خانه فناوری تهران

جزوه تخصصی | بخش چهارم وبینار «مسیر آموزش بعد از پایتون»

از تحلیل داده تا یادگیری ماشین و هوش مصنوعی

نقشه راه کاربردی Data Analysis، Data Science، Machine Learning، Deep Learning به محصول

تحلیل داده | یادگیری ماشین | ی عصبی | مدل‌های مولد | پروژه عملی

تدوین آموزشی بر پایه محتوای بخش پایانی وبینار

ارائه علمی: آقای علی پیله‌ور

ویژه دانشجویان و علاقه‌مندان پایتون

www.fanavari.co
@fanavari.co

خلاصه جامع جلسه

مروری بر محتوا

این جلسه به معرفی و دسته‌بندی حوزه‌های کلیدی علوم داده، شامل تحلیل داده، علم داده و یادگیری ماشین می‌پردازد. سخنران ضمن تشریح وظایف هر حوزه، از پاک‌سازی و مصورسازی داده تا مدل‌سازی پیش‌بینانه، بر اهمیت دانش تخصصی و نقش مکمل ابزارهای هوش مصنوعی مانند ChatGPT تأکید می‌کند. همچنین کتابخانه‌های پایتون، مدل‌های یادگیری ماشین، شبکه‌های عصبی مصنوعی و کاربردهای صنعتی، مالی و پزشکی مورد بحث قرار می‌گیرند.

موضوعات و محورهای کلیدی

1. **تفاوت و ارتباط میان تحلیل داده، علم داده و یادگیری ماشین**
   شناخت خروجی هر حوزه و روابط هم‌پوشان آن‌ها در یک پروژه واقعی.
2. **اهمیت پیش‌پردازش، پاک‌سازی و مصورسازی**
   از دریافت CSV و SQL تا یافتن داده گمشده و گزارش مدیریتی.
3. **نقش و محدودیت‌های AI در پروژه‌های حساس**
   کمک به کدنویسی و تحلیل، در کنار ضرورت آزمون، امنیت و قضاوت متخصص.
4. **کتابخانه‌های کاربردی پایتون**
   scikit-learn، Matplotlib، SciPy، Pandas، NumPy و PyTorch.
5. **یادگیری ماشین، شبکه‌های عصبی و مقیاس**
   رگرسیون، دسته‌بندی، شبکه‌های عصبی، داده پیچیده و انتخاب الگوریتم مناسب.

درباره متن: محتوای سخنران بازنویسی و دسته‌بندی شده و مثال‌ها، نمونه‌کدها و توضیحات تکمیلی برای افزایش ارزش آموزشی اضافه شده‌اند. نکات اصلاحی یا موارد نیازمند راستی‌آزمایی به‌صورت شفاف مشخص شده‌اند.

راهنمای مطالعه

فهرست مطالب

این جزوه از مبانی تحلیل داده آغاز می‌شود و به طراحی مدل، کاربردهای یادگیری عمیق، توسعه محصول و نقشه راه یادگیری می‌رسد. هر فصل را می‌توان جداگانه خواند؛ پیشنهاد می‌شود ترتیب فصل‌ها حفظ شود.

01 نقشه مفهومی حوزه داده و هوش مصنوعی

02 جمع‌آوری، ساختاردهی و پاک‌سازی داده‌ها

03 تحلیل اکتشافی، شاخص‌ها و مصورسازی

04 آمار کاربردی؛ پلی میان داده و نتیجه‌گیری

05 ابزارها و کتابخانه‌های تخصصی پایتون

06 مبانی یادگیری ماشین؛ از مثال تا مدل

07 رگرسیون، دسته‌بندی و خوشه‌بندی

08 ارزیابی، خطاها و نقش واقعی هوش مصنوعی کمکی

09 شبکه‌های عصبی مصنوعی و یادگیری عمیق

10 بینایی ماشین و پردازش تصویر

11 صوت، متن و معماری ترنسفورمر

12 از مدل تا محصول: API، بک‌اند و استقرار

13 ایجنت‌ها، ابزارهای آماده و توسعه اختصاصی

14 توسعه و بومی‌سازی مدل‌های فارسی

15 مطالعات موردی: فروشگاه، صنعت، بانک و سلامت

16 نقشه راه یادگیری، بازار کار و تمرین پایانی

فصل 01

نقشه مفهومی حوزه داده و هوش مصنوعی

۱. از داده خام تا تصمیم‌گیری هوشمند

در بخش پایانی وبینار، مدرس مسیر یادگیری را از **تحلیل داده** آغاز کرد، سپس به **علم داده** و **یادگیری ماشین** رسید و در ادامه **شبکه‌های عصبی، یادگیری عمیق، مدل‌های مولد و ایجنت‌ها** را معرفی کرد. این مسیر برای آموزش، ترتیب مفیدی دارد؛ بااین‌حال در دنیای حرفه‌ای مرز حوزه‌ها همیشه خطی و جدا از یکدیگر نیست.

جمع‌آوری داده ← پاک‌سازی و تحلیل ← مدل‌سازی و ارزیابی ← ارائه سرویس و پایش

حوزه	سؤال اصلی	خروجی نمونه
تحلیل داده (Data Analysis)	چه اتفاقی افتاده و چه الگویی دیده می‌شود؟	گزارش فروش، شاخص، داشبورد
علم داده (Data Science)	با داده و روش علمی چه بینشی یا مدلی می‌توان ساخت؟	تحلیل آماری، آزمایش، مدل پیش‌بینی
یادگیری ماشین (Machine Learning)	آیا الگوریتم می‌تواند رابطه‌ای را از نمونه‌ها بیاموزد؟	پیش‌بینی مقدار، طبقه‌بندی یا خوشه‌بندی
یادگیری عمیق (Deep Learning)	چگونه مدل‌های چندلایه الگوهای پیچیده را یاد بگیرند؟	تحلیل تصویر، گفتار، متن و داده‌های پیچیده
مهندسی محصول هوشمند	چگونه مدل به یک خدمت مطمئن و قابل استفاده تبدیل شود؟	API، اپلیکیشن، پایش، اتصال به سامانه

۲. تفاوت اصطلاحات را درست بفهمیم

- **هوش مصنوعی (AI)** چتری گسترده‌تر برای ساخت سیستم‌هایی است که وظایف مرتبط با توانایی‌های شناختی را انجام می‌دهند.
- **یادگیری ماشین (ML)** زیرمجموعه‌ای از AI است؛ مدل از داده و تجربه آموزشی الگو یاد می‌گیرد.
- **یادگیری عمیق (DL)** دسته‌ای از روش‌های یادگیری ماشین بر پایه شبکه‌های عصبی چندلایه است.
- **علم داده** یک حوزه بین‌رشته‌ای است و الزاماً محدود به یادگیری ماشین نیست؛ آمار، شناخت مسئله، مهندسی داده و ارتباط نتایج را هم شامل می‌شود.

نکته تکمیلی: در جلسه برای ساده‌سازی گفته شد علم داده «یک پله بالاتر» از تحلیل داده است. از نظر حرفه‌ای، بهتر است این‌ها را حوزه‌های هم‌پوشان بدانیم؛ تحلیل داده می‌تواند بخشی از علم داده باشد و بسیاری از پروژه‌های علم داده بدون مدل یادگیری ماشین هم ارزشمندند.

۳. مسیر یادگیری را بر اساس خروجی انتخاب کنید

برای شروع لازم نیست متخصص همه شاخه‌ها باشید. اگر خروجی مطلوبتان گزارش و تصمیم‌سازی است، تحلیل داده نقطه شروع مناسبی است. اگر هدف شما پیش‌بینی و ساخت مدل است، باید آمار و یادگیری ماشین را جدی‌تر دنبال کنید. اگر قصد دارید یک محصول تجاری عرضه کنید، در ادامه به مباحث بک‌اند، API، استقرار و امنیت نیز نیاز خواهید داشت.

فصل 02

جمع‌آوری، ساختاردهی و پاک‌سازی داده‌ها

۱. داده از کجا وارد پروژه می‌شود؟

مدرس توضیح داد که داده ممکن است از فایل **Excel** یا **CSV**، پایگاه داده، وب‌اپلیکیشن و سامانه‌های سازمانی وارد شود. اولین تصمیم فنی این است که بدانیم داده کجا قرار دارد، چه قالبی دارد، چه کسی مجاز به استفاده از آن است و چگونه باید آن را به‌صورت قابل تکرار دریافت کنیم.

منبع	روش نمونه دریافت	نکته مهم
Excel / CSV	Pandas	بررسی نوع ستون‌ها و کدگذاری متن
پایگاه داده	SQL و اتصال مجاز	کنترل دسترسی و محرمانگی
API سازمانی	درخواست HTTP	محدودیت نرخ و وضعیت خطا
وب‌سایت عمومی	API رسمی یا استخراج مجاز	رعایت شرایط استفاده و قوانین
حسگر و دستگاه	فایل ثبت رویداد یا جریان داده	زمان‌سنجی و کالیبراسیون

۲. مشکلات متداول کیفیت داده

مقدار گمشده (Missing Value): فیلدهایی که مقدار ندارند؛ مانند سن خالی یا قیمت ثبت‌نشده. **رکورد تکراری (Duplicate):** ثبت یک رویداد چند بار. **ناهمگونی قالب (Format):** تاریخ‌ها یا اعداد با الگوهای مختلف. **خطای تایپ و اندازه‌گیری:** مانند درج واحد اشتباه یا مقدار خارج از دامنه. **داده پرت (Outlier):** مشاهده‌ای بسیار متفاوت که ممکن است واقعی یا ناشی از خطا باشد.

- **گام ۱** شناخت ستون‌ها، واحدها و معنای هر رکورد
- **گام ۲** بررسی فراوانی مقادیر ناقص، تکراری و نامعتبر
- **گام ۳** تعریف تصمیم برای حذف، اصلاح یا جایگزینی مقادیر
- **گام ۴** ثبت تغییرات و نگهداری نسخه داده خام
- **گام ۵** کنترل کیفیت خروجی پیش از تحلیل یا آموزش مدل

۳. حذف‌کردن همیشه پاسخ درست نیست

برای مقدار گمشده می‌توان رکورد را حذف کرد، مقدار را براساس قواعد معتبر جایگزین کرد یا «گمشده‌بودن» را به‌عنوان ویژگی نگه داشت. انتخاب روش به علت فقدان داده، ماهیت ستون و هدف تحلیل بستگی دارد. حذف بدون بررسی ممکن است نتایج را منحرف کند.

> **احتیاط:** در داده پزشکی، بانکی یا صنعتی، اصلاح داده بدون مستندات و تأیید صاحب داده می‌تواند پیامد جدی داشته باشد. هیچ نتیجه‌ای نباید بر اساس داده دست‌کاری‌شده یا غیرقابل‌ردیابی به‌عنوان حقیقت قطعی ارائه شود.

۴. نمونه کوتاه Pandas

```
import pandas as pd

df = pd.read_csv("sales.csv")
print(df.head())
print(df.isna().sum())
print("Duplicate rows:", df.duplicated().sum())
print(df.describe(include="all"))
```

این کد تنها یک **بازرسی مقدماتی** است؛ نه یک فرایند کامل پاک‌سازی.

فصل 03

تحلیل اکتشافی، شاخص‌ها و مصورسازی

۱. هدف تحلیل داده چیست؟

در سخنرانی مثال داده شد که مدیر نباید برای فهم رفتار ۵۰ هزار رکورد، همه سطرها را جداگانه بخواند. تحلیلگر داده، اطلاعات را به شاخص‌ها، مقایسه‌ها و نمایش‌های قابل‌فهم تبدیل می‌کند تا تصمیم‌گیری بر پایه شواهد انجام شود.

در تحلیل اکتشافی داده‌ها (EDA) پرسش‌هایی از این دست مطرح می‌شود: فروش کدام محصول بیشتر بوده؟ تغییرات ماهانه چگونه است؟ کدام گروه مشتری رفتار متفاوتی دارد؟ آیا میان دو متغیر ارتباطی مشاهده می‌شود؟

۲. انتخاب نمودار متناسب با سؤال

نمودار	کاربرد آموزشی مناسب	خطای متداول
خطی	نمایش روند در طول زمان	نادیده‌گرفتن فاصله زمانی
میله‌ای	مقایسه گروه‌ها و دسته‌ها	استفاده از محور گمراه‌کننده
هیستوگرام	مشاهده توزیع متغیر عددی	انتخاب بازه‌های نامناسب
پراکندگی	بررسی ارتباط دو متغیر عددی	نتیجه‌گیری علّی صرفاً از همبستگی
جعبه‌ای	مقایسه میانه و پراکندگی گروه‌ها	حذف بی‌دلیل داده‌های پرت
ماتریس همبستگی	مرور رابطه متغیرهای عددی	تعبیر همبستگی به علت

۳. گزارش خوب فقط نمودار زیبا نیست

نمودار باید عنوان روشن، واحد مناسب، بازه زمانی، منبع داده و توضیح محدودیت داشته باشد. برای مثال، افزایش فروش هم‌زمان با تبلیغ **لزوماً** ثابت نمی‌کند تبلیغ علت افزایش بوده است؛ شاید فصل فروش یا عامل دیگری اثر گذاشته باشد.

۴. نمونه مصورسازی با Matplotlib

```
import pandas as pd
import matplotlib.pyplot as plt

sales = pd.DataFrame({
    "month": ["Jan", "Feb", "Mar", "Apr"],
    "revenue": [120, 145, 132, 181]
})

plt.plot(sales["month"], sales["revenue"], marker="o")
plt.title("Monthly revenue — illustrative data")
plt.xlabel("Month")
plt.ylabel("Revenue (arbitrary units)")
plt.tight_layout()
plt.show()
```

تمرین: سه سؤال مدیریتی درباره یک فروشگاه فرضی بنویسید؛ برای هر سؤال، شاخص و نمودار مناسب پیشنهاد دهید و مشخص کنید برای پاسخ‌دادن به آن چه داده‌ای لازم است.

فصل 04

آمار کاربردی؛ پلی میان داده و نتیجه‌گیری

۱. چرا آمار ضروری است؟

در جلسه نمونه‌ای مطرح شد: آیا می‌توان با مشاهده یک گروه ۵۰ تا ۶۰ نفری، درباره کل جامعه ادعا کرد؟ پاسخ به نحوه انتخاب نمونه، اندازه نمونه، تغییرپذیری داده و روش تحلیل بستگی دارد. آمار کمک می‌کند **میزان عدم‌قطعیت** را در نتیجه‌گیری لحاظ کنیم.

۲. مفاهیم پایه

- **جامعه آماری:** تمام افراد یا موارد مورد نظر مطالعه.
- **نمونه:** زیرمجموعه‌ای از جامعه که درباره آن داده جمع شده است.
- **میانگین و میانه:** دو شاخص از موقعیت مرکزی توزیع.
- **واریانس و انحراف معیار:** سنجه‌هایی برای پراکندگی داده.
- **احتمال و توزیع:** چارچوبی برای توصیف رخدادها و عدم‌قطعیت.
- **فاصله اطمینان:** بازه برآوردشده با یک روش آماری مشخص؛ نه تضمین قطعی درباره مقدار واقعی.
- **آزمون فرضیه:** روشی برای ارزیابی سازگاری داده با فرض آماری.

۳. خطاهای رایج در تفسیر

سوگیری نمونه‌گیری: افرادی که انتخاب شده‌اند نماینده کل جامعه نیستند. **همبستگی و علیت:** ارتباط مشاهده‌شده لزوماً به معنای اثر علّی نیست. **چندنمایشگری نتایج:** امتحان‌کردن تعداد زیادی تحلیل بدون کنترل مناسب ممکن است روابط تصادفی را مهم جلوه دهد. **اندازه اثر:** حتی نتیجه‌ای که از لحاظ آماری معنادار باشد، ممکن است اثر عملی کوچکی داشته باشد.

۴. آمار در چرخه یادگیری ماشین

آمار در نمونه‌گیری، تحلیل توزیع‌ها، تشخیص انحراف داده، ارزیابی مدل و گزارش عدم‌قطعیت استفاده می‌شود. برنامه‌نویسی به‌تنهایی نمی‌تواند این تصمیم‌ها را جایگزین کند.

> **نکته:** برای شروع بهتر است به‌جای حفظ فرمول‌های طولانی، ابتدا کاربرد میانگین، توزیع، نمونه‌گیری، احتمال شرطی و فاصله اطمینان را با مثال‌های واقعی یاد بگیرید.

فصل 05

ابزارها و کتابخانه‌های تخصصی پایتون

۱. چرا کتابخانه‌ها مهم‌اند؟

پایتون به دلیل اکوسیستم گسترده علمی و مهندسی، گزینه‌ای پرکاربرد در داده و هوش مصنوعی است. به‌جای پیاده‌سازی همه محاسبات از صفر، توسعه‌دهنده از کتابخانه‌های آزموده‌شده استفاده می‌کند و انرژی خود را بر تعریف مسئله، کیفیت داده و ارزیابی نتیجه می‌گذارد.

کتابخانه	وظیفه اصلی	نمونه کاربرد
NumPy	آرایه‌های چندبعدی و محاسبات عددی	ماتریس، میانگین و عملیات برداری
Pandas	جدول داده، تبدیل و پاک‌سازی	CSV، گروه‌بندی و مقدار گمشده
SciPy	محاسبات علمی و روش‌های عددی	بهینه‌سازی و محاسبات آماری
SymPy	ریاضیات نمادین	ساده‌سازی عبارت و حل نمادین
Matplotlib	رسم نمودار	نمودار خطی، پراکندگی، هیستوگرام
Seaborn	نمودارهای آماری سطح‌بالا	مقایسه توزیع‌ها و heatmap
scikit-learn	مدل‌های متداول ML و ارزیابی	رگرسیون، دسته‌بندی، Pipeline
PyTorch	مدل‌سازی شبکه عصبی و یادگیری عمیق	آموزش مدل و محاسبه گرادیان
TensorFlow / Keras	توسعه و اجرای مدل‌های DL	شبکه‌های عصبی و استقرار
Transformers	استفاده و توسعه مدل‌های مبتنی بر ترنسفورمر	پردازش متن و تصویر

۲. تفاوت NumPy و فهرست‌های پایتون

فهرست پایتون ساختاری عمومی است؛ آرایه NumPy برای بسیاری از عملیات عددی برداری و داده‌های هم‌نوع، پیاده‌سازی کارآمدتری ارائه می‌دهد. بااین‌حال «NumPy همیشه سریع‌تر است» گزاره دقیقی نیست؛ کارایی به حجم داده و نوع عملیات بستگی دارد.

```
import numpy as np

values = np.array([12, 18, 20, 30], dtype=float)
print("Mean:", values.mean())
print("Standard deviation:", values.std())
```

۳. انتخاب ابزار از مسئله شروع می‌شود

نیازی نیست همه کتابخانه‌ها را هم‌زمان یاد بگیرید. برای تحلیل داده، Pandas، NumPy و نمودارها را در اولویت بگذارید. برای ML به scikit-learn و اصول ارزیابی برسید. برای DL یکی از چارچوب‌های اصلی مثل PyTorch را عمیق‌تر یاد بگیرید.

فصل 06

مبانی یادگیری ماشین؛ از مثال تا مدل

۱. یادگیری ماشین چگونه مسئله را صورت‌بندی می‌کند؟

مدرس با مثال‌هایی از قیمت خانه، کیفیت محصول صنعتی و تصاویر پزشکی توضیح داد که در بسیاری از مسائل، **ورودی‌ها (Features)** و **خروجی هدف (Target)** داریم. هدف این است که مدل از نمونه‌های موجود رابطه‌هایی را بیاموزد تا برای داده جدید، خروجی معقول تولید کند.

مثال‌های آموزشی

- **قیمت ملک:** متراژ، منطقه و ویژگی‌های ملک ← قیمت تخمینی.
- **کیفیت صنعتی:** دما، فشار و سرعت فرآیند ← شاخص کیفیت.
- **تصویر پزشکی:** تصویر ← احتمال تعلق به یک دسته تعریف‌شده در داده برچسب‌خورده.
- **رفتار مشتری:** سوابق خرید ← احتمال خرید گروه محصول.

> **هشدار پزشکی:** مدل تشخیص تصویر در این جزوه فقط برای توضیح مفهوم طبقه‌بندی ذکر می‌شود. به‌کارگیری عملی برای تصمیم پزشکی مستلزم داده معتبر، ارزیابی بالینی، مجوزها و نظارت متخصص است.

۲. مراحل استاندارد ساخت مدل

تعریف مسئله: قرار است چه تصمیمی بهتر شود و خطای قابل قبول چقدر است؟

آماده‌سازی داده: برچسب، کیفیت، تعادل و نمایندگی نمونه‌ها بررسی شود.

تقسیم داده: داده آموزش، اعتبارسنجی و آزمون با روش متناسب جدا شود.

انتخاب مبنا: یک روش ساده و قابل فهم به‌عنوان baseline ساخته شود.

آموزش و تنظیم: مدل روی آموزش یاد بگیرد و بدون آلودگی داده تنظیم شود.

ارزیابی: معیارهایی متناسب با هزینه خطا و هدف کسب‌وکار بررسی شود.

استقرار و پایش: تغییر رفتار داده و کیفیت پیش‌بینی در زمان کنترل شود.

۳. دو نگاه مهم به یادگیری

نظارت‌شده (Supervised): نمونه‌ها خروجی معلوم دارند؛ مانند قیمت ملک یا برچسب دسته. **بدون نظارت (Unsupervised):** خروجی ازپیش‌برچسب‌خورده نداریم و به دنبال ساختارهای داده مانند گروه‌های مشابه هستیم.

۴. چرا مدل باید برای داده ندیده آزمون شود؟

حفظ‌کردن داده آموزش به معنی یادگیری قابل تعمیم نیست. اگر مدلی روی داده‌های قبلی عالی عمل کند اما در مواجهه با مشتری یا دستگاه جدید خطا داشته باشد، از منظر کاربرد عملی قابل اتکا نیست.

فصل 07

رگرسیون، دسته‌بندی و خوشه‌بندی

۱. رگرسیون (Regression)

برای پیش‌بینی مقدار عددی استفاده می‌شود؛ مانند قیمت خانه، میزان مصرف انرژی یا مقدار یک ویژگی صنعتی. نمونه الگوریتم‌ها شامل رگرسیون خطی، درخت رگرسیون، جنگل تصادفی و روش‌های بردار پشتیبان رگرسیونی (SVR) هستند.

۲. طبقه‌بندی (Classification)

برای انتخاب یک کلاس یا برآورد احتمال تعلق به کلاس به کار می‌رود؛ مانند «ایمیل مربوط به فروش یا پشتیبانی» یا «تراکنش نیازمند بررسی». خروجی طبقه‌بندی معمولاً باید با هزینه انواع خطا ارزیابی شود.

۳. خوشه‌بندی (Clustering)

در خوشه‌بندی، هدف پیداکردن گروه‌هایی از مشاهده‌های مشابه بدون برچسب ازپیش‌تعیین‌شده است. مثلاً گروه‌بندی مشتریان براساس رفتار خرید؛ اما معنای تجاری خوشه‌ها باید جداگانه تفسیر شود.

مسئله	نمونه خروجی	معیار ارزیابی نمونه
رگرسیون	قیمت تخمینی ملک	MAE، RMSE و بررسی خطا
طبقه‌بندی	گروه درخواست مشتری	Precision، Recall، F1 و ماتریس سردرگمی
خوشه‌بندی	گروه‌های رفتاری	تحلیل پایداری و تفسیرپذیری خوشه‌ها

۴. یک نمونه کد آموزشی با scikit-learn

```
from sklearn.datasets import load_diabetes
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error

X, y = load_diabetes(return_X_y=True)
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)

model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
print("MAE:", mean_absolute_error(y_test, y_pred))
```

این داده آموزشیِ مشهور، برای یادگیری برنامه‌نویسی مدل است؛ خروجی آن یک ابزار تشخیص یا تصمیم‌گیری پزشکی محسوب نمی‌شود.

۵. چرا انتخاب الگوریتم فقط با یک دستور انجام نمی‌شود؟

مدرس بر فهم تفاوت مدل‌هایی مانند SVR و Random Forest تأکید داشت. انتخاب به تعداد نمونه، نوع ویژگی‌ها، نیاز به تفسیر، محدودیت زمان پاسخ، توزیع خطا و هزینه استقرار وابسته است. بهتر است چند روش مناسب با یک چارچوب ارزیابی منصفانه مقایسه شوند.

فصل 08

ارزیابی، خطاها و نقش واقعی هوش مصنوعی کمکی

۱. چرا دقت بالا کافی نیست؟

فرض کنید تنها یک درصد تراکنش‌ها واقعاً تقلبی باشند. مدلی که همیشه «سالم» اعلام کند، ۹۹ درصد Accuracy دارد اما هیچ تقلبی را پیدا نمی‌کند. بنابراین باید معیار ارزیابی را متناسب با هدف انتخاب کرد.

Precision می‌پرسد از هشدارهای مثبت، چند مورد درست بوده است؟ **Recall** می‌پرسد از موارد مثبت واقعی، چند مورد شناسایی شد؟ در پروژه‌های حساس، هزینه مثبت کاذب و منفی کاذب ممکن است بسیار متفاوت باشد.

۲. خطای نشت داده (Data Leakage)

اگر اطلاعاتی که در زمان پیش‌بینی واقعاً در دسترس نیست، هنگام آموزش وارد مدل شود، ارزیابی ظاهراً عالی ولی گمراه‌کننده خواهد بود. نمونه رایج: استفاده از داده آینده برای پیش‌بینی گذشته، یا انجام پردازش مبتنی بر کل داده پیش از تفکیک آموزش و آزمون.

۳. بیش‌برازش و کم‌برازش

- **Overfitting:** مدل جزئیات و نویز داده آموزش را بیش از حد یاد می‌گیرد و روی داده جدید ضعیف عمل می‌کند.
- **Underfitting:** مدل آن‌قدر ساده یا ناکافی است که الگوی اصلی را هم یاد نمی‌گیرد.
- **Drift:** داده واقعی در طول زمان تغییر می‌کند و مدل ممکن است نیازمند ارزیابی مجدد یا بازآموزی شود.

۴. آیا ChatGPT جای متخصص را می‌گیرد؟

در سخنرانی تأکید شد که ابزارهای مولد کد می‌توانند نوشتن، توضیح‌دادن و آزمودن کد را سریع‌تر کنند؛ اما تشخیص مناسب‌بودن داده، انتخاب معیار، طراحی آزمون و تفسیر ریسک همچنان به قضاوت تخصصی و کنترل مستقل نیاز دارد.

> **دقت علمی:** در متن گفتار به پژوهش‌هایی منسوب به نشریات علمی اشاره شد، اما اطلاعات کتاب‌شناختی کافی برای شناسایی دقیق آن‌ها ارائه نشده است. بنابراین در این جزوه نتیجه کمی یا نقل‌قولی از آن پژوهش‌ها به‌عنوان منبع تأییدشده بازنشر نشده است.

۵. چک‌لیست پیش از پذیرش خروجی تولیدشده با AI

- ☐ کد را فهمیده‌ام و می‌توانم توضیح دهم.
- ☐ نتیجه را روی داده آزمون مستقل بررسی کرده‌ام.
- ☐ نشت داده و سوگیری‌های احتمالی را بررسی کرده‌ام.
- ☐ اطلاعات حساس بی‌مجوز وارد ابزار خارجی نکرده‌ام.
- ☐ معیار موفقیت برای استفاده عملی تعریف شده است.

فصل 09

شبکه‌های عصبی مصنوعی و یادگیری عمیق

۱. شبکه عصبی مصنوعی چیست؟

شبکه عصبی مصنوعی (ANN) ساختاری محاسباتی با لایه‌ها و واحدهای پردازشی است. در آموزش، پارامترهایی مانند **وزن‌ها (Weights)** و **بایاس‌ها (Biases)** تغییر می‌کنند تا مدل بتواند رابطه ورودی و خروجی را بهتر تقریب بزند. این نام از نورون زیستی الهام گرفته است، ولی شبکه عصبی توصیف دقیقی از مغز انسان نیست.

ویژگی‌های ورودی ← لایه‌های پردازش ← خروجی مدل

۲. یادگیری عمیق چه تفاوتی دارد؟

اصطلاح یادگیری عمیق معمولاً به شبکه‌هایی با لایه‌های متعدد یا نمایش‌های پیچیده اشاره می‌کند. این روش‌ها در بسیاری از مسائل تصویر، متن و گفتار موفق بوده‌اند؛ اما لزوماً برای تمام مسائل بهترین انتخاب نیستند.

۳. آیا مدل‌های سنتی ML با داده بزرگ کار نمی‌کنند؟

در متن جلسه برای توضیح مفهومی، محدودیت مقیاس برخی الگوریتم‌های قدیمی مطرح شد. تفسیر دقیق‌تر این است که **مقیاس‌پذیری به الگوریتم، پیاده‌سازی، ابعاد داده، سخت‌افزار و هدف وابسته است**. بسیاری از مدل‌های غیرعمیق روی داده‌های بسیار بزرگ نیز استفاده می‌شوند و برخی شبکه‌های عصبی برای داده کم عملکرد ضعیفی دارند.

معیار	مدل‌های رایج غیرعمیق	شبکه‌های عمیق
داده جدولی محدود	اغلب نقطه شروع مناسب	گاهی هزینه بیشتر بدون مزیت واضح
تصویر، گفتار و متن پیچیده	ممکن است نیازمند استخراج ویژگی باشند	اغلب توانمند در یادگیری نمایش
توضیح‌پذیری	در برخی روش‌ها بالاتر	معمولاً به ابزارهای تفسیر نیاز دارد
هزینه محاسباتی	اغلب پایین‌تر، نه همیشه	می‌تواند بسیار بالا باشد
نیاز به داده برچسب‌خورده	متغیر	در آموزش از صفر اغلب زیاد؛ انتقال یادگیری کمک می‌کند

۴. PyTorch و TensorFlow

در سخنرانی، PyTorch به‌عنوان چارچوب ترجیحی مدرس برای آموزش DL معرفی شد و از TensorFlow نیز نام برده شد. هر دو ابزار شناخته‌شده‌اند. انتخاب یکی از آن‌ها باید با هدف پروژه، منابع تیم و محیط استقرار هماهنگ باشد.

نکته اجرایی: ابتدا مدل مبنا بسازید و انتخاب شبکه عمیق را به داده و نتیجه ارزیابی وابسته کنید.

فصل 10

بینایی ماشین و پردازش تصویر

۱. پردازش تصویر و بینایی ماشین

مدرس به کاربردهایی مانند تشخیص شیء، دسته‌بندی تصویر و بخش‌بندی اشاره کرد. بخشی از پردازش تصویر با روش‌های کلاسیک مانند فیلتر و عملیات هندسی انجام می‌شود؛ بخشی دیگر به روش‌های یادگیری ماشین و شبکه عصبی متکی است.

۲. سه مسئله متفاوت

وظیفه	سؤال	خروجی
Classification	این تصویر متعلق به کدام دسته است؟	برچسب یا احتمال کلاس
Object Detection	چه اشیایی و کجای تصویر هستند؟	جعبه‌های مکانی و کلاس‌ها
Segmentation	هر بخش/پیکسل متعلق به چیست؟	نقشه ناحیه‌ها

۳. شبکه کانولوشنی (CNN)

CNN خانواده‌ای از معماری‌های شبکه عصبی است که از عملیات کانولوشن برای پردازش ساختار مکانی داده، به‌ویژه تصویر، استفاده می‌کند. در جلسه این خانواده به‌عنوان یکی از روش‌های مهم بینایی ماشین معرفی شد.

امروزه روش‌های دیگری ازجمله معماری‌های مبتنی بر ترنسفورمر نیز در بینایی ماشین وجود دارند؛ بنابراین CNN تنها انتخاب نیست.

۴. از آزمایش تا محصول

درباره هر مسئله باید معلوم باشد داده‌ها چگونه برچسب‌گذاری شده‌اند، آیا افراد/دستگاه‌های جدید در آزمون نمایندگی دارند، حساسیت خطا چقدر است و کدام خروجی نیازمند تأیید انسانی خواهد بود.

تمرین: سه سناریو تعریف کنید: کنترل کیفیت بسته‌بندی، شمارش اشیا در انبار و تشخیص نوع یک تصویر آزمایشگاهی. برای هر مورد تعیین کنید دسته‌بندی، تشخیص شیء یا بخش‌بندی مناسب‌تر است و چه مجوزها یا ریسک‌هایی وجود دارد.

فصل 11

صوت، متن و معماری ترنسفورمر

۱. تبدیل گفتار و متن به یکدیگر

یکی از مثال‌های جلسه، سامانه‌ای بود که **صدا را به متن** تبدیل می‌کند. در آموزش یک مدل تشخیص گفتار، نمونه‌های صوت و متن متناظر می‌توانند نقش داشته باشند. امروز بسته به کاربرد، مدل‌های آماده و روش‌های مختلفی برای تبدیل گفتار به متن (STT)، تبدیل متن به گفتار (TTS) و پردازش گفتار وجود دارد.

۲. ترنسفورمر چیست؟

Transformer خانواده‌ای از معماری‌های شبکه عصبی مبتنی بر مکانیزم‌های توجه (Attention) است که در بسیاری از مدل‌های زبانی و چندوجهی استفاده می‌شود. مدل‌های مبتنی بر ترنسفورمر می‌توانند بسته به طراحی و آموزش برای تولید متن، طبقه‌بندی، خلاصه‌سازی و کارهای دیگر به‌کار روند.

۳. کاربردهای چندوجهی (Multimodal)

- متن ← متن: پاسخ‌گویی و خلاصه‌سازی.
- گفتار ← متن: رونوشت جلسه و زیرنویس.
- متن ← گفتار: خواندن متن با صوت طبیعی.
- تصویر ← متن: توصیف محتوا یا استخراج اطلاعات.
- متن ← تصویر: تولید تصویر.
- متن و تصویر ← متن: پاسخ به پرسش درباره عکس.

این‌ها «نوع وظیفه» هستند؛ همه آن‌ها الزاماً با یک معماری واحد یا یک مدل واحد انجام نمی‌شوند.

۴. آیا می‌توان مدل اختصاصی ساخت؟

بسته به بودجه و هدف، سه راه قابل بررسی است: استفاده از مدل آماده، **تنظیم تکمیلی (Fine-tuning)** مدل مناسب روی داده مجاز، یا آموزش مدل از ابتدا. مسیر سوم عموماً به داده، محاسبات و تخصص بیشتری نیاز دارد.

> **نکته فنی:** در سخنرانی ترنسفورمر به‌صورت آموزشی به‌عنوان معماری بسیار رایج توضیح داده شد. این معماری «آخرین معماری ممکن» یا تنها پایه همه ابزارهای هوش مصنوعی نیست و تعمیم درصدیِ بدون منبع معتبر مناسب نیست.

فصل 12

از مدل تا محصول: API، بک‌اند و استقرار

۱. چرا داشتن یک فایل مدل کافی نیست؟

در مثال سخنرانی، مدلی برای برآورد قیمت ملک آموزش می‌بینند. اگر قرار باشد بانک یا پلتفرم دیگری از آن استفاده کند، باید راهی امن، قابل اندازه‌گیری و پایدار برای ارسال ورودی و دریافت خروجی وجود داشته باشد.

اپلیکیشن مشتری ← API و احراز هویت ← سرویس مدل ← پاسخ و گزارش

۲. اجزای یک محصول مدل‌محور

API: قرارداد دریافت درخواست و ارائه پاسخ. **بک‌اند:** مدیریت کاربران، سهمیه استفاده، پرداخت، منطق کسب‌وکار و اتصال‌ها. **مدل:** محاسبه پیش‌بینی. **پایگاه داده:** نگهداری اطلاعات و رخدادها طبق مجوز. **پایش:** سنجش خطا، تأخیر، هزینه و تغییر عملکرد مدل.

۳. مدل‌های کسب‌وکار

در سخنرانی دو مسیر مطرح شد: فروش مستقیم خدمت به مصرف‌کننده، یا ارائه API به سازمان‌ها (B2B). در مدل B2B می‌توان قرارداد مبتنی بر تعداد درخواست، اشتراک، سطح خدمات و پشتیبانی تعریف کرد؛ اعداد و قیمت‌های مطرح‌شده در مثال شخصی سخنران، تعرفه عمومی بازار نیستند.

۴. مثال فرضی سرویس پیش‌بینی

```
{
  "property_area": 95,
  "bedrooms": 2,
  "parking": true,
  "region_id": "R12"
}
```

پاسخ نمونه می‌تواند مقدار تخمین، نسخه مدل و هشدار محدودیت داشته باشد. آنچه سرویس را به محصول واقعی تبدیل می‌کند صرفاً پاسخ عددی نیست؛ دقت مستند، امنیت، پاسخ‌گویی، نگهداری و قرارداد خدمات هم اهمیت دارند.

پیوند با دیگر بخش‌های وبینار: آموزش Django یا دیگر ابزارهای توسعه بک‌اند با دوره یادگیری ماشین در تعارض نیست؛ برای عرضه مدل در قالب سرویس، این مهارت‌ها اغلب مکمل یکدیگرند.

فصل 13

ایجنت‌ها، ابزارهای آماده و توسعه اختصاصی

۱. ایجنت چه نسبتی با مدل زبانی دارد؟

در جلسه گفته شد یک مدل زبانی خام عمدتاً ورودی دریافت می‌کند و خروجی تولید می‌کند؛ اما برای اجرای کار در سامانه‌های واقعی باید به ابزار، مجوز، داده و جریان کاری مناسب متصل شود. در طراحی ایجنت، مدل می‌تواند بر اساس قواعد سامانه درخواست استفاده از یک ابزار را پیشنهاد دهد و برنامه میزبان آن درخواست را کنترل و اجرا کند.

۲. دو رویکرد رایج

رویکرد	مزیت احتمالی	محدودیت احتمالی
ابزارهای بصری مانند n8n	توسعه سریع، مشاهده مراحل، اتصال آماده	محدودیت یکپارچه‌سازی یا منطق بسیار اختصاصی
توسعه کدنویسی با Python و کتابخانه‌ها	کنترل بیشتر، تست و توسعه سفارشی	هزینه توسعه، عملیات و نگهداری بیشتر
روش ترکیبی	استفاده از سرعت ابزار بصری و انعطاف کد	نیاز به قرارداد روشن میان اجزا

۳. نمونه آموزشی ایجنت اطلاعات سازمانی

فرض کنید یک کارمند در تلگرام از وضعیت سفارش می‌پرسد. جریان ممکن است این‌گونه باشد: دریافت پیام ← شناسایی درخواست ← کنترل هویت و مجوز ← دریافت وضعیت از API سفارش ← تولید پاسخ قابل اتکا ← ثبت رخداد. در هیچ مرحله‌ای نباید مدل بدون احراز هویت به اطلاعات سایر مشتریان دسترسی داشته باشد.

۴. هزینه و محرمانگی بخشی از انتخاب معماری‌اند

انتخاب سرویس خارجی، مدل میزبانی‌شده در سازمان یا مدل متن‌باز باید با بررسی هزینه کل مالکیت انجام شود: هزینه هر درخواست، زیرساخت، پشتیبانی، کیفیت زبان فارسی، امنیت، محدودیت‌های قانونی، تأخیر و سطح کنترل بر داده.

> **نکته:** «ساخت مدل اختصاصی» همیشه اقتصادی‌تر از API آماده نیست. گاهی هزینه GPU، توسعه، به‌روزرسانی و پایش از مبلغ مصرف سرویس آماده بیشتر می‌شود؛ در برخی پروژه‌های محرمانه یا پرمصرف، میزبانی داخلی می‌تواند منطقی‌تر باشد.

فصل 14

توسعه و بومی‌سازی مدل‌های فارسی

۱. فرصت‌ها و مسئله زبان فارسی

در پایان سخنرانی، فرصت توسعه مدل‌های متناسب با زبان فارسی و نیازهای سازمان‌های داخل ایران مطرح شد. نمونه‌هایی شامل پردازش پیام مشتری، پالایش نظرها، تبدیل گفتار فارسی به متن و پاسخ‌گویی سازمانی بود.

۲. مدل آماده را از کجا بیابیم؟

Hugging Face یکی از بسترهای شناخته‌شده برای انتشار و دسترسی به مدل‌ها، داده‌ها و ابزارهای یادگیری ماشین است. پیش از انتخاب مدل باید مجوز استفاده، اندازه، زبان‌های پشتیبانی‌شده، کیفیت مستند، محدودیت منابع و شرایط محرمانگی را مطالعه کرد.

۳. مراحل یک پروژه بومی‌سازی

1. تعریف وظیفه دقیق (مثلاً دسته‌بندی پیام فارسی).
2. گردآوری داده مجاز و باکیفیت، همراه با بررسی محرمانگی.
3. تعیین داده آزمون مستقل و معیار موفقیت.
4. ارزیابی مدل آماده به‌عنوان مبنا.
5. اجرای تنظیم تکمیلی یا بهبود پرامپت/بازیابی، اگر لازم باشد.
6. سنجش کیفیت روی گویش‌ها، نگارش‌ها و نمونه‌های واقعی.
7. استقرار محدود، بازخورد انسانی و پایش مستمر.

۴. Fine-tuning با آموزش از صفر تفاوت دارد

Fine-tuning یعنی ادامه آموزش یک مدل پیش‌آموزش‌دیده برای وظیفه یا داده هدف. بسته به روش، ممکن است کل پارامترها یا بخش محدودی از آن‌ها به‌روزرسانی شود؛ صرفاً تغییر «چند لایه آخر» تعریف همیشگی آن نیست.

تمرین پروژه: یک سرویس دسته‌بندی پیام‌های پشتیبانی فارسی طراحی کنید. سه برچسب نمونه تعیین کنید، درباره خطر خطای طبقه‌بندی توضیح دهید و مشخص کنید در چه شرایطی پیام باید به اپراتور انسانی ارجاع شود.

فصل 15

مطالعات موردی: فروشگاه، صنعت، بانک و سلامت

مطالعه موردی ۱. فروشگاه اینترنتی

مسئله: فروشگاه داده سفارش و رفتار مشتری دارد. **تحلیل داده:** پرفروش‌ترین محصولات، روند درآمد، سبد خرید و موجودی. **علم داده/ML:** پیش‌بینی تقاضا یا پیشنهاد محصول پس از تعریف معیار موفقیت. **محصول نهایی:** داشبورد، API پیش‌بینی و سامانه پایش خطا.

مطالعه موردی ۲. نگهداری تجهیزات صنعتی

مسئله: اطلاعات دما، فشار، سرعت و زمان تعمیر ثبت شده است. **تحلیل داده:** یافتن تغییرات و الگوها. **مدل‌سازی:** برآورد احتمال خرابی یا زمان باقی‌مانده تا تعمیر، تنها در صورت وجود داده معتبر. **کاربرد عملی:** ارائه هشدار کمکی به متخصص، با کنترل نرخ هشدار کاذب و ملاحظات ایمنی.

مطالعه موردی ۳. بانک و املاک

مسئله: برآورد ارزش ملک برای تحلیل یک پرونده مالی. **تحلیل:** بررسی ویژگی املاک و کیفیت معاملات گذشته. **مدل:** قیمت تخمینی همراه با خطا و دامنه کاربرد. **سرویس:** API دارای قرارداد دسترسی، ثبت نسخه مدل و پایش. تصمیم سرمایه‌گذاری نباید صرفاً متکی به یک پیش‌بینی خودکار باشد.

مطالعه موردی ۴. سلامت

مسئله: دسته‌بندی تصویر یا پردازش گزارش پزشکی. **چالش:** کیفیت برچسب، سوگیری نمونه، تفاوت مراکز درمانی و هزینه خطا. **نتیجه آموزشی:** معیارهای آماری و اعتبارسنجی تخصصی از پیاده‌سازی صرف کد مهم‌ترند.

> **منبع این مطالعات:** حوزه‌های مثال‌زده‌شده در سخنرانی مبنا قرار گرفته‌اند، اما ساختار مسئله، مراحل اجرایی و توصیه‌ها برای آموزش بازنویسی شده‌اند و شرح مستند نتایج پروژه واقعی اشخاص یا شرکت‌ها محسوب نمی‌شوند.

فصل 16

نقشه راه یادگیری، بازار کار و تمرین پایانی

۱. ترتیب پیشنهادی یادگیری

مرحله ۱: پایتون، توابع، فایل‌ها، محیط مجازی و اصول Git

مرحله ۲: SQL مقدماتی، NumPy و Pandas، کیفیت داده و EDA

مرحله ۳: آمار، مصورسازی و گزارش‌نویسی قابل دفاع

مرحله ۴: مدل‌های پایه ML با scikit-learn و ارزیابی درست

مرحله ۵: انتخاب تخصص: داده جدولی، تصویر، گفتار، متن یا مدل‌محور

مرحله ۶: در صورت نیاز PyTorch، ترنسفورمر و تنظیم تکمیلی

مرحله ۷: API، استقرار، امنیت، تست و نمونه‌کار قابل ارائه

۲. ساخت رزومه حرفه‌ای

در پرسش‌وپاسخ وبینار، GitHub و LinkedIn به‌عنوان ابزارهای مهم معرفی شدند. برای هر پروژه قابل ارائه، بهتر است هدف، داده مورد استفاده (یا وضعیت عمومی/ساختگی آن)، مراحل، معیار ارزیابی، محدودیت‌ها و روش اجرا را در README بنویسید.

داشتن صفحه GitHub و LinkedIn به‌تنهایی تضمین استخدام نیست. کیفیت پروژه، توان توضیح‌دادن تصمیم‌های فنی، مهارت ارتباطی و میزان تقاضای واقعی بازار نیز اهمیت دارند.

۳. یک پروژه نهایی پیشنهادی

عنوان: تحلیل فروش و پیش‌بینی تقاضا با داده ساختگی.

- داده سفارش شامل تاریخ، محصول، تعداد و مبلغ ایجاد کنید.
- داده را با Pandas بخوانید و خطاهای عمدی را شناسایی کنید.
- پنج شاخص و سه نمودار معنادار تولید کنید.
- یک مدل ساده برای پیش‌بینی مقدار فروش بسازید.
- داده آزمون را مستقل نگه دارید و خطای مدل را گزارش کنید.
- برای خروجی، README و یک API اختیاری تعریف کنید.

۴. چک‌لیست آمادگی دانشجو

- ☐ فرق تحلیل داده، علم داده و یادگیری ماشین را می‌توانم توضیح دهم.
- ☐ می‌توانم داده را از فایل یا پایگاه داده مجاز دریافت کنم.
- ☐ می‌توانم داده گمشده، تکراری و نامعتبر را گزارش کنم.
- ☐ می‌دانم کدام نمودار برای چه پرسشی مناسب است.
- ☐ مفهوم نمونه‌گیری و تفاوت همبستگی و علیت را می‌دانم.
- ☐ برای یک مسئله رگرسیون یا طبقه‌بندی، مدل مبنا می‌سازم.
- ☐ تفاوت داده آموزش و آزمون و خطر نشت داده را می‌شناسم.
- ☐ درباره هزینه، امنیت و استقرار سرویس مدل فکر کرده‌ام.
- ☐ یک پروژه مستند و قابل تکرار در GitHub دارم.

۵. اصطلاحات ضروری

اصطلاح	معادل / مفهوم
Dataset	مجموعه داده
Feature / Target	ویژگی ورودی / متغیر هدف
EDA	تحلیل اکتشافی داده
Data Leakage	نشت داده
Regression / Classification	رگرسیون / طبقه‌بندی
Overfitting	بیش‌برازش
Neural Network	شبکه عصبی
Fine-tuning	تنظیم تکمیلی مدل
Inference	اجرای مدل آموزش‌دیده برای نتیجه‌گیری
Deployment	استقرار در محیط عملیاتی

۶. جمع‌بندی

پایتون نقطه شروع است، نه مقصد نهایی. تحلیل داده کمک می‌کند شواهد را بفهمیم؛ آمار به نتیجه‌گیری مسئولانه معنا می‌دهد؛ یادگیری ماشین امکان ساخت الگوهای پیش‌بینی را فراهم می‌کند؛ شبکه‌های عصبی بعضی مسائل پیچیده را قابل حل می‌کنند و مهندسی نرم‌افزار، این توانایی‌ها را به خدمات واقعی تبدیل می‌کند. ابزارهای هوش مصنوعی مسیر کار را سریع‌تر می‌کنند، اما کیفیت محصول به تعریف درست مسئله، داده قابل اعتماد، آزمون دقیق و مسئولیت‌پذیری تیم وابسته است.