1. از مهدی متشکرم. یه کم دقیق تر بشیم روی موضوع. البته قبلش یخورده چیزهایی که لازمه از قبل بدونیم رو میگم که مقاله بهشون اشاره نکرده حالا به علت این که مثلا جا نداشته.
2. اسلاید شروع ۵. خب در مورد روش های تفسیرپذیری میدونیم که یه سری ها وابسته به روش خاص لرنینگ مثلا فلان الگوریتم دیپ هستند و بعضی مستقل از متد. این روش مستقل از متد یادگیریه. یعنی در واقع اگناستیک به مدل هست. اگناستیک در لغت معنی لاادری به نمیدانم هست. در دیکشنری ها اصولا فقط معنی مذهبی اش نوشته که یعنی کسی که میگه نمیدانم خدا هست یا نیست میشه. یه چیزی بین خدا باور و خداناباور.
3. بعد به جناب ارسطو میرسیم که طی یه نقل قولی میگن که کل از جمع جزء بیشتره! چرا؟
4. طبق تعریف تعامل وقتی فیچرها با یکدیگر تعامل دارند پیش بینی کل رو نمیشه گفت جمع تک تک تاثیرها چون تاثیر هر جزء به مقدار یک جز دیگه بستگی داره. جمله سنگین شد! یه مثال بزنم.
5. اگر به جدول نگاه کنید و مثلا فرض کنید که سایز ۱۰۰ هزار تا اثر دارد و مکان ۵۰ هزار تا. ضرایبش مهم نیست ها. تفاوت ها رو اگر دقت کنید میبینید که وقتی سایز خونه از کوچک به بزرگ تغییر میکنه مستقل از این که مکان خونه کجا باشه همواره ۱۰۰ هزار تا تغییر داریم و همچنین برای مکان همواره ۵۰ هزارتا تغییر داریم. دقت کنید کمی با کوریشیلن فرق داره.
6. حالا تو این اسلاید مشخص هست که تغییر سایز خونه مستقل از مکان نیست. یعنی اگر مکان خوب باشه تغییر سایز ۲۰۰ هزار تا تاثیر داره و اگر مکان بد باشه تغییر سایز ۵۰ هزار تا. حالا برای اندازه گیری اش هم یه روشی داریم به نام h-statistics به نام اقای فریدمن که خارج از موضوع مقاله است.
7. اسم الگوریتم ما ارکیپلاگو هستش که معنی یه سری جزیره میده که به نوعی ارتباط فیچرها با هم دیگه رو جزیره جزیره در نظر میگیره.
8. ما در مقاله مون دو تا ماژول مهم داریم که توضیحش میدیم. البته نه کاملا دقیق. چون خود مقاله ۳۰ صفحه اس و قطعا تو ۱۰ ۱۵ دقیقه نمیشه کل مدل رو دقیق توضیح داد ولی کلیت و ایده های اصلی مقاله رو مطرح میکنیم. اولی کامپونتت نسبت دادن تعامل به فیچرها هست و دومی پیدا کردن این تعامل ها است.
9. خب کامپوننت یا ماژول اولی یعنی مثلا ما یک مجموعه از فیچرها را انتخاب کردیم مثلا ۵ تا از ۱۰ تا فیچر را داریم و میخواهیم ببینیم تعامل بین آنها به چه صورت است؟! فرمولی که دارن خیلی ساده است که فقط یه تفاضل داره. ایکس استار معنی سمپل ما هست و ایکس بار به معنای بیس لاین یا اون خط جدا کننده ما که خنثی یا نوترال هست. در واقع داره تاثیر سمپل ما را از بیس لاینش جدا میکنه. بهش میگن جزایر جدا. که نامگذاری هم به همین خاطره.
10. خب مثال اگر بخوام بزنم مثلا عبارت نه خیلی بد با مبنای سه تایی. اگر بخواهیم خیلی و بد بصورت جدا در نظر بگیریم تعامل به چه صورته؟ مشخصه که نات ارتباط خیلی زیادی با اینها داره و طبق تعریف با تغییر نات تفاوت زیادی در پیش بینی ما صورت میگیره اما وقتی که از نات به سمت جای خالی بریم که میتونه یه چیز کلی باشه. میبنیم که این کار تغییرات زیادی در پیش بینی به عمل نمیآورد مشخصا در این حالت خیلی و بد ارتباط زیادی با اندرلاین پیدا نمیکنند که این مثال اهمیت بیس لاین رو مشخص میکنه که در واقع بیس لاین چیزی هست که تغییرات زیادی رو بوجود نمیاره.
11. خب حالا این که آیا این روش اول ما از اصول پیروی میکند یا نه خیلی مسئله مهمی است. در واقع یک سری ویژگی داریم که هر روشی که برای تفسیرپذیری میاد باید اینها رو اعمال کنه. به طور خلاصه جوابش اینه که بله این روش همه اصول رو رعایت میکنه و اثباتش به طور خیلی مفصل در ضمیمه مقاله هست. یک سری اصولی که وجود دارند به این شکل هستند.
12. اگر بخوام یک اصل رو مثلا توضیح بدم که کلیت قضیه دستتون بیاد. مثلا اصل کامل بودن میگه که جمع تمام نسبت دهی های ما برابر تفاضل پیش بینی بیس لاین با پیش بینی سمپل ما هست. این اصل میگه که هر فیچر چقدر در واقع در پیش بینی تاثیر گذار بوده.
13. کامپوننت دوم این الگوریتم یک الگوریتم تشخیص فیچرها هست که ماژول اول فیچرها رو ارزیابی میکرد ولی ماژول دوم فیچرها رو درمیاره که کدوم ها میتونن تعامل داشته باشند.
14. پس در واقع مسئله ما این هست که چگونه تعامل بین فیچرها در فضای فیچراسپیس رو پیدا کنیم؟ در حالت اولیه اش یه امیدریاضی رو تعریف میکنند که به نوعی میاد یه روش جستجوی گرافی ما رو هرس میکنه. در حالت های اولیه اش محاسبه امیدریاضی اش هم کار سختی نیست ولی وقتی تعداد فیچرها بالا میره در واقع امیدریاضی غیرقابل محاسبه میشه و در واقع اردرش خیلی بالا میره که همین باعث میشه که از فرمول پایین برای پیدا کردن اون مجموعه متغیرهایی که لازم هستند استفاده کنیم. این فرمولی هم که نوشته یه تخمینی از امیدریاضی ای هست که در بالا داشتیم.
15. متد دوم یک متد است که خود مقاله هم یه پاراگراف بیشتر توضیح نداده و گفته که به نوعی میاد و مجموعه ویژگی هایی که اشتراک دارند را با هم ترکیب میکند (یونیون) و بعد از اون این ها رو جوری جدا میکنه که اشتراک نداشته باشند و زیرمجموعه هایی که بدست میان همه اعضا با همدیگه تعامل داشته باشند.
16. در اخر هم کد گیت هاب مقاله برای پیاده سازی روش های مختلفش هست که شاید بد نباشه ببینید. کد رسمی مال خود شخصی هست که مقاله رو نوشته.
17. بقیه مقاله رو حسین بهمون توضیح میده که بهش واگذار میکنم.