Tu _C

به نام خدا

آموزشکده ملی مهارت میناب

تهدیدات بخش چهارم

نام و نام خانوادگی : فردین صداقت

واحد درسی : مباحث ویژه

رشته : مهندسی کامپیوتر

مدرس : مهندس محمد احمدزاده

فروردین ۱۴۰۴

clips™

# بخش چهارم Data Preprocessing

## A- چرا Data cleaning در علم داده اهمیت دارد؟

تمیزسازی داده ها نه تنها برای تجزیه و تحلیل داده ها مهم است، همچنین برای مدیریت عمومی کسب و کار یا حاکمیت داده نیز مهم می باشد. منابع کلان داده های پر و پر برده و دائماً در حال تغییر هستند. بنابراین نگهداری منظم پایگاه های داده به شما کمک تا همه چیز را اضافه کنید. اگر کیفیت داده های شما از سطح پایین تری برخوردار باشد، نتیجه هرگونه تجزیه و تحلیل با این داده ها نیز ناقص خواهد بود. حتی اگر در هر مرحله دیگر از روند تجزیه و تحلیل داده ها به درستی دنبال کنید، اگر داده های شما نامرتبط باشد تفاوتی ایجاد نمی کند.

## B- Missing Values چگونه مدیریت می شوند؟

### حذف اشیاء یا ویژگی های داده؟

اگر در یک مجموعه داده تنها چند نمونه دارای مقادیر مفقود شده باشد، حذف آنها ممکن است منید باشد. یک راه مرتبط این است که ویژگی های دارای مقادیر مفقود شده مستعد حذف باشیم.

### برآورد مقادیر مفقود شده:

یک مجموعه داده را در نظر بگیرید که دارای نقاط داده بسیار مشابه هستند. در این وضعیت، اغلب از مقادیر همسایگی نقاط نزدیک نه نقطه با مقدار مفقود شده، برای برآورد مقدار از دست رفته استفاده می شود.

### نادیده گرفتن مقادیر گم شده در هنگام تجزیه و تحلیل:

اگر یک یا هر دو نمونه دارای مقادیر مفقود شده برای برخی از ویژگی ها باشند، می توان شباهت را فقط با استفاده از ویژگی هایی که مقادیر مفقود شده ندارند محاسبه کرد.

## C- outlier چیست و چگونه می توانیم آنها را تشخیص دهیم؟

برخی اوقات در مجموعه داده ها، داده هایی ثبت شده اند که به طور چشمگیری با سایر موارد مغایرت هستند. آنها خود داده یک یا چند ویژگی متمایز می کنند و می توانند باعث ایجاد نامتجاری در نتایج بدست آمده از طریق الگوریتم ها و سیستم های تحلیل شوند. در مدل های با دقت داده های دور افتاده می توانند مزایای آموزش را تحریب کنند که این امر می تواند موجب به طولانی شدن زمان آموزش و یا منجر به ایجاد مدل نهایی با دقت کم شود.

تشخیص outliers:

- مرتب سازی
- با استفاده از نمودار
- با استفاده از نمره Z

## D- Data Transformation چه کاربردی دارد؟

پس از استخراج اطلاعات مورد نیاز از مکان های مورد نظر URL داده ها در قالب بدون ساختار خواهند بود. سپس می توان آنرا به صورت ساختار یافته مانند CSV، صفحه گسترده یا Pdf برای ارائه یا ذخیره سازی تبدیل کرد.

## E- one hot encoding و Label encoding چه تفاوتی با هم دارند؟

one hot encoding روشی برای تبدیل متغیر های دسته ای به فرمت دودویی (binary) است. این تکنیک ستون های جدیدی ایجاد می کند که حاوی مقادیر (صفر و یک) برای دسته موجود متغیر اصلی است. هر دسته از ستون اصلی به یک صورت ستون جداگانه نمایش داده می شود که مقدار ۱ حضور آن در دسته و مقدار ۰ نبود آنرا نشان می دهد.

زمانی که متغیر دسته‌ای دارای ترتیب طبیعی (مانند "پایین"، "متوسط" و "بالا") باشد بهتر است از Label encoding استفاده کنید. این روش به هر دسته یک عدد صحیح منحصربفرد اختصاص می‌دهد.

F- چرا Feature Selection در Model-building اهمیت دارد؟

روش های انتخاب ویژگی (Feature selection models) به منظور مواجه با داده های ابعاد بالا، به ملوله‌ای جدایی ناپذیر از فرایند یادگیری مبدل شده‌اند. یک انتخاب ویژگی صحیح می‌تواند منجر به بهبود یادگیرنده استقرایی از جمله گوناگون ازجمله سرعت یادگیری، ظرفیت تعمیم و سادگی مدل استنتاج شده می‌شود.

در مباحث علمی یک متخصص داده بایستی خود ویژگی های مورد نیاز را از میان دادگان استخراج کند. حتی در برخی موارد بایستی به دنبال دیتاست جدید بگردد و داده ها را جمع آوری کند.

G- Duplicate Data چگونه در پایگاه داده حذف می شود؟

برای حذف داده های تکراری از پایگاه داده ابتدا باید مشخص بکنیم که معیار تکراری بودن چیست (مثلاً تکرار در بستن ایمیل یا شماره تلفن) سپس با استفاده از دستورات SQL مانند Group by یا ROW-NUMBER() داده های تکراری شناسایی می شوند. بعد از شناسایی معمولاً فقط یکی از رکوردها (مثلا اولین مورد) نگه داشته و بقیه حذف می شوند. این کار ممکن است با یک دستور Delete همراه با شرط انجام شود.

H- Irrelevant Data چه مشکلاتی را در پیش بینی های Machine Learning ایجاد می کند؟

داده های بی ربط یا غیرضروری آنهایی هستند که در واقع مورد نیاز نیستند و در چارچوب مسأله‌ای که سعی داریم آنرا حل کنیم، مناسب نیستند. برای مثال اگر ما داده های مربوط به سلامت عمومی

مردم را تجزیه تحلیل می کنیم، شاید ذهن داده بی ربط است. به عنوان مثال دیگر، اگر می خواهید داده های مربوط به نعل تازه را تجزیه و تحلیل کنید، اما مجموعه داده شما قدیمی تر است، باید آن مشاهدات بی ربط را حفظ بکنید. این امر می تواند تجزیه و تحلیل را کار آمدتر بکند و سرگرمی را از هدف اصلی شما به حواس پرتی برساند.

۱- چرا Data Imputation برای پر کردن missing Values کاربرد دارد؟

1- نامناسب گاهی با کتابخانه های یادگیری ماشین در زبان برنامه نویسی پایتون؛

2- ایجاد اختلال در مجموعه داده

3- تاثیر گذاری بر مدل نهایی

4 تمایل به بازیابی تمامی نمونه های مجموعه داده

ل- چگونه می توانیم Normality را در داده های عددی بررسی بکنیم؟

برای بررسی نرمال بودن داده های عددی می توان از دو روش کلی استفاده کرد:

روش های مثل هیستوگرام و Q-Q plot که شکل توزیع داده را نشان می دهند اگر شبیه منحنی زنگوله ای باشد یا نقاط Q-Q روی خط باشد، داده نرمال نیست.

روش های آماری مانند آزمون Shapiro-Wilk، Kolmogorov-Smirnov و Anderson-Darling که با بررسی P-Value نتیجه می دهند اگر $P > 0.05$، داده نرمال فرض می شود.

ترکیب این روش ها ارزیابی دقیق تری از نرمال بودن ارائه می دهد.

clips™