Subject :

Year :   Month :   Date :

به نام خدا

تمرین ۴: چرا Data cleaning در علم داده اهمیت دارد!

درس: مباحث ویژه

استاد: احمدزاده

اعضای گروه: سمیرا صالحی - سمانه بهاری

تمرین بهمن : ۱۴۰۲

رشته: کامپیوتر

۱۴۰۲/۱۱/۲۰

(بخش ۴)

SUBJECT:

Year: Month: Day:

Page: ( )

۱- چرا Datacleaning در علم داده اهمیت دارد؟

Datacleaning یا پاکسازی داده ها یکی از مراحل کلیدی در علم داده است که دلیل زیر اهمیت زیادی دارد. افزایش دقت مدل های نادرست ناقص یا سازماری می توانند باعث ایجاد نتایج نامعتبر و می شود ۲- کاهش هزینه های داده های نامناسب می توانند هزینه های اضافی را در مراحل بعدی فرآیند تحلیل و مدل سازی ایجاد کنند با پاکسازی داده ها در مراحل اولیه این هزینه ها کاهش می یابد.

۲- Missingvalues چگونه مدیریت می شوند؟

مدیریت مقادیر گمشده (Missingvalues) یکی از مراحل مهم در پاکسازی داده قبل از تحلیل است. روش های مختلفی برای مدیریت این مقادیر موجود دارد که بسته به نوع داده و نیاز تحلیل می توان آن ها استفاده کرد در ادامه به چند روش متداول برای مدیریت مقادیر گمشده اشاره می شود. حذف داده ها اگر درصد کمی از داده های گمشده باشند. می توان ردیف های کامل مجموعه های داده را حذف کرد حذف ستون ها اگر یک ستون دارای مقدار گمشده زیاد باشد و اطلاعات مهمی ارائه دهد می توان آن ستون را حذف کرد.

Senobar

(بخش ۴)

Page: 1 / 1

Year: Month: Day:

۳- outliers است و چگونه می‌توانید آن‌ها را شخص دهید؟

outliers داده‌های پرت مقادیری هستند که به طور قابل توجهی با سایر داده‌ها در یک مجموعه متفاوت هستند این مقادیر می‌توانند ناشی از خطاهای اندازه‌گیری ورود اشتباه داده‌ها در تحلیل داده‌ها باشد و تأثیر زیادی بر نتایج مدل‌های تحلیل داشته باشد.

۴- DataTransformer چه کاربردی دارد؟

DataTransformation یکی از مراحل کلیدی در فرآیندهای داده پردازی و تحلیل داده است این فرآیند به دلایل مختلفی کاربرد دارد که برخی از آن‌ها اشاره می‌کنم بعدش

و کیفیت داده‌های خام معمولاً شامل خطا ناهماهنگی و اطلاعات ناقص هستند با استفاده از تکنیک‌های تبدیل داده می‌توان داده‌ها را تصحیح پاک‌سازی و یکپارچه کرد تا کیفیت آن‌ها بهبود یابد.

۵- one-Hot Encoding و Encoding-Label چه تفاوتی دارند؟

تفاوت‌های Label و one-Hot در یادگیری ماشین برای تبدیل ویژگی‌های کلامی (Categorical Features) به فرمت عددی در نظر گرفته می‌شود دو تکنیک متداول

در این زمینه LabelEncoder هستند در ادامه به تفاوت‌ها و ویژگی‌های هر یک می‌پردازیم

Senobar

۶- چرا Model building, feature selection اهمیت دارد؟

Feature selection یا انتخاب ویژگی یکی از مراحل کلیدی در فرآیند ساخت مدل های یادگیری ماشین است. اهمیت آن به دلایل مختلفی برمی گردد. بهبود دقت مدل: با انتخاب ویژگی های مهم و مفید می توان دقت مدل را افزایش داد و ویژگی های غیرضروری یا زائد می توانند باعث کاهش کیفیت پیش بینی ها شوند.

۷- Duplicate Data چگونه در پایگاه داده ها حذف می شود؟

حذف Duplicate Data (داده های تکراری) از پایگاه داده یکی از وظایف مهم در مدیریت داده ها و پردازش داده های بزرگ است به دلیل بازخوانی وارد کردن نادرست یا داده های نادرست داده ها داده های تکراری می توانند به وجود بیایند. در زیر به مراحل و تکنیک های متداول برای شناسایی و حذف داده های تکراری در پایگاه داده ها اشاره می شود.

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

وجود Irrelevant Data (داده های بی ربط) در داده های آموزشی مدل های یادگیری ماشین می تواند مشکلات و چالش های زیادی را در پیش بینی ها ایجاد کند. در زیر به برخی از مشکلات اصلی ناشی از وجود داده های بی ربط اشاره می شود.

۱۴۰۲/۱۱/۲۰

بخش (۲)

Year: Month: Day: Page: ( )

۹- چرا DataImputation برای برگردن Missing Values کاربرد دارد؟

DataImputation با برگردن مقادیر گمشده Missing Values به عنوان یک

تکنیک در علم داده و یادگیری ماشین کاربرد دارد دلایل آن به شرح زیر است. افزایش

دقت مدل - مقادیر گمشده می توانند باعث کاهش کیفیت و دقت مدل های پیش بینی شوند با

برگردن این مقادیر می توان عملکرد مدل را بهبود بخشید و اطمینان حاصل کرد که تمام داده

ها در فرآیند آموزش و ارزیابی در نظر گرفته می شوند.

۱۰- چگونه می توانید Normality را در داده های عددی بررسی کنید؟

بررسی نرمال بودن (Normality) داده های عددی یکی از مراحل مهم تحلیل داده ها

است به ویژه زمانی که قصد داریم از آزمون های آماری مبتنی بر فرض نرمال بودن

استفاده کنیم. برای بررسی نرمال بودن داده های عددی می توانیم از روش ها و آزمون

های زیر استفاده کنیم - توزیع هیستوگرام - یک هیستوگرام از داده ها رسم کنید و به

شکل توزیع آن توجه کنید اگر داده ها به طور تقریبی به شکل زنگوله ای Normal Distrib

قرار بگیرند احتمالاً نرمال هستند. نتیجه گیری ترکیب چند روش می تواند به شما دید بهتری

درباره نرمال بودن داده های تان بدهد با این حال همیشه باید همگی حجم نمونه و نوع داده های

مورد توجه داشته باشید و براساس آن نتیجه گیری کنید

Senobar

۴- outliers چیست و چگونه می‌توانیم آن‌ها را تشخیص دهیم؟

outlier ها یا داده های پرت هستند برخی هستند که به طور قابل توجهی با سایر داده ها در یک مجموعه متفاوت هستند این ها برخی می توانند ناشی از خطاهای اندازه گیری ورود داده یا اشتباهات

outlier ها در تحلیل داده ها بسیار مهم است زیرا می توانند تاثیر زیادی بر نتایج مدل ها و تحلیل ها داشته باشند

۵- Data transformation چه کاربردی دارد؟

Data Transformation یکی از مراحل کلیدی در فرآیندهای داده پردازی و تحلیل داده است این فرآیند به دلایل مختلفی کاربرد دارد که به برخی از آن ها اشاره می کنیم

۱- بهبود کیفیت داده های خام معمولاً شامل خطا ناهماهنگی و اطلاعات ناقص هستند با استفاده از تکنیک های تغییر داده می توان داده ها را تصحیح یا یکسان سازی و یکپارچه کرد تا کیفیت آن ها بهبود یابد

۶- Label Encoding و One-Hot Encoding چه تفاوتی دارند؟

تفاوت های Label و One-Hot در یادگیری ماشین برای تبدیل ویژگی های کلامی (categorical features) به فرمت عددی در نظر گرفته می شود دو تکنیک متداول در این زمینه Label Encoding هستند در ادامه به تفاوت ها و ویژگی های هر یک می پردازیم