مباحث ویژه

بخش چهارم

Subject: استاد احمدزاده

Year. Month. Date. دکتر قادری

Data PreProcessing

A: چرا Data cleaning در علم داده اهمیت دارد؟ زیرا داده‌های باکیفیت بالا امکان ایجاد تحلیل دقیق‌تر و تصمیم‌گیری بهتری که کاهش خطاها و ریسک‌ها را افزایش به صورت بهبود عملکرد کسب و کار را فراهم می‌کند.

B: Missing Values چگونه مدیریت می‌شوند؟ ۱- حذف داده‌های گمشده (حذف لیست‌وار - حذف پیمانه‌وار) ۲- جایگزینی داده‌های گمشده (جایگزینی با میانگین، میانه یا ...) تحلیل حساسیت کنیم

بهترین روش برای مدیریت داده‌های گمشده بستگی به نوع داده‌ها، میزان داده‌های گمشده؛ برای داده‌های کم و تصادفی مناسب حذف، اما روش‌های پیشرفته‌تر مانند تحلیل چندگانه برای داده‌های بیشتر توصیه می‌شود.

C: Outliers چیست و چگونه می‌توانند آن‌ها را تشخیص و تحلیل کنید؟ داده‌های پرت داده‌ای است که در مقادیر غیرعادی از بقیه مقادیر داده در یک نمونه تصادفی از یک جمعیت مشاهده می‌شود تا حدی فاصله این داده با بقیه داده‌ها غیرعادی است که آنالیزگر را به این وا می‌دارد که بررسی کند چه چیزی در داده‌ها به نظر غیرعادی می‌آید.

روش اساسی برای تشخیص مدل همه داده به صورت نمودار مطرح شده، ارزیابی شکلی داده به خیر های اصلی آن به صورت گرافیکی در آن در ارزیابی داده‌های مربوط به مشاهدات غیرعادی که از بقیه حجم داده جدا افتاده.

D: Data transformation چه تأثیری دارد؟ برای بهبود کیفیت، معیار، قابلیت تجزیه و تحلیل داده استفاده می‌شوند. این فرآیند به یکپارچه‌سازی داده از منابع مختلف کمک کرده و کارایی سیستم‌ها را افزایش می‌دهد. همچنین تصمیم‌گیری را تسهیل می‌کند.

Almas

Subject:

Year. Month. Date.

Encoding Techniques (One-Hot Encoding و Label Encoding) : ۶

1. Label Encoding

در این روش هر دسته به یک عدد صحیح متفاوت داده می‌شود. برای مثال اگر ویژگی با مقادیر قرمز، سبز و آبی داشته باشیم، ممکن است به ترتیب مقادیر ۰، ۱ و ۲ داده شود.

معایب: این روش می‌تواند الگوریتم‌های یادگیری ماشین را دچار مشکل کند زیرا ممکن است عدد بزرگتر به معنای اهمیت بیشتر از اعداد کوچکتر تفسیر شود.

2. One-Hot Encoding

در این روش هر دسته به یک متغیر باینری جداگانه تبدیل می‌شود. برای مثال اگر ویژگی رنگ با مقادیر قرمز، سبز و آبی به صورت سه ستون باینری نمایش داده می‌شود:

- قرمز: [0, 0, 1]
- سبز: [0, 1, 0]
- آبی: [1, 0, 0]

مزایای این روش جلوگیری از حفظ رابطه نادرست بین دسته‌ها به صورت خلاصه Label Encoding برای ویژگی‌هایی که ترتیب ندارند مناسب است و به جای One-Hot Encoding برای ویژگی‌هایی که ترتیب دارند بهتر است.

Almas

Subject:
Year. Month. Date.

F. چرا feature selection در Model-building اهمیت دارد؟ ۱. بهبود عملکرد ۲. کاهش پیچیدگی مدل ۳. جلوگیری از اورفیتینگ ۴. افزایش قابلیت تفسیر

G. چگونه Duplicate Data در پایگاه داده‌ها حذف می‌شود؟
۱. استفاده از دستورات SQL ۲. استفاده از توابع گروه‌بندی ۳. ابزارها و نرم‌افزارهای مدیریت داده ۴. بررسی و پاکسازی
به طور کلی حذف داده‌های تکراری به بهبود کیفیت داده‌ها و کاهش حجم پایگاه داده کمک می‌کند.

H. Irrelevant Data چه مشکلاتی را در روش‌های Machine Learning ایجاد می‌کند؟ ۱. کاهش دقت مدل ۲. آلودگی سیگنال ۳. افزایش زمان آموزش ۴. کاهش تعمیم‌پذیری ۵. خطا بیشتر در پیش‌بینی

I. چرا Data Imputation برای پر کردن Missing values کاربرد دارد؟ ۱. حفظ اطلاعات ۲. بهینه‌سازی ۳. کاهش انحراف ۴. ایجاد تعادل. در نتیجه یک روش موثر برای مدیریت داده‌های ناقص است که به حفظ یکپارچگی و کیفیت داده‌ها کمک می‌کند.

Subject:

Year. Month. Date.

آ: چگونه می توانید Normality را در داده های عددی بررسی کنید؟ ۱- هیستوگرام ۲- نمودار جعبه ای ۳- نمودار پراکنش

ج: آزمون های آماری: ۱- آزمون شاپیرو-ویلک ۲- آزمون کولموگروف-اسمیرنوف این روش ها در آزمون های تحلیل بکار می روند تا نرمالیته داده های عددی را به صورتی برای شما توضیحات مناسب را مورد تحلیل های آماری و مدل سازی بکنیم.