به نام خدا

تمرین چهارم داده کاوی

حدیث غفوری 9825413

سوال 1

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

۱۲تاداده داریم.

روش depth-equal binning :

گام اول sort کردن داده هاست که داده ها قبلا sort شده اند.

گام دوم پیداکردن سایز هر bin است:

12/3 = 4

گام اخرساختن هر Bin:

Bin1: [5,10,11,13] Bin2: [15,35,50,55] Bin3: [72, 92, 204, 215]

روش width-equal binning :

range = max(values) - min(values)

range = 215 – 5 = 210

bin width = range / num_bins

num_bins = 3

bin width = 210/3 = 70

Bin1 در بازه ی ۵ تا ۷۵

Bin2 در بازه ی ۷۵ تا ۱۴۵

Bin3 در بازه ی ۱۴۵تا ۲۱۵

پس درنهایت جواب:

Bin1: [5, 10, 11, 13, 15, 35, 50, 55, 72] Bin2: [92] Bin3: [204,215]

سوال 2

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

میانه داده ها را پیدا میکنیم:به علت زوج بودن تعداد داده ها (۲۰تا)میانه برابر میانگین دو عدد وسط است یعنی ۱۰مین و ۱۱مین مقادیر که :

(10 + 10)/2 = 10 = median

سپس مقدار Q1 را محاسبه میکنیم:

Q1 برابر میانه ی نیمه ی پایینی داده هاست که شامل ۱۰مقدار اول است:

0, 0, 2, 5, 8, 8, 8, 9, 9, 10

که میانه این داده ها برابر میانگین دو مقدار وسط است:

(8+8)/2 = 8

سپس مقدار Q3 را محاسبه میکنیم:

Q3 برابر میانه ی نیمه ی بالایی داده هاست که شامل ۱۰مقدار دوم است:

10, 10, 11, 12, 12, 12, 14, 15, 20, 25

که میانه این داده ها برابر میانگین دو مقدار وسط است:

(12+12)/2 = 12

گام دوم:محاسبه ی مقدار IQR = Q3-Q1 = 12 – 8 = 4:

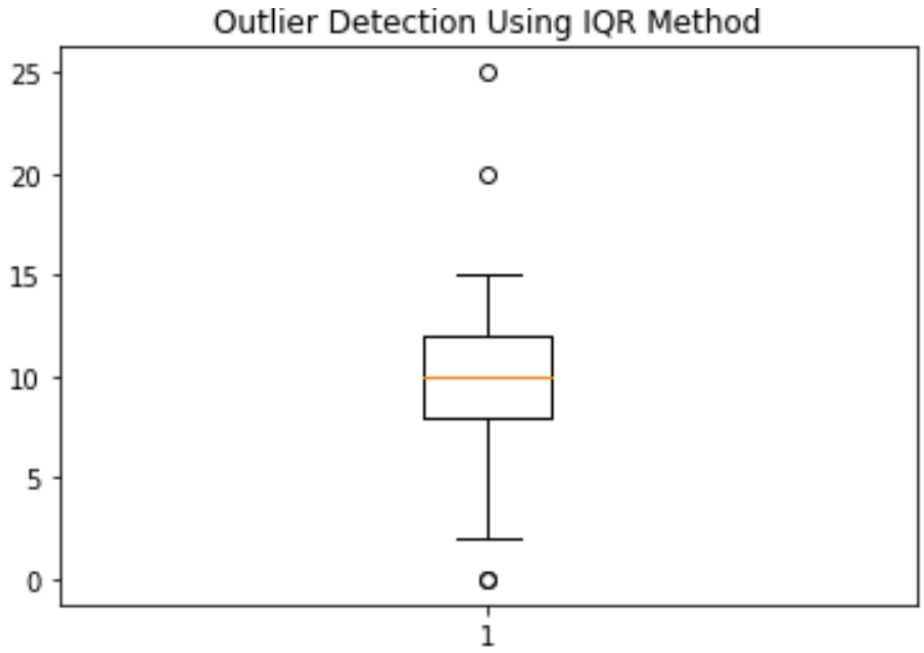
گام سوم:محاسبه ی outlier threshold

Upper bound = Q3 + (1.5 * IQR) = 12 + (1.5*4) = 18

lower bound = Q1 - (1.5 * IQR) = 8 – 6 = 2

مقادیر بزرگتراز ۱۸ وکوچکتر از۲ داده های اوتلایر میشوند پس اوتلایرها برابر:

0, 0, 2,20, 25 است.



سوال 3

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

گام اول sortکردن داده‌هاست که چون داده ها مرتب شده اند نیازی نیست.

گام دوم پارتیشن کردن داده ها با عمق ۳ است:

Bin 1: 13, 15, 16 Bin 2: 16, 19, 20 Bin 3: 20, 21, 22

Bin 4: 22, 25, 25 Bin 5: 25, 25, 30 Bin 6: 33, 33, 35

Bin 7: 35, 35, 35 Bin 8: 36, 40, 45 Bin 9: 46, 52, 70

گام سوم محاسبه ی میانگین هر Bin است:

Bin 1: 142/3, 142/3, 142/3 Bin 2: 181/3, 181/3, 181/3 Bin 3: 21, 21, 21

Bin 4: 24, 24, 24 Bin 5: 262/3, 262/3, 262/3 Bin 6: 332/3, 332/3, 332/3

Bin 7: 35, 35, 35 Bin 8: 401/3, 401/3, 401/3 Bin 9: 56, 56, 56

This method smooths a sorted data value by consulting to its ”neighborhood”. It performs local smoothing.

(ب) چگونه می توانید مقادیر پرت را در داده ها تعیین کنید؟

نقاط پرت در داده ها ممکن است با خوشه بندی شناسایی شوند، جایی که مقادیر مشابه در گروه ها یا "خوشه ها" سازماندهی می شوند. مقادیری که خارج از مجموعه خوشه ها قرار می گیرند ممکن است مقادیر پرت در نظر گرفته شوند. روش دیگر، ترکیبی از کامپیوتر و بازرسی انسانی می‌تواند در جایی که توزیع داده‌های از پیش تعیین‌شده اجرا می‌شود تا به رایانه اجازه دهد تا نقاط پرت احتمالی را شناسایی کند، استفاده شود. پس از آن می توان با بازرسی انسانی با تلاش بسیار کمتری نسبت به تأیید کل مجموعه داده های اولیه، این نقاط دور از دسترس احتمالی را تأیید کرد.

(ج) چه روش های دیگری برای هموارسازی داده ها وجود دارد؟

روش‌های دیگری که می‌توان برای هموارسازی داده‌ها مورد استفاده قرار داد، شامل اشکال جایگزین binning مانند هموارسازی با میانه‌های bin یا هموارسازی با bin boundaries است.

از equiwidth bins می‌توان برای پیاده‌سازی هر یک از اشکال binning استفاده کرد، که در آن محدوده بازه مقادیر در هر bin ثابت است. روش‌های دیگر به جز binning شامل استفاده از تکنیک‌های regression برای هموارسازی داده‌ها با فیت کردن آن‌ها به تابعی مانند رگرسیون خطی یا چندگانه است. همچنین، تکنیک‌های classification را می‌توان برای پیاده‌سازی سلسله مراتب مفاهیم استفاده کرد که می‌تواند داده‌ها را با جمع کردن مفاهیم سطح پایین‌تر به مفاهیم سطح بالاتر صاف کند. روش (rolling-up)