Discovering Knowledge in Data_ An Introduction to Data Mining (2nd ed.)

دوره مقدماتی آموزش داده کاوی

- جلسه اول

۱. معرفی دوره و چگونگی تدریس و آموزش

۲. معرفی افراد

۳. شروع مقدمات با طرح موضوع مفهوم داده کاوی و مراحل داده کاوی

– **معرفی دوره** : شما توی دوره مقدماتی مثبت دیتا ثبت نام کردین، این دوره کلا آموزشش با خودم هست و در این دوره سعی میشه پیش نیازهایی که احتیاج دارید و یکسری مفاهیم پایه ای رو براتون بگیم. اول از همه اینکه شک نکنید این دوره بسیار دوره شیرینی خواهد بود و اصلا حوصلتون سر نمیره. هدف من در درجه اول یادگیری شماست و اینکه هزینه ای که میکنید در نهایت در آینده برای شما کسب هزینه کنه. دوم اینکه کسانی که بتونن توی این دوره و دوره پیشرفته جزو سه نفر اول دوره بشن، در پروژه های واقعی بکارگیری میشن و آموزش داده میشن. از همین اول بهتون بگم تلاشتون رو زیاد کنید و برای هزینه ای که کردین احترام قائل باشید. نفر اول این دوره که بعدا معیارهای نفر اول شدن رو هم براتون توضیح میدم به صورت رایگان توی دوره پیشرفته ما شرکت میکنه. در کنار آموزش مفاهیم سعی میشه نرم افزار **R** که یکی از قدرتمندترین نرم افزارها برای داده کاوی هست رو هم آموزش بدیم. هر هفته یه جلسه ۲ تا ۳ ساعته کارگاه **R** دارید. یه سوالی که از من میپرسن اینه که چرا پایتون کار نمیکنین به جای **R**؟ و جواب من اینه که اولا از نظر من **R** خیلی قدرتمنده و توی فیلد دیتا چیزی کمتر از پایتون نداره ولی دلیلی که من دارم برای انتخابش این امر هست که مصورسازی بسیار قوی داره و چیزی که الان به عنوان دانشجوهای پایه شما بهش احتیاج دارد همین دیدن اتفاقاتی هست که رخ میده، فهمیدن نمودارهاست و این خیلی نکته مهمی هست از نظر من. بچه ها تفکر من این نیست که شما رو مثه دانشگاه درس بدم، مفاهیم رو باید بفهمید، باید درک کنید چی به چیه، اینجور بدردرتون میخوره و بهتون قول میدم در آینده نه چندان دور برای همتون پول ساز خواهد بود.

– **معرفی افراد** : خب من خیلی حرف زدم الان نوبت شماست. لطفا خودتون، رشته تحصیلیتون رو بگید تا هم من باهاتون آشنا بشم و هم شماها باهم آشنا بشید.

–**داده کاوی چیست؟** داده کاوی فرایند کشف الگوها و روندهای مفید در مجموعه داده های بزرگ است. میدونین تفاوت اختراع و کشف چیه؟ اختراع یک حرکت نو آورانه است که قبلا نبوده و ابتکار یک نفر هست ولی کشف چیزی هست که بوده ولی به چشم کسی نیومده... یعنی چیزی که وجود داره ولی نمیبینیش... دقیقا داده کاوی هم همین کار رو میخواد بکنه. شما یکسری عدد دارید، نگاهشون که میکنید هیچ چیزی دستگیرتون نمیشه و اگر هم بشه چیز خاصی نیست... ولی وقتی وارد داده ها میشین اگر از متدهای درست

Commented [h1]: طراحی یک mind map که با جلو رفتن دوره کاملش میکنیم. این ایده ی خیلی خوبی میتونه باشه. هر روز به مفاهیم نگاه میکنیم و سعی میکنیم یادمون بیاد که چکار میکرد این مفهوم. تعریفش چی بود. چکار میخواستیم بکنیم باهاش و بقیه چیزا

Commented [h2]: اینکه بعد از هر کلاس یه کوئیز ازشون گرفته میشه که آیا مفاهیم رو فهمیدن یا نه گفته بشه یا شایدم گفته نشه و آخر جلسه بگیم که سورپرایز بشن!!! توی سوالات پرسشنامه اینکه قبلا با برنامه نویسی آشنا بودن یا نه آورده بشه

استفاده کنید به یه کشفیاتی میرسین که از چشم همه حتی کسایی مستقیما با اون داده ها سر و کار داشتن پنهون مونده.. پس با داده کاوی قراره چیزهای ببینیم که وجود دارن ولی از دید پنهون شدند.

مراحل داده کاوی: مثل همه فرآیندهای علمی داده کاوی هم دارای یه فرآیند هست که طبق اون اگر جلو بریم میتونیم بگیم که تونستیم به یه نتیجه درست برسیم. بذارید یه مثال ملموس براتون بزنم. من قبلا یعنی قبل دانشگاه تابستونا میرفتم نقاشی ساختمون کار میکردم. این بنده خدا اوستای ما همیشه یه حرف خوبی بهم میزد... میگفت یه نقطه رو شروع قرار بده و یه نقطه رو پایان ، از وسط کار شروع نکن، اینجور می فهمی چکار کردی و کجا رو رنگ کردی... خلاصه یعنی توی کار گم نشید. این فرآیند داده کاوی هم دقیقا همونه.. یعنی باید یاد بگیریم که از یک جا شروع کنیم و یک جا تموم کنیم. توی کار نیفتیم. حالا این فرآیند داده کاوی مراحل زیر رو داره:

٤. **فهم حوزه ی منبع داده ها**: این حوزه هدفش چیه، میخواد به کجا برسه، درآمدش از کجاس، بازار فروشش کجاست و کلا تمام روابط کاری این حوزه رو باید درک کنید.
٥. **فهم داده ها**: داده های ذخیره شده از چه نوعی هستند؟ نقاط ضعف و قوت داده هاشون چیه؟ حجم داده ها چقدره؟ چندتا مشخصه داره؟ کدوم مشخصه ها برای مدیران اون حوزه اهمیتش بیشتره؟ کدوم از دیتاها بد جمع شده و باید خیلی مورد بررسی قرارش داد و کلا کیفیت داده ها مورد بررسی قرار بگیره
٦. **پیش پردازش داده‌ها**: داده ها حتما و حتما دچار مشکلاتی هستند، مثل گمشده بودن داده ها، همبستگی داده ها با هم، بی نظم بودن در صورتی که از چند دیتابیس مختلف جمع شده باشند، هم مقیاس نبودن و ... برای رسیدن به یک پردازش خوب و نتیجه قابل اطمینان داده ها قبل از استفاده باید مورد بررسی دقیق قرار گیرند. این مرحله یکی از مهمترین مراحل داده کاوی محسوب میشه.
٧. **مدل کردن داده‌ها**: در این قسمت بسته به نوع دیتا و خروجی مد نظر خود یک مدل را برای داده ها انتخاب کرده و آن را برازش میدهیم. این نکته بسیار حائز اهمیت است که انتخاب متد درست منجر به خروجی اطلاعات و مفاهیم درست می گردد.
٨. **فاز ارزیابی** : در این قسمت بر روی داده ها مدل برازش داده شده و خروجی هم حاصل شده، حال باید این سوال را از خودمان بپرسیم که مدل برازش داده شده مدل درستی است؟ خطای حاصل از این مدل چقدر است؟ چه پیشنهادی برای کم شدن خطا داریم؟
٩. **فاز توسعه** : در این فاز باید این امر بررسی شود که آیا مدل برازش شده برای جامعه دیگر نیز قابل توسعه است؟ و یا فقط در چارچوب همین مسئله جواب میدهد؟

یکسری کج فهمی ها در خصوص داده کاوی هست که اینجا باید براتون توضیح بدم :

- اول اینکه داده کاوی یک فرآیند هست و به صورت اتوماتیک انجام نمیشه، یعنی یه مرحله رو باید انجام بدی ببینی چه خبره، اگر همه چیز اوکی بود بعد بری مرحله بعد، حتی بعضی اوقات باید از مرحله های جلوتر برگردی مرحله های عقب تر بعضی از نقوص رو برطرف کنی
- داده کاوی به نظارت انسان احتیاج دارد زیرا که هر مرحله باید مورد بررسی قرار بگیره وگرنه خروجی اشتباه هست.
- عدم تجزیه و تحلیل از تجزیه و تحلیل اشتباه بهتر است. اگر شما در یک سیستم بخواهید با استفاده از داده کاوی مشاوره بدهید به مدیریت دستگاه، اگر مشاوره ندهید بهتر است تا مشاوره اشتباه دهید.

دوره مقدماتی آموزش داده کاوی

- جلسه دوم (مفاهیم پایه پایتون)

Commented [I3]: اینجا رو میتونیم بسپاریم به هیراد تا یک کارگاه یک روزه برگزار کنه...

-در این قسمت قرار به توضیح کدهای پایتون نیست و تنها برخی مفاهیم پایه پایتون رو مورد بررسی قرار می دهیم، دوستانی که اصلا با پایتون آشنایی ندارند، میتونن از سایت W3school و یا سایت های مرتبط آموزش ببینند. اگر هم ظرفیتتون به ۲۰ نفر برسه، کارگاه یک روزه براتون برگزار میکنیم. ولی خب کدهایی که اینجا داریم باهاشون کار میکنیم، کدهای مربوط به کار با دیتا اونم درسطح ابتدایی هست.

- کامنت

کامنت قطعه ای از کد است که اجرا نمیشود. در نوشتن کامنت نباید زیاد وسواسی باشیم، از طرفی نداشتن کامنت در کدها، باعث سردرگمی شما در مراجعات بعدی خود به کدها میشه. از طرفی استفاده افراطی از کامنت در کدها، بیشتر باعث سردرگمی و بی نظمی کدها می شود.

```
# This is a comment
```

- وارد کردن پکیج ها در پایتون

در حالی که بسیاری از کارها را می‌توان در پایتون "خارج از جعبه" (out of the box - یعنی پس از نصب پایتون در اختیار هست و احتیاجی به نصب ماژول و یا کتابخانه ای نداریم) انجام داد، بیشتر کارهایی که می‌خواهیم انجام دهیم نیازمند وارد کردن بسته‌ها هستند. بسته ها حاوی کدهای طراحی شده ویژه ای هستند که ما را قادر می سازد تا کارهای پیچیده علم داده را بدون

نوشتن کد خود انجام دهیم. به عنوان مثال، در فصل ۶، باید یک مدل طبقه بندی و درخت رگرسیون (classification and regression tree) بسازیم. به جای اینکه بفهمیم چگونه یک مدل CART را از ابتدا بسازیم، بسته ای را وارد می کنیم که حاوی آن کد است. هنگامی که بسته وارد شد، می توانیم کد را برای ایجاد یک مدل CART اجرا کنیم.

دو مورد از مهمترین پکیج ها (ماژول و یا همان بسته) در پایتون که جز پرکاربردترین پکیج ها در علوم داده هستند، pandas و numpy هستند. برای وارد کردن این دو پکیج در پایتون باید دوخط زیر تایپ و اجرا کنیم:

```
import pandas as pd
import numpy as np
```

توجه داشته باشید که ما بسته ها را با استفاده از دستور import وارد می کنیم. کد as با استفاده از نامی که ما می توانیم تعیین کنیم، بسته را تغییر نام می دهد. ما بسته ها را تغییر نام می دهیم تا کار با آنها آسان تر شود.

برای استفاده از دستورات موجود در بسته‌های panda و numpy، باید نام بسته‌ها را قبل از نام فرمان بیان کنیم. به عنوان مثال برای استفاده از دستور read_csv()، باید

```
pandas.read_csv()
```

را تایپ کنیم. همچنین می‌توانیم یک نام مستعار به بسته بدهیم. در مورد بالا، بسته pandas را با استفاده از as pd به pd تغییر می دهیم و با استفاده از as np بسته numpy را به np تغییر می دهیم. پس دستور بالا به شکل زیر بازنویسی می‌گردد:

```
pd.read_csv()
```

همچنین ما میتوانیم قسمتی از یک پکیج را در پایتون بازخوانی کنیم:

```
from  "module name" import Func1,Func2,…
```

برای مثال از ماژول sklearn.tree توابع DecisionTreeClassifier و export_graphviz را فراخوانی میکنیم:

```
from sklearn.tree import DecisionTreeClassifier, export_
graphviz
```

- وارد کردن داده ها در پایتون

اکنون در مورد نحوه وارد کردن یک مجموعه داده به پایتون بحث خواهیم کرد. در این متن از دستور read_csv() با استفاده از ساختار زیر استفاده می کنیم:

```
your_name_for_the_data_set = pd.read_csv ("the_path_to_
the_file")
```

برای مثال جهت خواندن داده ها در فایل (bank_marketing)از کد زیر استفاده میکنیم:

```
bank_train=pd.read_csv("C:/Users/Data_Science/Data/bank_market
ing")
```

- دسترسی به رکوردها و متغیرها در پایتون

در بررسی های متفاوت ممکن است بخواهید یک رکورد خاص را بررسی کنید. به عنوان مثال، چگونه به یک رکورد در مجموعه داده `bank_train` دسترسی پیدا کنیم؟ ما از متد `loc` که در بسته Pandas فراهم شده برای انجام این کار استفاده خواهیم کرد.

دوره مقدماتی آموزش داده کاوی

- جلسه سوم (پیش پردازش دیتا)

۱. چرا به پیش پردازش دیتا احتیاج داریم؟

۲. مدیریت داده های از دست رفته

۱- چرا به پیش پردازش دیتا احتیاج داریم

بسیاری از داده های خام موجود در پایگاه های داده ، پردازش نشده ، ناقص و نویزی هستند، به عنوان مثال ، پایگاه های داده ممکن است حاوی :

- اطلاعات منسوخ شده باشد (برای مثال فرض کنید یک خط تولید از یک کارخانه کاملا از روند تولید خارج شده باشد در این صورت، دیتای مربوط به آن منسوخ شده است)

- مقادیر از دست رفته (**Missing value**)
- داده های پرت (**Outlier**)
- داده های خارج از فرم و نامناسب برای ورود به مدل (داده هایی که باید با استفاده از عملیات هایی آن ها را تبدیل به داده های خوش فرم کنیم)
- داده هایی که با عقلانی نیستند.

اغلب داده هایی که با آن ها سر و کار داریم با این مشکلات مواجه هستند، برای اینکه درست برنامه ریزی نشده اند و اپراتور (چه ماشین و چه انسان) دارای خطای ثبت اطلاعات است. حال چرا باید پیش پردازش داده ها انجام شود؟ جواب یک اصطلاح است : **GIGO (garbage in garbage out)**

Commented [h4]: شاید بشه اینجا یه شکلی چیزی بیاریم برای بیان بهتر مفهوم

اگر به مدل خود آشغال وارد کنید، حتما مدل به شما آشغال تحویل میدهد. پس باید دیتا به صورت تمیز و پیش پردازش شده به مدل داده شود، تا به خروجی درست و قابل اطمینان دست پیدا کنیم. باید حواسمان باشد که گاهی اوقات تا ۶۰ درصد فرآیند داده کاوی در همین مرحله خلاصه میشود و این مرحله مهمترین مرحله داده کاوی است.

برای مثال به جدول زیر که یک قسمت از دیتابیس یک فروشگاه اینترنتی هست دقت کنید تا مشکلاتش رو با هم بررسی کنیم:

Customer ID	Zip	Gender	Income	Age	Marital Status	Transaction Amount
1001	10048	M	78,000	C	M	5000
1002	J2S7K7	F	-40,000	40	W	4000
1003	90210	….	10,000,000	45	S	7000
1004	6269	M	50,000	0	S	1000
1005	55101	F	99,999	30	D	3000

- متغیر **Customer id** کلید اصلی جدول و نشان دهنده شماره مشتری است که مشکل خاصی از نظر پیش پردازش دیتا در آن وجود ندارد
- متغیر **Zip** متغیر کد پستی مشتریان است، در این جا فرم های متفاوت برای این متغیر میتواند مشکل ساز باشد، برای مثال مشتری ۱۰۰۲ از کانادا و مشتری ۱۰۰۴ از انگلیس هستند و بقیه از آمریکا، اگر پایه ریزی این متغیر براساس کدهای ۵ رقمی آمریکا باشد، آنگاه این متغیر بد فرم است.
- متغیر جنسیت یا **Gender** دارای مقدار گمشده برای مشتری شماره ۱۰۰۳ است.

- متغیر درآمد از چند بعد دچار ایراده، اول اینکه مشتری شماره ۱۰۰۲ درآمد منفی براش ثبت شده که این میتونه خطای برنامه نویسی پایگاه داده باشه که اجازه ثبت مقدار منفی در این فیلد رو میده، دوم اینکه داشتن درآمد ۱۰ میلیون دلار در سال برای فردی (مشتری ۱۰۰۳)که در منطقه هیلس هست مقداری دور از انتظاره، سوم اینکه داشتن درآمد ۹۹۹۹۹ برای مشتری ۱۰۰۵ غیرعادیه، برای اینکه مگه میشه کسی دقیقا این درآمد رو داشته باشه و از طرفی این عدد برخی اوقات شیطنت مدیران پایگاه داده برای خالی نموندن فیلدها هم به حساب میاد. مشکل آخر و مهمترین مشکل اینکه درآمد افراد در این فیلد بر مبنای یک واحد پولی ثابت نیست، مشتری ۱۰۰۲ بر اساس دلار کانادا، مشتری ۱۰۰۴ براساس پوند انگلیس و بقیه مشتری ها براساس دلار آمریکا در آمد خود رو اعلام کردند.
- متغیر **Age** و یا همان سن مشتری دارای دو مشکل است، اول اینکه سن افراد به دو صورت گروه بندی شده و عددی وارد شده است و دیگر اینکه عدد غیرممکن صفر برای مشتری شماره ۱۰۰۴ ثبت شده است.
- متغیر **Marital status** و یا وضیعت تاهل هم دچار مشکل کدگذاری استو برای مثال ما نمیدانم کد **S** نشان دهنده مجرد **(single)** و یا جدا شده **(separated)** است.
- متغیر میزان تراکنش یا **Transaction amount** با شرط اینکه مقدار این تراکنش ها با یک واحد پولی باشد، مشکلی ندارد.

۲-مدیریت داده های از دست رفته

اول از همه باید گفت که هر چه اطلاعات بیشتر باشد، با فرض انتخاب روش درست تحلیل، خروجی تحلیل قابل اعتمادتر می باشد. مشکل داده های گمشده تقریبا در همه ی پایگاه های داده ای هست و برای حل مشکل اون باید کاری بکنیم.