سؤال ۱ — تحلیل کوواریانس و کورلیشن در دیتاست واقعی (Diabetes Dataset)

با استفاده از sklearn.datasets.load_diabetes :

دیتاست را بارگذاری کرده و آن را به یک DataFrame پاندا تبدیل کنید.

فقط ۱۰ ویژگی عددی را نگه دارید.

ماتریس کوواریانس ۱۰×۱۰ را محاسبه و چاپ کنید.

سه جفت ویژگی با بیشترین کوواریانس مثبت و سه جفت با بیشترین کوواریانس منفی را پیدا کنید.

ماتریس کورلیشن را نیز محاسبه کنید.

پاسخ تحلیلی بنویسید:

چرا برخی زوج ویژگی‌ها کوواریانس زیاد دارند ولی همبستگی ضعیف؟

چه تفاوت بنیادی بین مقیاس‌پذیری covariance و correlation وجود دارد؟

سؤال ۲ — اثر تغییر مقیاس روی covariance و correlation (Wine Dataset)

از دیتاست sklearn.datasets.load_wine استفاده کنید.

دیتاست را بارگذاری کنید و دو ستون "alcohol" و "magnesium" را استخراج کنید.

مقدار covariance و Pearson correlation این دو ستون را محاسبه کنید.

ستون magnesium را در عدد ۲۰ ضرب کنید.

دوباره covariance و correlation را محاسبه کنید.

تحلیل کنید:

چرا 20× covariance شده است؟

چرا correlation تغییری نکرده؟

با فرمول ریاضی نشان دهید correlation نسبت به تغییر واحد مقاوم است.

سؤال ۳ — تشخیص رابطه‌های پنهان با ماتریس همبستگی (Breast Cancer Dataset)

از دیتاست sklearn.datasets.load_breast_cancer:

دیتاست را بارگذاری کنید و از ۸ ویژگی اول استفاده کنید.

ماتریس correlation را محاسبه و یک heatmap ترسیم کنید.

سه زوج ویژگی با بالاترین correlation مثبت و سه زوج با بیشترین correlation منفی را پیدا کنید.

تحلیل:

آیا وجود correlation بالا بین دو ویژگی به معنی علیت است؟

مثال واقعی‌ای از یک همبستگی گمراه‌کننده بنویسید (spurious correlation).

سؤال ۴ — اثر نویز بر covariance و correlation

یک دیتاست دوبعدی با رابطه خطی بسازید:

$$X \sim N(0,1), \quad Y = 5X + N(0,\sigma^2)$$

این تولید داده را برای سه مقدار σ انجام دهید:

σ = 0.5

σ = 5

σ = 20

برای هر حالت:

covariance

correlation

scatter plot را رسم و مقایسه کنید.

چرا correlation با افزایش نویز شدیداً افت می‌کند؟

چرا covariance همچنان بزرگ باقی می‌ماند؟

بخش ۲ — انواع فاصله‌ها

سؤال ۵ — مقایسه انواع فاصله‌ها در داده واقعی (Iris Dataset)

با استفاده از sklearn.datasets.load_iris:

دیتاست را بارگذاری کنید.

میانگین هر کلاس (setosa، versicolor، virginica) را محاسبه کنید.

بین میانگین سه کلاس، فاصله‌های زیر را محاسبه کنید:

Euclidean

Manhattan

Chebyshev

Cosine distance

Mahalanobis distance (با covariance کل دیتاست)

تحلیل کنید:

کدام فاصله کلاس‌ها را بهتر جدا می‌کند؟ چرا؟

چرا Euclidean در داده‌هایی با scale متفاوت اشتباه می‌کند؟

نقش covariance در Mahalanobis چیست؟

سؤال ۶ — رفتار غیرمنتظره فاصله کسینوسی در بردارهای موازی

یک دیتاست مصنوعی بسازید:

$$A = 100X,\ B = 250X,\ C = \text{random noise}$$

X یک بردار ۲۰۰ بعدی با توزیع N (۰,۱) باشد.

فاصله اقلیدسی بین A و B را حساب کنید. فاصله کسینوسی بین A و B را حساب کنید.

فاصله‌های Euclidean و Cosine بین A و C را حساب کنید.

چرا cosine(A,B)=0 ولی Euclidean(A,B) بسیار بزرگ است؟

مفهوم "جهت" و "مقیاس" را در تحلیل فاصله‌ها توضیح دهید.

سؤال ۷ — تشخیص Outlier با فاصله Mahalanobis (Wine Dataset)

از load_wine:

دیتاست را بارگذاری کنید و فقط ویژگی‌های عددی را حفظ کنید.

یک نقطه outlier اضافه کنید (مثلاً همه فیچرها = ۸ برابر میانگین).

فاصله Mahalanobis همه نقاط را از میانگین حساب کنید.

نقاطی که distance > σ۳ دارند را برچسب outlier بزنید.

تحلیل:

چرا Mahalanobis برای یافتن Outlier مناسب‌تر از Euclidean است؟

نقش covariance در تشخیص جهت‌های واقعی پراکندگی چیست؟

بخش ۳ — PCA روی دیتاست جدولی

سؤال ۸ — PCA کامل روی داده جدولی + جدول explained variance و loadings (Wine Dataset) با استفاده از load_wine():

دیتاست را بارگذاری کرده و به DataFrame تبدیل کنید.

داده را با StandardScaler استاندارد کنید.

PCA را روی تمام ویژگی‌ها اجرا کنید.

یک جدول (DataFrame) با ستون‌های زیر بسازید:

PC index

eigenvalue (explained_variance_)

explained_variance_ratio

cumulative_explained_variance

یک جدول جداگانه بسازید که loadings (eigenvectors) را نشان دهد:

سطرها: نام ویژگی‌ها

ستون‌ها: PC1, PC2, ..., PC13

مشخص کنید برای رسیدن به ۹۰٪ واریانس چند مؤلفه لازم است.

تحلیل کنید:

کدام ویژگی‌ها بیشترین وزن را روی PC1 دارند؟

آیا این ویژگی‌ها با هم همبسته‌اند؟

اگر بعد را روی k مؤلفه کاهش دهیم، کدام اطلاعات از دست می‌رود؟

بخش ۴ — PCA روی تصویر

سؤال ۹ — فشرده‌سازی تصویر با PCA (Digits Dataset)

با استفاده از sklearn.datasets.load_digits:

مجموعه تصاویر ۸×۸ را بارگذاری کنید (شکل داده: ۱۷۹۷ × ۶۴).

PCA را روی کل دیتاست اعمال کنید.

چند تصویر نمونه را با k های مختلف ۵ و ۱۵و ۳۰ و ۴۰ مؤلفه بازسازی کنید.

برای هر k، خطای بازسازی (MSE) را محاسبه کنید.

نمودار MSE مقابل k را رسم کنید.

چرا تصاویر با k کم، تار می‌شوند؟

نقش مؤلفه‌های اصلی اول در ساختار تصویر چیست؟

سؤال ۱۰ — حذف نویز تصویر با PCA (Digits Dataset)

یک تصویر تصادفی از digits را انتخاب کنید.

به داده نویز Gaussian با $\sigma^2=50$ اضافه کنید.

PCA را روی کل دیتاست اجرا کنید.

تصویر را با مؤلفه‌هایی که ٪۹۰ واریانس را پوشش می‌دهند بازسازی کنید.

مقایسه کنید:

تصویر اصلی

تصویر نویزی

تصویر denoised

MSE و PSNR را محاسبه کنید.

چرا PCA نویز را حذف می‌کند؟

نویز معمولاً در کدام مؤلفه‌ها قرار می‌گیرد؟

بخش ۵ — مسائل ترکیبی

سؤال ۱۱ — اثر PCA بر ساختار فاصله‌ها (Wine Dataset)

دیتاست را بارگذاری و استاندارد کنید.

میانگین سه کلاس را محاسبه کنید.

فاصله Euclidean و Mahalanobis بین مراکز کلاس‌ها را در فضای اصلی حساب کنید.

PCA را اجرا و ابعاد را به ۲ کاهش دهید.

دوباره فاصله‌ها را محاسبه کنید.

آیا ترتیب نزدیکی کلاس‌ها تغییر کرد؟

چرا Mahalanobis نسبت به کاهش بعد پایدارتر است؟

چه ویژگی‌هایی هنگام کاهش بعد حذف می‌شوند؟

سؤال ۱۲ — تأثیر همبستگی ویژگی‌ها بر PCA و Mahalanobis (Diabetes Dataset)

با دیتاست load_diabetes:

دیتاست را بارگذاری کنید.

ماتریس correlation را محاسبه و ویژگی‌هایی که $|corr| > ۰.۵$ دارند فهرست کنید.

PCA را فقط روی این زیرمجموعه اجرا کنید.

بارهای PC1 و PC2 را بررسی کنید و تحلیل کنید:

آیا ویژگی‌های همبسته در یک مؤلفه خوشه شده‌اند؟

دو نمونه دلخواه انتخاب کرده و فاصله Mahalanobis آنها را:

قبل از PCA

بعد از PCA (در فضای PCها)حساب کنید.

تحلیل:

چرا Mahalanobis همبستگی بین ویژگی‌ها را بهتر از Euclidean مدیریت می‌کند؟

چرا PCA در داده‌های همبسته مؤثرتر است؟