(تمارین درس مباحث ویژه)

بخش چهارم

مدرس: مهندس احمدزاده

دانشجویان:

عباس کارگرجهرمی / مهدی فرازنده شهرکی

دانشکده ملی مهارت میناب

بهمن ۱۴۰۳

A. چرا Data Cleaning در علم داده اهمیت دارد؟

Data Cleaning(پاک‌سازی داده) یکی از مهم‌ترین مراحل در فرآیند علم داده است، چون:

- داده‌های خام معمولاً ناقص، اشتباه یا ناسازگار هستند.
- مدل‌های یادگیری ماشین و تحلیل‌های آماری بر اساس داده‌ها آموزش می‌بینند، بنابراین داده‌های اشتباه باعث ایجاد مدل‌های نادرست می‌شوند.
- پاک‌سازی داده باعث افزایش دقت، کیفیت و قابلیت اعتماد نتایج می‌شود.
- مثال: اگر داده‌های فروش یک محصول دارای ورودی‌های تکراری یا غلط باشند، پیش‌بینی فروش بسیار نادرست خواهد بود.

B. Missing Values چگونه مدیریت می‌شوند؟

Missing Values(مقادیر گمشده) به داده‌هایی اشاره دارد که در برخی ستون‌ها یا ردیف‌ها ثبت نشده‌اند. روش‌های مدیریت آن:

۱. حذف ردیف‌ها یا ستون‌ها :وقتی مقدار گمشده زیاد نباشد.

۲. جایگزینی با مقدار میانگین/میانه/مد :برای داده‌های عددی.

۳. استفاده از مدل‌های پیش‌بینی) مانند **KNN** یا رگرسیون (برای تخمین مقادیر گمشده.

۴. استفاده از **Data Imputation** که در ادامه توضیح داده خواهد شد.

- انتخاب روش مناسب به نوع داده و درصد missing values بستگی دارد.

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

Outliers(داده‌های پرت) داده‌هایی هستند که به‌طور غیرعادی با بقیه داده‌ها تفاوت دارند.
راه‌های تشخیص:

۱. بصری‌سازی با نمودارهایی مثل **Boxplot** یا**Scatterplot**

۲. محاسبه :**Z-Score** اگر مقدار z بیشتر از ۳ یا کمتر از -۳ باشد.Outlier →

۳. روش:**IQR (Interquartile Range)**

- Q1 = چارک اول، Q3 = چارک سوم

- داده‌هایی خارج از بازه Outlier = $Q1-1.5\times IQR, Q3+1.5\times IQR$ $Q1 - 1.5\times IQR\ ,\ Q3 + 1.5\times IQR$

D. Data Transformation چرا کاربرد دارد؟

Data Transformation به معنای تبدیل شکل یا مقیاس داده‌هاست. اهمیت آن:

- نرمال‌سازی **(Normalization)** برای یکنواخت‌کردن بازه ویژگی‌ها در مدل‌هایی مثل KNN یا SVM.
- استانداردسازی **(Standardization)** برای حذف تأثیر مقیاس.
- تبدیل‌های لگاریتمی یا ریشه دوم برای کاهش چولگی داده.
- در مجموع: باعث می‌شود مدل بهتر یاد بگیرد و سریع‌تر به نتیجه برسد.

E. Encoding Techniques چه تفاوتی دارند؟

ویژگی	Label Encoding	One-Hot Encoding
تعریف	تبدیل دسته‌ها به اعداد ترتیبی	ایجاد ستون مجزا برای هر دسته
مثال	Red=0, Green=1, Blue=2	Red → [1,0,0], Green → [0,1,0]
مشکل	به مدل حس ترتیب می‌دهد (که ممکن است اشتباه باشد)	افزایش حجم داده
کاربرد	در مدل‌هایی که با ترتیب مشکلی ندارند	در مدل‌های حساس به مقادیر عددی

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

Feature Selection یعنی انتخاب مهم‌ترین ویژگی‌ها (ستون‌ها) از میان کل ویژگی‌ها.

اهمیت:

- کاهش پیچیدگی مدل و زمان پردازش
- جلوگیری از Overfitting (یادگیری بیش‌ازحد جزئیات غیرمفید)
- افزایش دقت مدل
- کاهش هزینه‌های ذخیره‌سازی و محاسباتی
- روش‌ها: روش‌های آماری، مدل‌های درختی، روش‌های مبتنی بر یادگیری.

G. Duplicate Data چگونه در پایگاه داده‌ها حذف می‌شود؟

Duplicate Data داده‌هایی هستند که چند بار تکرار شده‌اند. برای حذف آن‌ها:

- استفاده از دستور **drop_duplicates()** در پایتون (پانداس)
- در SQL: استفاده از دستور SELECT DISTINCT یا DELETE با شرایط خاص
- گاهی لازم است معیار تکرار **ID**، **timestamp**، و **...** تعریف شود تا ردیف‌های تکراری واقعی شناسایی شوند.

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

داده‌های بی‌ربط ویژگی‌هایی هستند که هیچ تأثیری در خروجی مدل ندارند یا حتی گمراه‌کننده‌اند.
مشکلات:

- **کاهش دقت مدل**
- **افزایش زمان آموزش**
- **افزایش احتمال Overfitting**
- **باعث می‌شوند مدل الگوهای اشتباهی را یاد بگیرد.**
  مثال: اگر بخواهیم قیمت خانه را پیش‌بینی کنیم، رنگ ماشین صاحب‌خانه داده‌ای بی‌ربط است.

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

Data Imputation **به فرآیند جایگزینی مقادیر گمشده با مقادیر تخمینی گفته می‌شود.**
کاربرد:

- **جلوگیری از حذف داده‌های ارزشمند**
- **حفظ ساختار داده‌ها برای الگوریتم‌ها**
- **روش‌ها:**
  - **استفاده از میانگین/میانه/مد**
  - **مدل‌های یادگیری مثل KNN ، Random Forest، یا رگرسیون برای پیش‌بینی مقدار گمشده**

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

Normalityیعنی بررسی اینکه آیا داده‌ها از توزیع نرمال پیروی می‌کنند یا نه.

روش‌ها:

۱. نمودارهای بصری :مانند Histogram یاQ-Q Plot

۲. آزمون‌های آماری:

- **Shapiro-Wilk Test**
- **Kolmogorov–Smirnov Test**
- **Anderson-Darling Test**

۳. محاسبه چولگی **(Skewness)** و کشیدگی**(Kurtosis)**

- مقدار Skewness نزدیک صفر → توزیع نرمال است