BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY
COMPUTING FACULTY
SCHOOL OF RESEARCH AND POSTGRADUATE STUDIES

MASTER OF SCIENCE DEGREE IN SOFTWARE ENGINEERING
MACHINE LEARNING

Individual Assignment-II

By:
Sisay Negash (BDU1500286)

Submitted to:
Gebeyehu B. (Dr. of Eng) Associate Professor

May 4, 2023
Bahir Dar, Ethiopia

1. What are the types of attributes that describe the height of a person in centimeters?

- Nominal
- Ordinal
- Interval-scaled
- ratio-scaled

Answer

The attribute that measures a person's height in centimeters can be classified as **ratio-scaled** due to its possession of a true zero point, where a value of zero denotes the total lack of the measured attribute. Hence, a height of 0 cm signifies the absence of height. Furthermore, ratio scale permits the calculation of meaningful ratios between values via multiplication or division. For instance, a person who has a height of 180 cm is twice as tall as an individual who stands at 90 cm.

2. In the Olympic games, three types of medals are awarded: bronze, silver, or gold. To describe these medals, which types of the attributes should be used? why?

- Nominal
- Ordinal
- Interval-scaled
- ratio-scaled

Answer

The types of attributes that should be used to describe the medals awarded in the Olympic Games are **ordinal** attributes. This is because the medals have a clear and predefined order, with gold being the highest, silver being second, and bronze being third. The ranking of the medals is meaningful and significant, but the differences in the quality or value between the medals are not necessarily equal. Therefore, the medals cannot be described as interval-scaled or ratio-scaled attributes. Nominal attributes are not appropriate either, as there is an inherent order to the medals that cannot be ignored. Ordinal data, on the other hand, refers to data that has a natural ordering or rank. In other words, it is data that can be ordered or arranged in a sequence. Examples of ordinal data include student grades (A, B, C, etc.), levels of education (high school, college, graduate school), and star ratings (1 star to 5 stars).

3. It is important to define or select similarity measures in data analysis. However, there is no commonly accepted subjective similarity measure. Result ca vary depending on the similarity measures used. Nonetheless, seemingly different similarity measures may be equivalent after some transformation. Suppose we have the following two-dimensional data set.

	A1	A2
X1	1.5	1.7
X2	2	1.9
X3	1.6	1.8
X4	1.2	1.5
X5	1.5	1.0

A. Consider the data as two-dimensional data points. Given a new data point, x= (1.4,1.6) as a query, rank the database points based on similarity with the query using Euclidean distance, Manhattan distance, supremum distance, and cosine similarity.

Answer

To rank the database points based on similarity with the query using different distance measures and cosine similarity, we need to calculate the distance/similarity between the query point and each of the database points. Here are the calculations:

Euclidean distance:

d(x, x1) = sqrt((1.4 - 1.5)^2 + (1.6 - 1.7)^2) = 0.1414
d(x, x2) = sqrt((1.4 - 2)^2 + (1.6 - 1.9)^2) = 0.6708
d(x, x3) = sqrt((1.4 - 1.6)^2 + (1.6 - 1.8)^2) = 0.2828
d(x, x4) = sqrt((1.4 - 1.2)^2 + (1.6 - 1.5)^2) = 0.2236
d(x, x5) = sqrt((1.4 - 1.5)^2 + (1.6 - 1)^2) = 0.6082

Ranking based on Euclidean distance: x1, x4, x3, x5, x2

Manhattan distance:

d(x, x1) = |1.4 - 1.5| + |1.6 - 1.7| = 0.2
d(x, x2) = |1.4 - 2| + |1.6 - 1.9| = 0.9
d(x, x3) = |1.4 - 1.6| + |1.6 - 1.8| = 0.4
d(x, x4) = |1.4 - 1.2| + |1.6 - 1.5| = 0.3
d(x, x5) = |1.4 - 1.5| + |1.6 - 1| = 0.7

Ranking based on Manhattan distance: x1, x4, x3, x5, x2

Supremum distance:

d(x, x1) = max(|1.4 - 1.5|, |1.6 - 1.7|) = 0.1
d(x, x2) = max(|1.4 - 2|, |1.6 - 1.9|) = 0.6
d(x, x3) = max(|1.4 - 1.6|, |1.6 - 1.8|) = 0.2
d(x, x4) = max(|1.4 - 1.2|, |1.6 - 1.5|) = 0.2
d(x, x5) = max(|1.4 - 1.5|, |1.6 - 1|) = 0.6

Ranking based on Supremum distance: x1, x3, x4, x2, x5 or x1, x4, x3, x2, x5; (x3=x4 and x2=x5)

Cosine similarity:

cos(x, x1) = (1.4*1.5 + 1.6*1.7) / (sqrt(1.4^2 + 1.6^2) * sqrt(1.5^2 + 1.7^2)) = **0.99999**
cos(x, x2) = (1.4*2 + 1.6*1.9) / (sqrt(1.4^2 + 1.6^2) * sqrt(2^2 + 1.9^2)) = **0.99575**
cos(x, x3) = (1.4*1.6 + 1.6*1.8) / (sqrt(1.4^2 + 1.6^2) * sqrt(1.6^2 + 1.8^2)) = **0.99997**
cos(x, x4) = (1.4*1.2 + 1.6*1.5) / (sqrt(1.4^2 + 1.6^2) * sqrt(1.2^2 + 1.5^2)) = **0.99903**
cos(x, x5) = (1.4*1.5 + 1.6*1) / (sqrt(1.4^2 + 1.6^2) * sqrt(1.5^2 + 1^2)) = **0.96536**

Ranking based on Cosine similarity: x1, x3, x4, x2, x5

B. Normalize the data set to make the norm of each data point equal to 1. Use Euclidean distance on the transformed data to rank the data points.

Answer

To normalize the data set, we need to calculate the norm of each data point and divide each value by the norm. Then we can use Euclidean distance on the transformed data to rank the data points.

Calculating the norm of each data point:
Norm(x1) = sqrt(1.5^2 + 1.7^2) = 2.225
Norm(x2) = sqrt(2^2 + 1.9^2) = 2.758
Norm(x3) = sqrt(1.6^2 + 1.8^2) = 2.408
Norm(x4) = sqrt(1.2^2 + 1.5^2) = 1.920
Norm(x5) = sqrt(1.5^2 + 1^2) = 1.802

Transforming the data set:
x1` = (1.5/2.225, 1.7/2.225) = (0.674, 0.764)
x2` = (2/2.758, 1.9/2.758) = (0.725, 0.688)
x3` = (1.6/2.408, 1.8/2.408) = (0.664, 0.747)
x4` = (1.2/1.920, 1.5/1.920) = (0.625, 0.781)
x5` = (1.5/1.802, 1/1.802) = (0.832, 0.555)

	A1	A2
x1`	0.674	0.764
x2`	0.725	0.688
x3`	0.664	0.747
x4`	0.625	0.781
x5`	0.832	0.555

Using Euclidean Distance on transformed data:
Distance with query point (1.4, 1.6). When we normalized the given query points, we get a value of x=(0.658, 0.752).

x1_new = sqrt((0.674-0.658)^2 + (0.764-0.752)^2) = 0.020
x2_new = sqrt((0.725-0.658)^2 + (0.688-0.752)^2) = 0.0927
x3_new = sqrt((0.664-0.658)^2 + (0.747-0.752)^2) = 0.0078
x4_new = sqrt((0.625-0.658)^2 + (0.781-0.752)^2) = 0.0462
x5_new = sqrt((0.832-0.658)^2 + (0.555-0.752)^2) = 0.2628

Ranking based on smallest distance: **x3_new, x1_new, x4_new, x2_new, x5_new**