

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

علوم داده - تحلیل انجمنی، خوشه بندی، ارزیابی مدل

تألیف :

محمد صدیقی منش، شروین سوری، علی صدیقی منش، فاطمه عباسی

سال ۱۳۹۹

سرشناسه: صدیقی منش، محمد

عنوان و نام پدید آور: علوم داده - تحلیل انجمنی، خوشه بندی، ارزیابی مدل - تألیف: محمد صدیقی منش - شروین سوری - علی

صدیقی منش - فاطمه عباسی

مشخصات نشر: قم، الهام نور، ۱۳۹۹.

مشخصات ظاهری: ۱۲۸ ص، مصور، نمودار

شابک: ۹۷۸-۶۰۰-۶۹۴۷-۷۷-۸

وضعیت فهرست نویسی: فیا.

موضوع: رپید ماینر (منبع الکترونیکی)، داده کاوی، نرم افزار

موضوع: RapidMiner (Electronic resource)-Data Mining- Software

شناسنامه افزوده: صدیقی منش، محمد، ۱۳۶۶-

رده بندی کنگره: QA ۷۶ / ۹

رده بندی دیویی: ۰۰۶ / ۳۱۲

شماره کتابشناسی ملی: ۶۰۴۶۶۰۴

انتشارات الهام نور

نام کتاب: علوم داده - تحلیل انجمنی، خوشه بندی، ارزیابی مدل

تألیف: محمد صدیقی منش، شروین سوری، علی صدیقی منش، فاطمه عباسی

تیراژ / چاپخانه: ۲۰۰ نسخه / شریعت

قطع / تعداد صفحات / صحافی: وزیری / ۱۲۸ صفحه / براتی

نوبت چاپ / قیمت: اول ۱۳۹۹ / ۵۰۰۰ تومان

شابک: ۹۷۸-۶۰۰-۶۹۴۷-۷۷-۸

کلیه حقوق محفوظ و مخصوص مؤلفین می باشد.

۶	فصل ۱- تحلیل وابستگی.....
۷	۱- ۱- مقدمه.....
۱۰	۱- ۲- کاوش قوانین وابستگی.....
۱۳	۱- ۲- ۱- مجموعه های آیتم.....
۱۸	۱- ۲- ۲- ایجاد قانون.....
۲۰	۱- ۳- الگوریتم استقرایی.....
۲۳	۱- ۳- ۱- چگونگی کار آن.....
۲۷	۱- ۴- الگوریتم رشد الگوی تکراری.....
۲۸	۱- ۴- ۱- چگونگی کار آن.....
۳۴	۱- ۴- ۲- چگونگی اجرا.....
۴۰	۱- ۵- نتیجه گیری.....
۴۱	فصل ۲- دسته بندی (خوشه بندی).....
۴۲	۲- ۱- مقدمه.....
۴۳	۲- ۲- خوشه بندی برای توصیف داده ها.....
۴۴	۲- ۳- گروه بندی برای پیش پردازش.....
۴۵	۲- ۴- انواع فنون خوشه بندی.....
۴۹	۲- ۵- سوابق بخش بندی مشتری.....
۵۱	۲- ۶- خوشه بندی K-MEANS.....

۵۳	۲-۶-۱- نحوه کارکرد
۶۰	۲-۶-۲- نمونه های خاص
۶۲	۲-۶-۳- ارزیابی خوشه ها
۶۳	۲-۶-۴- نحوه اجرا
۶۹	۲-۷-۷- خوشه بندی DBSCAN
۷۲	۲-۷-۱- این الگوریتم چگونه کار میکند
۷۴	۲-۷-۲- پارامترهای بهینه سازی
۷۵	۲-۷-۳- نمونه های خاص: تغییر تراکم ها
۷۷	۲-۷-۴- نحوه اجرا
۸۲	۲-۸-۱- نقشه های خود-سازمانده
۸۴	۲-۸-۱- چگونه کار میکند
۸۹	۲-۸-۲- نحوه کارکرد
۹۵	۲-۸-۳- مدل تصویری
۹۸	۲-۸-۴- مختصات موقعیت
۹۹	۲-۹- نتیجه گیری
۱۰۰	فصل ۳- ارزیابی مدل
۱۰۱	۳-۱- مقدمه
۱۰۳	۳-۲- ماتریس درهم ریختگی
۱۰۷	۳-۳- ROC و AUC
۱۱۲	۳-۴- lift curves

- ۳-۵- چگونگی اجرا..... ۱۱۶
- ۳-۶- نتیجه گیری..... ۱۲۴
- مقایسه الگوریتم‌های علم داده..... ۱۲۶

www.Ketab.ir

اهداف یادگیری این فصل

پس از خواندن این فصل باید بتوانید مفاهیم کلیدی زیر را درک کنید:

- تحلیل وابستگی چیست؟
- الگوریتم‌های تحلیل وابستگی

داستان وابستگی آجو و پوشک در یک چرخه تحلیلی به یک داستان قدیمی تبدیل شده است (پاور، ۲۰۰۲). انواع مختلف از این داستان وجود دارند اما مسأله بنیادین این است که یک شرکت فروشگاه زنجیره‌ای پی‌برد مشتریانی که پوشک خریداری می‌کنند، مایل به خرید آجو نیز هستند. رابطه آجو و پوشک وابستگی غیرمعمول، ناشناخته و عجیبی را نمایان کرد که می‌تواند از پردازش داده‌های تراکنش (معامله) یک سوپرمارکت به دست بیایند. سوپرمارکت چگونه تعیین کرد که یک چنین رابطه‌ای بین محصولات وجود دارد؟ پاسخ: علم داده (در دهه ۲۰۰۰ این علم داده‌کاوی نامیده می‌شد). به خصوص، تحلیل وابستگی.

تحلیل وابستگی احتمال رخداد همزمان بین یک آیتم و آیتم دیگر را می‌سنجد. هدف ایجاد این کلاس از الگوریتم‌های علم داده، پیش‌بینی رخداد یک آیتم نمی‌باشد، (همانند چیزی که الگوریتم‌های طبقه‌بندی یا رگرسیون انجام می‌دهند)، بلکه هدف آن یافتن الگوهای قابل استفاده در رخداد مشترک آیتم‌ها است. یادگیری قوانین وابستگی (انجمنی) شاخه‌ای از فرآیندهای یادگیری نظارتی است که الگوهای پنهان در داده‌ها را به شکل قوانین ساده و قابل شناسایی، کشف می‌کند.

الگوریتم‌های وابستگی به طور گسترده در آنالیز معاملات خرده‌فروشی، موتورهای توصیه‌گر (پیشنهاد دهنده) و آنالیز جریان کلیک آنلاین در صفحات وب و غیره مورد استفاده قرار گرفتند. یکی از کاربردهای پرطرفدار این تکنیک، تحلیل پورتفولیو یا سبد بازار نامیده شده است که رخداد مشترک یک آیتم خرده‌فروشی با آیتم دیگر در معامله خرید خرده‌فروشی مشابه را پیدا می‌کند (آگراوال و همکاران، ۱۹۹۳). اگر الگوها در داده‌های معامله به ما بگویند که شیرخشک

و پوشک معمولاً همراه با یکدیگر در خرید مشابه خریداری می‌شوند، یک خرده‌فروش می‌تواند از مزیت این وابستگی برای قیمت‌گذاری گروهی، چینش محصول و حتی بهینه‌سازی فضای قفسه در طرح فروشگاه سود ببرد. به طور مشابه، در یک محیط کسب و کار آنلاین، این اطلاعات می‌توانند برای فروش مکمل بلادرنگ، توصیه‌ها، پیشنهادات سبد خرید و راهکارهای بازاریابی پس از خرید مورد استفاده قرار گیرند. نتایج تحلیل وابستگی معمولاً شناخته شده هستند برای مثال یک برگر با سیب زمینی سرخ شده یا شیرخشک با پوشک؛ به هر حال، روابط غیرمعمول کشفیات با ارزش هستند که کسب و کارها می‌توانند از آن سود ببرند. حرکت نزولی این است که تحلیل وابستگی ممکن است همچنین روابط فریبده بین آیتم‌ها به دست دهد. در هنگام کار با داده‌های حاوی میلیاردها تراکنش، تراکنشهای با تمامی انواع احتمالات با ترکیبات عجیب و غریب مجموعه‌های آیتم (برای مثال بسته نیکوتین و سیگار) می‌توانند یافت شوند. این مستلزم مهارت تحلیلی و دانش کسب و کار برای استفاده موفقیت‌آمیز از نتایج یک تحلیل وابستگی است. نتیجه مدل یک تحلیل وابستگی می‌تواند به صورت یک مجموعه‌ای از قوانین همانند مورد زیر ارائه شود:

$$\{A \text{ آیتم}\} \leftarrow \{B \text{ آیتم}\}$$

این قانون نشان می‌دهد که براساس تاریخچه تمامی تراکنشها (معاملات)، وقتی که آیتم A در یک معامله یا سبد خرید یافت شود، احتمال زیادی برای رخداد آیتم B در همان معامله وجود دارد. آیتم A مقدم یا فرض قبلی قانون و آیتم B پیامد یا نتیجه قانون است. مقدم و پیامد قانون می‌توانند حاوی بیش از یک آیتم باشند همانند آیتم A و آیتم C. برای کاوش این نوع قوانین از داده‌ها، نیاز است تا معاملات خرید مشتری قبلی تحلیل شوند. در یک کسب و کار خرده‌فروشی،

میلیونها خرید در یک روز با هزاران واحد نگهداری موجودی انجام می‌شوند که برای یک ایتیم منحصر به فرد هستند.

فروش مکمل: مشتریانی که این محصول را خریدند ... را نیز خریدند

یک وبسایت تجارت الکترونیک را در نظر بگیرید که محصولات پرطرفدار را آنلاین به فروش می‌رساند. یکی از اهداف مدیریت یک کسب و کار الکترونیک، افزایش ارزش سفارش میانگین یک بازدید است. بهینه‌سازی اندازه سفارش حتی بیشتر حیاتی می‌باشد وقتی که کسب و کار برای اکتساب ترافیک از طریق بازاریابی موتور جستجو، تبلیغات آنلاین و بازاریابی رابطه‌مند هزینه پرداخت کرده باشد. کسب و کار برای افزایش ارزش سفارش میانگین به وسیله فروش مکمل و پیش فروشی محصولات مرتبط به مشتری براساس چیزی که آنها خریدند یا در حال حاضر در حال خرید در معامله کنونی هستند، تلاش می‌کند [یک معادل فست فودی رایج: «آیا سیب زمینی سرخ کرده با برگر می‌خواهید؟». نیاز است تا کسب و کار با سنجش سود پیشنهادی یک محصول مرتبط نهایی در برابر ریسک آزدگی یک مشتری دقیق باشد کسی که قبلا خریدی انجام داده است. در یک کسب و کار جایی که محصولات محدود (برای مثال صنعت فست فود) وجود دارند، فروش مکمل یک محصول با محصول دیگر ساده بوده و به طور کامل در این کسب و کار جا افتاده است. اما وقتی که تعداد محصولات منحصر به فرد به هزاران یا میلیونها محصول می‌رسد، تعیین یک مجموعه محصولات وابسته هنگامی که مشتریان به دنبال یک محصول هستند، کاملاً چالش برانگیز می‌باشد.

برای یادگیری بهتر درباره وابستگی محصول، درک داده‌های تاریخچه خرید مفید می‌باشد. اطلاعات درباره چگونگی اینکه یک محصول وابستگی به محصول دیگر ایجاد می‌کند، متکی بر

این حقیقت است که هر دو محصول همراه با یکدیگر خریداری شده‌اند، سپس می‌توان فرض کرد که ضرورت این محصولات به طور همزمان برای مشتری بروز پیدا می‌کند. اگر دو محصول همراه با یکدیگر به طور مکرر توسط تعداد زیادی از مشتریان خریداری شده باشند، سپس احتمال قوی برای یک الگوی وابستگی در این محصولات وجود دارد. در آخرین تراکنش جدید، اگر یک مشتری یکی از آن محصولات مرتبط را انتخاب نماید، سپس احتمال زیادی وجود دارد که محصول دیگر توسط این مشتری در همان خرید انتخاب گردد.

ورودی کلیدی برای آنالیز وابستگی لیستی از خریدهای قبلی با اطلاعات محصول است. براساس آنالیز این خرید (معاملات)، تکرارپذیرترین زوج محصولات می‌توانند تعیین شوند. یک آستانه نیاز خواهد بود تا برای عبارت «تکراری» تعریف شود زیرا که ظهور اندک یک زوج محصول بعنوان یک الگو احراز نمی‌شود. نتیجه یک آنالیز وابستگی یک مجموعه قانون است که می‌گوید: «اگر محصول A خریداری شده باشد، احتمال زیادی وجود دارد که محصول B در یک خرید مشابه خریداری شود». این مجموعه قانون می‌تواند برای ارائه توصیه‌های فروش مکمل روی صفحه محصول، محصول A استفاده شود. آنالیز وابستگی مفهوم در وای ابزارهای وب است که می‌گوید «مشتریانی که این را خریدند را نیز خریدند»

۱-۲- کاوش قوانین وابستگی

تحلیل وابستگی بنیادین تنها با رخداد یک ایتِم با ایتِم دیگر سر و کار دارد. تحلیل بیشتر پیشرفته می‌تواند تعداد رخداد، قیمت و توالی رخداد و غیره را در نظر بگیرد. روش برای یافتن قوانین وابستگی (انجمنی) از طریق توالی داده‌ها شامل مراحل ترتیبی زیر است:

مرحله ۱: آماده‌سازی داده‌ها به فرمت معاملاتی. یک الگوریتم وابستگی نیاز دارد تا داده‌های ورودی به فرمت معاملاتی $t_x = \{i_1, i_2, i_3\}$ فرمت شوند.

مرحله ۲: لیست کوتاهی از مجموعه‌های آیتم که به طور مکرر رخ می‌دهند. مجموعه‌های آیتم ترکیباتی از آیتم‌ها هستند. یک الگوریتم وابستگی تحلیل را به آیتم‌های با تکرار حداکثر محدود می‌کند به طوری که مجموعه قانون نهایی استخراج شده در مرحله بعد بیشتر معنی‌دار می‌باشد.

مرحله ۳: ایجاد قوانین انجمنی مرتبط از مجموعه‌های آیتم. در نهایت، الگوریتم قوانین را براساس معیار بهره ایجاد و فیلتر می‌کند.

برای شروع، یک وبسایت رسانه‌ای همانند BBC یا یاهو نیوز با طبقه‌هایی همانند اخبار، سیاست، مالی، سرگرمی، ورزش و هنر را در نظر بگیرید. یک session یا تراکش در این مثال یک بازدید برای وبسایت است جایی که کاربر مشترکی به محتوی از طبقه‌های مختلف در یک دوره session معین دسترسی پیدا می‌کند. یک session جدید معمولاً بعد از ۳۰ دقیقه عدم فعالیت آغاز می‌شود. session ها کاملاً مشابه با تراکش‌ها در یک مدل آجر و ملات سنتی هستند و صفحات دسترسی شده می‌توانند به محصولت خریداری شده مرتبط شوند. در سایتهای اخبار آنلاین، آیتم‌ها عبارتند از بازدیدها برای طبقاتی همانند اخبار، مالی، سرگرمی، ورزش و هنر. داده‌ها می‌توانند همانطور که در جدول ۱-۱ نشان داده شدند با لیستی از session ها و طبقات رسانه‌ای دسترسی شده در طول یک session معین جمع‌آوری شوند. هدف در این مسئولیت علم داده یافتن وابستگی‌ها بین طبقات رسانه‌ای است.