دانشگاه ملی مهارت

آموزشکده میناب

تمرینات سری اول

نام و نام خانوادگی : مرضیه نجفی _ هدا مهرانی پور

واحد درسی : مباحث ویژه

رشته : مهندسی کامپیوتر

مدرس : محمد احمد زاده

فروردین ۱۴۰۴


بسم الله الرحمن الرحيم

۱. چرا Data Cleaning در علم داده اهمیت دارد؟

- اهمیت : Data Cleaning یا پاکسازی داده‌ها، فرآیندی برای شناسایی و اصلاح داده‌های نادرست، ناقص، نامربوط، تکراری یا قالب‌بندی نشده است. این فرآیند برای اطمینان از کیفیت داده‌ها و به دست آوردن نتایج دقیق و قابل اعتماد در تحلیل‌ها و مدل‌سازی‌ها ضروری است.
- دلایل اهمیت :
- بهبود دقت مدل : داده‌های تمیز، دقت مدل‌های یادگیری ماشین را افزایش می‌دهند.
- کاهش خطا : از بروز خطاها و اشتباهات در تصمیم‌گیری‌ها جلوگیری می‌کند.
- افزایش قابلیت اطمینان : نتایج تحلیل‌ها و گزارش‌ها قابل اعتمادتر می‌شوند.
- صرفه‌جویی در زمان و هزینه : با جلوگیری از تحلیل داده‌های نادرست، در زمان و هزینه صرفه‌جویی می‌شود.

۲. Missing Values چگونه مدیریت می‌شوند؟

روش‌ها:

- حذف : حذف ردیف‌ها یا ستون‌هایی که دارای مقادیر از دست رفته هستند (در صورتی که تعداد Missing Values کم باشد).
- جایگزینی (Imputation) :
- میانگین/میانه/مد : جایگزینی با میانگین برای داده‌های نرمال، میانه برای داده‌های غیر نرمال، و مد برای داده‌های دسته‌ای.
- رگرسیون : استفاده از مدل رگرسیون برای پیش‌بینی مقادیر از دست رفته.
- روش‌های پیشرفته : مانند k-Nearest Neighbors (KNN) یا الگوریتم‌های یادگیری ماشین.

۳. Outliers چیست و چگونه می‌توانید آن‌ها را تشخیص دهید؟

- تعریف : Outliers یا داده‌های پرت، مقادیری هستند که به طور قابل توجهی با سایر داده‌ها متفاوت هستند.

روش‌های تشخیص :

روش‌های گرافیکی :

- Box Plot : نمایش توزیع داده‌ها و شناسایی داده‌های خارج از بازه IQR.
- Scatter Plot : نمایش رابطه بین دو متغیر و شناسایی داده‌های دور از الگو.

روش‌های آماری :

- Z-Score : محاسبه فاصله داده‌ها از میانگین (مقادیر Z-Score بزرگتر از ۳ یا کوچکتر از -۳ به عنوان Outlier در نظر گرفته می‌شوند).
- IQR (Interquartile Range) : محاسبه بازه بین چارک اول و سوم و شناسایی داده‌های خارج از ۱.۵ برابر IQR.

روش‌های مبتنی بر فاصله : مانند DBSCAN.

۴. Data Transformation چرا کاربرد دارد؟

دلایل کاربرد:

- بهبود عملکرد مدل : برخی از الگوریتم‌ها به داده‌های نرمال یا مقیاس‌بندی شده بهتر پاسخ می‌دهند.
- تفسیرپذیری : تبدیل داده‌ها می‌تواند تفسیر نتایج را آسان‌تر کند.
- یکنواختی : تبدیل داده‌ها به یک مقیاس یکنواخت، مقایسه و ترکیب داده‌ها را آسان‌تر می‌کند.

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

Label Encoding:

- کاربرد : تبدیل مقادیر دسته‌ای به اعداد صحیح.
- مشکل : ایجاد ترتیب اشتباه در داده‌ها (برای داده‌های اسمی مناسب نیست).

One-Hot Encoding

- کاربرد : ایجاد ستون‌های باینری برای هر دسته (مقدار ۱ برای وجود دسته و ۰ برای عدم وجود).
- مزیت : جلوگیری از ایجاد ترتیب اشتباه در داده‌ها.

نکته : برای داده‌های اسمی مناسب است.

۶. چرا Feature Selection در Model-building اهمیت دارد؟

دلایل اهمیت:

- کاهش پیچیدگی مدل : مدل ساده‌تر، تفسیرپذیرتر و سریع‌تر است.
- بهبود دقت مدل : حذف ویژگی‌های نامربوط یا تکراری می‌تواند دقت مدل را افزایش دهد.
- کاهش overfitting : جلوگیری از overfitting با انتخاب ویژگی‌های مهم‌تر.

۷. ❊Duplicate Data چگونه در پایگاه داده‌ها حذف می‌شود؟❊

روش‌ها:

- شناسایی : استفاده از SQL Queries برای شناسایی ردیف‌های تکراری بر اساس یک یا چند ستون.

- حذف : استفاده از SQL Queries برای حذف ردیف‌های تکراری (مانند استفاده از DISTINCT یا ROW_NUMBER()).
- ابزارها : استفاده از ابزارهای Data Quality برای شناسایی و حذف داده‌های تکراری.

۸. Irrelevant Data چه مشکلاتی را در پیش‌بینی‌های Machine Learning ایجاد می‌کند؟

مشکلات :

- کاهش دقت مدل : داده‌های نامربوط می‌توانند مدل را گمراه کنند و دقت آن را کاهش دهند.
- افزایش overfitting : مدل ممکن است به داده‌های نامربوط بیش از حد توجه کند و overfitting رخ دهد.
- افزایش زمان آموزش : داده‌های نامربوط می‌توانند زمان آموزش مدل را افزایش دهند.

۹. چرا Data Imputation برای پر کردن Missing Values کاربرد دارد؟

- کاربرد : Data Imputation یا جایگزینی داده‌ها، فرآیندی برای جایگزینی مقادیر از دست رفته با مقادیر تخمینی است.

دلایل کاربرد :

- حفظ اطلاعات : جلوگیری از حذف اطلاعات مفید در ردیف‌ها یا ستون‌هایی که دارای Missing Values هستند.
- بهبود دقت مدل : جایگزینی مقادیر از دست رفته با مقادیر مناسب می‌تواند دقت مدل را افزایش دهد.
- سازگاری با الگوریتم‌ها : برخی از الگوریتم‌های یادگیری ماشین نمی‌توانند با Missing Values کار کنند.

۱۰. چگونه می‌توانید Normality را در داده‌های عددی بررسی کنید؟

روش‌ها:

روش‌های گرافیکی:

- Histogram : نمایش توزیع داده‌ها و بررسی تقارن آن.
- Q-Q Plot : مقایسه توزیع داده‌ها با توزیع نرمال.

روش‌های آماری :

- Shapiro-Wilk Test : آزمون آماری برای بررسی Normality (مقادیر p-value بزرگتر از ۰.۰۵ نشان‌دهنده Normality است).
- Kolmogorov-Smirnov Test : آزمون آماری دیگر برای بررسی Normality.
- Skewness و Kurtosis : محاسبه مقادیر Skewness و Kurtosis (مقادیر نزدیک به صفر نشان‌دهنده Normality است).

منابع

۱. Data Cleaning: کتاب "Data Quality" (Olson)، SearchDataManagement.

۲. Missing Values: مقاله "A review of missing data" (Enders)، Towards Data Science.

۳. Outliers: مقاله "Anomaly Detection" (Chandola et al.)

۴. Data Transformation: کتاب "Feature Engineering" (Zheng & Casari)، Scikit-learn.

۵. Encoding: Scikit-learn, Towards Data Science.

۶. Feature Selection: کتاب "Feature Engineering" (Zheng & Casari)، Scikit-learn.

۷. Duplicate Data: SQL Server، MySQL، PostgreSQL (مستندات).

۸. Irrelevant Data: تجربه و دانش تخصصی.

۹. Data Imputation: همون Missing Values.

۱۰. Normality: آماردان، Scipy.