۱- دو مجموعه داده به صورت زیر بسازید:

- توزیع نرمال با $\sigma = 1, \mu = 0$
- توزیع یکنواخت در بازه‌ی 3- تا 3

۲- برای هر دیتاست،

- یک هیستوگرام با bins=10
- همان هیستوگرام با bins=50
- یک boxplot رسم نمایید

۳-. تفاوت ها را تحلیل کنید:

- چرا در توزیع نرمال داده‌ها وسط تجمع دارند؟
- چرا توزیع یکنواخت boxplot تخت‌تر است؟
- چرا bins کوچک/بزرگ شکل هیستوگرام را تغییر می‌دهد؟

تمرین ۴، اثبات تجربی قضیه حد مرکزی(Central Limit Theorem)

۱- با استفاده از Numpy یک دیتاست به تعداد N=10000 نمونه از توزیع یکنواخت در بازه [0,10] بسازید.

۲- یک loop تعریف کنید که 500 بار تکرار شود.

۳- در هر تکرار، یک نمونه تصادفی با اندازه n=30 از دیتاست گرفته شود.

۴- میانگین این ۳۰ نمونه را محاسبه کنید و در یک آرایه ذخیره کنید (آرایه نهایی شامل ۵۰۰ میانگین هست)

۵- تحلیل و مصورسازی:

- هیستوگرام ۵۰۰ میانگین ذخیره شده را رسم کنید
- توضیح دهید که شکل هیستوگرام به چه توزیعی شباهت دارد و چرا؟

۶- اگر اندازه نمونه در مرحله ۳ را به n=2 کاهش دهیم، و مجدد هیستوگرام را ترسیم کنیم، چه تغییر در شکل مشاهده می‌شود؟ (اثر اهمیت اندازه نمونه)

تمرین ۵، ترکیب توزیع‌های گوسی(Gaussian Mixture)

۱- با استفاده از Numpy یک دیتاست به تعداد N=1000 نمونه از ترکیب دو توزیع یکنواخت و نرمال را به صورت زیر درست کنید:

- ۵۰۰ نمونه اول: $\mu_1 = -2, \sigma_1 = 1$
- ۵۰۰ نمونه دوم: $\mu_1 = 5, \sigma_1 = 1.5$

۲- مصورسازی داده ها:

- **هیستوگرام** دیتاست ترکیبی را رسم کنید. شکل حاصل را توصیف کنید (باید دو قله داشته باشد).
- **محاسبه کنید :**میانگین و واریانس کل دیتاست ترکیبی.
- **پرسش چالشی :**آیا میانگین کل دیتاست (که محاسبه کردید) برابر با میانگین‌های وزنی توزیع‌های سازنده است؟ ($\mu_{total} = (\mu_1 + \mu_2)/2$)

تمرین ۶، پیاده‌سازی و تحلیل فاصله اقلیدسی(Euclidean Distance)

هدف: درک نحوه محاسبه و حساسیت فاصله اقلیدسی به مقیاس داده‌ها.

۱- پیاده‌سازی تابع فاصله:

- یک تابع به صورت دستی) بدون استفاده از توابع آماده (NumPy برای محاسبه **فاصله اقلیدسی** بین دو بردار P و Q در فضای n=50 بُعدی بنویسید.
  (راهنمایی: $Distance\ (P,Q) = \sqrt{\Sigma_i^n (P-Q)^2}$)

۲-. محاسبه در فضای ۲بُعدی:

- یک نقطه تست $\mathrm{P_{test}} = (4,7)$ را در نظر بگیرید.
- فاصله داده تا دو نقطه مرجع $R_a = (1,1), R_b = (8,1)$ را محاسبه کنید

۳-. تأثیر مقیاس‌گذاری:(Scaling)

- فرض کنید داده‌های شما از دو فیچر تشکیل شده‌اند: X1 (با مقادیر ۰ تا ۱۰) و X2 (با مقادیر ۰ تا ۱۰۰۰)
- فاصله اقلیدسی بین $\mathrm{P_{new}} = (5,500)$ و $Q_{new} = (6,510)$ را محاسبه کنید.
- **توضیح دهید :**چرا در این مثال، فیچر X2 که مقیاس بزرگتری دارد (**تقریباً تمام سهم** در محاسبه فاصله اقلیدسی را به خود اختصاص می‌دهد؟ (این توضیح نشان می‌دهد که چرا **نرمال‌سازی** داده‌ها قبل از استفاده از فاصله اقلیدسی در یادگیری ماشین ضروری است)

تمرین ۷، مقایسه آماری دامنه (Range) و انحراف معیار (Standard Deviation)

هدف :درک تفاوت بین معیارهای پراکندگی مبتنی بر آماره‌های رتبه‌ای (Range) و معیارهای مبتنی بر همه داده‌ها(Std Dev)

۱- **ساخت دیتاست‌ها :**دو مجموعه داده A و B با N=100 نمونه بسازید:

- **A:** توزیع نرمال با میانگین صفر و واریانس ۲
- **B:** توزیع نرمال با میانگین صفر و واریانس ۲ که در آن ۱۰ داده اول به صورت عمدی به مقادیر ناهنجار (20- و 20+) تغییر داده شده‌اند.

۲- **محاسبه و مقایسه :**برای هر دو مجموعه A و B ، موارد زیر را محاسبه کنید:

- دامنه (Range)
- انحراف معیار

۳- معمولاً انحراف معیار/واریانس بر دامنه ترجیح داده می‌شود یا اینکه به عنوان مکمل در نظر گرفته می‌شود؟

تمرین ۸، تولید داده‌های چندمتغیره و تحلیل کوواریانس(Multivariate Data)

هدف: درک مفهوم تولید داده‌های چندمتغیره با همبستگی مشخص و تحلیل ماتریس کوواریانس.

۱- تولید داده‌های مستقل:

- با استفاده از تابع np.random.multivariate_normal، یک مجموعه داده ۲بُعدی (N=500 نمونه) بسازید که میانگین آن $\mu = [0, 0]$ باشد و **هیچ همبستگی‌ای** بین دو ویژگی وجود نداشته باشد (ماتریس کوواریانس: $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$)

۲- تولید داده‌های همبسته:

- یک مجموعه داده ۲بُعدی دیگر با همان میانگین $\mu = [0, 0]$ بسازید، اما این بار با **همبستگی قوی مثبت** (ماتریس کوواریانس: $\Sigma_2 = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$)

۳- تحلیل و مصورسازی:

- برای هر دو مجموعه داده، یک **نمودار پراکندگی (Scatter Plot)** رسم کنید.
- به صورت تجربی **ماتریس کوواریانس** هر دو مجموعه داده را با استفاده از NumPy تابع **np.cov** محاسبه کنید.

۳- **توضیح دهید:**

- چگونه شکل ابر داده (Data Cloud) در نمودار پراکندگی برای Σ_1 و Σ_2 متفاوت است؟
- مقدار 0.8 در ماتریس کوواریانس چه معنایی در رابطه با متغیرها دارد؟

تمرین ۹، تاثیر تغییر واحد بر آماره‌های توصیفی(Scaling)

هدف :نشان دادن اینکه چگونه تغییر مقیاس (که در پیش‌پردازش داده رایج است) بر آماره‌های مختلف تأثیر می‌گذارد.

۱- **دیتاست اصلی :**فیچر **طول کاسبرگ (Sepal Length)** از دیتاست Iris را انتخاب کنید و آن را X نامگذاری کنید.

۲- **محاسبه آماره‌ها برای X:** میانگین، واریانس و دامنه (رنج) را محاسبه کنید

۳- یک متغیر جدید به نام Y به صورت زیر بسازید:

$$Y = 2X + 5$$

(یعنی هر داده را در ۲ ضرب کرده و ۵ واحد به آن اضافه کنید)

۴- **محاسبه آماره‌ها برای Y:** میانگین، واریانس و دامنه (رنج) را محاسبه کنید

۵- مقادیر دو متغیر را نسبت به هم مقایسه کنید و میزان تغیرات را گزارش دهید.

تمرین۱۰، طبقه‌بندی داده‌ها با احتمال گوسی

هدف:

- ساخت سه کلاس با آمار مختلف
- محاسبه احتمال گوسی برای هر کلاس
- انتخاب کلاس با بیشترین احتمال
- رنگ‌آمیزی بر اساس دسته(A,B,C)

سه کلاس به صورت زیر تعریف کنید:

Class	Average (μ)	Variance (σ^2)
A	**0**	**1**
B	**3**	**0.5**
C	**-3**	**2**

۱- با Numpy از هر کلاس ۲۰۰ داده تولید کنید

۲- همه داده‌ها را در یک آرایه کنار هم قرار دهید (۶۰۰ نمونه)

۳- برای هر داده، مقدار احتمال گوسی نسبت به هر کلاس را حساب کنید:

$$P(x|Class) = e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

۴- برای هر داده کلاس با احتمال بیشینه را انتخاب کنید. (آیا داده‌ایی بوده که در دسته بندی جدید، به عضویت دسته دیگه ایی در بیاید؟)

۵- داده ها را به سه رنگ (بر اساس کلاس درست شده در قسمت ۴) رسم کنید:
A=blue | B=black | C=red

تمرین ۱۱، طبقه‌بندی داده‌های تست iris

هدف تست مدل مبتنی بر iris

۱- ابتدا داده های iris_dataset.csv رو فراخوانی کنید

۲- ۴ فیچر اول را به عنوان ورودی (X) در نظر بگیرید و فیچر خروجی را Y (لیبل هر کلاس) در نظر بگیرید

۳- مقدار میانگین، واریانس، range، max و min برا برای هر فیچر در کلاسها بدست بیاورید

۴- با در نظر گرفتن تنها اطلاعات مرتبط به فیچر سوم ((petal length (cm)) برای هر کلاس، داده های تست را به عضویت هر کلاس در بیاورید (از احتمال ساخته شده از میانگین و واریانس استفاده کنید).

۵- دقت تصمیم گیری چقدر هست؟

Precision=100*(Correct Answers)/(Correct and Wrong Answers)

۶- علاوه بر فیچر شماره ۳، از فیچر شماره ۴ ((petal width (cm)) هم استفاده کنید. در اینجا، برای هر فیچر یک احتمال وجود دارد (پس برای دو فیچر، دو احتمال). میانگین این دو احتمال را به عنوان احتمال محاسبه شده برای هر کلاس استفاده کنید و دسته بندی را انجام دهید.

۷- دقت تصمیم گیری چقدر هست؟

۸- از تمام ۴ فیچر ورودی استفاده کنید و سپس، دو احتمال برتر از بین چهار احتمال را در نظر بگیرید برای تصمیم گیری. بعد از انتخاب دو احتمال برتر، از آنها میانگین گرفته، و این میانگین را در بین کلاس ها مقایسه کنید تا تشخیص دهید داده های تست به چه کلاسی مرتبط می‌شوند.

تمرین ۱۲، تمرین ترکیبی: تحلیل و طبقه‌بندی کلاس Versicolor (میانگین گلبرگ)

هدف :محاسبه آمار توصیفی، استفاده از توزیع نرمال برای تحلیل و پیدا کردن یک مرز تصمیم (Decision Boundary)برای یک کلاس خاص.

بخش ۱: محاسبه آماره‌ها

۱- فیچر طول گلبرگ (Petal Length) را از دیتاست Irisانتخاب کنید.

۲- مقدار میانگین و انحراف معیار این فیچر را فقط برای نمونه‌های متعلق به کلاس Iris-Versicolor محاسبه کنید.

بخش ۲: تحلیل توزیع و نقاط بحرانی

۱- بر اساس آمار محاسبه شده در بخش ۱، فرض کنید که طول گلبرگ‌های Versicolor از **توزیع نرمال** پیروی می‌کند.

۲- با استفاده از قانون تجربی توزیع نرمال (۶۸-۹۵-۹۹,۷)، محدوده‌ای را محاسبه کنید که **٪۹۵** از داده‌های Versicolor در آن قرار گیرند:

$$Lower\ Band = \mu - 2\sigma$$
$$Upper\ Band = \mu + 2\sigma$$

۳- یک **Boxplot**برای **طول گلبرگ** سه کلاس (Setosa, Versicolor, Virginica) رسم کنید.

بخش ۳: تعیین مرز تصمیم(Decision Boundary)

۱- اگر بخواهید یک قانون تصمیم (Rule) بسیار ساده برای شناسایی **Versicolor**بر اساس **طول گلبرگ** بنویسید (با این فرض که هر داده‌ای خارج از آن محدوده به کلاس دیگری تعلق دارد):

- **مرز تصمیم شما** برای حداقل و حداکثر طول گلبرگ Versicolor چیست؟

۲- ارزیابی دقت:

- چند نمونه از کلاس‌های **Setosa**و **Virginica**به طور نادرست در این محدوده **Versicolor** قرار می‌گیرند؟ (به این‌ها **خطای نوع اول** یا **False Positives**گفته می‌شود.)
- چند نمونه از خود کلاس **Versicolor**به طور نادرست **خارج** از این محدوده قرار می‌گیرند؟ (به این‌ها **خطای نوع دوم** یا **False Negatives**گفته می‌شود.)