

Exercise #3

Data analytics in management

Mahsa Choopannezhad – student number: 98207477

نکته و توضیح اضافه شده:

استاد من واقعا معذرت میخوام ولی مثل اینکه ۵ام که اومدم تمرین را اپلود کنم درنهایت تمرین اپلود نشده بوده و من این دو هفته هم درگیر کرونا و بیمارستان شدم و اصلا ندیده بودم که اپلود نشده تو سایت 😞
الان که دارم دوباره تمرین را اپلود میکنم میدونم که مدت زمان زیادی از ددلاین گذشته ولی به هرحال گفتم فرستادنش بهتر از نفرستادنش هست. امیدوارم که سالم و سلامت باشید.

عنوان مقاله انتخاب شده:

Learning Preferences with Side Information

Published Online: 8 May 2019

<https://doi.org/10.1287/mnsc.2018.3092>

۱ - خلاصه ای از مقاله را بنویسید. (دقت کنید که این خلاصه باید به زبان شما باشد و از ترجمه مستقیم چکیده خودداری کنید)

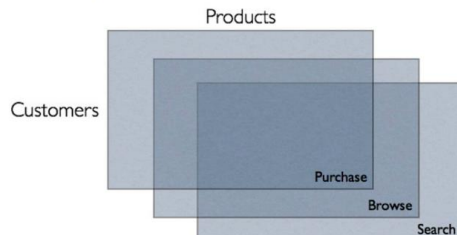
در قسمتی از مقاله اصلی معرفی شده برای تمرین (Data Analytics in Operations Management: A Review) در بخش ۵.۳ وقتی درباره متدهای مواجهه با دیتاهای کوچک صحبت میکند نسبت به این مقاله هم رفرنس میدهد و بیان میکند که: درسته که دسترسی به حجم زیادی از دیتا فراهم شده اما پارامترهایی که باید تخمین بزنیم و پیش بینی کنیم هم زیاد هستند. برای مثال یک خرده فروشی آنلاین را در نظر بگیریم که یک ماتریس متشکل از کاربران (در هر سطر) و محصولات (در هر ستون) را تشکیل میدهد و قصد دارد که خرید هر محصول توسط هر کاربر را پیش بینی کند و درایه های این ماتریس را پر کند و بعد بر اساس این ماتریس پیش بینی شده سیستم پیشنهاد دهنده خرید خود را بهینه کند. در حل کردن این مسئله در مقیاس بزرگی مثل آمازون که ابعاد این ماتریس حدود ۱۰۰ میلیون محصول و ۱۰۰ میلیون کاربر است و تعداد کمی تراکنش به ازای هر کاربر موجود است، چالش های جدی به وجود میاد. این مقاله برای حل این مشکل استفاده از اطلاعات جانبی را پیشنهاد میکند، برای مثال در نظر گرفتن اطلاعات دیگه ای مثل مشاهده کاربر از صفحات محصولات، اضافه کردن برخی از محصولات به کارت و بقیه اطلاعاتی که الزاما به خرید ختم نشده اند، اطلاعاتی هستند که در پیش

بینی ماتریس مورد نظر کمک کننده هستند. درواقع ایده اصلی مقاله این است که هر تعامل کاربر و سایت (از جمله همه این فعالیت هایی که تحت عنوان اطلاعات جانبی مقاله ازش یاد میکنند) یک ماتریس محصول- کاربر است و درواقع ماتریس اولیه که دنبال پیش بینی اون بودیم و ماتریس کاربران و محصولات خریداری شده بود یکی از این ماتریس ها است و درواقع هر ماتریس را میشود به عنوان بخشی از یک تنسور ۳ بعدی دید! (که خب خیلی زاویه نگاه جالبی هست!) مقاله در ادامه برای حل مشکل، یک الگوریتم نوآورانه ارائه میدهند تا بخش هایی از این تنسور داده را با استفاده از رگرسیون خطی و تجزیه مقادارهای منفرد^۱ (که درواقع تجزیه به عوامل و بردارهای متعامد فضااست به جورایی) بازسازی کنه.

۲- دستاورد اصلی مقاله چیست؟ از نظر شما کاری که انجام داده چرا مهم است؟

کار خیلی جالبی که مقاله انجام داده درواقع تو قدم اول نگاه متفاوت به مشکل پیش بینی ماتریس کاربر- محصول بوده که اومده در یک نگاه کلی تر در نظر گرفته که خرید به عنوان یک اکشن کاربر حساب میشه و اگر ما میخوایم این ماتریس را پیش بینی کنیم بیایم همه ماتریس های تعامل کاربر با محصولات را حساب کنیم و تو قدم بعدی برای بازسازی بخش های مختلف این تنسور تشکیل شده از دادهای نویزی استفاده کنیم. شکل ۱ از مقاله درواقع دایره نشون میدهند که چقدر نوع نگاه متفاوتی به قضیه کردند و درواقع ابعاد اینتراکشن های متفاوت یک کاربر را در نظر گرفتند و اینجوری به قضیه نگاه کردند که خب اگه مسئله را تو ابعاد وسیع تر و روی این تنسور حل کنیم درواقع مسئله کوچک ترمون را هم تو هریک از این ابعاد (که مسئله اولیه ماتریس خرید محصولات بود) حل کردیم.

Figure 1. (Color online) Customer-Product Interaction Matrices Represented as Slices of a Tensor



۳- قبل از این مقاله موضوع مورد بحث تا کجا پیش رفته بود؟ (این قسمت باید خیلی خالصه و در حد چند جمله باشد)

بدنه اصلی که تو ادبیات حوزه اش وجود داشته این هست که طراحی سیستم های پیشنهاد دهنده تو مقاله های قبلی همواره از دیتاسورس های مختلفی استفاده میکردند یا تو بعضی از کارها از داده های تبلیغات متفاوتی استفاده میکردند تا در نهایت ریکامندر با عملکرد بهتری را بهش دست پیدا کنند. نکته ای که مقاله را از بدنه ادبیات جدا میکنه این هست که چون ماتریسی دایره مسئله را جنرال نگاه میکنه، میگه درواقع میشه یه فریم ورکی ارائه داد که با همه منابع دیتا و تعاملات کاربر که دایره لاگ میندازه کار کرد و از اونها برای ریکامندر با این الگوریتم بازسازی بخش های تنسور استفاده کرد. یه بدنه ای از ادبیات هم به ادبیات روش های بازسازی تنسورها تخصیص داده شده.

¹ singular value decomposition

۴- رویکرد حل یا مدلسازی را به طور خلاصه توضیح دهید. همچنین اشاره کنید که نقش داده در این مساله چیست؟

مقاله با در نظر گرفتن مشکل پیش بینی خرید کاربران، مدلسازی را شروع میکند و بعد از تعریف و توضیح مفاهیم ماتریس و تنسور موردنظر، حدود خطا و گارانتی بازسازی را هم بحث میکند و بعد در قسمت آخر برای ارزیابی الگوریتم از دو روش استفاده میکند:

۱- از تنسورهای رندوم تولید شده استفاده میکند تا ارزیابی کند الگوریتم چقدر توانایی بازسازی تنسور را داشته است.

۲- از دیتاست واقعی موزیک یک سایت پخش کننده موزیک استفاده میکند.

۵- کاربردهای این مقاله چیست؟

هدف مقاله این است که از داده موجود استفاده کند و ویژگی های متفاوت رفتار کاربر را پیش بینی کند تا ورودی مناسبی را در اختیار تیم های عملیاتی و مارکتینگ قرار دهد و خوب به عنوان بارزترین کاربرد هم استفاده از این مدل برای بهینه کردن مدل های ریکامندر هست که تو صنعت خرده فروشی آنلاین کاربرد بسیار دارد. مثال دیگه استفاده از این مدل برای تولید پلی لیست های سایت های موزیک آنلاین مثل اسپاتیفای است.

۶- ساختار مقاله مورد نظر به چه صورت است؟ در این قسمت مسیری که مقاله طی کرده را بیان کنید به عنوان مثال :

الگوریتمی را ارایه داده، درجه پیچیدگی آن را بدست آورده، بعد آن را به کمک یک مثال عددی با دیگر الگوریتمهای

مطرح مقایسه کرده است)

به طور خلاصه مشکلی که به طور خاص مدلسازی با آن آغاز میشود این هست که ترجیحات مشتریان را روی محصولات با استفاده از داده های تراکنش خرید، پیش بینی کنیم. درواقع بعد از تعریف ماتریس و تنسور مورد نظر، هدف این است که تنسور سه بعدی مورد نظر را از مشاهدات نویزی که تو دیتا داریم، بازسازی کنیم.

۱- ماتریسها و بخش های تنسور را معرفی و تعریف میکند

۲- حدپایین خطا برای الگوریتم های متفاوت بازسازی مطرح میکند (درواقع با بررسی حدود خطا این موضوع را هم مطرح میکند که الگوریتم بیان شده بهینه ترین نیست ولی ممکن ترین راه حل برای مشکل است.)

۳- بیان الگوریتم

۴- بیان گارانتی برای بازسازی حتمی تنسور

۵- محاسبه پیچیدگی الگوریتم

۶- بحث درباره نویزهای متعادل و حل چالش هایی که با این فرض متعادل بودن نویز داده میشه داشت

۷- طراحی آزمایش برای ارزیابی الگوریتم (با استفاده از تنسورهای رندوم تولید شده و دیتاست واقعی موزیک و مقایسه ROC الگوریتم با دیگر الگوریتم های (Navie, matrix)

۷- آیا از دیتاستی استفاده کرده که قابل دانلود باشد؟ در این صورت دیتاست مورد نظر را بفرستید و نمره اضافه دریافت

کنید.

برای نشان دادن بهبود از دیتاست سایت Xiami.com استفاده شده است که یکی از سرویس های ارائه موزیک آنلاین است و سطح تعامل کاربران با سایت بالاست. دیتاستش الان مثل اینکه در دسترس نیست.

۸- چه مفاهیمی در این مقاله مطرح شده که دانش شما برای فهم آنها کافی نیست؟ (به حداقل سه مورد اشاره کنید).

- اون قسمت هایی که با یه سری قضیه آماری داره حد پایین خطا را حساب میکنه
- اون قسمتی که داره با فرض های Balanced Noise کار میکنه
- اون قسمتی که یه مقایسه ای با الگوریتم های convex انجام میده

۹- به نظر شما چه کاری میتوان در ادامه ی این مقاله انجام داد؟

یه قسمتی از نتیجه گیری هم به این موضوع اشاره کرده ولی درواقع موضوع مهمی که وجود داره تو حل این مشکل این هست که رفتار کاربر در طول زمان عوض میشه و درواقع زمان بعد ویژه و مهمی از مسئله است حالا مقاله کلا داره به بازسازی تنسورهای ۳ بعدی میپردازه درصورتی که اضافه کردن زمان تنسور را ۴ بعدی میکنه و شاید در ادامه بشه همین الگوریتم را توسعه داد و با یه سری فرض ها برای تنسورهای با ابعاد بالاتر هم تعمیم داد.