عنوان: **پیش نیاز ورود به دنیای داده کاوی**
نویسنده: محمد رجبی
تاریخ: ۱۳۹۳/۰۸/۳۰ ۲۲:۴۴
آدرس: www.dotnettips.info
گروه‌ها: Analysis Services, data mining, Microsoft SQL Server

علم داده کاوی از علوم مختلفی از جمله علم آمار، هوش مصنوعی، یادگیری ماشین، شناسائی الگو و پایگاه داده نشات گرفته است و این علوم ریشه‌های علم داده کاوی هستند. برای مثال الگوریتم هایی که یک مدل را یاد می‌گیرند یا الگویی را شناسائی می‌کنند؛ معمولا وجه مشترک یادگیری ماشین و شناسائی الگو با داده کاوی هستند.

در این قسمت پیش از درگیر شدن با جزئیات هر الگوریتم تمایل دارم خوانندگان محترم را با مطالبی که شاید کمتر در دنیای IT با آن درگیر بوده اند؛ آشنا کنم. این کار به این دلیل انجام می‌شود که برای مثال در کشف قوانین انجمنی یا دسته بند مبتنی بر قانون (مثال متداول آن تحلیل سبد خرید مشتری در هایپر مارکت است) خروجی به شکل مجموعه ای قانون «اگر الف؛ آنگاه ب» و ... بدست می‌آید. بنابراین برای تفسیر صحیح این مدل‌ها علاوه بر آشنائی با کسب وکار مربوطه؛ نیازی نسبی به آشنائی با سایر علوم نیز می‌باشد و بدین ترتیب از اتلاف انرژی و زمان و همچنین از بروز خطا در استدلالمان جلوگیری می‌کنیم. جمله معروفی با این مضمون در سایر فرهنگ‌ها وجود دارد که اعداد دروغ نمی‌گویند؛ ولی فردی دروغگو می‌تواند از اعداد سوء استفاده کند. بنابراین زمان مناسبی است که با بعضی مغالطات آشنا شویم.

اساس کار علمی به بیان ساده عبارت است از: به پرسش گرفتن همه چیز و دنبال کردن مدارک و شواهد به هر کجا که ما را رهنمون سازد؛ اینکار بوسیله آزمودن هر نظر و ایده ای، با انجام آزمایش روی آن‌ها و مشاهده نتایج بدست آمده و سپس توسعه دادن مواردی که از آزمایشات موفق بیرون آمده اند و رد کردن آنهایی که در آزمون شکست خورده اند، انجام می‌گیرد. روش علمی آنچنان قدرتمند است که در طی چهار قرن گذشته (قرن 16 میلادی) ما را از نخستین نگاهی که گالیله از درون تلسکوپ به دنیای دیگر انداخت، به گام گذاشتن بر روی ماه رسانده است و به ما اجازه داده تا به پهنه فضا و زمان بنگریم تا کشف کنیم که در کجا و در چه زمانی از عالم قرار داریم.

اجداد ما ستاره شناسان خانه به دوشی بودند که در گروه‌های کوچک زندگی می‌کردند، آسمان تقویم و راهنمای زندگی آنها بود، بقای شان به این وابسته بود که بدانند چگونه ستاره‌ها را بخوانند و بدین ترتیب بتوانند فرا رسیدن زمستان را پیش بینی کنند و زمان کوچ کردن را بدست آورند. در واقع نعمت **تشخیص الگو** باعث شانس بیشتر زنده ماندن و تولید مثل آنها بود و بدین ترتیب ژنهای تشخیص الگو را به نسل‌های آینده منتقل می‌کردند. آنها وقتی که ارتباط مستقیمی بین حرکت ستارگان و گردش فصلی حیات روی زمین پیدا کردند، نتیجه گرفتند که اتفاقاتی که آن بالا می‌افتد به ما در پائین مربوط می‌شود و آنرا به خود می‌گرفتند!؟ آنها توضیح منطقی دیگری برای اتفاق پیش آمده نداشتند. کلمه یونانی Dis-aster به معنی "ستاره شوم" حتی برای اقوام مختلف به معنای جنگ، قحطی، مریضی و ... تعبیر می‌شد. (در فرهنگ ما نیز جملاتی با این مضمون کم وجود ندارد، برای مثال:" قمر در عقرب است"، پس اتفاق بدی خواهد افتاد!. البته منظور قرار گرفتن ماه در برج عقرب است و ...).

می توان گفت استعداد انسان در تشخیص الگو شمشیری دو لبه است، ما انسان‌ها قادریم در تشخیص الگوهائی که اصلاً وجود ندارند نیز خیلی خوب عمل کنیم!، چیزی که به معنای "تشخیص الگوی اشتباه" است. ما عاشق خاص بودن هستیم و با داشتن این هدف همواره در تلاش برای فریب خود و دیگران هستیم. علم در مرز میان دانایی و جهالت گام بر می‌دارد، از نظر یک محقق هیچ شرمساری در ندانستن وجود ندارد، تنها شرمساری در آن است که تظاهر کنیم همه جواب‌ها را می‌دانیم. علم راهی است که انسان را از فریب خود و دیگران باز می‌دارد و امروزه به نیکی می‌دانیم هر چه علم بیشتر در اختیار ابنای بشر قرار گیرد، امکان سوء استفاده از آن کمتر خواهد شد. بدین ترتیب با دانستن ارزش‌های علمی تقاضا برای جهالت و تعصب کم خواهد شد.
ارزش‌های علمی مختصراً به شرح زیر هستند: قدرت سوال کردن، وقتی موضوعی را بررسی می‌کنید تنها چیزی که باید از خودتان بپرسید این است که واقعیت‌ها در این موضوع (فلسفه) چه هست و چه حقایقی در آن نهفته است. هیچگاه به خودتان اجازه ندهید که آنچه را دوست دارید، حقیقت داشته باشد (اگر یک ایده دلخواه در یک آزمایش خوب مردود شد، پس اشتباه است و از آن عبور کنید)، همچنین آنچه را که فکر می‌کنید حقیقت بودنش برای بشر سودمند است شما را منحرف نکند (برای خودتان فکر کنید و از خودتان بپرسید)، فقط و تنها به این که واقعیت چه هست بنگرید، در ضمن اگر مدرکی ندارید؛ قضاوت نکنید و مهمترین قانون؛ به یاد داشته باشید که شما انسان هستید و می‌توانید اشتباه کنید، همانطور که مهمترین دانشمندان در مواردی اشتباهاتی داشته اند.

منطق ابزاری علمی است که بکارگیری آن ذهن انسان را از خطای در تفکر باز می‌دارد، مبارزه با مغالطات و لغزش‌های اندیشه هدف علم منطق است. مغالطه منحصر به استدلال نیست، به بیان دقیق‌تر شکل هایی از استدلال است که نتیجه تابع مقدمه یا مقدمه هایش نیست. مغالطه ای که عمدی یعنی با آگاهی از عدم اعتبار انجام می‌شود اما به ظاهر معتبر و مجاب کننده و در واقع فریب دهنده مخاطب است سفسطه نامیده می‌شود. عدم اعتبار یک استدلال ممکن است به دلایل زیر باشد: ناشی از نادرستی یکی از مقدمات استدلال باشد و یا علی رغم درستی مقدمات؛ نظم و صورت استدلال نادرست باشد. برای آشنایی ذهن خواننده به معرفی نمونه ای از این مغالطات اشاره می‌شود؛ برای مثال این مغالطه بر این پیش فرض استوار است که هر زمان دو حادثه با یکدیگر اتفاق افتاد؛ می‌توان یکی را علت و دیگری را معلول آن به حساب آورد. برای مثال در تحقیقی به ارتباط مستقیم میان وجود داشتن چتر در ماشین به هنگام تصادفات رانندگی پرداخته شده و به این نتیجه رسیده اند زمانی که تصادفی رخ می‌دهد با احتمال بسیار بالاتری چتر در ماشین وجود دارد به نسبت حالتی که چتر در ماشین وجود ندارد؛ به همین دلیل چتر عامل تصادف است! برای اجتناب از این مغالطات باید قادر به تفکیک اصل علیت (Causality) و همبستگی (Correlation) باشیم. (در توضیح مثال فوق لغزندگی جاده عامل تصادف در روزی بارانی است نه چتر!).

همچنین استفاده از آمار و اطلاعات آماری علی رغم فوائد زیاد در اطلاع رسانی، می‌تواند لغزشگاهی باشد که زمینه ارتکاب برخی مغالطات را نیز فراهم کند در ادامه به معرفی تعدادی از این مغالطات آماری (**Statistical Fallacies**) می‌پردازیم:
مغالطه متوسط که می‌تواند با سوء استفاده از برخی اصطلاحات آماری مطابق با اهداف و اغراضی که موسسات ارائه دهنده اطلاعات آماری دنبال می‌کنند، متوسط یک مجموعه را کم یا زیاد اعلام کنند! به بیان دیگر کلمه متوسط در نوبت‌های مختلف به معانی متداولی استعمال می‌شود که عبارتند از:
میانگین (Average) یا معدل که برای چند عدد برابر است با مجموع آنها تقسیم بر تعدادشان.
میانه (Median) که یک مجموعه عددی را به دو نیم تقسیم می‌کند؛ نیمی که هر یک از اعداد آن بیشتر از میانه و نیمی که کمتر از میانه است.
نما (Mode) که در یک مجموعه؛ عددی است که بیش از دیگر اعداد تکرار شده است.
پس می‌توان نتیجه گرفت وقتی اعلام می‌شود که در یک جامعه آماری فلان عدد یک متوسط است هنوز اطلاع دقیقی داده نشده و باید صراحتا مشخص کنند کدامیک از معانی متوسط مورد نظر است.
باید در نظر داشته باشید این مغالطه زمانی استفاده می‌شود که دامنه تغییرات در میان جامعه آماری بسیار زیاد است، چنانچه دامنه تغییرات حداقل و حداکثر نسبت به تعداد افراد جامعه زیاد نباشد، مقادیر میانگین؛ میانه و نما تقریبا منطبق بر هم خواهند شد (برای مثال در محاسبه متوسط طول قد افراد یک کشور). اما در مواردی که تغییرات مذکور زیاد باشد باید با هوشیاری از وقوع این مغالطه جلوگیری نمود (از مصادیق و زمینه‌های بارز و مهم ارتکاب این مغالطه محاسبه متوسط حقوق و درآمد افراد است).

مغالطه نمودارهای گمراه کننده (Misleading Graph) استفاده از نمودار می‌تواند وسیله ای موثر در بیان مغالطه آمیز بودن اطلاعات آماری باشد. برای مثال نمودار رشد سود خالص شرکتی را در نظر بگیرید که در محور افقی آن بعد زمان و در محور عمودی مقادیر مالی درج شده است. با رسم نمودار مذکور سود خالص هر ماه به صورت واضح و آشکار مثلا رشدی ده درصدی را نمایش می‌دهد چنانچه شرکت مذکور اصول اخلاقی را رعایت نکند و برای جذابیت بیشتر و جذب سرمایه‌های بیشتر؛ قسمت هایی از نمودار را به گونه ای حذف کند که حاصل کار این شود که خواننده احساس کند سود خالص شرکت در عرض دوازده ماه به بالای کاغذ رسیده (یعنی به طور ضمنی افزایشی معادل صد در صد) و یا نسبت بین خطوط افقی و عمودی را بگونه ای تغییر دهد تا رشد ده درصدی را بسیار بزرگتر نشان داده شود (می تواند با تقلیل مقیاس واحد مالی به یک دهم به این هدف برسد) بدین ترتیب نمودار حاصل چنان جذاب می‌شود که هر کس با تماشای آن رگه‌های موفقیت و پیشرفت را در شرکت متقلب بوضوح مشاهده می‌کند.

مغالطه تصاویر یک بعدی (One Dimensional Pictures) از روش‌های تقلب دیگر می‌تواند باشد که باید توجه کرد آیا نسبت القا شده بوسیله تصاویر با نسبت اعداد مطابقت دارد یا خیر.

می دانیم آنچه پایه و اساس آمار استنباطی را تشکیل می‌دهد روش‌های نمونه گیری است که اتفاقا این روش‌ها منشاء برخی مغالطات و ترفندهای آماری نیز هست در این قسمت به معرفی تعدادی از این موارد می‌پردازیم:

نمونه ناکافی (Deficient Examples) چنانچه در روش نمونه گیری مقدار و نسبت «نمونه» به «جامعه آماری» به اندازه کافی

بزرگ باشد و به طرز صحیحی انتخاب شده باشد؛ غالبا می‌تواند معرف خوبی برای جامعه آماری باشد. اما چنانچه نمونه به اندازه کافی بزرگ نباشد؛ گرچه اطلاعاتی را در خصوص جامعه آماری در اختیارمان قرار می‌دهد ولیکن احتمال وقوع خطا در چنین حالتی بسیار زیاد است که این مغالطه دارای این شرایط است؛ البته باید توجه داشت که کافی یا ناکافی بودن تعداد نمونه‌ها نسبت به جامعه آماری امری نسبی است. بنابراین جهت اجتناب از بروز این مغالطه باید همواره در نظر داشت آیا تعداد نمونه‌ها در مقایسه با کل جامعه آماری راضی کننده و کافی است یا خیر.

نمونه غیر تصادفی (Deliberate Examples) برای بدست آوردن اطلاعات آماری در روش نمونه برداری؛ کافی بودن نمونه‌ها شرط لازم است و کافی نیست؛ یکی از مواردی که باید مورد توجه قرار داد تصادفی بودن نمونه‌ها می‌باشد. به بیان دیگر تنها کافی بودن نمونه‌ها یا فراوانی آنها برای تعمیم دادن حکمی به کل آن جامعه آماری کفایت نمی‌کند. تصادفی بودن نمونه‌ها بدین معناست که نمونه‌ها نباید نماینده و بیانگر دسته و گروه خاصی از جامعه آماری باشند. همچنین در روش نمونه برداری افراد جامعه آماری باید از شانس یکسانی برای انتخاب شدن در نمونه برداری برخوردار باشند از راه‌های تحقق این هدف تقسیم افراد جامعه آماری به دسته‌ها و طبقات مختلف و تعیین کردن درصد و نسبت هر یک از آنها به کل مجموعه می‌باشد بدین ترتیب در نمونه برداری نیز سعی می‌شود این نسبت لحاظ گردد؛ این روش اصطلاحا روش نمونه گیری تصادفی طبقه ای نامیده می‌شود روش‌های دیگری نیز به منظور اینکه کلیه افراد جامعه آماری از شانسی یکسان برای انتخاب شدن در نمونه برخوردار باشند وجود دارد مانند روش‌های نمونه گیری تصادفی ساده؛ نمونه گیری تصادفی خوشه ای و نمونه گیری تصادفی سیستماتیک.

عدم واقع نمائی نمونه‌ها (Unrealistic Examples) در نمونه برداری به صورت پرسش‌های شفاهی از جامعه آماری انسانی مسئله عدم واقع نمائی نمونه‌ها رخ می‌دهد بدین ترتیب همواره موجب بروز خطاهای جدی در بدست آوردن اطلاعات آماری دقیق است. این مشکل عملا به روش جمع آوری داده‌ها از طریق مصاحبه بر می‌گردد خواه به صورت نمونه ای یا سرشماری باشد.