مبانی علم داده چیست؟ مفاهیم مهم، کاربردها

اگر بخواهیم علم داده (Data Science) را بدون اصطلاحات پیچیده توضیح دهیم، می‌توان گفت: علم داده مجموعه‌ای از روش‌ها و مهارت‌ها برای تبدیل داده خام به **بینش، پیش‌بینی و تصمیم قابل استفاده** است.

در این حوزه از آمار، ریاضیات، برنامه‌نویسی، تحلیل داده، یادگیری ماشین و دانش تخصصی یک حوزه استفاده می‌شود. بنابراین علم داده فقط «کدنویسی با Python» یا «ساخت مدل هوش مصنوعی» نیست؛ بخش مهمی از آن از جایی شروع می‌شود که یک سؤال واقعی را به مسئله‌ای قابل حل با داده تبدیل می‌کنیم .نیز علم داده را حوزه‌ای میان‌رشته‌ای می‌داند که آمار، ریاضیات، برنامه‌نویسی، تحلیل پیشرفته، هوش مصنوعی و یادگیری ماشین را با دانش حوزه ترکیب می‌کند تا از داده بینش قابل اقدام استخراج شود .منبعIBM :

آیا داشتن داده برای ساخت مدل به تنهایی کافیه؟

یکی از مهم‌ترین نکاتی که یک فرد مبتدی باید بداند این است که داشتن داده به‌تنهایی دلیل کافی برای ساختن مدل نیست.

فرض کنید مدیر یک فروشگاه اینترنتی می‌پرسد:

«فروش ماه گذشته چقدر بوده است؟»

برای پاسخ، احتمالاً یک گزارش ساده کافی است.

اگر بپرسد:

«چرا فروش این ماه کاهش پیدا کرده؟»

باید وارد تحلیل تشخیصی شویم و عوامل مختلف را بررسی کنیم.

اگر بپرسد:

«کدام مشتریان احتمالاً در ماه آینده خرید نمی‌کنند؟»

مسئله به سمت پیش‌بینی می‌رود و ممکن است یادگیری ماشین ارزشمند باشد.

و اگر سؤال این باشد:

«برای کاهش ریزش مشتری چه اقدامی انجام دهیم؟»

حالا نتیجه تحلیل باید به تصمیم و اقدام تبدیل شود.

به همین دلیل می‌توان مسائل داده‌ای را به چهار سؤال اصلی تقسیم کرد:

نوع رویکرد	سؤال
Descriptive	چه اتفاقی افتاده؟
Diagnostic	چرا اتفاق افتاده؟
Predictive	چه اتفاقی احتمالاً خواهد افتاد؟
Prescriptive	چه اقدامی بهتر است؟

این چهار رویکرد در توضیحات AWS درباره انواع تحلیل داده نیز مطرح شده‌اند.

نتیجه عملی :قبل از انتخاب الگوریتم، اول مشخص کنید دقیقاً چه سؤالی قرار است پاسخ داده شود.

مبانی اصلی علم داده چیست؟

علم داده را بهتر است به چند مهارت به‌هم‌پیوسته تقسیم کنیم.

داده

داده می‌تواند عدد، متن، تصویر، صدا، ویدئو، اطلاعات تراکنش، رفتار کاربر یا داده سنسورها باشد.

داده ممکن است ساختاریافته باشد؛ مانند جدول مشتریان، یا غیرساختاریافته؛ مانند متن و تصویر.

آمار

آمار کمک می‌کند داده را توصیف کنیم، تفاوت‌ها را بسنجیم، الگوها را بررسی کنیم و درباره نتایج با در نظر گرفتن عدم قطعیت قضاوت کنیم.

برنامه‌نویسی

برای پردازش داده، خودکارسازی کارها، تحلیل و ساخت مدل استفاده می‌شود Python .یکی از زبان‌های رایج در علم داده است و R نیز در برخی پروژه‌های آماری کاربرد زیادی دارد IBM .نیز Python و R را از زبان‌های متداول مورد استفاده در علم داده معرفی می‌کند.

SQL

SQLبرای استخراج و کار با داده‌های موجود در پایگاه‌های داده اهمیت زیادی دارد.

مصورسازی داده

گاهی یک نمودار خوب می‌تواند الگویی را آشکار کند که در صدها ردیف جدول دیده نمی‌شود.

یادگیری ماشین

وقتی مسئله به پیش‌بینی، دسته‌بندی یا کشف الگوهای پیچیده مربوط باشد، می‌توان از Machine Learning استفاده کرد.

دانش حوزه

این بخش بسیار مهم است. یک مدل ممکن است از نظر فنی عملکرد خوبی داشته باشد، اما اگر مسئله کسب‌وکار اشتباه تعریف شده باشد، خروجی مدل ارزش چندانی ندارد.

چرخه علم داده؛ از سؤال تا تصمیم

یک پروژه واقعی علم داده معمولاً چنین مسیری دارد:

مدل‌سازی در صورت نیاز → ارزیابی → تفسیر → تصمیم و استفاده

تعریف مسئله → جمع‌آوری داده → پاک‌سازی → تحلیل اکتشافی →

AWSاین فرایند را در قالب OSEMN توضیح می‌دهد: دریافت داده، پاک‌سازی، اکتشاف، مدل‌سازی و تفسیر نتایج.

اما این چرخه الزاماً خطی نیست. ممکن است در مرحله تحلیل متوجه شویم داده کافی نداریم و دوباره به مرحله جمع‌آوری برگردیم.

یک مثال واقعی از پروژه علم داده

فرض کنید یک فروشگاه آنلاین می‌خواهد بداند:

کدام مشتریان احتمالاً در ۳۰ روز آینده خرید مجدد انجام نمی‌دهند؟

این سؤال، نقطه شروع پروژه است.

۱- تعریف مسئله

هدف نباید این باشد که:

»داده مشتریان را بررسی کنیم«.

هدف باید دقیق باشد:

»احتمال خرید مجدد هر مشتری را در بازه ۳۰ روز آینده پیش‌بینی کنیم«.

این تعریف مشخص می‌کند چه داده‌ای لازم داریم و خروجی مدل چه خواهد بود.

۲- تعیین ویژگی‌ها

برای هر مشتری ممکن است اطلاعاتی مانند این داشته باشیم:

ویژگی	مثال
تعداد سفارش	7
مبلغ کل خرید	12میلیون تومان
روزهای گذشته از آخرین خرید	42
تعداد بازدید	15
مدت عضویت	14ماه

ویژگی	مثال
خرید مجدد	بله/خیر

»خرید مجدد« می‌تواند Targetباشد و سایر اطلاعات، Featureهای مدل باشند.

۳- پاک‌سازی

ممکن است بعضی مشتریان چند بار ثبت شده باشند، تاریخ‌ها قالب یکسانی نداشته باشند یا برخی مقادیر خالی باشند.

پاک‌سازی داده شامل اصلاح خطاها، یکسان‌سازی قالب‌ها، مدیریت داده‌های ناقص و بررسی داده‌های غیرعادی است. کیفیت داده قبل از مدل‌سازی اهمیت زیادی دارد و در چرخه‌های استاندارد علم داده مرحله‌ای مستقل محسوب می‌شود.

Data Leakageچیست و چرا مهم است؟

یکی از اشتباهات مهم مبتدیان Data Leakageاست.

فرض کنید در اول خرداد می‌خواهیم پیش‌بینی کنیم مشتری در ماه آینده خرید خواهد کرد یا نه.

اگر اطلاعاتی که در ۲۰ خرداد به دست آمده را وارد داده‌های مدل کنیم، مدل عملاً بخشی از آینده را دیده است.

ممکن است نتیجه آزمایش بسیار خوب به نظر برسد، اما مدل هنگام استفاده واقعی چنین اطلاعاتی در اختیار ندارد.

بنابراین یک قانون مهم وجود دارد:

مدل نباید هنگام پیش‌بینی به اطلاعاتی دسترسی داشته باشد که در زمان واقعی پیش‌بینی هنوز در دسترس نیست.

این نکته یکی از تفاوت‌های مهم میان یک تمرین ساده و یک پروژه قابل اعتماد در دنیای واقعی است.

تحلیل اکتشافی داده‌ها چیست؟

قبل از ساخت مدل باید بدانیم با چه داده‌ای سروکار داریم.

در تحلیل اکتشافی یا **EDA** معمولاً مواردی مانند این بررسی می‌شوند:

- تعداد رکوردها
- نوع متغیرها
- داده‌های گمشده
- داده‌های تکراری
- میانگین و میانه
- پراکندگی
- داده‌های پرت
- روابط میان متغیرها
- توزیع داده‌ها

مصورسازی نیز در همین مرحله بسیار کاربردی است IBM .تحلیل اکتشافی را برای بررسی الگوها، توزیع‌ها، محدوده مقادیر و سوگیری‌های احتمالی داده مهم می‌داند.

یک نکته حرفه‌ای : هر نمودار باید یک سؤال را پاسخ دهد. ساختن نمودار صرفاً برای زیباتر شدن گزارش، تحلیل داده محسوب نمی‌شود.

آیا همیشه باید Machine Learning استفاده کنیم؟

خیر.

فرض کنید می‌خواهیم بدانیم:

»کدام مشتریان در ۹۰ روز گذشته خرید نکرده‌اند؟«

این سؤال ممکن است با یک Query ساده SQL پاسخ داده شود.

اما اگر سؤال این باشد:

»کدام مشتریان با توجه به رفتار قبلی احتمالاً در آینده خرید نمی‌کنند؟«

موضوع به پیش‌بینی نزدیک شده و استفاده از Machine Learning می‌تواند منطقی باشد.

پس:

اول مسئله → بعد داده → بعد روش

نه:

اول الگوریتم → بعد پیدا کردن مسئله برای آن.

مهم‌ترین روش‌های علم داده

روش مورد استفاده به نوع مسئله بستگی دارد.

Classification

مدل داده را در دسته‌های مشخص قرار می‌دهد؛ مثلاً:

اسپم / غیر اسپم

Regression

خروجی معمولاً یک مقدار عددی است؛ مثلاً:

قیمت احتمالی خانه

Clustering

داده‌ها بر اساس شباهت به گروه‌هایی تقسیم می‌شوند، بدون اینکه الزاماً دسته‌بندی از قبل مشخص شده باشد.

AWSنیزClassification ، Regressionو Clustering را از تکنیک‌های رایج علم داده معرفی می‌کند.

ارزیابی مدل؛ چرا Accuracy کافی نیست؟

یکی از اشتباهات رایج این است که تصور کنیم:

»هرچه Accuracy بیشتر باشد، مدل بهتر است«.

همیشه این‌طور نیست.

فرض کنید ۹۹ درصد تراکنش‌های یک سیستم سالم و فقط یک درصد تقلبی باشند. مدلی که تقریباً همه تراکنش‌ها را «سالم» اعلام کند، ممکن است Accuracy بسیار بالایی داشته باشد، اما در تشخیص تقلب عملکرد مناسبی نداشته باشد.

بسته به مسئله ممکن است معیارهایی مانند:

- Precision
- Recall
- F1-score
- MAE
- RMSE

مناسب‌تر باشند.

بنابراین معیار ارزیابی باید بر اساس هزینه واقعی خطا انتخاب شود، نه صرفاً یک عدد جذاب.

علم داده چه تفاوتی باData Analytics ، AIو Machine Learning دارد؟

این اصطلاح‌ها گاهی به جای یکدیگر استفاده می‌شوند، اما یکسان نیستند.

حوزه	تمرکز
Data Analytics	تحلیل داده برای پاسخ به سؤال
Data Science	فرایند گسترده‌تر حل مسئله با داده
Machine Learning	یادگیری الگو از داده برای پیش‌بینی یا تصمیم
Artificial Intelligence	ساخت سیستم‌هایی با قابلیت‌های هوشمند
Data Engineering	ساخت زیرساخت، جریان و سیستم‌های داده

IBMنیز علم داده را حوزه‌ای گسترده و میان‌رشته‌ای معرفی می‌کند و Machine Learning را یکی از حوزه‌های مرتبط با آن می‌داند.

مثال در یک فروشگاه

Data Engineer:
سیستم انتقال و ذخیره داده سفارش‌ها را ایجاد می‌کند.

Data Analyst:
بررسی می‌کند فروش ماه گذشته چرا کاهش یافته است.

Data Scientist:
مدلی برای پیش‌بینی ریزش مشتری می‌سازد.

ML Engineer:
مدل را برای استفاده در یک سیستم واقعی آماده و مستقر می‌کند.

مرز این نقش‌ها در همه شرکت‌ها یکسان نیست و ممکن است یک نفر چند نقش را بر عهده داشته باشد.

کاربردهای علم داده

کاربردهای علم داده بسیار گسترده‌اند، اما بهتر است آنها را با «مسئله» ببینیم نه فقط نام صنعت.

بنابراین ارزش علم داده از خود الگوریتم نمی‌آید؛ از بهبود یک تصمیم یا حل یک مسئله واقعی می‌آید.

مزایا و محدودیت‌های علم داده

مزایا

- کشف الگوهای پنهان
- پیش‌بینی روندها
- شناخت بهتر رفتار مشتری
- خودکارسازی برخی تحلیل‌ها
- کمک به تصمیم‌گیری
- تشخیص ناهنجاری
- بهینه‌سازی فرایندها

محدودیت‌ها

- وابستگی شدید به کیفیت داده
- احتمال سوگیری
- خطای مدل

- هزینه زیرساخت و نگهداری
- دشواری تفسیر برخی مدل‌ها
- خطر استفاده اشتباه از خروجی

علم داده «ماشین تولید حقیقت» نیست. مدل می‌تواند بر اساس داده ناقص یا سوگیرانه نتیجه‌ای تولید کند که ظاهراً دقیق اما در واقع نامناسب باشد AWS .نیز کیفیت داده، پیچیدگی منابع و سوگیری را از چالش‌های مهم این حوزه معرفی می‌کند.

چه زمانی نباید سراغ پروژه پیچیده علم داده برویم؟

این سؤال معمولاً در مقالات مقدماتی فراموش می‌شود.

اگر:

- داده کافی ندارید،
- کیفیت داده بسیار پایین است،
- مسئله مشخص نیست،
- نتیجه مدل هیچ تصمیمی را تغییر نمی‌دهد،
- یک گزارش ساده پاسخ سؤال را می‌دهد،
- هزینه ساخت و نگهداری مدل بیشتر از ارزش آن است،

ممکن است پروژه Machine Learning انتخاب مناسبی نباشد.

راه‌حل پیچیده‌تر لزوماً راه‌حل بهتر نیست.

برای یادگیری علم داده چه چیزهایی لازم است؟

برای شروع لازم نیست تمام ریاضیات دانشگاهی یا ده‌ها الگوریتم را یاد بگیرید.

مسیر پیشنهادی:

مرحله اول: مفاهیم داده

ابتدا بفهمید:

- داده چیست؟
- Featureچیست؟
- Targetچیست؟
- داده ساختاریافته چیست؟
- Missing Valueچیست؟
- Outlierچیست؟

مرحله دوم:Python

مفاهیم پایه مانند:

- متغیر
- شرط
- حلقه
- تابع
- لیست
- دیکشنری
- فایل

مرحله سوم: Pandas وNumPy

برای کار با داده‌های جدولی و محاسبات عددی.

مرحله چهارم: Visualization

نمودار و مهم‌تر از آن **تفسیر نمودار**.

مرحله پنجم: Statistics

حداقل:

- میانگین
- میانه
- واریانس
- انحراف معیار
- احتمال
- توزیع
- همبستگی
- نمونه‌گیری

مرحله ششم: SQL

برای استخراج و تحلیل داده از پایگاه‌های داده.

مرحله هفتم: Machine Learning

بعد از پایه‌ها:

- Regression
- Classification
- Clustering
- Model Evaluation
- Feature Engineering

اگر هدفم مشخص باشد، چه مسیری بهتر است؟

برای Data Analyst شدن

SQL → Excel/BI → Statistics → Visualization → Python/Pandas

برای Data Scientist شدن

Python → Statistics → SQL → Pandas → EDA → Machine Learning → پروژه

اگر کاملاً مبتدی هستید

مفاهیم داده → SQL → Statistics → Visualization → Pandas → Python → پروژهML →

یک منبع رایگان و معتبر برای شروع، دوره **Data Science for Beginners** مایکروسافت است که یک مسیر آموزشی ۱۰ هفته‌ای و ۲۰ درسی دارد و موضوعاتی مانند مفاهیم علم داده، اخلاق داده، داده، آمار، SQL، Visualizationو پروژه‌های عملی را پوشش می‌دهد .منبعMicrosoft :

۷ اشتباه رایج مبتدیان

۱ .شروع مستقیم باDeep Learning
قبل از فهم داده و آمار، مدل‌های پیچیده بیشتر باعث حفظ کردن کد می‌شوند.

۲ .تصور اینکه Data Science همان AI است
یادگیری ماشین و AI می‌توانند بخشی از پروژه باشند، اما کل علم داده نیستند.

۳ .استفاده از الگوریتم قبل از تعریف مسئله
ابتدا سؤال را مشخص کنید.

۴ .نادیده گرفتن کیفیت داده
داده نامناسب می‌تواند کل تحلیل را خراب کند.

۵ .اعتماد بهAccuracy
معیار باید متناسب با مسئله انتخاب شود.

۶ .بی‌توجهی بهData Leakage
داده آینده نباید به شکل مخفی وارد آموزش مدل شود.

۷ .یادگیری بدون پروژه
بعد از هر چند مفهوم، یک مسئله واقعی را اجرا کنید.

Cheat Sheetمبانی علم داده

اگر بخواهید کل مسیر را در یک نقشه کوتاه به خاطر بسپارید:

۱ .مسئله چیست؟

↓

۲ .چه داده‌ای لازم دارم؟

↓

۳ .آیا داده قابل اعتماد است؟

↓

۴ .داده را پاک و آماده کن

↓

۵ .داده را اکتشاف و مصورسازی کن

↓

۶ .آیا واقعاً به مدل نیاز است؟

↓

۷ .روش مناسب را انتخاب کن

↓

۸ .مدل را با معیار مناسب ارزیابی کن

↓

۹ .نتیجه را تفسیر کن

↓

۱۰ .نتیجه را به تصمیم تبدیل کن

↓

۱۱ .عملکرد راه‌حل را در دنیای واقعی بررسی کن

این نگاه از حفظ کردن نام الگوریتم‌ها مهم‌تر است؛ زیرا تقریباً هر پروژه واقعی علم داده با تغییر داده، سؤال یا محدودیت، نیاز به بازگشت به یکی از مراحل قبلی دارد.

پاسخ سریع به سؤال‌های مهم

علم داده چیست؟
حوزه‌ای میان‌رشته‌ای برای استخراج بینش و دانش قابل استفاده از داده با استفاده از آمار، برنامه‌نویسی، تحلیل و روش‌های محاسباتی است.

آیا علم داده همان هوش مصنوعی است؟
خیر AI .و Machine Learning می‌توانند در پروژه‌های علم داده استفاده شوند، اما علم داده گسترده‌تر است.

آیا Python برای علم داده ضروری است؟
ضروری به معنای مطلق نیست، اما Python یکی از زبان‌های بسیار رایج این حوزه است.

آیا باید ریاضی پیشرفته بلد باشم؟
برای شروع خیر. آمار و ریاضیات پایه را یاد بگیرید و با پیشرفت در Machine Learning مباحث پیشرفته‌تر را اضافه کنید.

آیا هر پروژه علم داده باید مدل Machine Learning داشته باشد؟
خیر. اگر تحلیل ساده مسئله را حل کند، ساخت مدل اضافی ضرورتی ندارد.

بهترین روش یادگیری چیست؟
ترکیب مفاهیم پایه با پروژه‌های کوچک واقعی؛ نه فقط تماشای دوره و حفظ کردن کد.

جمع‌بندی

مبانی علم داده را نباید با حفظ کردن ده‌ها الگوریتم شروع کرد.

هسته اصلی این حوزه را می‌توان در یک فرایند خلاصه کرد:

مسئله → داده → بررسی کیفیت → تحلیل → انتخاب روش → ارزیابی → تفسیر → تصمیم

Python، SQL، آمار، Pandasو Machine Learning ابزارهای این مسیر هستند؛ اما هیچ‌کدام به‌تنهایی علم داده نیستند.

اگر تازه شروع کرده‌اید، مهم‌ترین مهارتی که باید تقویت کنید این است که بتوانید یک سؤال مبهم را به مسئله‌ای دقیق و قابل بررسی با داده تبدیل کنید.

همچنین به یاد داشته باشید که هر پروژه‌ای به Machine Learning نیاز ندارد. گاهی یک تحلیل ساده، SQL Queryیا نمودار خوب پاسخ مسئله را می‌دهد. ارزش واقعی علم داده زمانی ایجاد می‌شود که نتیجه تحلیل بتواند درک بهتر، پیش‌بینی بهتر یا تصمیم بهتری ایجاد کند.