داده‌کاوی از دریچه علم آمار

مقدمه‌ای بر ارتباط بین داده کاوی و آمار

چرا داده‌کاوی بخشی از علم آمار نیست؟

پروفسور فریدمن[1]، استاد دانشکده آمار دانشگاه استنفورد، در مقاله‌ای خرد به نام "چه ارتباطی بین داده‌کاوی و علم آمار وجود دارد؟" ، تشریح می‌کند که چرا داده‌کاوی بخشی از علم آمار نیست؛ وی معتقد است اگر داده‌کاوی بخشی از علم آمار باشد، آنگاه حداقل می‌بایست موارد ذیل به وقوع بپیوندد:

- مقالات حوزه داده‌کاوی در نشریات آمار به چاپ برسند.
- مباحث داده‌کاوی به دانشجویان مقطع کارشناسی رشته آمار تدریس شوند.
- مباحث تحقیقاتی مرتبط با داده‌کاوی به دانشجویان تحصیلات تکمیلی رشته آمار تدریس شود.
- شناساندن مشاغل و مفاخر داده‌کاوی، به دانشجویان آمار صورت پذیرد.

بدیهی است که چنین فعالیت‌هایی صورت نمی‌پذیرد، لذا نمی‌توان داده‌کاوی را بخشی از علم آمار دانست. علاوه بر آن، فریدمن معتقد است که مباحثی همچون الگوشناسی، شبکه‌های عصبی، یادگیری ماشین، مدل‌های گرافیکی و مصورسازی داده‌ها در داده‌کاوی مطرح می‌شوند که متعاقباً در رشته آمار نادیده گرفته شدند.

تفاوت و ارتباط بین علم آمار با داده‌کاوی:

پایه و اساس داده‌کاوی در سه شاخه اصلی ریشه دارد که مهم‌ترین آن‌ها آمار است. در واقع بدون آمار، داده‌کاوی معنا نخواهد داشت، چرا که آمار زیربنای اکثر فناوری‌هایی است که داده‌کاوی بر اساس آن‌ها بنا شده است. لازم به‌ذکر است که هوش مصنوعی و یادگیری ماشین، دو شاخه دیگر مرتبط با داده‌کاوی هستند.

آمار شاخه‌ای از علم ریاضی است که به جمع آوری، توضیح و تفسیر داده‌ها می‌پردازد. این علم در مقایسه با داده‌کاوی، قدمت بیشتری دارد و جزء روش‌های کلاسیک داده‌کاوی محسوب می‌شود. وجه اشتراک تکنیک‌های آماری و داده‌کاوی، بیشتر در تخمین و پیش‌بینی است. البته از آزمون‌های آماری در ارزیابی نتایج داده‌کاوی نیز استفاده می‌شود. در کل اگر تخمین و پیش‌بینی جزء وظایف داده‌کاوی در نظر گرفته شوند، در نتیجه تحلیل‌های آماری، داده‌کاوی را بیش از یک قرن اجرا کرده است! می‌توان تحلیل‌های آماری از قبیل فاصله اطمینان، رگرسیون و غیره را مقدمه و پیش زمینه داده‌کاوی دانست که به‌تدریج در زمینه‌ها و رشته‌های دیگر رشد و توسعه پیدا کرد. در واقع تکنیک‌های آماری توسط داده‌کاوی به‌کار برده می‌شوند و برای کشف موضوعات و ساختن مدل‌های پیش‌گویانه مورد استفاده قرار می‌گیرند. به‌طورکلی روش‌های آماری روش‌های قدیمی‌تری هستند که به حالت‌های احتمالی

[1] Jerome H. Friedman

مربوط می‌شوند. داده‌کاوی جایگاه جدیدتری دارد که به هوش مصنوعی، یادگیری ماشین، سیستم‌های اطلاعات مدیریت و متدلوژی پایگاه داده مربوط می‌شود.

چرا به استفاده از داده‌کاوی علاقه‌مندتریم؟

داده کاوی شباهت زیادی به تحلیل‌های آماری دارد، ولی از جهات زیادی با آمار متفاوت است و مزیت‌های زیادی نسبت به آمار دارد. همان‌طور که گفتیم، روش‌های آماری جزء روش‌های قدیمی داده‌کاوی است لذا با پیشرفت علم و فناوری، دچار محدودیت‌هایی خواهد بود. امروزه حجم اطلاعات و مخصوصاً سرعت رشد آنها به قدری زیاد شده است که آمارشناسان و تحلیل‌گران، در بررسی و تحلیل پایگاه‌های داده در زمینه‌های مختلف ناتوانند. بعضی از پایگاه داده‌ها به قدری بزرگ و پیچیده شده‌اند که تحلیل روابط و استخراج اطلاعات مفید پنهان شده در آنها، از ظرفیت ذهنی بشری فراتر رفته است.

تکنیک‌های داده‌کاوی و تکنیک‌های آماری در مباحثی چون تعیین مقدار هدف برای پیش‌گویی و ارزشیابی مناسب داده‌های دقیق، خوب عمل می‌کنند، اما چرا نسبت به داده‌کاوی و استفاده از آن علاقه بیشتری از خود نشان می‌دهیم؟ برای این علاقه‌مندی، چندین دلیل عمده وجود دارد:

۱. روش‌هایی مانند شبکه‌های عصبی مصنوعی در داده‌کاوی، با ارائه تکنیک نزدیک‌ترین همسایه، روش‌های قوی‌تری برای داده‌های واقعی به ما می‌دهند و البته استفاده از آنها برای کاربرانی که تجربه کمتری دارند، راحت‌تر است.

۲. روش‌های داده‌کاوی با اطلاعات کمتر، بهتر کار می‌کنند و برای حجم وسیعی از داده‌ها مناسب‌ترند. تحلیل داده‌های کلان و توسعه روش‌های کلاسیک آماری برای تحلیل چنین داده‌هایی، موضوع جدیدی را تحت عنوان تحلیل داده‌های نمادین به وجود آورده است؛ در حالی‌که روش‌های آماری بیشتر زمانی که تعداد داده‌ها کمتر است و اطلاعات بیشتری در مورد داده‌ها می‌توان بدست آورد، استفاده می‌شوند؛ به عبارت دیگر، روش‌های آماری با مجموعه داده‌های کوچک‌تر سروکار دارند و به‌کار بردن آن‌ها در مجموعه داده‌های عظیم، احتمال خطا را زیاد می‌کند.

۳. یکی از محدودیت‌های اصلی آمار، داشتن فرض اولیه در مورد داده‌هاست. داده‌های جمع‌آوری شده نوعاً خیلی از فرض‌های آماری را در نظر نمی‌گیرند مانند مستقل بودن مشخصه‌ها، مشخص بودن توزیع داده‌ها، داشتن کمترین هم‌پوشانی در فضا و زمان و غیره. تخلف از هر یک از فرض‌ها، می‌تواند مشکلات بزرگی ایجاد کند و در نهایت درستی یا نادرستی نتایج نهایی، به درست بودن فرض اولیه وابسته است؛ در مقابل روش‌های یادگیری ماشین از هیچ فرضی در مورد داده‌ها استفاده نمی‌کند. در واقع در آمار ما فرضیه‌ای مطرح می‌کنیم و با استفاده از تحلیل های آماری به اثبات یا رد فرضیه می پردازیم، اما داده‌کاوی به فرضیه احتیاجی ندارد؛ ابزار داده‌کاوی فرض می کند که شما خود هم نمی‌دانید به دنبال چه می‌گردید! و این نکته ای است که باعث می شود کارآمدی داده‌کاوی در مواقع بروز مشکل نمایان شود.

برای مثال ما در آمار فرض می کنیم که فردی که یک چکش خریده است، حتماً یک بسته میخ هم خواهد خرید، سپس با استفاده از روش‌های آماری مشخص می کنیم که این ارتباط وجود دارد یا خیر؛ اما داده‌کاوی بدون توجه به اینکه چنین فرضی داشته باشیم، با کاوش میان داده ها، اگر ارتباطی مخفی معنی‌داری وجود داشته باشد، آن را به اطلاع ما می رساند. به طور مثال در یک فروشگاه سخت‌افزار ممکن است بین خرید ابزار توسط مشتریان با تملک خانه شخصی یا نوع خودرو، سن، شغل، میزان درآمد یا فاصله محل اقامت آنها با فروشگاه، رابطه‌ای برقرار شود.

۴. یکی دیگر از محدودیت‌های آمار این است که فقط می تواند از داده‌های عددی استفاده کند ولی داده‌کاوی از داده های غیر عددی نیز استفاده می کند.

در جدول زیر، به طور خلاصه به مقایسه داده‌کاوی و آمار پرداخته‌ایم.

داده کاوی	آمار
اکتشافی	تایید کننده
مجموعه داده کلان	مجموعه داده کوچک
تعداد زیادی از متغیرها	تعداد کمی از متغیرها
استنتاجی	استقرایی
مخصوص داده های عددی و غیر عددی	مخصوص داده های عددی
پاکسازی داده ها را انجام می‌دهد	باید داده‌های دقیق (تمیز) در اختیار آن قرار گیرد

جدول مقایسه آمار و داده‌کاوی

کاربردهای علم آمار در داده‌کاوی:

در سال‌های اخیر پژوهش‌های مختلفی در زمینه کاربرد داده‌کاوی در علم آمار صورت گرفته و چندین کارگاه با عنوان کاوش داده‌های رسمی برگزار شده است که اولین کارگاه در سال ۲۰۰۲ در فنلاند و دومین کارگاه در سال ۲۰۰۴ در ایتالیا بوده است. این امر نشان‌دهنده آن است که این موضوع مورد توجه سازمان‌های ملی آمار و محققان آماری قرار گرفته است. در ادامه به بررسی این موضوع مهم می‌پردازیم:

داده‌کاوی معمولاً استراتژی‌های زیر را در داده ها بکار می برد: توضیح و تفسیر، تخمین، پیش‌بینی، دسته بندی، خوشه بندی، وابسته سازی و ایجاد رابطه، قواعد انجمنی، ترتیب و مصورسازی.

روش‌های آماری در مباحث تخمین وپیش بینی کاربرد دارند. در تحلیل آماری، تخمین وپیش بینی عناصری از استنباط‌های آماری هستند. استنباط‌های آماری شامل روش‌هایی برای تخمین وتست فرضیات درباره جمعیتی از ویژگی‌ها براساس اطلاعات حاصل از نمونه است. یک جمعیت شامل مجموعه ای از عناصر از قبیل افراد یا داده‌هایی که در یک مطالعه خاص آمده است. لذا در این‌جا دو مبحث تخمین و پیش‌بینی را تشریح می‌کنیم:

الف- تخمین:

در تخمین به دنبال تعیین مقدار یک مشخصه خروجی مجهول هستیم. مشخصه خروجی در مسائل تخمین بیشتر عددی هستند تا قیاسی؛ بنابراین مواردی که بصورت قیاسی هستند باید به حالت عددی تبدیل شوند، مثلاً موارد بلی و خیر به صفر و یک تبدیل می شوند. روشهای آماری مورد استفاده دراین بخش، بطورکلی شامل تخمین نقطه و فاصله اطمینان می‌باشند که از برآوردکننده‌های نقطه‌ای و فاصله‌ای کمک گرفته می‌شود.

لازم به‌ذکر است زمانی به سراغ تخمین می‌رویم که مقدار واقعی پارامترها ناشناخته باشد. به‌طور مثال زمانی که مقدار واقعی میانگین یک جامعه مشخص نیست. در برخی موارد تعیین میانگین مجموعه ای از داده ها مهم است، مثلاً میانگین تعداد نفراتی که در یک روز به بانک مراجعه می کنند یا متوسط مقدار پولی که افراد در یک شعبه خاص از بانک واریز می کنند.

ب- پیش بینی:

هدف از انجام پیش بینی تعیین ترکیب خروجی با استفاده از رفتار موجود است؛ یعنی رسیدن به یک نتیجه به‌وسیله اطلاعات موجود از داده ها. مشخصه های خروجی در این روش هم می‌توانند عددی باشند و هم قیاسی. این استراتژی در بین استراتژی های داده‌کاوی از اهمیت خاصی برخوردار است و مفهوم کلی‌تری را نسبت به موارد دیگر دارد. خیلی از تکنیک‌های نظارتی داده‌کاوی که برای دسته بندی و تخمین مناسب هستند، عملاً کار پیش‌بینی انجام می دهند.

آنچه از کتاب‌های آماری و تحت عنوان پیش بینی برمی آید، رگرسیون و مباحث مربوط به آن است. هدف از تحلیل رگرسیون، یکی از موارد ذیل است:

- بدست آوردن رفتار متغیر y توسط متغیر x ، یعنی اینکه متغیر y با تغییر x در نمونه ها چه رفتاری از خود نشان می دهد.
- پیش بینی بر اساس داده‌ها برای نمونه های آینده که هدف اصلی در داده کاوی از طریق روش‌های آماری است. مثلاً از روی اطلاعاتی مثل داشتن کارت اعتباری یک فرد جدید، نوع جنسیت او، سن فرد و میزان درآمد سالیانه او، بتوان حدس زد که این فرد از بیمه عمر استفاده می کند یا خیر. و یا اینکه با داشتن اطلاعات در مورد داشتن یا نداشتن کارت اعتباری، بیمه عمر و سن فرد، بتوان جنسیت فرد را تعیین کرد.
- تحلیل حساسیت، یعنی تعیین اینکه اگر x به اندازه خاصی تغییر کند، y تا چه اندازه تغییر خواهد کرد.

روش‌های مختلف رگرسیون برای داده کاوی وجود دارد. رگرسیون خطی بیشترین کاربرد را دارد و همچنین مشتقات آن حائز اهمیت است. در این مقاله به سه مورد از انواع رگرسیون که در داده‌کاوی کاربرد دارند، اشاره می‌کنیم:

۱) رگرسیون خطی:

روش رگرسیون خطی[۱] یک تکنیک یادگیری نظارتی است که به وسیله آن تغییرات یک متغیر وابسته به‌وسیله ترکیب خطی از یک یا چند متغیر مستقل مدل می‌شود. یکی از هدف‌های اصلی بسیاری از پژوهش‌های آماری ایجاد وابستگی هایی است تا پیش بینی یک یا چند متغیر را بر حسب سایرین ممکن می سازد. مثلاً مطالعاتی انجام می‌شود تا فروش‌های بالقوه یک محصول جدید را بر حسب قیمت آن پیش بینی کند.

۲) رگرسیون لجستیک:

روش رگرسیون لجستیک[۲]، یک تکنیک رگرسیون غیرخطی است و در حالتی که نتایج خروجی به صورت باینری[۳] هستند، مورد توجه قرار می گیرد. در کل زمانی نتایج خروجی به صورت باینری هستند، رگرسیون خطی خیلی کارا نیست، در این حالت استفاده از این تکنیک مناسب تر است. در واقع دلیل استفاده از رگرسیون لجستیک آن است که در رگرسیون خطی علاوه بر اینکه نتایج خروجی باید به صورت عددی باشد، متغیر ها هم باید به صورت عددی باشد، بنابراین حالت‌هایی که به صورت قیاسی هستند باید به حالت عددی تغییر شکل پیدا کنند. مثلاً جنسیت افراد از حالت زن و مرد، به ترتیب به حالت‌های صفر و یک تغییر پیدا می کند، اما اساس رگرسیون خطی در این حالت ایراد پیدا می کند و ارزش قیدی که بر روی متغیر وابسته قرار می گیرد توسط معادله رگرسیون در نظر گرفته نمی‌شود. در واقع چون رگرسیون خطی، معادله یک خط را ترسیم می کند، نمی تواند حالت مثبت و منفی یا به عبارتی صفر و یک را در نظر بگیرد، لذا با اعمال تغییراتی از رگرسیون لجستیک استفاده می‌کنیم.

[۱] Linear Regression
[۲] Logistic Regression
[۳] Binary

۳) رگرسیون سلسله مراتبی یا چندسطحی:

این نوع از رگرسیون، یکی از ابزارهای تحلیل داده‌های پیچیده به شمار می‌رود. مدل‌های رگرسیون چند سطحی برای حالت‌هایی که همپوشانی در سطوح مختلف وجود دارد مفید است؛ برای مثال اطلاعات آموزشی ممکن است شامل اطلاعاتی از قبیل اطلاعات فردی دانش آموزان، اطلاعات سطح کلاس و همچنین اطلاعات درباره مدرسه باشد.

در روش رگرسیون چند سطحی یا سلسله مراتبی محدودیتی برای تعداد سطوح تغییر که می تواند انجام شود، وجود ندارد. روش‌های بیزی در تخمین پارامترهای مجهول کمک می‌کند، هرچند که محاسبات پیچیده‌ای دارد. از ذکر توضیحات بیشتر در مورد انواع رگرسیون که در داده‌کاوی استفاده می‌شوند، خودداری می‌کنیم و بد نیست اشاره‌ای به الگوریتم نایو بیز[۱] هم داشته باشیم.

الگوریتم نایو بیز:

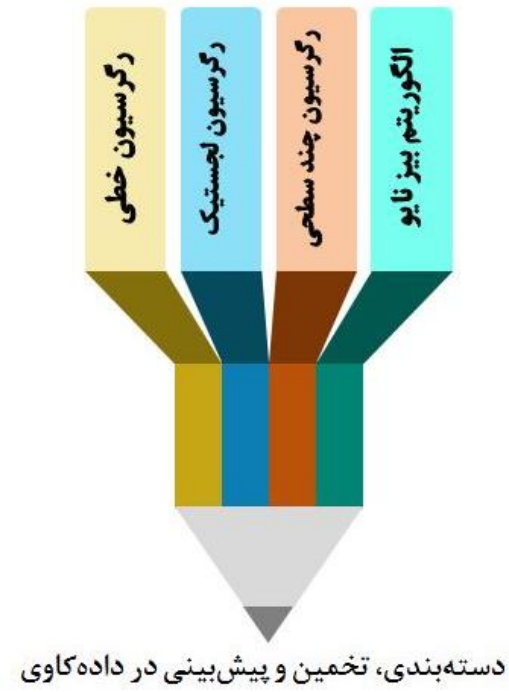
این الگوریتم روشی برای دسته بندی پدیده‌ها بر مبنای احتمال وقوع یا عدم وقوع یک پدیده است و بر اساس تئوری بیز بنا شده است که از مهم‌ترین تئوری‌ها در احتمالات و آمار مهندسی است. این الگوریتم یکی از روش‌های ساده یادگیری نظارتی است،که در آن فرض می شود تمام متغیرهای ورودی به یک اندازه مهم هستند و مستقل از هم می‌باشند و نیز اگر یکی از شرایط هم برقرار نباشد، این روش در شرایطی کاربرد دارد. البته این الگوریتم برای رسیدن به یک نتیجه خوب نیاز به تعداد زیادی رکورد دارد و تا حدودی جانبدارانه عمل می‌کند. مثالی ساده از کاربرد این الگوریتم را بررسی می‌نمائیم.

نامی مشترک میان دخترها و پسرها را در نظر بگیرید و فرض کنید که مجموعه‌ای از این نام در پایگاه داده بیمارستانی وجود دارد که در پرونده هر یک مشخص شده است که کدام یک پسر و کدام یک دختر هستند. حال اگر نوزادی متولد شود و همین نام برای او انتخاب گردد، شما می‌توانید به کمک این الگوریتم و پایگاه داده موجود، جنسیت نوزاد را پیش‌بینی کنید و در واقع نوزاد را در یکی از دو دسته دخترها یا پسرها قرار دهید.

در پایان بد نیست بدانید که آنالیز واریانس تک متغیره[۲]، آنالیز واریانس چند متغیره[۳]، آنالیز کوواریانس[۴]، آمار غیرپارامتریک[۵]، زنجیره مارکوف[۶]، تحلیل سری‌های زمانی و چندین روش آماری دیگر نیز در داده‌کاوی کاربردهای گسترده‌ای دارند و این خود تاکیدی بر اهمیت و نقش بی بدیل علم آمار در داده‌کاوی است.

---

[۱] Naive Bayes Algorithm
[۲] ANOVA
[۳] MANOVA
[۴] ANCOVA
[۵] Nonparametric Statistics
[۶] Marcov Chain



منابع و مراجع:

۱. “Data Mining & Statistics: What’s the Connection?”, Jerome H. Friedman, Department of Statistics and Stanford Linear Accelerator Center, Stanford University.

۲. Yoav Benjamini, Moshe Leshno, “Statistical Methods for Data Mining” A chapter in the book: “Data Mining and Knowledge Discovery Handbook” Maimon, O. Rocach, L. Springer. October ۲۰۱۱.

۳. مقاله"داده‌کاوی و کاربرد آن در آمار رسمی"- نویسندگان: علی اصغر حائری مهریزی و حسین حسنی- فصلنامه گزیده مطالب آماری- سال ۱۶- شماره ۴- زمستان ۱۳۸۴.

٤. https://www.stats.ox.ac.uk

٥. https://www.indiana.edu/

٦. https://www.dadekavan.ir/

۷. https://www.simplilearn.com