به نام خدا

تمرین اول درس مبانی داده کاوی

امیررضا حسینی ۹۸۲۰۳۶۳

سوالات تئوری

سوال ۱) یکی از این کاربردها میتواند برای شرکت بیمه‌ای جهت پیشبینی اینکه آیا این مریض زنده میماند که آن را بیمه کند یا نه بدست بیاورد مثلا میتوان فهمید اکثر مریض‌هایی که عمل پیوند قلب انجام داده‌اند بیشتر از دوسال پس از عمل زنده نمانده‌اند پس در نتیجه برای این شرکت به صرفه نیست تا آنان را بیمه کند و اگر مریض جدیدی با شرایط مشابه بیایند نباید آن را بیمه کند. با اینکار میتواند سود خود را بیشینه کند.

کاربرد دیگر میتواند برای دسته‌بندی بیماران جهت اورژانسی بودن یا نبودن وضعیت آنان با توجه به شرایطشان دانست. با دانستن خصوصیاتی در مورد آنان میتوان آن‌ها را در دسته اورژانسی بودن یا نبود قرار داد و برای آنان شرایط خاص در نظر گرفت.

دسته بندی بیماران: با توجه به ویژگی های مختلف بیماران، می توان آن ها را به گروه های مختلفی دسته بندی کرد. به عنوان مثال، با در نظر گرفتن ویژگی های سن، جنسیت، سابقه بیماری قبلی، شغل، قد، وزن و ... می توان بیماران را به دسته های مختلفی مانند بیماران قلبی، بیماران دیابتی، بیماران سرطانی و ... تقسیم کرد.

توصیه‌های بهبود سلامتی بیماران :با تحلیل داده‌های موجود در مورد سن، جنسیت، سابقه بیماری، قد، وزن و ... می‌توان الگوهای خاصی در مورد سلامتی بیماران شناسایی کرد. بر اساس این الگوها، می‌توان به مدیران بیمارستان توصیه‌هایی برای بهبود سلامتی بیماران داد. به عنوان مثال، برای بیمارانی که دارای وزن بالا هستند، می‌توان توصیه کرد تا با توجه به وضعیت سلامتی، رژیم غذایی مناسبی را انتخاب کنند. همچنین برای بیمارانی که دارای سابقه بیماری قلبی هستند، می‌توان توصیه کرد که فعالیت‌های بدنی روزانه را افزایش دهند و روی رژیم غذایی مناسب و کم‌نمک تمرکز کنند.

سوال ۲)

همه موارد ذیل not binary هستند.

سن بر حسب سال : discrete چون مقادیر گسسته هستند و بین آنها را نمیتوانند کسب کنند. Quantitatively – ratio به دلیل مطلق بودن و نداشتن مقادیر منفی

روشنایی که با نورسنج اندازه گیری می شود: continuous چون یک طیف پیوسته را شامل میشود. Quantitatively – ratio به دلیل مطلق بودن و نداشتن مقادیر منفی

روشنایی که با نظر افراد بیان می شود: discrete چون مقادیر گسسته هستند و بین آنها را نمیتوانند کسب کنند. Qualitative – ordinal به دلیل وجود ترتیب معنادار و اینکه میتوان شدت روشنایی را فهمید.

زاویه اندازه گیری شده با وسیله اندازه گیری (نقاله و ...): continuous چون مقادیر پیوسته هستند. Quantitavely – ratio به دلیل مطلق بودن و نداشتن مقادیر منفی

مدال های اهدایی در مسابقات المپیک: discrete چون مقادیر گسسته هستند و بین آنها را نمیتوانند کسب کنند. Qualitative – ordinal به دلیل وجود ترتیب معنادار

ارتفاع از سطح دریا: continuous دارای مقادیر پیوسته میباشد. Quantitative – intervalبه دلیل محل قرارگیریمان (زیردریا یا روی هوا) میتواند منفی یا مثبت باشد.

تعداد بیماران یک بیمارستان: discrete چون دارای مقادیر گسسته و شمارش اعداد طبیعی میباشد. Quantitavely – ratio به دلیل مطلق بودن و نداشتن مقادیر منفی

شماره ISBN: discerete چون دارای مقادیر گسسته‌ای است که به کتاب‌ها تخصیص داده میشود. Qualitativite – nominal طبق اسلایدها اگر تمامی شماره ISBN کتاب‌ها را از اول به هم بریزیم و مجددا تخصیص بدهیم مشکلی پیش نخواهد آمد (اگرچه میتوان آن را به فضای شامل ترتیب نیز برد و ordinal کرد)

سوال ۳)

محاسبه میانگین:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

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

محاسبه میانه: به دلیل زوج بودن تعداد داده‌ها، میانگین دو داده وسط را در نظر میگیریم.

$$\frac{9+15}{2} = 12$$

محاسبه مد: داده‌ای که بیشترین تکرار را داشته باشد برابر ۱۲ است.

محاسبه انحراف معیار:

$$s_N = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

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

$$= \frac{(10 - 17)^2 + \ldots + (14 - 17)^2}{12}$$

$$= \frac{3900}{12}$$

$$= 325$$

$$\sigma = \sqrt{325}$$

$$= 18.02775637732$$

محاسبه شاخص Z-score:

$$z_i = \frac{x_i - \bar{x}}{s}$$

$$Z_1 = \frac{10 - 17}{18.03} = -0.38824$$

$$Z_2 = \frac{7 - 17}{18.03} = -0.55463$$

$$Z_3 = \frac{20 - 17}{18.03} = 0.16639$$

$$Z_4 = \frac{12-17}{18.03} = -0.27732$$

$$Z_5 = \frac{75-17}{18.03} = 3.21686$$

$$Z_6 = \frac{15-17}{18.03} = -0.11093$$

$$Z_7 = \frac{9-17}{18.03} = -0.4437$$

$$Z_8 = \frac{18-17}{18.03} = 0.055463$$

$$Z_9 = \frac{4-17}{18.03} = -0.72102$$

$$Z_{10} = \frac{12-17}{18.03} = -0.27732$$

$$Z_{11} = \frac{8-17}{18.03} = -0.49917$$

$$Z_{12} = \frac{14-17}{18.03} = -0.16639$$

سوال ۴)

فرمول عمومی برای فاصله‌ها (رابطه فاصله Minkowski)

$$d(p,q) = \sqrt[r]{\sum_{i=1}^{n} |q_i - p_i|^r}$$

محاسبه فاصله اقلیدسی: (r=2)

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

$$d(p,q) = \sqrt{(22-20)^2 + (1-0)^2 + (42-36)^2 + (10-8)^2} = \sqrt{45} = 6.7$$

محاسبه فاصله Manhattan: با توجه به فرمول داخل کتاب داریم: (r=1)

$$d(p,q) = \sum_{i=1}^{n} |q_i - p_i|$$

$$d(p,q) = |22-20| + |1-0| + |42-36| + |10-8| = 11$$

محاسبه فاصله Minkowski وقتی r=3

$$d(p,q) = \sqrt[3]{\sum_{i=1}^{n} |q_i - p_i|^3}$$

$$d(p,q) = \sqrt[3]{|22-20|^3 + |1-0|^3 + |42-36|^3 + |10-8|^3} = \sqrt[3]{233} = 6.15$$

محاسبه فاصله *supremum*

$$\lim_{r\to\infty} \sqrt[r]{\sum_{i=1}^{n} |q_i - p_i|^r} = \max_{p} |x_i - x_j| = 6$$