Subject :
Year : Month : Date :

به نام خدا

تمرین ۴: چرا Data ethics در علم داده اهمیت دارد؟

درس: مباحث ویژه

استاد: احمدزاده

اعضای گروه: میرا صالحی - سمانه نصاری

تمرین پس: ۱۴۰۳

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 تواند باعث ایجاد نتایج بهبود می یابد ۲- کاهش هزینه های داده های نامناسب می تواند هزینه های اضافی را در مراحل بعدی فرآیند تحلیل و مدل سازی ایجاد کند 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 هاست 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.

Sonobar

(بخش ۴)

Page: Year: Month: Day:

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

outliers داده‌های پرت مقادیری هستند که به‌طور قابل توجهی با سایر داده‌ها در یک مجموعه متفاوت هستند. این مقادیر می‌توانند ناشی از خطاهای اندازه‌گیری، ورود اشتباه داده‌ها یا تغییرات واقعی در فرایند باشند. outliers در تحلیل داده‌ها بسیار مهم است زیرا می‌توانند تأثیر زیادی بر نتایج مدل‌های تحلیلی داشته باشند.

۴- DataTransformation چه کاربردی دارد؟

DataTransformation یکی از مراحل کلیدی در فرآیندهای داده‌پردازی و تحلیل داده است. این فرآیند به دلایل مختلفی کاربرد دارد که برخی از آن‌ها اشاره می‌کنم. همچنین کیفیت داده‌های خام معمولاً شامل خطاها، ناهماهنگی‌ها، اطلاعات ناقص هستند و با استفاده از تکنیک‌های تبدیل داده می‌توان داده‌ها را تمیز، یکپارچه‌سازی و کیفیت آن‌ها را بهبود داد.

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

تفاوت‌های Label و one-Hot در یادگیری ماشین برای تبدیل ویژگی‌های کلامی (Categorical) به فرمت عددی در نظر گرفته می‌شود و دو تکنیک متداول در این زمینه LabelEncoder هستند. در ادامه به تفاوت‌ها و ویژگی‌های هر یک می‌پردازیم.

Senobar

۶- چرا Model Building Tools اهمیت دارد؟

Feature Selection و Feature Extraction به افزایش دقت و بهینگی یکی از مراحل کلیدی در فرآیند ساخت مدل های یادگیری ماشین است و اهمیت آن به دلایل مختلفی برمی گردد. ۱- بهبود دقت مدل: با انتخاب ویژگی های مهم و مفید می توان دقت مدل را افزایش داد و ویژگی های غیرضروری یا زائد را حذف نمود تا مدل بهتر یاد بگیرد و پیش بینی های دقیق تری انجام دهد.

۷- DuplicatedData چگونه در پایگاه داده ها حذف می شود؟

حذف DuplicateData (داده های تکراری) از پایگاه داده ها یکی از وظایف مهم در مدیریت داده ها و پیش از پیش داده های بزرگ است. دلیل یادآوری وارد کردن نادرست یا ادغام نادرست داده ها، داده های تکراری می تواند به وجود بیاید. چند روش برای شناسایی و حذف داده های تکراری در پایگاه داده ها اشاره می شود:

۸- IrrelevantData چه مشکلاتی را در پیش بینی های ایجاد می کند؟

Machine Learning

وجود IrrelevantData (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)

SUBJECT

Year: Month: Day: Page: ( )

۹- چرا Values Missing برای برگردن DataImputation کاربرد دارد؟

Values-Missing را با برگردن مقادیر گمشده DataImputation 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

پر کردن 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 متکی 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 هر دو 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 بر اساس آن نتیجه گیری کنید

Sanobar

۱۴۰۲/۱۰/۱۰

بخش (۱۴)

SUBJECT

Year Month Day Page

- outliers چیست و چگونه می‌توانیم آن‌ها را شناسایی کنیم؟

outliers داده‌های پرت هستند که به طور قابل توجهی با سایر داده‌ها متفاوت هستند. این داده‌ها می‌توانند ناشی از خطاهای اندازه‌گیری، ورود اشتباه

outliers در تحلیل داده‌ها مهم است زیرا می‌توانند تأثیر زیادی بر نتایج مدل‌ها و تحلیل‌ها داشته باشند

- DataTransformation چیست و کاربرد دارد؟

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

۱. بهبود کیفیت داده‌ها: داده‌های خام معمولاً شامل خطاها، ناهماهنگی‌ها و اطلاعات ناقص

هستند. با استفاده از تکنیک‌های تغییر داده می‌توان داده‌ها را تصحیح یا به

شکلی درآورد که کیفیت آن‌ها بهبود یابد

- Encoding (Label Encoding و One-Hot Encoding) چه تفاوتی دارند؟

تفاوت‌های Label Encoding و One-Hot Encoding برای تبدیل ویژگی‌های

کلامی (Categorical Features) به فرمت عددی در خط‌مشی گرفته می‌شود و در تکنیک متداول

در این زمینه Label Encoding هستند داده‌ها به تفاوت‌ها و ویژگی‌های هر یک

می‌پردازیم