داده‌کاوی

Data Mining

فصل اول:
معرفی داده‌کاوی
و کشف دانش

استاد مدرس:
دکتر سعید عبادالله‌ی

فصل اول

معرفی داده‌کاوی و کشف دانش

مقدمه

تاریخچه

در اواسط قرن بیستم کلیه اطلاعاتی که تا قرن دوم میلادی از خورشید، گردش زمین، ماه، شب و روز و... بدست آمده بود توسط پژوهشگران آمریکایی به شکل مجموعه‌ای از داده‌های عددی و سمبولیک جهت کاوش درآمد.

✓ خروجی الگوریتم داده‌کاوی به صورت مجموعه‌ای از روابط بودکه پس از تفسیر بدین شکل داریم:

- شیءای که زمین نامیده شده گرد است.
- شیء زمین به دور شیءای که خورشید نامیده شده می‌گردد.
- شیءای که ماه نامیده شده به دور زمین می‌گردد.

❑ زمینه‌های متعدد برای داده‌کاوی: پزشکی، بورس اوراق بهادار، هواشناسی، بازاریابی، تشخیص کلاهبرداری‌های بانکی و بیمه‌ای، تجارت الکترونیک، بیوالکترونیک و ...

❑ به دلایل مختلف اعم از سرعت پردازشی، جهل داده‌ای، چگونگی تحلیل و... کشف نظم‌های پیچیده موجود در نهان داده‌های حجیم مشکل است؛

در واقع علم داده‌کاوی بعنوان قادر ساختن انسان برای پردازش عمیق حجم عظیمی از داده‌ها تعریف می شود!

مقدمه

مثالهایی از کاربردهای داده کاوی:

*از بررسی تغییرات بورس برای بررسی روند تغییرات آن جهت سرمایه‌گذاری موفق استفاده می‌شود.

*با استفاده از داده های محیطی می‌توان الگوهایی برای تغییرات نرخ فقر بر اساس فاصله‌های شهری از بزرگراه‌های اصلی تعیین کرد.

توجه: *ارتباط بین یک دسته از اهداف برای تعیین اینکه کدام زیردسته از داده های محیطی خودهمبسته یا مرتبط هستند باید بررسی شود!*

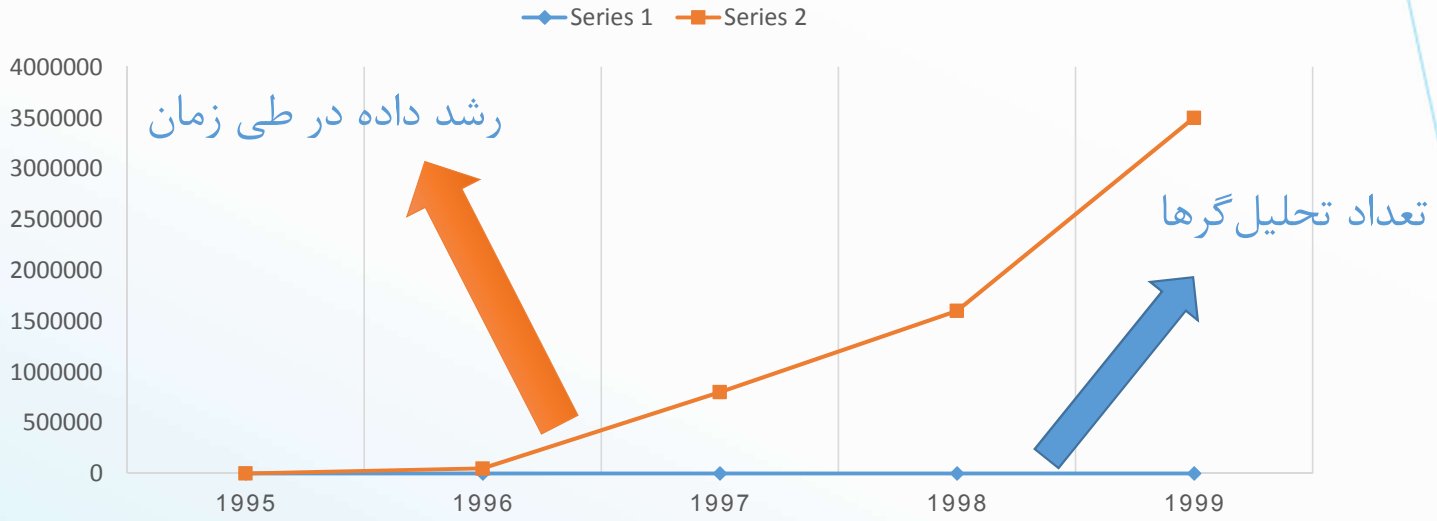
*با استفاده از داده کاوی در بازی هاکی می توان اتفاقاتی که به گل ختم می شوند را بررسی کرد!

*وب‌کاوی می تواند اطلاعاتی راجع به توزیع اطلاعات بطور عمومی بر روی WWW، مشخصه‌بندی و دسته‌بندی صفحه‌های وب و از اینگونه اطلاعات بدست بدهد!

باید در نظر داشته باشیم که در بسیاری از کاربردها چندین دسته از داده‌ها موجود است، مثلا در وب‌کاوی داده‌های نوشته و داده‌های چندرسانه(ویدیو و تصویر)، گراف‌های داده مانند گراف وب و داده‌های نقشه بر روی وب‌سایتها وجود دارد که بررسی آنها خود یک چالش جداگانه است!

انگیزه‌های کاوش داده

قیاس رشد حجم داده با رشد تعداد تحلیلگران داده



با توجه به نمودار : – با گذشت زمان **تعداد تحلیلگرها** در مقایسه **با رشد داده** تقریبا **ثابت** بوده است.

– **حجم داده** با گذشت زمان درحال **رشد انفجارگونه** است.

تفاوت بین این دو نمودار (فضای خالی بین دو نمودار) = **شکاف‌داده‌ای** بین دو نمودار

فاصله بین این دو نمودار نشان‌دهنده: افزایش تعداد داده‌ها نسبت به تعداد افرادی که بتوانند این داده‌ها را تحلیل کنند؛

پس **نیاز به ابزار مکانیزه‌ای برای تحلیل داده**، روز به روز در حال افزایش است!

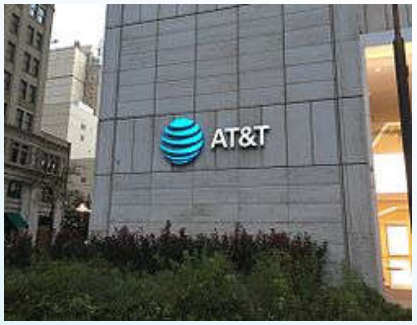
انگیزه‌های کاوش داده

مثالهایی برای تبیین سرعت رشد داده‌ها:

مرکز ستاره‌شناسی VLBI دارای ۱۶ تلسکوپ بزرگ است که هر یک با سرعت ۱ گیگابیت برثانیه داده ستاره‌شناسی را بر مبنای مشاهدات یک ماه ۲۵ روزه تولید می‌کنند.

- ✓ شرکت مخابراتی AT & T با میلیاردها تماس در روز سروکار دارد. چنین داده عظیمی را نمی‌توان ذخیره نمود. تجزیه و تحلیل آن می بایست به صورت برخط بر روی جریان داده باشد.
- ✓ تیم جمع‌آوری وب کتابخانه‌ملی‌آمریکا در ماه می ۲۰۰۸ اعلام نموده که کتابخانه‌ای با بیش از ۸۲,۶ ترابایت داده گردآوری نموده است.

و ...



اولین انگیزه کاوش داده: رشد روز افزون داده

انگیزه‌های کاوش داده
۱) انگیزه‌های تجاری
۲) انگیزه‌های علمی

انگیزه‌های کاوش داده

1) انگیزه‌های تجاری

سه منبع برای جمع‌آوری داده‌های تجاری عبارتند از:

- داده‌های وب و داده‌های تجارت‌الکترونیک
- خرید و فروش‌های موجود در فروشگاه‌های خواروبار فروشی/ سوپر مارکت‌های زنجیره‌ای
- تراکنش‌های بانکی / تراکنش‌های کارت‌های اعتباری

داده‌های وب و داده‌های تجارت الکترونیک

منظور از داده‌ها تراکنش‌هایی است که همه روزه در اینترنت انجام شده و ثبت می‌شوند.
داده‌های وب یا داده‌های تجارت الکترونیک داده‌هایی مهم و شامل اطلاعات زیادی هستند؛ تعدادی از این داده‌ها عبارت اند از: خرید و فروش بلیط‌های هواپیما، قطار، پرداخت قبوض و ...

خرید و فروش‌های موجود در فروشگاه‌های خواروبار فروشی/ سوپر مارکت‌های زنجیره‌ای

منظور خرید و فروش‌هایی است که همه‌روزه در فروشگاه‌های خواروبار فروشی و سوپرمارکت‌های زنجیره‌ای انجام می‌شود شامل کالاهایی در سبدخرید مشتری‌های مختلف قرار می‌گیرد.

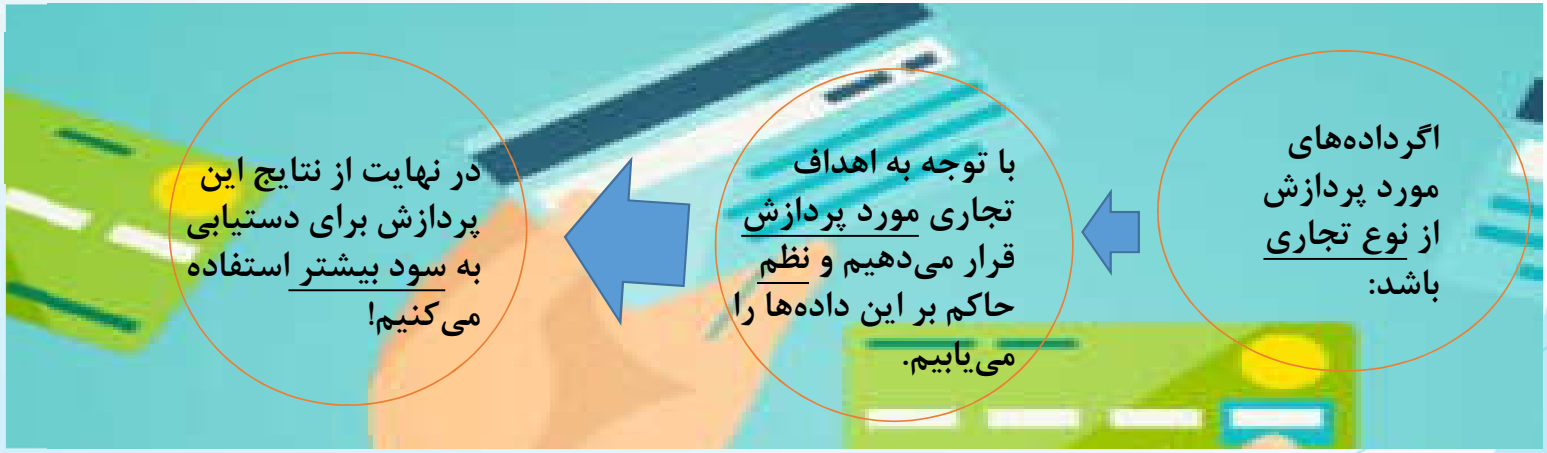
انگیزه‌های کاوش داده

تراکنش‌های بانکی / تراکنش‌های کارت‌های اعتباری

- منظور: داده‌های مربوط به عملیات بانکی که همه روزه توسط مشتریان مختلفی که به یک بانک مراجعه می‌کنند، انجام شده و ثبت می‌شود.

- تراکنش‌های بانکی / تراکنش‌های کارت‌های اعتباری دارای حجم عظیمی از داده‌ها به صورت روزانه هستند.
مثلا اطلاعات مربوط به واریز پول، برداشت پول و... در حساب مشتری

همچنین تراکنش‌های مربوط به کارت‌های اعتباری نیز در این مجموعه داده قرار می‌گیرند؛به عنوان مثال:
ثبت اطلاعات مربوط به مشتری و کالاهایی که در حال خرید آن با کارت‌های اعتباری است!



هدف اصلی از پردازش داده‌های تجاری ← دستیابی به سود بیشتر

انگیزه‌های کاوش داده

۲) انگیزه‌های علمی

نیاز به داده‌هایی داریم که ماهیت علمی داشته باشند.

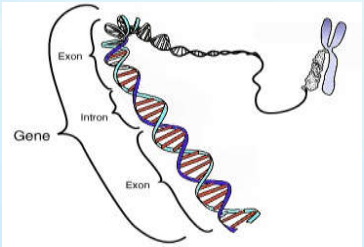
چهار منبع عمده برای جمع‌آوری داده‌های علمی در حجم‌های بالا وجود دارند که عبارتند از:

1) تصاویر ارسالی از طریق ماهواره‌ها
2) تصاویر ارسالی از تلسکوپ‌ها
3) داده‌های دنباله ژنی
4) داده‌های حاصل از شبیه‌سازی‌های علمی

تصاویر ارسالی از طریق ماهواره‌ها و تلسکوپ‌ها

- داده‌ها و تصاویر ارسالی از طریق ماهواره‌ها شامل تصاویرمخابره‌شده به زمین از حسگرهای نصب شده بر روی ماهواره‌ها مورد نظر است.
- انتقال این داده‌ها و تصاویر از طریق سیگنال‌های ماهواره‌ای که در مدار زمین قرار دارد است.
- تصاویر تلسکوپ‌ها در هر روز نیز شامل اطلاعات بسیار زیادی هستند.

داده‌های دنباله ژنی



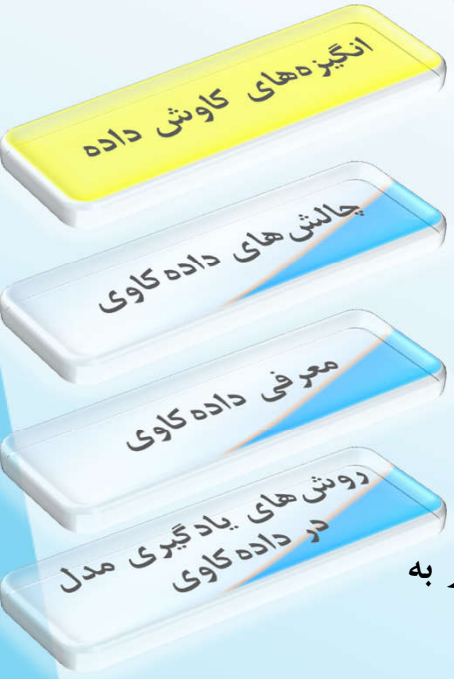
- داده‌های دنباله ژنی، حاصل توصیف دنباله ژنی افراد مختلف می باشد.
- استخراج دنباله‌ژنی مربوط به یک بافت‌خاص (مثلا کبد، معده، خون، ...) توسط تکنیک‌های میکرو آرایه صورت می‌گیرد.
- حاوی اطلاعات و ویژگی‌های بسیار است.
- عموما این اعداد و ویژگی‌ها در تشخیص بیماری افراد، کمک قابل ملاحظه‌ای می‌کنند.

انگیزه‌های کاوش داده

داده‌های حاصل از شبیه سازی‌های علمی

- منظور از شبیه‌سازی علمی: مدل‌سازی یک سیستم در ابعاد کوچکتر
- شبیه سازی با دانستن قوانین حاکم بر آن سیستم و محیط
- مزیت: امکان انجام آزمایش‌های مختلفی روی سیستم، درنتیجه ⬅ تولید داده‌های زیاد

به عنوان مثال شبیه‌سازی بدن یک انسان و امتحان داروهای مختلف روی آن، شبیه‌سازی آزمایش‌هسته‌ای، شبیه‌سازی زلزله
اگر این آزمایش‌ها هزینه سنگینی داشته باشند، شبیه‌سازی و نتایج حاصل از آن بسیار با ارزش خواهند بود.



داده‌های علمی لزوما به سود بیشتر منجر نمی‌شود، اما:
منجر به ایجاد دانش جدید، دستاوردهای جدید و نیز باعث خدمت بیشتر به افراد و یا کشف حقایق خواهد شد!

- ✓ به عنوان مثال پردازش تصاویر ارسالی از حسگرهای روی ماهواره‌ها و نیز تصاویر ارسالی از طریق تلسکوپ‌ها منجر به نتایج جالبی از جمله کشف یک کهکشان جدید یا سیاره جدید، محاسبه عمر یک کهکشان و...
- ✓ پردازش داده‌های دنباله‌ژنی مربوط به یک بافت‌خاص از یک شخص، برای یاری رساندن در تشخیص‌بیماری

انگیزه اصلی در پردازش داده‌های علمی ⬅ کمک به بسط و گسترش مرزهای دانش بشری در یک حوزه خاص

چالش‌های داده‌کاوی

مهم‌ترین نقاط ضعف روش‌های داده‌کاوی:
وجود داده، صحت داده و کافی بودن ویژگی‌ها

- **وجود داده:** اصولا داده‌ای برای کاوش وجود داشته باشد و این گونه نباشد که داده در محیط مورد کاوش استخراج و یا ثبت نشده باشد. متأسفانه این مشکل در بسیاری از محیط‌های واقعی وجود دارد.

- **صحت داده:** داده جمع‌آوری شده صحیح بوده و نادرستی در آن وجود نداشته باشد.
به عنوان مثال نباید جنسیت شخصی با نام «محسن» زن وارد شده باشد و یا اشتباه‌های دیگری که دلیل وجودی آنها خطا در ورود داده است، رخ دهد.
توجه شود که داده اشتباه با داده دارای نویز کاملا متفاوت است!

- **کافی بودن ویژگی‌ها:** ویژگی‌های اخذ شده برای هر رکورد یا شیء برای یادگیری مدل و یا کشف نظم حاکم بر داده مؤثر، مناسب و کافی باشند.
به عنوان مثال اگر هدف ما یادگیری یک مدل دسته‌بندی‌کننده برای تشخیص بیماری دیابت در یک کلینیک است، ثبت ویژگی قندخون بسیار مهم است در حالی‌که وجود ویژگی میزان تحصیلات اهمیتی ندارد!

توجه کنید چنانچه هرکدام از مشکلات سه گانه فوق در کل داده‌ها موجود باشند هیچ یک از الگوریتم‌های داده‌کاوی، هرقدر هم که توانا باشند، نخواهند توانست نظم حاکم بر داده را تحت هیچ شرایطی بیابند!

در داده‌کاوی می‌توان چالش‌ها را به دو گروه **اولیه** و **ثانویه** تقسیم نمود.

چالش‌های داده‌کاوی

چالش‌های اولیه

چالش‌های اولیه که انگیزه مهم به کارگیری فرآیند داده‌کاوی به جای روش‌های سنتی تحلیل داده‌ها هستند عبارتند از:

۱)حجیم بودن داده‌ها
۲) ابعاد بالای داده‌ها
۳) طبیعت توزیع شده
۴)طبیعت ناهمگن داده‌ها

- **حجم‌بالای‌داده:**

کار کردن الگوریتم‌های داده‌کاوی با تعداد زیادی از رکوردها و پردازش حجم زیادی از داده‌ها ← روش‌های سنتی نمی‌توانند این رکوردها را پردازش کنند.

هر چه تعداد رکوردها بیشتر ← کارکرد علم داده‌کاوی درخشان‌تر

- مانند داده‌های بسیارزیاد در سرشماری، داده‌های مخابراتی، تراکنش‌های بانکی و ....

- **ابعاد بالای داده‌ها:**

بعد یا فیلد یا ویژگی (خصیصه)

Attribute →	Data mining
Feature →	Machine learning
Dimension →	Data warehouse
Variable →	Statistics

هر چه تعداد ویژگی‌ها بیشتر ← تحلیل داده‌ها مشکل‌تر ← اثربخشی و توان بالقوه‌ای الگوریتم‌های داده‌کاوی بیشتر

- مانند: شماره‌دانشجویی، مقطع‌تحصیلی، محل‌تولد، تعداد واحد گذرانده و .... برای یک دانشجو

چالش‌های داده‌کاوی

- **طبیعت توزیع‌شده داده‌ها:**

به دلیل طبیعت توزیع‌شده داده‌ها و وجود داده‌ها در منابع پراکنده ⬅ **نیاز به روش‌های داده‌کاوی وجود دارد.**

این روش‌ها باید قادر باشند داده‌هایی را که در مکان‌های مختلف ذخیره شده‌اند به گونه‌ای مدیریت کنند که:

دانش نهفته را از نهان این داده‌های پراکنده و توزیع شده استخراج کنند!

- **مانند داده‌های به دست آمده از ترکیب اطلاعات چند سنسور، تصاویر ارسالی از طریق ماهواره‌ها و تلسکوپ‌ها و ...**

چالش‌های داده‌کاوی

چالش‌های اولیه

- **طبیعت ناهمگن داده‌ها:**

وجود انواع مختلفی از ویژگی‌ها درانباره داده‌ای که به عنوان مخزن فرایند داده‌کاوی عمل می‌کند.
هر ویژگی محدوده مقادیر مشخص و ویژه‌ای اختیار می‌کند.
کمینه و بیشینه مقادیر مربوط به بعضی از ویژگی‌ها با هم تفاوت زیادی دارند.

مسائل مربوط به ویژگی‌ها:

- ✓ بعضی از ویژگی‌ها، حوزه مقداری بسیار وسیع و بعضی دیگر حوزه محدودی دارند. در این مواقع می‌توان از مباحث نرمال‌سازی برای برخورد با این مشکل استفاده کرد.
- ✓ بعضی از ویژگی‌ها عددی (صحیح یا حقیقی) هستند، بعضی دودویی هستند، بعضی دیگر از ویژگی‌ها اسمی هستند (مثل رنگ چشم) که در مورد آنها تنها می‌توان گفت آیا با هم مساوی هستند یا خیر(مثل رنگ چشم)،
- ✓ گونه‌ای دیگر از ویژگی‌ها به این شکل‌اند که در مورد آنها علاوه بر مساوی یا نامساوی بودن می‌توان بزرگتر و کوچکتر بودن را نیز تعیین نمود (مثل سطح تحصیلات)
- ✓ برخی دیگر از ویژگی‌ها علاوه‌بر مساوی یا نامساوی بودن و تعیین کوچکتر و یا بزرگ‌تری، می‌توان از عملگرهای جمع و تفریق نیز استفاده نمود (مثل تاریخ‌های تقویم) و در نهایت در مورد گروهی دیگر از ویژگی‌ها علاوه بر مساوی یا نامساوی، کوچکتر و یا تفاوت بودن، همچنین جمع و تفریق آنها، می‌توان عملگرهای ضرب و تقسیم را نیز به کار برد.(مثل قد و وزن)

چالش‌های داده‌کاوی

چالش‌های ثانویه

در قیاس با چالش‌های اولیه اهمیت کمتردارند اما اهمیت این چالش‌ها کم نیست و یا حل مشکلات مربوط به آنها کار ساده و یا کم تأثیری نیست.

- کیفیت داده(Data Quality): مربوط به زمانی است که کیفیت داده‌ها پایین است.
  به عنوان نمونه داده‌های شامل نویز، داده پرت(outlier)، داده گمشده (Missing value) و داده‌تکرارشده (Duplicate Data) ⬅ کیفیت داده‌ها پایین
- مالکیت و توزیع داده(Data ownership and Distribution):
  به دلایل گوناگون مانند توزیع‌شدگی ممکن است نتوانیم کل داده‌ها را یک‌جا در مالکیت داشته باشیم و فرآیند کاوش را روی آنها انجام دهیم.
- حفظ حریم‌شخصی داده‌ها(Privacy Preservation):
  در فرآیند کاوش داده می بایست بتوان بدون دسترسی به همه داده‌ها و با دیدن تنها بخش محدودی از آن برای حفظ و رعایت حریم‌شخصی داده‌ها، فرآیند داده‌کاوی را پیش برد.
  ✓ *تفاوت حفظ حریم‌شخصی داده‌ها با توزیع‌شدگی و عدم مالکیت داده در این است که در توزیع‌شدگی و عدم‌مالکیت داده ممکن است برای یادگیری مدل از همه داده‌ها استفاده شود ولی در اینجا ممکن است به بخشی از داده‌ها اصلا دسترسی وجود نداشته باشد، یعنی باید بتوانیم مدل خود را با همان داده‌های در دسترس بسازیم!*
- داده‌های جریانی(Streaming Data):
  به داده‌هایی گفته می‌شود که سرعت‌تولید آنها بالا است به گونه‌ای که فرصت تحلیل آنها و ساخت مدل وجود ندارد چرا که حین انجام‌عملیات کاوش مرتبا داده‌های جدیدی تولید می‌شوند بنابراین سیستم باید به صورت برخط باشد تا بتواند خودش را تصحیح کند و قادر باشد مدل به روزی را در اختیار قرار دهد.
  مانند ماهواره‌های فضایی که تعداد آنها چندین میلیون است و در هر ثانیه تصویری برای مرکز مخابره می‌کنند یا داده‌های کاربران وب در موتور جست و جو گوگل

معرفی داده‌کاوی

برای درک کامل معنای داده‌کاوی می‌بایست ابتدا تعریف درستی از معانی کلمات داده، اطلاعات و دانش را داشته باشیم.

- داده: به هر گونه سیمبل، عدد، رقم، کاراکتر، رشته و یا سیگنال که معنای خاصی را به ذهن القاء نکند داده گفته می‌شود. داده پایه ای‌ترین مفهوم در داده‌کاوی است که مبرا از هرگونه پردازشی می‌باشد.
- اطلاعات: چنان چه در کنار عدد، کاراکتر و یا هر عنصر داده‌ای رشته‌ای به عنوان توصیف‌کننده داده وجود داشته باشد، داده ابتدایی به اطلاعات تبدیل خواهد شد. می توان به صورت خلاصه برای تعریف اطلاعات از عبارت داده درباره داده (Data about Data) استفاده نمود.
- دانش: وجود یک رابطه میان دو عنصر اطلاعاتی مبین دانشی در آن زمینه است. در تعریف ساده دانش می‌توان از عبارت جالب اطلاعات درباره اطلاعات(Information about Information) استفاده کرد.
- خرد: عالی‌ترین سطح بینش است که توسط علائم و نمادهای قراردادی تبیین می شود. تعریف ساده آن دانش درباره دانش (Knowledge about Knowledge) می‌باشد.

مثال:

عدد ۵۹ ← داده ← می‌تواند شماره پلاک یک خانه، تعداد درخت‌های یک باغ، سن و یا وزن یک انسان و ...

مجموعه سن ۵۹ ← یک عنصر اطلاعاتی

«اگر سن پدر تنها پسر یک خانواده ۵۹ سال تمام باشد آن پسر از سربازی معاف می‌شود» ← یک عنصر دانش ← قوانین «اگر آن گاه» شکل خاصی برای نمایش دانش هستند

معرفی داده‌کاوی

برای مفهوم خرد داریم:
یک پزشک متخصص و یک دانشجوی پزشکی که تازه فارغ التحصیل شده است هر دو مجموعه بسیار زیادی از قوانین اگر آن گاه مربوط به دانش پزشکی را در ذهن خود نگهداری می‌کنند. اما در برخورد با یک بیمار مشابه ممکن است تشخیص‌های بسیار متفاوتی بدهند. این تفاوت ریشه در فرادانش (Meta Knowledge) و یا همان خردی دارد که پزشک متخصص در طی سالیان متمادی کسب کرده است. این فرادانش به گونه‌ای است که امکان کنترل و هدایت قوانین و یا همان دانش ابتدایی را برای پزشک به وجود می‌آورد!

- ✓ همزمان با افزایش ارزش معنایی ← حجم آنها کاهش می یابد.
- ✓ بدیهی است که بتوان حجم بالایی داده را با تنها چند قانون توصیف و تبیین نمود یعنی کاری که اصلی‌ترین هدف در فرآیند داده کاوی است!

سلسله‌مراتب ارزشی برای معانی داده، اطلاعات، دانش و خرد

معرفی داده‌کاوی

تعریف داده‌کاوی:

استخراج خودکار دانش جدید و مفید از منابع داده‌ای حجیم موجود طی یک فرایند غیر بدیهی مشخص داده‌کاوی نامیده می‌شود.

هدف اصلی در داده‌کاوی کشف دانش است، این دانش نظمی خواهد بود که در داده‌ها وجود دارد!

منشا علمی

منشأ علمی علم داده‌کاوی از علوم مختلفی از جمله علم آمار، هوش‌مصنوعی، یادگیری‌ماشین، شناسایی الگو و پایگاه داده نشأت گرفته است؛در واقع این علوم ریشه‌های علم داده‌کاوی هستند.

الگوریتم‌های موجود در هوش مصنوعی و علم آمار کمک شایانی به داده‌کاوی می‌کنند. مباحث موجود در یادگیری ماشین و شناسایی الگو نیز با مباحثی که در داده‌کاوی هستند همپوشانی قابل ملاحظه‌ای دارند.

به عنوان مثال الگوریتم‌هایی که یک مدل را یاد می‌گیرند یا الگویی را شناسایی می‌کنند، به خصوص اگر داده‌های مورد پردازش عددی یا متنی باشند (سمبولیک نباشند) معمولا وجه مشترک یادگیری ماشین و شناسایی الگو با داده‌کاوی هستند!

معرفی داده‌کاوی

مراحل داده‌کاوی :

فرآیند داده کاوی شامل سه مرحله است: **آماده‌سازی‌داده، یادگیری مدل، ارزیابی و تفسیرمدل**

آماده سازی داده

اولین و مهم‌ترین مرحله در فرآیند داده‌کاوی: آماده‌سازی داده

هدف : در این مرحله تأمین ورودی مناسب برای مرحله حیاتی یادگیری مدل

- ✓ استخراج داده پردازش نشده از کل منابع داده‌ای موجود (که ممکن است توزیع شده نیز باشد)
- ✓ سپس پردازش اولیه در مرحله ای مستقل
- ✓ خروجی در مرحله آماده‌سازی داده: داده پیش‌پردازش‌شده که امکان یادگیری مدل از روی آن وجود دارد.

اولین گام : استخراج داده از منابع داده‌ای موجود

در این گام می‌بایست داده‌ها که در منابع مختلفی پراکنده شده‌اند، به صورت متمرکز در یک محل جمع آوری شده و یک انباره‌داده مرکزی ایجاد شود. دلیل اصلی این گرداوری: در اغلب موارد داده به صورت متمرکز در یک مکان وجود ندارد.

به علاوه داده‌ها در بخش‌های مختلف ممکن است در فرمت‌های گوناگونی نیز ذخیره شده باشند!

دومین گام :پیش‌پردازش داده‌های استخراج‌شده

مهم‌ترین رسالت این گام: زدودن مشکلات مختلفی که احتمالا در داده وجود دارند ← می‌توان در مرحله یادگیری مدل نظم واقعی در داده را پیدا کرد!

یادگیری مدل

با استفاده از الگوریتم‌های متنوع و با توجه به ماهیت داده، نظم‌های مختلف موجود در داده را شناسایی و در فرمتی مشخص ارائه کنیم.

برای یادگیری‌مدل می‌بایست روش‌های آن را به درستی شناخت تا بتوان در جای مناسب، روش درست را انتخاب نمود و به کار بست!

معرفی داده‌کاوی

ارزیابی و تفسیر مدل

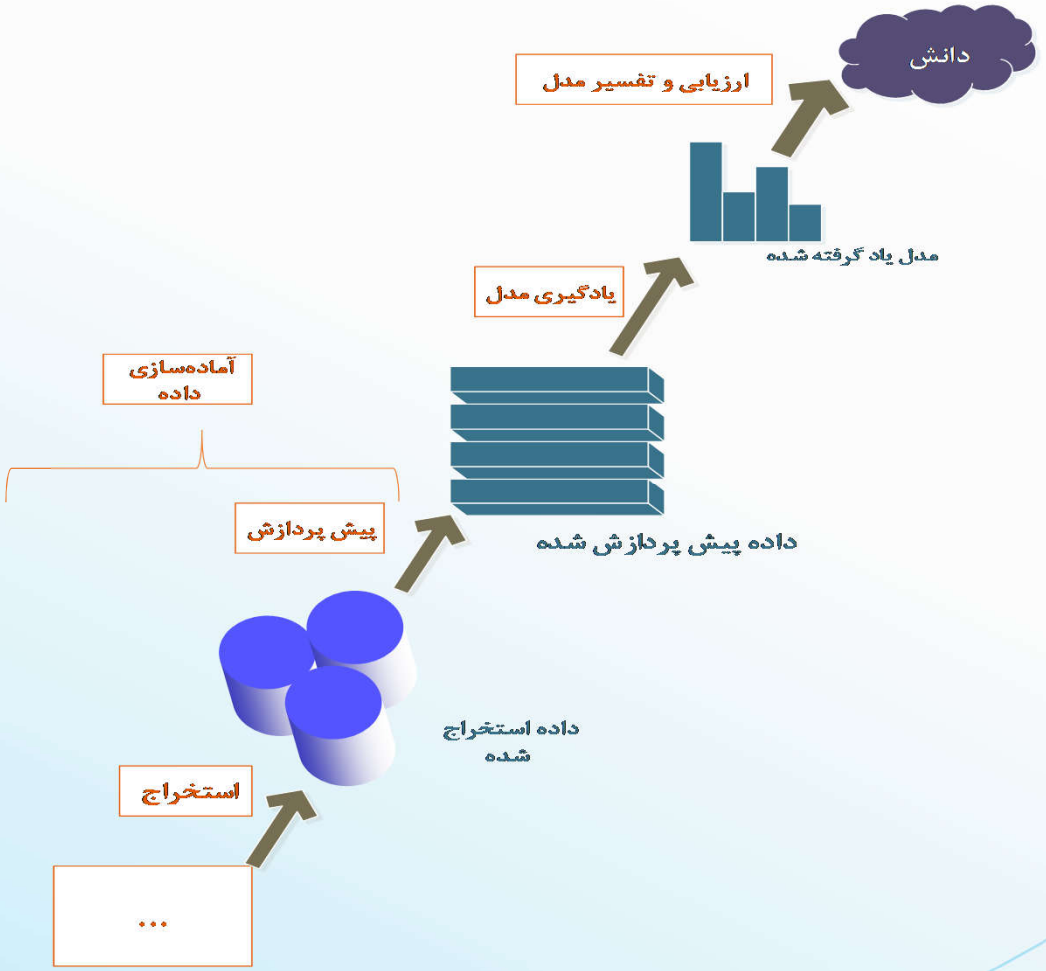
ارزیابی و تفسیر دانش تولید شده در مرحله قبل

هدف: تعیین میزان صحت دانش تولیدشده است تا بتوان به آن اعتماد نمود و به صورت علمی از آن استفاده کرد!

تفسیر مدل به معنای آن است که توجیهی معنایی جهت تبیین منطق دانش تولیدشده ارائه نماییم.

- **درصورت قابل تفسیر بودن دانش تولید شده، انجام این کار بسیار ساده است (به عنوان مثال زمانی که دانش به صورت درخت یا مجموعه‌قوانین باشد).**
- **در مقابل امکان تفسیر دانش برای مواقعی که دانش به صورت غیرقابل تفسیر باشد (مانند دانش تولید شده توسط شبکه‌های‌عصبی و ماشین‌بردار پشتیبان) بسیار مشکل‌تر و شاید غیرممکن خواهد بود!**

معرفی داده‌کاوی



روش‌های یادگیری مدل در داده‌کاوی

روش‌های مختلف کاوش‌داده: دو گروه روش‌های پیش‌بینی و روش‌های توصیفی

روش‌های پیش بینی

از مقادیر بعضی از ویژگی‌ها برای پیش‌بینی کردن مقدار یک ویژگی مشخص استفاده می‌کنند.
در متون علمی مختلف روش‌های پیش‌بینی با نام روش‌های با ناظر(Supervised Methods) نیز شناخته می‌شوند.
روش‌های دسته‌بندی، رگرسیون و تشخیص انحراف سه روش یادگیری مدل در داده‌کاوی با ماهیت پیش بینی هستند.

دسته بندی (Classification)

در الگوریتم‌های دسته‌بندی مجموعه داده اولیه به دو مجموعه داده با عنوان مجموعه‌داده‌های‌آموزشی (Train Dataset) و مجموعه‌داده‌آزمایشی (Test Dataset) تقسیم می شود.

- ساخت مدل با استفاده از مجموعه داده‌های‌آموزشی
- اعتبارسنجی و محاسبه دقت مدل با استفاده از مجموعه‌داده آزمایشی

هر رکورد شامل یک مجموعه از ویژگی‌ها است،یکی از این ویژگی‌ها، ویژگی دسته نامیده می‌شود. در الگوریتم‌های دسته‌بندی چون ویژگی دسته مربوط به هر رکورد مشخص است بنابراین جزء الگوریتم‌های با ناظر محسوب می‌شوند!

الگوریتم‌های با ناظر شامل دو مرحله با عنوان مرحله آموزش(یادگیری) و مرحله ارزیابی هستند.
مرحله آموزش: مجموعه داده‌های آموزشی به یکی از الگوریتم‌های دسته‌بندی داده می‌شود تا براساس مقادیر سایر ویژگی‌ها برای مقادیر ویژگی دسته، مدل ساخته شود. شکل مدل ساخته شده به نوع الگوریتم یادگیرنده بستگی دارد.
مثلا:
اگر الگوریتم یادگیرنده الگوریتم درخت‌تصمیم (Decision Tree) باشد مدل ساخته شده یک درخت تصمیم خواهد بود.
اگر الگوریتم یادگیرنده یک دسته بندی مبتنی بر قانون(Rule-Based Classifier) باشد مدل ساخته شده یک مجموعه قانون خواهد بود.

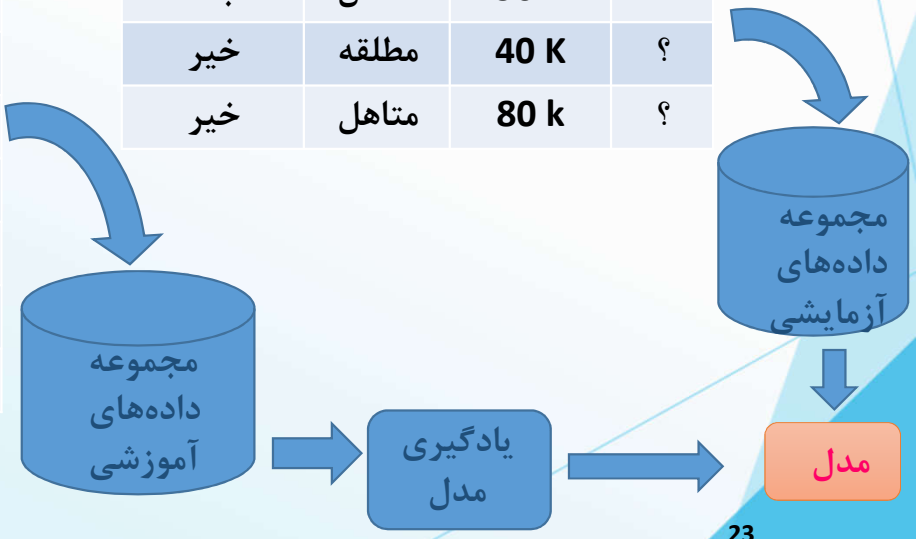
روش‌های یادگیری مدل در داده‌کاوی

از مجموعه داده‌های آزمایشی در مرحله آموزش و ساخت مدل استفاده نمی‌شود!

- **مثالی برای دسته‌بندی:**

تقلب	مالیات بر درآمد	وضعیت تاهل	بازپرداخت
؟	75 K	مجرد	خیر
؟	50 K	متاهل	بله
؟	150 K	مجرد	خیر
؟	90 K	متاهل	بله
؟	40 K	مطلقه	خیر
؟	80 k	متاهل	خیر

ردیف	بازپرداخت	وضعیت تاهل	مالیات بر درآمد	تقلب
۱	بله	مجرد	125 K	خیر
۲	خیر	متاهل	100 K	خیر
۳	خیر	مجرد	70 K	خیر
۴	بله	متاهل	120 K	خیر
۵	خیر	مطلقه	95 K	بله
۶	خیر	متاهل	60 k	خیر
۷	بله	مطلقه	220 K	خیر
۸	خیر	مجرد	85 K	بله
۹	خیر	متاهل	75 K	خیر
۱۰	خیر	مجرد	90 K	بله



روش‌های یادگیری مدل در داده‌کاوی

- پزشکی: هدف ساخت یک مدل برای دسته هپاتیت، دیابت و سالم می‌باشد.
یک بیمار جدید با ویژگی‌های شامل سن، قد ، وزن، فشارخون و... به مدل به کدام یک از این سه دسته تعلق دارد؟
امکان تشخیص دسته‌جدید در کاربرد دسته‌بندی وجود نخواهد داشت.

- بازاریابی‌مستقیم(Direct Marketing): هدف، کاهش هزینه‌پست بسته‌های تبلیغاتی از طریق پیداکردن مصرف‌کننده‌هایی است که احتمال خرید یک گوشی‌تلفن‌همراه جدید توسط آنها نسبت به سایرین بیشتر است.
دو نوع دسته متفاوت: عنوان خریدار و غیرخریدار
در هر رکورد به غیر از دسته اطلاعات سبک زندگی شامل نوع کار، محل سکونت، میزان درامد و ...وجود دارد.
براساس مجموعه داده‌های آموزشی برای ویژگی‌های دسته خریدار و غیرخریدار مدل‌نهایی ساخته می‌شود.

- تشخیص‌کلاهبرداری(Fraud Detection): هدف، پیش‌بینی موارد کلاهبرداری در تراکنش‌های کارتهای اعتباری است.
ویژگی دسته: دو مقدار کلاهبرداری یا عادی. زمانهای خرید، نوع خرید، پرداخت به موقع، اطلاعات حساب و ..در هر رکورد وجود دارد.
براساس مجموعه داده‌های‌آموزش، مدلی را برای وضعیت‌های کلاهبرداری و عادی می‌سازد.

روش‌های یادگیری مدل در داده‌کاوی

- **میزان ماندگاری یا از دست دادن مشتری(Customer Attrition/Churn):**
  هدف پیش‌بینی آن است که آیا احتمال دارد که یک مشتری به سمت رقیب ما برود یا خیر؟

از میان داده‌های مربوط به مشتریان گذشته و حاضر یک مجموعه ویژگی استخراج می‌کنیم.
ویژگی دسته: دو مقدار وفادار و بی وفا
فرکانس تماس مشتری، تماس با کدام شعب، میزان درامد مشتری، وضعیت تاهل، .. در رکورد مشتری میتواند وجود داشته باشد.
در مرحله آموزش الگوریتم براساس مجموعه داده‌های آموزشی مدلی برای وفاداری و بی‌وفایی می‌سازد. حال ویژگی‌های مربوط به مشتری‌جدید را به مدل می‌دهیم. مدل بر اساس آنها تصمیم می‌گیرد که آیا این فرد وفادار خواهد بود یا بی‌وفا.

- **دسته بندی اجرام آسمانی:** تعدادی از ویژگی‌ها: بزرگی توده نورانی، هیستوگرام شدت نور تصویر و... به ازای تصویر مربوط به هر کهکشان
  ویژگی دسته : «مراحل شکل‌گیری»دارای سه مقدار جوان، میانسال و پیر
  به ازای هر تصویر یک رکورد داریم و هر رکورد یکی از این سه مقدار ویژگی دسته را به خود می‌گیرد.
  به ازای تصویر مربوط به یک کهکشان جدید ویژگی‌های آن را استخراج کرده فراهم کرده و در اختیار مدل ساخته شده قرار می‌دهیم.
  مدل بر اساس انها یکی از سه دسته جوان میان‌سال و یا پیر را به تصویر کهکشان جدید نسبت می‌دهد.

روش‌های یادگیری مدل در داده‌کاوی

رگرسیون (Regression)

- ✓ پیش‌بینی مقدار یک متغیرپیوسته براساس مقادیر سایر متغیرها بر مبنای یک مدل وابستگی‌خطی یا غیرخطی رگرسیون نامیده می‌شود.
  در واقع یک بردار به عنوان ورودی داریم که به یک متغیر خروجی y نگاشت شده است.

- ✓ هدف محاسبه y یا همان $F(\vec{x})$ است که از روی تخمین تابع مقدار آن محاسبه می‌شود. باید به ازای یک بردار $\vec{x}$ ، مقدار دقیق y قابل محاسبه باشد.
- ✓ یک کاربرد از نوع پیش‌بینی با ناظر است. رگرسیون هم دارای دو مرحله: آموزش و ارزیابی
- ✓ روش‌های موجود در رگرسیون بیشتر مبتنی بر ریاضیات آماری هستند.
- ✓ یک نوع خاصی از رگرسیون: پیش‌بینی سری‌های زمانی
  در مسائل پیش‌بینی سری‌های زمانی یکی از متغیرهای اصلی زمان می‌باشد. در این مسائل یک مجموعه از X ها و y ها به صورت یک تابع‌ریاضی وجود دارند. هدف این است که به ازای یک X جدید مقدار آن را پیش‌بینی کنیم.

به عنوان مثال: پیش‌بینی تغییرات قیمت سهام ایران خودرو با داشتن نمودار از سه سال پیش تا به امروز!
به این مسأله پیش‌بینی سری زمانی گفته می‌شود که نوع خاصی از رگرسیون است؛ رگرسیون لزوما سری زمانی نیست.

مثال‌هایی از رگرسیون عبارتند از :

- ❑ پیش‌بینی میزان فروش یک محصول جدید براساس میزان فروش محصولات گذشته، مشخصات محصولات گذشته و میزان تبلیغات انجام شده برای آنها.
- ❑ پیش‌بینی سرعت باد به عنوان تابعی از دما، رطوبت و فشار هوا.
- ❑ مسائل مربوط به پیش بینی سری‌های زمانی از قبیل: بورس اوراق بهادار، تغییرات جوی آب و هوا و...

روش‌های یادگیری مدل در داده‌کاوی

تشخیص انحراف(Anomaly Detection)

اخرین کاربرد مهم یادگیری با ناظر در داده‌کاوی: تشخیص انحراف

موارد استفاده: هنگامی که تنها نمونه‌های با یک برچسب یکسان، که معمولا وضعیت نرمال را نشان می‌دهد در دسترس باشند و امکان مالکیت بر داده‌ها با تمامی برچسب‌های موجود به دلایل مختلف وجود نداشته باشد.

بنابراین الگوریتم برای وضعیت نرمال و با توجه به یک آستانه‌مشخص مدل می‌سازد و هر گونه تخطی از آن آستانه را به عنوان وضعیت غیرنرمال در نظر می‌گیرد و هشدار می‌دهد!

دو نمونه از کاربردهای تشخیص انحراف عبارتند از: کشف کلاهبرداری‌های کارت‌های اعتباری و تشخیص نفوذ به شبکه‌های کامپیوتری.

مثال: یک شرک بیمه را در نظر بگیرید که یک سری افراد ادعای خسارت کرده‌اند.

– شرکت پرونده افرادی که به دروغ ادعای خسارت کرده‌اند، رکوردی از آنها در سیستم نگهداری نکرده است.

– اما به افرادی که ادعای خسارت آنها درست ارزیابی شده، در سیستم ثبت کرده است.

حال در چنین شرایطی چون فقط نمونه‌های دسته اول را داریم نمی‌توانیم از دسته‌بندی استفاده کنیم اما می‌توانیم از تشخیص انحراف استفاده نموده و برای رکوردهای با دسته اول یک مدل نرمال را بسازیم.

از مهم‌ترین نقاط قوت روش‌های تشخیص انحراف امکان تشخیص کلاه‌برداری‌ها و یا نفوذهایی است که قبلا رخ نداده و یا به اصطلاح جدید می‌باشند.

امکان تشخیص هیچ گونه کلاهبرداری و یا نفوذ جدیدی توسط روش‌های دسته‌بندی وجود ندارد زیرا روش‌های دسته‌بندی تنها قادر به تشخیص دسته‌هایی هستند که در مرحله آموزش نمونه ای از آنها به الگوریتم ارائه شده باشد!

روش‌های یادگیری مدل در داده‌کاوی

روش‌های توصیفی

این روش‌ها الگوهای قابل‌توصیفی را پیدا می‌کنند که روابط حاکم بر داده‌ها را بدون در نظر گرفتن هر گونه برچسب و یا متغیرخروجی تبیین نمایند.
در متون علمی مختلف روش‌های توصیفی با نام روش‌های بدون ناظر (Unsupervised Method) نیز شناخته می‌شوند.
روش‌های خوشه بندی، کاوش قوانین انجمنی و کشف الگوهای ترتیبی به روش یادگیری مدل در داده کاوی با ماهیت توصیفی هستند.

خوشه بندی (Clustering)

- در مسائل خوشه‌بندی یک مجموعه رکورد داریم که هر کدام یک مجموعه از ویژگی‌ها را دارا هستند.
- یک معیار مشابهت میان آنها تعریف می‌کنیم، این معیار مشابهت در مسائل مختلف متفاوت است!

به عنوان مثال اگر ویژگی‌ها پیوسته باشند می‌توان فاصله اقلیدسی را به عنوان معیار مشابهت در نظر گرفت، به این ترتیب هر رکورد را به صورت یک نقطه در فضای چندبعدی در نظر می‌گیریم، هر بعد، نماینده یکی از ویژگی‌های مسأله است.

نکته مهم: در مسائل خوشه‌بندی هیچ گونه دسته خاصی وجود ندارد، در واقع ویژگی‌دسته نداریم و فقط براساس معیار شباهت گروه‌بندی و خوشه‌بندی داده‌ها صورت می‌پذیرد!

درخوشه‌بندی رکوردهایی که بیشترین شباهت را به یکدیگر دارند (با توجه به معیار شباهت تعریف شده) در یک خوشه قرار می‌گیرند.

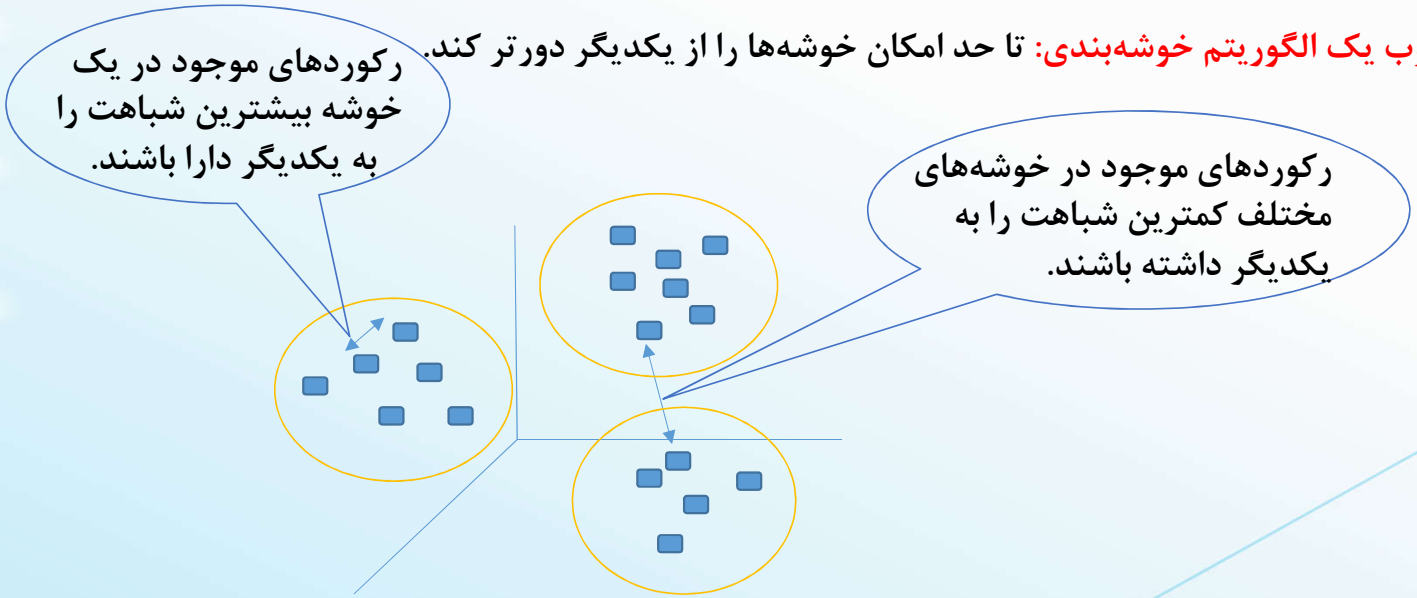
از آنجایی که برای الگوریتم‌های خوشه‌بندی ویژگی دسته تعریف نمی‌شود و رکوردها برچسب خاصی ندارند، بنابراین جزء الگوریتم‌های بدون ناظر محسوب می‌شوند.

روش‌های یادگیری مدل در داده‌کاوی

- خروجی الگوریتم‌های خوشه‌بندی دوباره تحلیل خواهد شد تا در صورت امکان نظمی در خوشه ها آشکار شود.
نکته مهم قابل توجه : خوشه‌بندی همیشه براساس ویژگی‌های‌ورودی‌نمونه‌ها انجام می‌شود.

به عنوان مثال: خوشه‌بندی رکوردهای مربوط به دانشجویان یک دانشکده
هر خوشه بیانگر رکوردهایی باشد که از جنبه‌های مختلف به یکدیگر شبیه هستند.
مثلا یک وضعیت: دو خوشه نشانگر دانشجویان زرنگ و تنبل

- هدف در همه الگوریتم‌های خوشه بندی: کمینه کردن فاصله درون خوشه‌ای و بیشینه نمودن فاصله بین خوشه‌ای می‌باشد.
- عملکرد خوب یک الگوریتم خوشه‌بندی: تا حد امکان خوشه‌ها را از یکدیگر دورتر کند.



روش‌های یادگیری مدل در داده‌کاوی

کاربردهای خوشه بندی

- **پزشکی:** ما قصد داریم بفهمیم به چند دلیل مختلف بیماران مربوط به کلینیک تلف شده‌اند.

کل رکوردها(شامل سن، وزن، قد، فشار خون، ..) به یک الگوریتم خوشه‌بندی ارائه می‌شود. حال اگر مثلا الگوریتم برای این رکوردها سه خوشه ایجاد کرده باشد و در هر خوشه رکوردهای مشابه را قرار داده باشد، رکوردهای موجود در هریک از خوشه‌ها را دوباره بررسی می‌کنیم تا در صورت امکان برای هر خوشه نظمی را پیدا کنیم. مثلا ممکن است پس از بررسی به این نتیجه برسیم:

- افراد موجود در خوشه اول: مبتلا به ایدز
- افراد موجود در خوشه دوم: مبتلا به سرطان
- افراد موجود در خوشه سوم: مبتلا به آنفولانزای مرغی

توجه شود که:

- این سه نظم بدون در نظر گرفتن هر گونه برچسب برای رکوردها یافته شده‌اند.
- ممکن است پس از بررسی این سه خوشه هیچ نتیجه ویژه‌ای نتوان گرفت پس ویژگیهای موثرتری را باید پیدا کنیم.

- **قطعه‌بندی بازار(Market Segmentation):** هدف در مثال قطعه‌بندی کردن مشتریان یک فروشگاه به زیرمجموعه‌های مجزا به گونه‌ای است که در هرقسمت از فروشگاه کالاهایی که توسط یک مجموعه مشتریان خریداری می‌شوند، قرار بگیرند!

هدف از خوشه‌بندی خریدها آن است که کشف کنیم چند نوع **عادت خرید** مختلف داریم!

توجه: رکوردها دسته خاصی ندارند؛ کل این رکوردها به عنوان ورودی به الگوریتم خوشه بندی ارائه می‌شوند!

معیارمشابهت: وجود کالاهای مشابه در سبدها!

خوشه‌های محتمل: پروتینی، لبنیاتی، تنقلاتی

روش‌های یادگیری مدل در داده‌کاوی

خوشه‌بندی

خوشه‌بندی اسناد(Document Clustering): هدف پیدا کردن گروه‌هایی از اسناد مشابه براساس تعداد رخداد کلمات و اصطلاحات مهم موجود در آنها می‌باشد.

به عنوان مثال فرض کنید یک مجموعه مقاله داریم، از میان چکیده این مقاله‌ها عبارات و کلمات کلیدی مهم را استخراج می‌کنیم.

یک مجموعه رکورد جمع آوری می‌نماییم که هر رکورد نماینده یکی از مقاله‌ها باشد.

هر رکورد مجموعه مشخصی از ویژگی‌ها را دارد که همان کلمات مهم موجود در مقاله‌ها هستند.

- **تعداد رخداد هرکدام از این کلمات را در هر یک از مقاله‌ها می‌یابیم و به عنوان مقدار ویژگی‌ها برای هر یک از رکوردها در جدول ثبت می‌کنیم.**
- **سپس مجموعه رکوردهای حاصله را به الگوریتم خوشه‌بندی ارائه می‌نماییم.**
  فاصله بین رکوردها: تعداد رخداد کلمات مهم در هر یک از اسناد
- **تعدادی خوشه براساس فاصله رکوردها ایجاد می‌شود.**

به عنوان مثال ممکن است چهار خوشه ایجاد شده باشد، پس از بررسی مشخص خواهد شد اسنادی که در یک خوشه قرارگرفته‌اند مرتبط با چه موضوعاتی هستند!

خوشههای محتمل: سیاسی، ورزشی، فرهنگی، مالی

مزیت خوشه‌بندی اسناد در کاربردهای بازیابی اطلاعات است. یک نمونه دیگر از کاربرد خوشه‌بندی اسناد، در تعیین خوشه یک سند جدید با توجه به خوشه‌های بافته شده می‌باشد!

روش‌های یادگیری مدل در داده‌کاوی

کشف قوانین انجمنی (Association Rule Mining)

در این کاربرد به دنبال پیدا کردن یک مجموعه از قوانین وابستگی یا انجمنی هستیم که براساس آن قوانین بگوییم وجود کدامیک از مجموعه اشیاء بر وجود چه مجموعه اشیاء دیگری اثرگذار است.

به عنوان مثال در یک سوپرمارکت اگر کسی کالای **x** را خرید آنگاه کالای **y** را هم می‌خرد. این قوانین وابستگی اتفاق و وقوع یک شیء را براساس وقوع سایر اشیا پیش‌بینی می‌کنند.

در مثال سوپرمارکت کالاهای هر مشتری در سبد خرید مربوط به آن مشتری قرار می‌گیرند.

هدف در کاوش قوانین انجمنی: یافتن نظم حاکم بر سبدها

به ازای هر سبد یک قانون پیدا می‌شود و بررسی خواهد شد که این قانون در چه تعداد از سبدها صدق می‌کند.

در نهایت یک مجموعه قانون که در بیشترین تعداد از سبدها صدق می‌کنند به عنوان مجموعه قوانین انجمنی خروجی ارائه می‌شوند !

در این کاربرد هم یک مجموعه رکورد داریم که هیچ کدام برچسب خاصی ندارند و از این رو در رده الگوریتم‌های توصیفی یا بدون‌ناظر جای خواهد گرفت.

در کاوش قوانین انجمنی فرض بر این است که یک جدول از تراکنش‌ها داریم (مثلا هر رکورد یک سبد خرید را نشان می‌دهد)

هدف: یک تحلیل ستونی یا مبتنی بر ویژگی روی رکوردها صورت پذیرد!

توجه: تحلیلی که در خوشه‌بندی رخ می‌دهد، یک تحلیل سطری است!

روش های یادگیری مدل در داده‌کاوی

کشف قوانین انجمنی

ردیف	تراکنش خرید
۱	نان، شیر، زغال
۲	نان، نوشابه
۳	نوشابه، زغال، مسواک، شیر
۴	نوشابه، نان، مسواک، شیر
۵	شیر، مسواک، زغال

قوانین انجمن کشف شده:

{زغال} ⟵ {شیر}

{نوشابه} ⟵ {شیر، مسواک}

روش‌های یادگیری مدل در داده‌کاوی

کشف قوانین انجمنی

چند نمونه از انواع مختلف کاربردهای کشف قوانین انجمنی :

- **ارتقای بازاریابی و فروش(Market and Sale Promotion):**

فرض کنید قانون کشف‌شده به این صورت باشد: {چیپس‌سیب‌زمینی} ⟶ {نان شیرینی حلقوی و ...}

یعنی: « اگر یک نفر نان‌شیرینی‌حلقوی و یک سری کالاهای دیگر را بخرد انگاه چیپس‌سیب‌زمینی را هم می‌خرد. »

سه نکته این قانون:

1. اگر نان‌شیرینی‌حلقوی و سایر کالاهای موجود در سمت چپ قانون را برای فروشگاه بیاوریم، می‌توانیم میزان فروش چیپس سیب‌زمینی را افزایش دهیم!

2. اگر فروشگاه فروش نان‌شیرینی‌حلقوی را متوقف کند فروش چه محصولات دیگری هم تحت تأثیر قرار می‌گیرد. اگر فروش نان شیرینی حلقوی را قطع کنیم، از لحاظ فروش چیپس‌سیب‌زمینی و کالاهای دیگر موجود در سمت چپ قانون هم ضرر می‌کنیم!

3. با توجه به سمت چپ قانون، می‌توانیم دریابیم که چه چیزهایی را باید همراه با نان‌شیرینی‌حلقوی بفروشیم تا میزان فروش چیپس‌سیب‌زمینی را تقویت کنیم!

روش‌های یادگیری مدل در داده‌کاوی

کشف قوانین انجمنی

- مدیریت قفسه در سوپرمارکت (Supermarket Shelf Management):
  - ✓ هدف: شناسایی کالاهایی در یک سوپرمارکت است که همراه با یکدیگر و توسط تعداد زیادی از مشتریان خریداری شده‌اند!
  - ✓ پس از پیدا کردن قوانین، کالاهایی که معمولا با یکدیگر خریداری می‌شوند را در چیدمان فروشگاه در کنار هم قرار دهیم، به این صورت میزان فروش افزایش می‌یابد و رضایت مشتریان، بیشتر جلب خواهد شد!

- مدیریت انبار (Inventory management):

به عنوان مثال یک تعمیرکار سیار خودرو مانند سایپایدک می‌بایست اطلاعات مربوط به تعمیر تمام خرابی‌هایی را که در گذشته برای وی رخ داده است ثبت نموده و آنها را مورد پردازش قرار دهد و بداند که معمولا کدام قطعات بیشتر استفاده می‌شوند تا در انبار خود ذخیره کند.

براساس کشف قوانین انجمنی تعمیرکار سیار مذکور خواهد توانست روابط میان قطعات مورد استفاده برای رفع خرابی در خودروها را یافته و از این روابط برای مدیریت مؤثر انبار قطعات خود بهره‌مند شود.

روش‌های یادگیری مدل در داده‌کاوی

کشف الگوهای ترتیبی(Sequential Pattern Discovery)

در اینجا به دنبال کشف الگوهای ترتیبی هستیم که وابستگی‌های ترتیبی محکمی را در میان وقایع مختلف نشان می‌دهند!

به عنوان مثال به دنبال پیدا کردن قانونی به صورت (D E) ⟶ (A B)(C) هستیم.

این قانون می‌گوید: اگر A و B به هر ترتیبی اتفاق افتادند (الگوهایی که در یک پرانتز قرار می‌گیرند ترتیب وقوعشان مهم نیست و با هر ترتیبی می‌توانند رخ دهند) و بعد از آنها C اتفاق افتاد، آنگاه D و E اتفاق می افتند (اینکه ابتدا D یا E اتفاق می‌افتند هم مهم نیست). قوانینی که در اینجا مطرح هستند در مورد مسائلی که در آنها زمان و ترتیب اهمیت دارد، قابل ارائه می باشند.

این کارکرد نیز ازجمله کارکردهای توصیفی است. این کاربرد مانند کاوش قوانین انجمنی می‌باشد، با این تفاوت که در کاوش قوانین انجمنی زمان و ترتیب زمانی مطرح نیست.

به عنوان مثال در مورد کالاهای موجود در سبد خرید ترتیب اهمیتی ندارد و اینکه کدامیک از کالاها را زودتر در سبد قرار می‌دهیم معنای خاصی را به تحلیل‌گر انتقال نمی‌دهد. اما در کشف الگوهای ترتیبی زمان و ترتیب دارای اهمیت ویژه‌ای است.

روش‌های یادگیری مدل در داده‌کاوی

کشف الگوهای ترتیبی

تعدادی از **کاربردهای** مربوط به **کشف الگوهای‌ترتیبی** به قرار زیر می باشند:

- **تغییرات ارزش سهام در بورس اوراق بهادار:** یکی از جدیدترین کاربردهای کشف الگوهای ترتیبی در یافتن ارتباط میان افت و خیزهای ارزش سهام شرکت‌های مختلف در بورس اوراق بهادار است.
  به عنوان مثال یکی از الگوهای ترتیبی نمونه در این مثال بدین‌گونه خواهد بود: چنان چه امروز ارزش سهام های بانک‌های ملی و ملت رشد بالایی داشته باشند در این صورت ارزش سهام بانک تجارت در دو روز بعد رشد زیادی را تجربه خواهد نمود.

روش‌های یادگیری مدل در داده‌کاوی

- **کشف عملیات خرابکارانه در شبکه: نفوذ یا هک که در یک شبکه اتفاق می‌افتد به صورت ناگهانی نیست، بلکه ابتدا پورت‌های شبکه مورد پویش قرار می گیرند، پس از آن به بعضی از این پورت‌ها بسته‌هایی ارسال خواهد شد.**
  به همین ترتیب در ادامه عملیاتی انجام می‌شود که در نهایت رویداد نفوذ رخ می‌دهد یعنی نفوذی که در شبکه با اهداف خرابکارانه اتفاق می‌افتد، دارای مراحل ترتیبی است، به همین دلیل با استفاده از روش‌های کشف الگوهای ترتیبی می‌توانیم در زمان مناسب، نفوذ را قبل از خطرساز شدن آن کشف و دفع نماییم!

روش‌های یادگیری مدل در داده‌کاوی

- **دنباله‌های تراکنش‌های فروش: در اینجا دو مثال ارائه می کنیم:**

✓ **در مثال اول محیط یک کتاب‌فروشی را در نظر بگیرید. اگر کسی وارد این کتاب‌فروشی شد و ابتدا کتاب پایگاه داده‌ها را خرید و پس از مدتی اقدام به خرید کتاب یادگیری‌ماشین نمود، آنگاه در آینده‌ای نزدیک کتاب داده‌کاوی کاربردی را نیز خواهد خرید؛ بدیهی است که در اینجا نیز زمان نقش تعیین‌کننده‌ای را ایفا می‌نماید!**

✓ **در مثال دوم محیط یک فروشگاه‌لباس‌ورزشی را در نظر بگیرید. اگر کسی وارد این فروشگاه شد و ابتدا کفش خرید و بعد از آن راکت و توپ تهیه نمود آنگاه بزودی یک لباس ورزشی از این فروشگاه خواهد خرید!**