تمرین چهارم درس مبانی داده کاوی

(بهار ۴۰۲)

مهلت تحویل تمرین: ۱ خرداد ماه

سوالات تئوری

سوال ۱- اعداد زیر (مربوط به میزان فروش) را در نظر بگیرید. با استفاده از روش های depth-equal و width-equal این داده ها را سبدبندی کنید.

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215.

سوال ۲- یک نظرسنجی بر روی یک نمونه تصادفی از ۲۰ دانشجوی سال دوم دانشگاه استنفورد انجام شد. از آنها پرسیده شد: "چند کتاب درسی دارید؟" پاسخ های آن ها عبارت بود از:

0, 0, 2, 5, 8, 8, 8, 9, 9, 10, 10, 10, 11, 12, 12, 12, 14, 15, 20, 25

با توجه به پاسخ دانشجویان، داده های outlier را با استفاده از روش IQR یافته و نمودار آن را رسم کنید.

سوال ۳- داده های زیر که به صورت صعودی مرتب شده اند، به عنوان مقادیر سن به شما داده شده است:

(تمرین ۳,۳ (سوال ۳ فصل سوم) کتاب آقای هان)

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 23, 23, 23, 25, 25, 25, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

- از smoothing به وسیله سبد بندی کردن برای smooth کردن داده ها استفاده کنید و عمق سبد بندی را برابر ۳ در نظر بگیرید. گام های مورد استفاده را نشان دهید. به نظر شما تاثیر این تکنیک برای داده های ارائه شده به چه صورتی است؟
- چگونه داده های پرت را در میان داده ها تشخیص می دهید؟
- چه روش های دیگری برای smooth کردن این داده ها وجود دارد؟

سوالات عملی

سوال ۴- از دیتاست data که در اختیارتان قرار داده شده است برای حل سوالات زیر استفاده نمایید:

۱- دیتاست را با استفاده از کتابخانه pandas خوانده و تبدیل به دیتافریم نمایید.
۲- با استفاده از تابع describe خلاصه آماری این دیتاست را به دست آورید.
۳- در این دیتاست مقدار صفر به معنای داده مفقود است. تعداد داده های صفر را در ستون های شماره ۱ تا ۵ نمایش دهید.
۴- در ستون های شماره ۱ تا ۵، مقدار صفر را با NaN جایگزین کنید.
۵- قسمت ۳ را با استفاده از تابع isnull تکرار کنید.
۶- ۱۵ ردیف اول دیتافریم را نشان دهید.
۷- یک دیتافریم از ستون های شماره ۱ تا ۵ دیتافریم اصلی ایجاد کنید. رکوردهای دارای مقدار مفقود را حذف کنید. ابعاد دیتاست را قبل و بعد از این تغییر نمایش دهید.
۸- یک دیتافریم دیگر از ستون های شماره ۱ تا ۵ دیتافریم اصلی ایجاد کنید. رکوردهای دارای مقدار مفقود را با مقدار میانگین دیتاست جایگزین کنید. تعداد مقادیر مفقود در هر ستون را با تابع isnull نشان دهید.
۹- یک دیتافریم دیگر از ستون های شماره ۱ تا ۵ دیتافریم اصلی ایجاد کنید. رکوردهای دارای مقدار مفقود را با مقدار میانگین هر ستون جایگزین کنید. از تابع SimpleImputer از کتابخانه sklearn استفاده کنید. در انتها تعداد مقادیر مفقود در هر ستون را با تابع isnull نشان دهید.
۱۰- استراتژی های دیگر برای جایگزینی مقادیر مفقود شده چیست؟ خروجی هر روش را نمایش دهید.
۱۱- کدام الگوریتم ها در برابر missing values مقاوم هستند؟ توضیح دهید.

سوال ۵- (اختیاری) از دیتاست heart diagnose که در اختیارتان قرار داده شده است برای حل سوالات زیر استفاده نمایید:

۱- هیستوگرام ویژگی resting blood pressure را رسم کرده و مشخص کنید که آیا کجی دارد؟ از چه نوعی است؟ مقدار عددی آن را نمایش دهید.

۲- به کمک لگاریتم طبیعی سعی کنید کجی را برطرف کنید. ویژگی نرمال شده را در یک متغیر جدید ذخیره کرده و هیستوگرام و مقدار عددی کجی آن را نشان دهید.

۳- به کمک جذر گرفتن سعی کنید کجی را برطرف کنید. ویژگی نرمال شده را در یک متغیر جدید ذخیره کرده و هیستوگرام و مقدار عددی کجی آن را نشان دهید.

۴- یک روش جدید برای برطرف کردن کجی پیدا کنید و به کمک آن کجی را برطرف کنید. ویژگی نرمال شده را در یک متغیر جدید ذخیره کرده و هیستوگرام و مقدار عددی کجی آن را نشان دهید.

۵- ۳ روش انجام شده را باهم مقایسه کنید.

نحوه تحویل: سوالات تئوری را به صورت تایپ شده و در قالب یک فایل PDF تحویل دهید. به علاوه هر یک از سوالات عملی را در قالب یک فایل ipynb به همراه نتایج قرار داده و فایل را به صورت Qn نام گذاری نمایید که n شماره سوال مربوطه می باشد. در انتها فایل های پایتون را به همراه فایل PDF تماما در قالب یک فایل zip نامگذاری شده به صورت NAME_STUDENTID در سامانه درس بارگذاری کنید. برای سوالات عملی توضیحات خود را به صورت Markdown در فایل پایتون بنویسید.

"When the whole world declares war against you, love will be your only shield in all moments of this war." - Based on the life of Severus Snape