به نام خدا

تمرین اول داده کاوی

حدیث غفوری 9825413

سوال ۱

فرض کنید داده هایی در مورد مراجعان یک بیمارستان در دسترس است. این داده ها می تواند شامل سن، جنسیت، سابقه بیماری قبلی، شغل، قد، وزن و ... باشد. با درنظرگیری داده های موجود چهار نمونه مسئله با استفاده از تسک های داده کاوی مانند پیش بینی، دسته بندی و ... معرفی و توصیف نمایید.

مساله ۱:

پیش بینی سن بیماران جدید به کمک روش رگرسیون

مساله ۲:

دسته بندی کردن افراد با روش KNN براساس بیماری هاشون

مساله ۳:

با توجه به ویژگی های جسمانی مثل وزن و قد و ... پیش بینی کنیم آیا افراد بیماری های قلبی یا مثلا قند خون داردند یا نه

مساله ۴:

پیش بینی وزن افراد بر اساس ویژگی های قد، سن، شغل

سوال ۲

Binary or not

Discrete or Continuous

qualitative (nominal or ordinal) or quantitative (interval or ratio)

1.سن بر حسب سال

باینری نیست، گسسته، ratio

2.روشنایی که با نورسنج اندازه گیری می شود

باینری نیست، پیوسته، ratio اگر عددی که نمایش میدهد از صفر باشد و فرض میکنیم مقدار منفی نداریم و پایه عددی مان از صفر حساب میشود.

پیوسته با فرض اینکه در یک بازه میتواند عدد اعشاری هم نشان بدهد.

3.روشنایی که با نظر افراد بیان می شود

اگر افراد بگویند مثلا اتاق تاریک است یا روشن است یعنی با دوحالت بیان کنند میشه باینری و گسسته و ordinal
اگه بگویند که روشنایی چندین حالت مثل تاریک، کم نور، متوسط، پرنور داریم: باینری نیست، گسسته و ordinal

4.زاویه اندازه گیری شده با وسیله اندازه گیری (نقاله و ...)

باینری نیست، گسسته

اگر برای زاویه مثل زاویه در نقاله یک زاویه اولیه و base ای درنظر بگیریم نوع داده ratio میشه.

ولی اگه زاویه منفی مثل محورهای مختصات داشته باشیم، interval میشه.

5.مدال های اهدایی در مسابقات المپیک

چون مدال های المپیک یک ترتیب معنایی و یک رتبه بندی دارند پس ordinal میشوند.

باینری نیست، گسسته، ordinal

6.ارتفاع از سطح دریا

باینری نیست، گسسته، ratio

7.تعداد بیماران یک بیمارستان

باینری نیست، گسسته، ratio

8.شمارهISBN

شابک یا (ISBN)، شماره استاندارد بین‌المللی کتاب است که برای هر کتاب منحصربه‌فرد می‌باشد. این استاندارد در سال ۱۹۶۶ در کشور انگلیس توسط صنف کتابداران پایه‌گذاری شد. شابک یا ISBN را در زبان انگلیسی به صورت مختصر با ISBN نشان می‌دهند که مخفف International Standard Book Number است.

پس شابک یک داده ی گسسته است، باینری نیست و nominal است.

سوال ۳

10,7,20,12,75,15,9,18,4,12,8,14

میانگین:

$$\frac{10+7+20+12+75+15+9+18+4+12+8+14}{12}=\frac{204}{12}=17$$

میانه:

اعداد را به ترتیب صعودی مینویسیم و عدد وسطی را انتخاب میکنیم

[4, 7, 8, 9, 10, 12, 12, 14, 15, 18, 20, 75]

ما ۱۲تا عدد داریم پس تعداد زوج است و باید میانه را به کمک میانگین گرفتن روی زوج وسط بدست بیاوریم یعنی میانیگین عضو ششم و هفتم

میانه = $\frac{12+12}{2}$ = 12

مد:

عددی که بیشترین تکرار را دارد که 12 است.

انحراف معیار:

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

σ = انحراف معیار جمعیت
N = تعداد اعضای جمعیت
x_i = اندازه هر عضو از جمعیت
μ = میانگین جمعیت

تعداد اعضای جمعیت: N=12

مربع اختلاف اعضا از میانگین:

$(4-17)^2 = 169$ $(7-17)^2 = 100$ $(8-17)^2 = 81$ $(9-17)^2 = 64$

$(10-17)^2 = 49$ $(12-17)^2 = 25$ $(14-17)^2 = 9$ $(15-17)^2 = 4$

$(18-17)^2 = 1$ $(20-17)^2 = 9$ $(75-17)^2 = 3364$

$$\sqrt{\frac{169+100+81+64+49+25+25+9+4+1+9+3364}{12}} = \sqrt{\frac{3,900}{12}} = \sqrt{325} = 18.02$$

پس مقدار انحراف معیار عدد 18.02 است.

شاخص zscore:

μبرابر میانگین جمعیت است.

σ برابر انحراف معیار جمعیت است.

$$z = \frac{x - \mu}{\sigma}$$

Z(4) = (4 - 17) / 54.09 = -0.72142	Z(7) = (7 - 17) / 18.02 = -0.55494	Z(8) = (8 - 17) / 18.02 = -0.49945	Z(9) = (9 - 17) / 18.02 = -0.44395
Z(10) = (10 - 17) / 18.02 = -0.38846	Z(12) = (12 - 17) / 18.02 = -0.27747	Z(14) = (14 - 17) / 18.02 = -0.16648	Z(15) = (15 - 17) / 18.02 = -0.11099
Z(18) = (18 - 17) / 18.02 = 0.055494	Z(20) = (20 - 17) / 18.02 = 0.16648	Z(75) = (75 - 17) / 18.02 = 3.21865	

سوال ۴

تمرین (سوال 6فصل دوم) کتاب آقای هان را حل نمایید.

Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):

(a) Compute the Euclidean distance between the two objects.

این فاصله از فرمول زیر محاسبه میشود:

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2}.$$

$$\sqrt{(22-20)^2 + (1-0)^2 + (42-36)^2 + (10-8)^2} = \sqrt{45} = 6.7082$$

(b) Compute the Manhattan distance between the two objects.

این فاصله از فرمول زیر محاسبه میشود:

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{in} - x_{jn}|.$$

|22 - 20| + |1 - 0| + |42 - 36| + |10 - 8| = 11

(c) Compute the Minkowski distance between the two objects, using q D 3.

این فاصله از فرمول زیر محاسبه میشود:

$$d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdot + |x_{ip} - x_{jp}|^h}$$

$$\sqrt[3]{|22 - 20|^3 + |1 - 0|^3 + |42 - 36|^3 + |10 - 8|^3} = \sqrt[3]{233} = 6.1534$$

(d) Compute the supremum distance between the two objects.

این فاصله از فرمول زیر محاسبه میشود:

$$d(i, j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|.$$

که بیشترین اختلاف حاصل از قدرمطلق عدد 6 است.