Course: CSC522, Automated Learning and Data Analysis
Homework 1
Student: Xusheng Xiao

---

1. **Problem 1 (6 points).** Please classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning then in doubt.

   (a) Time in terms of AM or PM
   Discrete and interval.

   (b) Brightness as measured by a light meter
   Continuous and ratio.

   (c) Brightness as measured by peoples judgments
   Discrete and ordinal. People may specify it like "dark, bright, very bright". Since it still has an way to order the brightness, it should be "discrete and ordinal"

   (d) Angles as measured in degrees between 0 and 360.
   Discrete and radio.

   (e) Bronze, Silver, and Gold medals as awarded at the Olympics
   Discrete and nominal.

   (f) Height above sea level.
   Continuous and ratio.

   (g) Number of patients in a hospital
   Discrete and radio.

   (h) ISBN numbers for books. (Look up the format on the Web.)
   Discrete and nominal.

   (i) Ability to pass light in terms of the following values: opaque, translucent, transparent.
   Discrete and ordinal.

   (j) Military rank
   Discrete and ordinal.

   (k) Distance from the center of campus.
   Continuous and ratio.

   (l) Density of a substance in grams per cubic centimeter.
   Continuous and ratio.

   (m) Coat check number. (When you attend an event and you give your coat in exchange to a number that you can use to claim your coat when you leave.)
   Discrete and nominal.

2. **Problem 2 (5 points).** For the following vectors, x and y, calculate the indicated similarity or distance measures.

   (a) X=(1,1,1,1), Y=(2,2,2,2): cosine, correlation coefficient, Euclidean distance

   cosine distance:

   $x \cdot y = 1*2+1*2+1*2+1*2 = 8$

   $||x|| = \sqrt{1*1+1*1+1*1+1*1} = 2$

   $||y|| = \sqrt{2*2+2*2+2*2+2*2} = 4$

   $cos(x,y) = \frac{x \cdot y}{||x||||y||} = \frac{8}{2*4} = 1$

   correlation coefficient:

   $\bar{x} = \frac{1}{4}\sum_{k=1}^{4} x_k = 1, \bar{y} = \frac{1}{4}\sum_{k=1}^{4} y_k = 2$

   $covariance(x,y) = s_{xy} = \frac{1}{4-1}\sum_{k=1}^{4}(x_k - \bar{x})(y_k - \bar{y}) = 0$

   $standard\_deviation(x) = s_x = \sqrt{\frac{1}{4-1}\sum_{k=1}^{4}(x_k - \bar{x})^2} = 0$

   $standard\_deviation(y) = s_y = \sqrt{\frac{1}{4-1}\sum_{k=1}^{4}(y_k - \bar{y})^2} = 0$

   $corr(x,y) = \frac{covariance(x,y)}{standard\_deviation(x) * standard\_deviation(y)} = \frac{s_{xy}}{s_x * s_y} = \frac{0}{0}$

   Euclidean distance $= \sqrt{(1-2)^2+(1-2)^2+(1-2)^2+(1-2)^2} = 2$

   (b) X=(0,1,0,1), Y=(1,0,1,0): cosine, correlation coefficient, Euclidean distance, Jaccard coefficient

   cosine distance:

   $x \cdot y = 0*1+1*0+0*1+1*0 = 0$

   $||x|| = \sqrt{0*0+1*1+0*0+1*1} = \sqrt{2}$

   $||y|| = \sqrt{1*1+0*0+1*1+0*0} = \sqrt{2}$

   $cos(x,y) = \frac{x \cdot y}{||x||||y||} = \frac{0}{\sqrt{2}*\sqrt{2}} = 0$

   correlation coefficient:

   $\bar{x} = \frac{1}{4}\sum_{k=1}^{4} x_k = 0.5, \bar{y} = \frac{1}{4}\sum_{k=1}^{4} y_k = 0.5$

   $covariance(x,y) = s_{xy} = \frac{1}{4-1}\sum_{k=1}^{4}(x_k - \bar{x})(y_k - \bar{y}) = -\frac{1}{3}$

   $standard\_deviation(x) = s_x = \sqrt{\frac{1}{4-1}\sum_{k=1}^{4}(x_k - \bar{x})^2} = \sqrt{\frac{1}{3}}$

   $standard\_deviation(y) = s_y = \sqrt{\frac{1}{4-1}\sum_{k=1}^{4}(y_k - \bar{y})^2} = \sqrt{\frac{1}{3}}$

$corr(x,y) = \dfrac{covariance(x,y)}{standard\_deviation(x) * standard\_deviation(y)} = \dfrac{s_{xy}}{s_x * s_y} = -1$

Euclidean distance $= \sqrt{(0-1)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2} = 2$

Jaccard coefficient $= \dfrac{f_{11}}{f_{01} + f_{10} + f_{11}} = \dfrac{0}{2+2+0} = 0$

(c) X=(0,-1,0,1), Y=(1,0,-1,0): cosine, correlation, Euclidean distance

cosine distance:

$x \cdot y = 0*1 + -1*0 + 0*-1 + 1*0 = 0$

$||x|| = \sqrt{0*0 + -1*-1 + 0*0 + 1*1} = \sqrt{2}$

$||y|| = \sqrt{1*1 + 0*0 + -1*-1 + 0*0} = \sqrt{2}$

$cos(x,y) = \dfrac{x \cdot y}{||x||||y||} = \dfrac{0}{2} = 0$

correlation coefficient:

$\bar{x} = \dfrac{1}{4}\sum_{k=1}^{4} x_k = 0, \bar{y} = \dfrac{1}{4}\sum_{k=1}^{4} y_k = 0$

$covariance(x,y) = s_{xy} = \dfrac{1}{4-1}\sum_{k=1}^{4}(x_k - \bar{x})(y_k - \bar{y}) = 0$

$standard\_deviation(x) = s_x = \sqrt{\dfrac{1}{4-1}\sum_{k=1}^{4}(x_k - \bar{x})^2} = 0$

$standard\_deviation(y) = s_y = \sqrt{\dfrac{1}{4-1}\sum_{k=1}^{4}(y_k - \bar{y})^2} = 0$

$corr(x,y) = \dfrac{covariance(x,y)}{standard\_deviation(x) * standard\_deviation(y)} = \dfrac{s_{xy}}{s_x * s_y} = \dfrac{0}{0}$

Euclidean distance $= \sqrt{(0-1)^2 + (-1-0)^2 + (0-(-1))^2 + (1-0)^2} = 2$

(d) X=(1,1,0,1,0,1), Y=(1,1,1,0,0,1): cosine, correlation coefficient, Jaccard coefficient

cosine distance:

$x \cdot y = 1*1 + 1*1 + 0*1 + 1*0 + 0*0 + 1*1 = 3$

$||x|| = \sqrt{1*1 + 1*1 + 0*0 + 1*1 + 0*0 + 1*1} = 2$

$||y|| = \sqrt{1*1 + 1*1 + 1*1 + 0*0 + 0*0 + 1*1} = 2$

$cos(x,y) = \dfrac{x \cdot y}{||x||||y||} = \dfrac{3}{2*2} = \dfrac{3}{4}$

correlation coefficient:

$\bar{x} = \dfrac{1}{6}\sum_{k=1}^{4} x_k = \dfrac{2}{3}, \bar{y} = \dfrac{1}{6}\sum_{k=1}^{4} y_k = \dfrac{2}{3}$

$covariance(x,y) = s_{xy} = \dfrac{1}{6-1}\sum_{k=1}^{6}(x_k - \bar{x})(y_k - \bar{y}) = \dfrac{1}{15}$

$standard\_deviation(x) = s_x = \sqrt{\frac{1}{6-1}\sum_{k=1}^{6}(x_k - \bar{x})^2} = \sqrt{\frac{4}{15}}$
$standard\_deviation(y) = s_y = \sqrt{\frac{1}{6-1}\sum_{k=1}^{6}(y_k - \bar{y})^2} = \sqrt{\frac{4}{15}}$
$corr(x,y) = \frac{covariance(x,y)}{standard\_deviation(x) * standard\_deviation(y)} = \frac{s_{xy}}{s_x * s_y} = \frac{1}{4}$

Jaccard coefficient $= \frac{f_{11}}{f_{01} + f_{10} + f_{11}} = \frac{3}{1+1+3} = \frac{3}{5}$

(e) X=(2,-1,0,2,0,-3), Y=(-1,1,-1,0,0,-1): cosine, correlation coefficient
cosine distance:
$x \cdot y = 2 * -1 + -1 * 1 + 0 * -1 + 2 * 0 + 0 * 0 + -3 * -1 = 0$
$||x|| = \sqrt{2 * 2 + -1 * -1 + 0 * 0 + 2 * 2 + 0 * 0 + -3 * -3} = \sqrt{18}$
$||y|| = \sqrt{-1 * -1 + 1 * 1 + -1 * -1 + 0 * 0 + 0 * 0 + -1 * -1} = 2$
$cos(x,y) = \frac{x \cdot y}{||x||||y||} = \frac{0}{18 * 2} = 0$

correlation coefficient:
$\bar{x} = \frac{1}{6}\sum_{k=1}^{4} x_k = 0, \bar{y} = \frac{1}{6}\sum_{k=1}^{4} y_k = -\frac{1}{3}$
$covariance(x,y) = s_{xy} = \frac{1}{6-1}\sum_{k=1}^{6}(x_k - \bar{x})(y_k - \bar{y}) = 0$
$standard\_deviation(x) = s_x = \sqrt{\frac{1}{6-1}\sum_{k=1}^{6}(x_k - \bar{x})^2} = \sqrt{\frac{18}{5}}$
$standard\_deviation(y) = s_y = \sqrt{\frac{1}{6-1}\sum_{k=1}^{6}(y_k - \bar{y})^2} = \sqrt{\frac{2}{3}}$
$corr(x,y) = \frac{covariance(x,y)}{standard\_deviation(x) * standard\_deviation(y)} = \frac{s_{xy}}{s_x * s_y} = 0$

3. **Problem 3 (6 points).** Show that the set difference metric given by $D(A,B) := size(A-B) + size(B-A)$ satisfies the metric axioms given on pages 70/71 of our textbook. Here, A and B are sets, and $A-B$ indicates the set difference.

Theorem 1 *$D(A,B) := size(A-B) + size(B-A)$ satisfies the metric axioms, where A and B are sets, and A-B indicates the set difference.*

Proof. To prove $D(A,B) := size(A-B)+size(B-A)$ satisfies the metric axioms, we need to show three properties of metrics hold for $D(A,B)$.

(a) **Positivity.**

i. $D(A, A) \geq 0$ for all A. Since $A - B$ indicates the set difference, $A - A = \emptyset \Rightarrow size(A - A) = 0$.
As a result, $D(A, A) := size(A - A) + size(A - A) = 0 + 0 = 0 \Rightarrow D(A, A) \geq 0$.

ii. $D(A, B) = 0$ only if $A = B$. Suppose $A \neq B$, then $A - B \neq \emptyset \Rightarrow size(A - B) > 0, size(B - A) > 0$.
As a result, $D(A, B) := size(A - B) + size(B - A) > 0 \Rightarrow D(A, B) \neq 0$.
Since $D(A, A) \geq 0$ for all A, $D(A, B) = 0$ only if $A = B$.

(b) **Symmetry.**
$D(A, B) = D(B, A)$ for all x and y. Since $D(A, B) = size(A - B) + size(B - A)$, $D(B, A) = size(B - A) + size(A - B) = D(A, B)$.

(c) **Triangle Inequality.**
$D(A, C) \leq D(A, B) + D(B, C)$ for all set A, B, and C.
By De Morgan, $D(A, B) = size(A - B) + size(B - A) = size(A) + size(B) - 2size(A \bigcap B)$
$D(B, C) = size(B - C) + size(C - B) = size(B) + size(C) - 2size(B \bigcap C)$
$D(A, C) = size(A-C)+size(C-A) = size(A)+size(C)-2size(A \bigcap C)$.
$D(A, B)+D(B, C)-D(A, C) = size(A)+size(B)-2size(A \bigcap B)+size(B)+size(C)-2size(B \bigcap C)-size(A)-size(C)+2size(A \bigcap C)$
$\Rightarrow D(A, B) + D(B, C) - D(A, C) = 2size(B) - 2size(A \bigcap B) - 2size(B \bigcap C) + 2size(A \bigcap C)$
By De Morgan, $size(B)+size(A \bigcap B \bigcap C) \geq size(A \bigcap B)+size(B \bigcap C)$.
Since $size(A \bigcap C) \geq size(A \bigcap B \bigcap C)$, $size(B)+size(A \bigcap C) \geq size(A \bigcap B)+size(B \bigcap C)$.
$\Rightarrow 2size(B) - 2size(A \bigcap B) - 2size(B \bigcap C) + 2size(A \bigcap C) \geq 0$
$\Rightarrow D(A, B) + D(B, C) - D(A, C) \geq 0 \Rightarrow D(A, C) \leq D(A, B) + D(B, C)$
Thus, $D(A, C) \leq D(A, B) + D(B, C)$ for all set A, B, and C.

Since all these three properties hold for $D(A, B) := size(A-B)+size(B-A)$, $D(A, B)$ satisfies the metric axioms.

■

4. **Problem 4 (6 points).** Describe how you would create visualizations to .display information that describes

(a) Computer networks. Be sure to include both the static aspects of the network, such as connectivity, and the dynamic aspects, such as traffic.

(b) The distribution of specific plant and animal species around the world for a specific moment in time.

(c) The use of computer resources such as processor time, main memory, and disk for asset of benchmark database programs.

In your answers, please address the following issues:

- Representation: how will you map objects, attributes, and relationships to visual elements?
- Arrangement: are there special considerations that need to be taken into account with respect to how visual elements are displayed? E.g. choice of viewpoint, use of transparency, etc.
- Selection: how will you handle a large number of attributes and data objects?

5. **Problem 5 (5 points).**

   (a) Describe how a box plot can give information about whether the value of an attribute is symmetrically distributed. What can you say about the symmetry of the distribution of the attributes shown in Figure 3.11 on page 115 of our textbook.

   (b) Compare sepal length, sepal width, petal length, and petal width using Figure 3.12 on page 115.