Date: Sat Sun Mon Tue Thu Wed Fri Subject: ----------

بخش ۴

درس : مباحث ویژه

موضوع : Data Preprocessing

نام دانشجو :

نرگس توکلی پور

کبری رحمی زاده

نام استاد: محمد احمدزاده

Date: / / Sat Sun Mon Tue Thu Wed Fri Subject: ----------

A. چرا Data cleaning در علم داده اهمیت دارد؟

Data cleaning یا پاک‌سازی داده‌ها اهمیت زیادی در علم داده دارد چون داده‌های نادرست، ناقص یا نامرتب می‌توانند نتایج تحلیل را به شدت تحت تاثیر قرار بدهند. وقتی داده‌ها تمیز نباشند، مدل‌های یادگیری ماشین ممکن است دقت پایینی داشته باشند و تحلیل ما به اشتباه بیفتند. بنابراین با پاک‌سازی داده‌ها باعث می‌شود نتایج دقیق‌تر و قابل اعتمادتر بشه و به تصمیم‌های بهتری برسیم.

B. Missing values چجوری مدیریت می‌شوند؟

مدیریت مقادیر گمشده در داده‌ها یکی از مراحل مهم در پیش‌پردازش داده‌هاست.

چند روش برای مدیریت این مقادیر وجود دارد:

1. حذف داده‌ها: ردیف یا ستون‌هایی که داراي مقادیر گمشده را حذف کنیم

2. جایگزینی: با میانگین، میانه یا مد پر کنیم.

3. پیش‌بینی: با مدل‌های یادگیری ماشین مقادیر را پیش‌بینی کنیم.

Date: / / Sat Sun Mon Tue Thu Wed Fri Subject: ----------

4. روش‌های خاص: روش‌هایی مثل k-Nearest یا multiple Imputation

وجود دارند که می‌توانند مقادیر گمشده را به شکل بهتری پر کنند.

5. نمودار: یک دسته جداگانه برای مقادیر گمشده تعریف کنیم.

c. outliers چیست و چگونه می‌توانید آن را تشخیص دهید؟

outliers یا نقاط پرت به داده‌هایی گفته می‌شود که به طور قابل توجهی از سایر داده‌ها در یک مجموعه داده فاصله دارند. این نقاط معمولاً می‌توانند نشان‌دهنده خطا در داده‌ها، تغییرات طبیعی یا حتی موارد خاص و منحصر به فرد باشند.

برای تشخیص outliers می‌توان از روش‌های زیر استفاده کرد:

1. نمودار جعبه‌ای (BoxPlot): نقاط خارج از محدوده جعبه.

2. نمودار پراکندگی (scatterPlot) نقاط دور از سایر نقاط.

3. معیارهای آماری: مانند z-score یا IQR.

Date: / / Sat Sun Mon Tue Thu Wed Fri Subject: ----------

D. Data transformation چه کاربردهایی دارد؟ تبدیل داده‌ها به فرایند تغییر شکل، ساختار یا فرمت داده‌ها اشاره دارد که به منظور آماده‌سازی آن‌ها برای تجزیه و تحلیل یا استفاده در سیستم‌های مختلف انجام می‌شود. این کار چندین کاربرد مهم دارد:

1. بهبود کیفیت داده‌ها: خطاها و نواقص را اصلاح می‌کند.

2. یکپارچه‌سازی داده‌ها: داده‌ها را به فرمت استاندارد تبدیل می‌کند.

3. تحلیل بهتر: سازماندهی داده‌ها برای تحلیل ساده‌تر.

4. افزایش کارایی: سرعت پردازش را بالا می‌برد.

5. پشتیبانی از تصمیم‌گیری: اطلاعات بهتری برای تصمیم‌گیرندگان فراهم می‌کند.

Date: / / Sat Sun Mon Tue Tha. Wed Fri Subject: ----------

E. Encoding (One-Hot Label Encoding)

Encoding Techniques چه تفاوتی دارند؟

تفاوت اصلی بین one-Hot Encoding و Lebel Encoding به نوع نمایش داده‌ها دسته‌ای (categorical Data) مربوط می‌شود:

Lebel Encoding: هر دسته به یک عدد منحصر به فرد تبدیل می‌شود.

one-Hot Encoding: هر دسته به یک بردار باینری تبدیل می‌شود.

F. چرا feature selection در building اهمیت دارد؟

feature selection یا انتخاب ویژگی‌ها در ساخت مدل خیلی مهم است. به چند دلیل:

1. کاهش ابعاد: مدل را ساده‌تر و سریع‌تر می‌کند.

2. بهبود دقت: ویژگی‌های نامربوط را حذف می‌کند و دقت مدل را افزایش می‌دهد.

3. جلوگیری از overfitting: کمک می‌کند تا مدل بهتر روی داده‌های جدید عمل کند.

4. تفسیرپذیری: فهمیدن تاثیر ویژگی‌ها را راحت‌تر می‌کند.

Date: Sat Sun Mon Tue Thu Wed Fri Subject:

6. Duplicate Data چگونه در پایگاه داده‌ها حذف می‌شود؟

حذف داده‌های تکراری در پایگاه داده‌ها معمولاً با استفاده از چند روش انجام می‌شود.

1. استفاده از دستور Distinct: برای انتخاب داده‌های یکتا یک جدول می‌توان از دستور SelectDistict استفاده کرد. این دستور فقط رکوردهای منحصربه‌فرد را برمی‌گرداند

2. استفاده از Group By: می‌توان داده‌ها را براساس یک یا چند ستون گروه‌بندی کرده و رکوردهای تکراری را حذف کرد

3. استفاده از JOIN: می‌توان با استفاده از یک JOIN بین خود جدول و یک زیرپرسش، رکوردهای تکراری را شناسایی و حذف کرد.

7. Irrelevent Data چه مشکلاتی را در پیش‌بینی‌های Machine Learning ایجاد می‌کنند؟ می‌توانند باعث کاهش دقت مدل‌های Machine Learning شوند. این داده‌های مربوط می‌توانند الگوهای واقعی را پنهان کنند و باعث ایجاد نویز در پیش‌بینی‌ها شوند

Date: / /

Sat Sun Mon Tue Thu Wed Fri

Subject: ----------

مدل ها ممکن است نتوانند به درستی یاد بگیرند و عملکردشان ضعیف شود.

أ. چرا Data Imputation برای پرکردن Missing values کاربرد دارد؟

Data Imputation یک تکنیک مهم در علم داده ها است که استفاده (missing values) برای پرکردن مقادیر گمشده می شود وقتی که در یک داده بعضی از مقادیر موجود نیستند.

این می تواند باعث مشکلاتی در تحلیل مدل سازی شود.

دلایل آن برای پرکردن مقادیر:

1. دقت مدل را افزایش می دهد.

2. اطلاعات مهم را حفظ می کند.

3. سوگیری نتایج را کاهش می دهد.

4. حجم داده ها را افزایش می دهد.

Date: Sat Sun Mon Tue Thu Wed Fri Subject:

ژ. چگونه می توانید Normality را در داده‌های عددی بررسی کنید؟

برای بررسی نرمالیتی های داده های عددی می توان از روش های زیر استفاده کرد:

1. هیستوگرام: بررسی شکل توزیع.

2. نمودار Q-Q: مقایسه نقاط داده با توزیع نرمال.

3. آزمون شاپیرو-ویلک: Pvalue کمتر از 0.05 نشانه عدم نرمالیتی.

4. آزمون کولموگروف-اسمیرنوف: بررسی انطباق با توزیع نرمال.

5. آزمون اندرسون-دارلینگ: مشابه آزمون های قبلی.