**یادگیری ماشین**

**بخش پنجم**

**علیرضا افرنگ**

**۱. Line Chart (نمودار خطی) - چرا برای نمایش رابطه‌های خطی استفاده می‌شه؟**

**تعریف:**

نمودار خطی داده‌ها رو با نقاطی که با خط به هم وصل شدن نشون می‌ده. معمولاً محور افقی (X) یه متغیر مستقل (مثل زمان) و محور عمودی (Y) یه متغیر وابسته (مثل مقدار) رو نشون می‌ده.

**چرا برای رابطه‌های خطی؟**

* **پیوستگی و روند**: خطوط توی این نمودار نشون‌دهنده یه جریان پیوسته‌ست. اگه بخوای ببینیم یه مقدار چطور با یه متغیر دیگه (مثلاً زمان) تغییر می‌کنه، خطوط بهمون کمک می‌کنن روند رو ببینی (مثلاً صعودی، نزولی یا ثابت).
* **رابطه خطی**: وقتی داده‌ها یه الگوی خطی دارن (یعنی تغییراتشون به صورت تدریجی و قابل پیش‌بینی باشه)، خطوط این رابطه رو واضح نشون می‌دن. مثلاً اگه دما در طول روز به مرور بالا بره، یه خط صاف یا منحنی ملایم این رو خوب نشون می‌ده.
* **مقایسه در طول زمان**: بیشتر برای داده‌های زمانی استفاده می‌شه (مثل فروش ماهانه، دمای روزانه)، چون خطوط می‌تونن تغییرات کوچک و پیوسته رو به خوبی نشون بدن.

**مثال:**

فرض کنیم می‌خوایم دمای هوا رو در ۲۴ ساعت روز نشون بدیم:

* محور X: ساعت‌ها (۰ تا ۲۳)
* محور Y: دما (مثلاً ۱۵ تا ۳۰ درجه)
* خطوط نشون می‌دن که دما از صبح تا ظهر بالا می‌ره و شب دوباره پایین میاد. این روند خطی یا نزدیک به خطی رو فقط خطوط می‌تونن واضح نشون بدن.

**محدودیت:**

اگه داده‌ها پراکنده باشن یا هیچ روند مشخصی نداشته باشن، نمودار خطی گمراه‌کننده می‌شه، چون خطوط به اشتباه یه پیوستگی رو نشون می‌دن که وجود نداره.

**۲. Bar Chart (نمودار میله‌ای) - چرا برای مقایسه داده‌های گروهی کاربرد داره؟**

**تعریف:**

نمودار میله‌ای از میله‌های جدا از هم برای نمایش داده‌ها استفاده می‌کنه. هر میله یه دسته یا گروه خاص رو نشون می‌ده، و ارتفاعش (یا طولش توی حالت افقی) مقدار اون گروه رو مشخص می‌کنه.

**چرا برای مقایسه داده‌های گروهی؟**

* **مقایسه مستقیم**: میله‌ها کنار هم قرار می‌گیرن و بهمون اجازه می‌دن به سرعت ببینیم کدوم گروه بزرگ‌تر یا کوچک‌تره. مثلاً مقایسه فروش چند محصول توی یه ماه.
* **دسته‌بندی مجزا**: هر میله یه گروه مستقل رو نشون می‌ده (مثل شهرها، افراد، یا محصولات)، و چون میله‌ها جدا هستن، هیچ پیوستگی یا رابطه خطی بینشون فرض نمی‌شه.
* **وضوح در تفاوت‌ها**: ارتفاع میله‌ها تفاوت‌های عددی رو خیلی واضح نشون می‌ده، مخصوصاً وقتی مقادیر خیلی با هم فرق دارن یا تعداد گروه‌ها کمه.
* **انعطاف‌پذیری**: می‌تونیم چند سری داده رو کنار هم بذاریم (مثلاً میله‌های چند رنگ برای سال‌های مختلف) و مقایسه رو چندجانبه کنی.

**مثال:**

فرض کنیم می‌خوایم فروش ۵ محصول رو توی یه فروشگاه مقایسه کنی:

* محور X: نام محصولات (مثلاً A، B، C، D، E)
* محور Y: مقدار فروش (مثلاً ۱۰۰، ۲۰۰، ۱۵۰، ۳۰۰، ۵۰)
* میله‌ها بهمون نشون می‌دن که محصول D بیشترین فروش رو داره و E کمترین. این مقایسه بین گروه‌های مجزا رو نمودار خطی نمی‌تونه به این وضوح نشون بده.

**محدودیت:**

اگه بخوایم روند پیوسته رو نشون بدیم (مثلاً تغییرات فروش در طول زمان)، میله‌ها مناسب نیستن چون پیوستگی رو منتقل نمی‌کنن و فقط مقادیر گسسته رو نشون می‌دن.

**۳. Scatter Plot (نمودار نقطه‌ای) - چرا برای نمایش رابطه‌های غیرخطی استفاده می‌شه؟**

**تعریف:**

نمودار نقطه‌ای داده‌ها رو به صورت نقاط جداگانه توی یه صفحه دوبعدی نشون می‌ده. محور X و Y معمولاً دو متغیر مستقل یا وابسته رو نشون می‌دن، و موقعیت هر نقطه رابطه بین این دو رو مشخص می‌کنه.

**چرا برای رابطه‌های غیرخطی؟**

* **عدم فرض پیوستگی**: برعکس نمودار خطی، نقطه‌ها به هم وصل نیستن، پس هیچ رابطه خطی یا پیوسته‌ای رو تحمیل نمی‌کنن. این باعث می‌شه برای داده‌هایی که الگوی غیرخطی دارن (مثلاً پراکنده یا منحنی پیچیده) مناسب باشه.
* **نمایش پراکندگی**: اگه بخوای ببینیم دو متغیر چطور به هم ربط دارن (مثلاً قد و وزن افراد)، نقاط می‌تونن خوشه‌ها، الگوها، یا پراکندگی رو نشون بدن بدون اینکه خطی بودن رو فرض کنن.
* **رابطه‌های پیچیده**: وقتی رابطه بین دو متغیر خطی نیست (مثلاً نمایی، لگاریتمی، یا تصادفی)، نقاط این پیچیدگی رو بهتر نشون می‌دن. مثلاً ممکنه یه خوشه توی یه منطقه خاص شکل بگیره که با خط نمی‌شه نشونش داد.
* **شناسایی داده‌های پرت (Outliers)**: چون هر نقطه جداست، راحت‌تر می‌تونیم داده‌هایی که از بقیه دورن رو پیدا کنیم.

**مثال:**

رابطه بین ساعت مطالعه و نمره امتحان رو برای ۵۰ دانش‌آموز :

* محور X: ساعت مطالعه (۰ تا ۱۰)
* محور Y: نمره (۰ تا ۱۰۰)
* نقاط نشون می‌دن که تا یه حدی (مثلاً ۵ ساعت) نمره‌ها با مطالعه بیشتر می‌رن بالا، ولی بعدش پراکنده می‌شن و رابطه خطی نیست. این غیرخطی بودن رو خطوط نمی‌تونن خوب نشون بدن.

**محدودیت:**

اگه داده‌ها خیلی زیاد باشن یا بخوای روند کلی رو به صورت پیوسته ببینیم، نمودار نقطه‌ای شلوغ می‌شه و ممکنه به اندازه نمودار خطی واضح نباشه.

**D. Bubble Chart (نمودار حبابی) - چرا برای نمایش سه متغیر استفاده می‌شه؟**

**تعریف:**

نمودار حبابی یه نوع نمودار نقطه‌ایه (Scatter Plot) که اندازه حباب‌ها یه متغیر سوم رو نشون می‌ده. محور X و Y دو متغیر رو نشون می‌دن و اندازه حباب متغیر سوم رو.

**چرا برای سه متغیر؟**

* **نمایش همزمان سه بعد**: محور X و Y رابطه بین دو متغیر رو نشون می‌دن (مثل قد و وزن)، و اندازه حباب یه متغیر دیگه رو (مثلاً سن) اضافه می‌کنه. این باعث می‌شه بدون نیاز به سه‌بعدی کردن، سه متغیر رو تو یه صفحه ببینی.
* **مقایسه بصری**: اندازه حباب‌ها بهمون اجازه می‌ده به سرعت مقادیر بزرگ‌تر یا کوچک‌تر متغیر سوم رو تشخیص بدی.
* **انعطاف‌پذیری**: می‌تونیم با رنگ یا شفافیت حباب‌ها حتی متغیر چهارم رو هم نشون بدیم (مثلاً دسته‌بندی).

**مثال:**

فرض کنیم می‌خوایم فروش شرکت‌ها رو بررسی کنیم:

* X: تعداد کارمندها
* Y: درآمد سالانه
* اندازه حباب: تعداد محصولات
* یه حباب بزرگ‌تر یعنی اون شرکت محصولات بیشتری داره، و موقعیتش روی صفحه رابطه بین کارمندها و درآمد رو نشون می‌ده.

**محدودیت:**

اگه تعداد داده‌ها زیاد باشه، حباب‌ها روی هم می‌افتن و نمودار شلوغ می‌شه.

**E. Heatmap (نقشه حرارتی) - چرا برای نمایش رابطه‌های بین متغیرها کاربرد داره؟**

**تعریف:**

نقشه حرارتی یه ماتریسه که از رنگ‌ها برای نشون دادن مقدار یه متغیر توی تقاطع دو متغیر دیگه استفاده می‌کنه. معمولاً برای ماتریس همبستگی (Correlation Matrix) به کار می‌ره.

**چرا برای رابطه‌های بین متغیرها؟**

* **نمایش شدت رابطه**: رنگ‌ها (مثلاً تیره‌تر برای مقادیر بالا) بهمون نشون می‌دن که رابطه بین دو متغیر چقدر قویه (مثلاً همبستگی مثبت یا منفی).
* **مقایسه چند متغیر**: می‌تونیم همزمان رابطه همه جفت‌های متغیرها رو توی یه نگاه ببینی، بدون اینکه نیاز به نمودارهای جداگانه داشته باشی.
* **تشخیص الگو**: الگوهای مخفی توی داده‌ها (مثلاً خوشه‌بندی یا ارتباط قوی) با رنگ‌ها راحت‌تر دیده می‌شن.

**مثال:**

فرض کنیم یه دیتاست داریم با متغیرهای "دما"، "رطوبت"، و "فشار هوا":

* یه ماتریس همبستگی می‌سازم و توی Heatmap نشون می‌دم.
* رنگ قرمز تیره بین "دما" و "رطوبت" یعنی همبستگی قوی، و رنگ روشن بین "دما" و "فشار" یعنی رابطه ضعیف.

**محدودیت:**

برای داده‌های خیلی بزرگ یا غیرعددی خوب کار نمی‌کنه، چون فقط مقادیر عددی رو می‌تونه نشون بده.

**F. Pairplot - چرا برای تحلیل روابط بین متغیرها کاربرد داره؟**

**تعریف:**

Pairplot یه مجموعه از نمودارهاست که رابطه بین همه جفت‌های متغیرها رو توی یه دیتاست نشون می‌ده. معمولاً توی قطرش هیستوگرام و توی بقیه جاها Scatter Plot داره.

**چرا برای تحلیل روابط بین متغیرها؟**

* **بررسی همه‌جانبه**: همه ترکیب‌های ممکن بین متغیرها رو توی یه صفحه نشون می‌ده، پس می‌تونی سریع ببینی کدوم متغیرها به هم ربط دارن.
* **تشخیص نوع رابطه**: با Scatter Plot‌ها می‌تونی ببینی رابطه خطی، غیرخطی، یا اصلاً وجود نداره.
* **توزیع و پراکندگی**: هیستوگرام‌های قطری توزیع هر متغیر رو نشون می‌دن، که به تحلیل اولیه داده‌ها کمک می‌کنه.

**مثال:**

دیتاست گل‌های Iris رو در نظر بگیر (طول گلبرگ، عرض گلبرگ، طول کاسبرگ، عرض کاسبرگ):

* Pairplot نشون می‌ده که طول گلبرگ و عرض گلبرگ یه رابطه خطی دارن، ولی عرض کاسبرگ و طول گلبرگ پراکنده‌ست.

**محدودیت:**

اگه تعداد متغیرها زیاد باشه (مثلاً بیشتر از ۱۰ تا)، Pairplot خیلی شلوغ و غیرقابل خوندن می‌شه.

**G. Boxplot - چرا برای تشخیص Outliers استفاده می‌شه؟**

**تعریف:**

Boxplot (نمودار جعبه‌ای) توزیع داده‌ها رو با یه جعبه (برای چارک‌ها) و خطوط (برای حداقل و حداکثر) نشون می‌ده. نقاط خارج از محدوده هم به عنوان Outliers مشخص می‌شن.

**چرا برای تشخیص Outliers؟**

* **محدوده بین‌چارک‌ها (IQR)**: جعبه بین چارک اول (Q1) و سوم (Q3) رو نشون می‌ده، و خطوط (Whiskers) تا ۱.۵ برابر IQR کشیده می‌شن. هر داده‌ای خارج از این محدوده به صورت نقطه جدا نشون داده می‌شه و احتمالاً Outlierه.
* **مقایسه بصری**: نقاطی که از بقیه داده‌ها خیلی دورن رو سریع مشخص می‌کنه، بدون نیاز به محاسبات پیچیده.
* **توزیع کلی**: همزمان شکل توزیع (مثلاً چولگی یا تقارن) رو هم نشون می‌ده.

**مثال:**

فرض کن حقوق کارمندهای یه شرکت رو داری:

* جعبه نشون می‌ده اکثر حقوق‌ها بین ۲ تا ۵ میلیونه.
* یه نقطه توی ۲۰ میلیون یه Outlierه که سریع دیده می‌شه.

**محدودیت:**

فقط داده‌های عددی رو می‌تونه نشون بده و برای روابط بین متغیرها مناسب نیست.

**H. Histogram - چرا برای نمایش توزیع داده‌ها کاربرد داره؟**

**تعریف:**

Histogram داده‌ها رو به بازه‌های مشخص (Bins) تقسیم می‌کنه و ارتفاع هر میله تعداد داده‌ها توی اون بازه رو نشون می‌ده.

**چرا برای توزیع داده‌ها؟**

* **شکل توزیع**: نشون می‌ده داده‌ها نرمال (زنگوله‌ای)، چپ‌چولگی، راست‌چولگی، یا چندقله‌ای هستن.
* **تمرکز و پراکندگی**: می‌تونی ببینی داده‌ها دور چه مقادیری جمع شدن و چقدر پخش شدن.
* **سادگی**: بدون نیاز به تحلیل پیچیده، یه نگاه بهت می‌گه داده‌ها چطور پخش شدن.

**مثال:**

فرض کن سن افراد توی یه شهر رو داری:

* بازه‌ها: ۰-۱۰، ۱۰-۲۰، و غیره
* اگه میله ۲۰-۳۰ بلندتر باشه، یعنی بیشتر افراد توی این سنن.

**محدودیت:**

فقط برای یه متغیر عددی کار می‌کنه و نمی‌تونه روابط بین متغیرها رو نشون بده.

**I. چگونه می‌توانید یک 3D Plot را در Python ایجاد کنید؟**

**توضیح:**

برای ساخت نمودار سه‌بعدی توی پایتون، از کتابخونه matplotlib استفاده می‌کنم. ماژول mpl\_toolkits.mplot3d این امکان رو می‌ده.

مراحل:

**نصب و وارد کردن کتابخونه**:

import matplotlib.pyplot as plt

from mpl\_toolkits.mplot3d import Axes3D

import numpy as np

**ایجاد داده‌ها**: مثلاً سه متغیر X، Y، و Z:

x = np.linspace(-5, 5, 100)

y = np.linspace(-5, 5, 100)

X, Y = np.meshgrid(x, y)

Z = np.sin(np.sqrt(X\*\*2 + Y\*\*2))

**ساخت نمودار**:

fig = plt.figure()

ax = fig.add\_subplot(111, projection='3d') # فعال کردن حالت 3D

ax.plot\_surface(X, Y, Z, cmap='viridis') # رسم سطح

ax.set\_xlabel('X Axis')

ax.set\_ylabel('Y Axis')

ax.set\_zlabel('Z Axis')

plt.show()

**انواع 3D Plot:**

* **Surface Plot**: برای سطوح پیوسته (مثل بالا).
* **Scatter Plot 3D**: برای نقاط:

ax.scatter(X, Y, Z)

**Wireframe**: برای خطوط مشبک:

ax.plot\_wireframe(X, Y, Z)

**نکته:**

برای تعامل بهتر، می‌تونی از plotly هم استفاده کنی که نمودارهای 3D تعاملی می‌سازه.

**J. چرا Seaborn برای تجسم داده‌های پیشرفته استفاده می‌شه؟**

**تعریف:**

Seaborn یه کتابخونه توی پایتونه که روی matplotlib ساخته شده و ابزارهای پیشرفته‌تری برای تجسم داده‌ها ارائه می‌ده.

**چرا برای تجسم پیشرفته؟**

* **زیبایی و سادگی**: نمودارها به صورت پیش‌فرض قشنگ‌ترن (رنگ‌ها، تم‌ها، و استایل‌ها بهترن) و کد کمتری نیاز داره.
* **توابع آماده**: ابزارهایی مثل pairplot، heatmap، و violinplot رو با یه خط کد می‌سازه که توی matplotlib پیچیده‌تره.
* **ادغام با Pandas**: مستقیم با DataFrame‌ها کار می‌کنه و نیازی به تبدیل داده‌ها نیست.
* **تحلیل آماری**: قابلیت‌هایی مثل خط رگرسیون توی Scatter Plot یا نمایش توزیع‌ها رو خودکار اضافه می‌کنه.
* **انعطاف‌پذیری**: می‌تونی با ترکیبش با matplotlib، نمودارها رو سفارشی کنی.

**مثال:**

ساخت Pairplot با Seaborn:

import seaborn as sns

import matplotlib.pyplot as plt

sns.pairplot(data) # data یه DataFrameه

plt.show()

توی matplotlib این کار خیلی بیشتر طول می‌کشه.

#### محدودیت:

برای نمودارهای خیلی خاص یا پیچیده، ممکنه هنوز به matplotlib نیاز داشته باشی.

جمع بندی:

هرکدوم ازین ابزار ها برای کنترل داده ها قرار داده شده و با شناختن بهتر اونها خیلی راحت تر و بهتر میتونیم این کار رو انجام بدیم