Homework 1

699 A1, Spring 2025

Due: 2/3

**Problem 1 (10 points).** Consider the following two objects with 7 binary attributes:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| OID | A1 | A2 | A3 | A4 | A5 | A6 | A7 |
| O1 | P | N | P | P | N | P | N |
| O2 | P | N | N | P | N | N | P |

1. Calculate the distance between O1 and O2 assuming all attributes are symmetric attributes.
2. Calculate the distance O1 and O2 assuming all attributes are asymmetric attributes with P being more important than N.

Problem 2 (10 points)

Consider the following dataset with two objects.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Object | A1 | A2 | A3 | A4 |
| O1 | 1 | second | gold | Small |
| O2 | 4 | third | silver | large |

Here, all attributes are ordinal attributes and ranks of their values are shown below (lowest rank on the left):

A1: {1, 2, 3, 4, 5}

A2: {first, second, third} A3: {bronze, silver, gold}

A4: {small, medium, large, xlarge}

Calculate the distance between O1 and O2 using the method discussed in the class. Use the Euclidean distance measure.

**Problem 3 (10 points).** Consider the following dataset:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| OID | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 |
| O1 | 1 | 2 | 4 | 1 | 3 | 1 | 3 | 1 | 2 | 2 |
| O2 | 2 | 2 | 3 | 5 | 0 | 4 | 0 | 3 | 5 | 2 |
| O3 | 2 | 0 | 4 | 2 | 2 | 3 | 2 | 1 | 3 | 4 |

1. Calculate the cosine similarity between O1 and O2, *cosine*(O1, O2).
2. Calculate the cosine similarity between O1 and O3, *cosine*(O1, O3).
3. Is O1 closer to O2 or O3?

You must do all calculations yourself.

**Problem 4 (10 points).** Consider the following dataset, which has attributes of mixed types.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Object ID | A1 | A2 | A3 | A4 | A5 | A6 | A7 |
| O1 | 19 | 1 | No | No | Yes | Low | mild |
| O2 | 42 | 1 | Yes | No | Yes | High | cold |
| O3 | 28 | 0 | No | Yes | No | Low | hot |
| O4 | 35 | 0 | Yes | No | No | Middle | mild |
| O5 | 63 | 1 | No | No | No | High | hot |
| O6 | 27 | 0 | Yes | No | No | High | mild |
| O7 | 82 | 1 | No | Yes | No | Low | cool |
| O8 | 36 | 1 | No | No | Yes | High | mild |
| O9 | 12 | 0 | Yes | No | Yes | High | hot |

* + A1 is a numeric attribute.
  + A2 and A3 are symmetric binary attributes.
  + A4 and A5 are asymmetric binary attributes, where Yes is more important than No
  + A6 is a categorical (nominal) attribute.
  + A7 is an ordinal attribute. The order of values is {cold, cool, mild, hot}, where cold has the lowest rank and high has the highest rank.

Calculate the distance between O1 and O2, *d*(O1, O2), and the distance between O1 and O3, *d*(O1, O3), using the method that we discussed in the class. Is O1 closer to O2 or closer to O3? You must do all calculations yourself.

You must do all calculations yourself.

Include all answers in a single Word or PDF document and upload it to Blackboard. Use

*LastName\_FirstName*\_*hw*1.*docx* or *LastName\_FirstName*\_*hw*1.*pdf* as the file name. If you have

additional files, such as an Excel file or a R code file, then combine all of them into a single archive file and name it *LastName\_FirstName*\_*hw*1.*EXT*, where *EXT* is an appropriate archive file extension such as *zip* or *rar*.