Data Science

Problem sheet 1

Ex 1

Determine the type of the following attributes in two ways:
1: Continuous, discrete, binary? 2: Nominal, ordinal, quantitative (interval, ratio)?

Solution

1. **Altitude** Continuous, ratio
2. **Total number of rooms in a hotel** Discrete, ratio
3. **Military ranks** Discrete, ordinal
4. **Distance from the center of Heroes Square** Continuous, ratio
5. **International Standard Book Number (ISBN)** Discrete, nominal
6. **Degree: measurement of plain angle (between 0 and 360)** Continuous, ratio
7. **Degree of transparency: transparent, translucent, opaque** Discrete, ordinal
8. **Cloakroom ticket numbers** Discrete, interval
9. **Grades (from F to A+)** Discrete, ordinal
10. **Medals (bronze, silver, gold)** Discrete, ordinal
11. **Sex (male, female)** Binary, nominal
12. **Age (in years)** Discrete (maybe continuous), ratio
13. **pH (acidity or basicity of an aqueous solution)** Continuous, interval

Ex 2

Prove the following statements:

1. $L_1(x, y) = \sum_{i=1}^{d} |x_i - y_i|$ is a distance metric,
2. $L_2^2(x, y) = \sum_{i=1}^{d} (x_i - y_i)^2$ is **not** a distance metric.

Solution

1. Let's check the definition of a metric:
   - Non-negative: absolute value function is non-negative: $|a - b| \geq 0$ ✓
   - Symmetric: $|a - b| = |b - a|$ ✓
   - $L_1(x, y) = 0$ if and only if $x = y$:
     Since $|x_i - y_i| \geq 0 \; \forall i$, the sum of non-negative terms is 0 if and only if all the terms are 0: $|x_i - y_i| = 0 \; \forall i \iff x_i = y_i \; \forall i \iff x = y$ ✓
   - Triangle-inequality also holds ✓

     $$\begin{aligned} L_1(x, y) &= \sum_{i=1}^{d} |x_i - y_i| = \sum_{i=1}^{d} |x_i - z_i + z_i - y_i| \leq \sum_{i=1}^{d} |x_i - z_i| + \sum_{i=1}^{d} |z_i - y_i| \\ &= L_1(x, z) + L_1(z, y) \end{aligned}$$

     Here we used the fact that $|a + b| \leq |a| + |b|$.

2. Let's check the definition of a metric:
   - Non-negative: quadratic function is non-negative $(a - b)^2 \geq 0$ ✓
   - Symmetric: $(a - b)^2 = (b - a)^2$ ✓
   - $L_1(x, y) = 0$ if and only if $x = y$:
     Since $(x_i - y_i)^2 \geq 0 \; \forall i$, the sum of non-negative terms is 0 if and only if all the terms are 0: $(x_i - y_i)^2 = 0 \; \forall i \iff x_i = y_i \; \forall i \iff x = y$ ✓
   - The triangle-inequality **does not** hold ✗
     1-dimensional counterexample: Let $x = 0$, $y = 2$, and $z = 1$. Then

     $$4 = (0 - 2)^2 = L_2^2(0, 2) \not\leq L_2^2(0, 1) + L_2^2(1, 2) = 1 + 1 = 2$$

     2-dimensional counterexample: Let $x = (0, 0)$, $y = (2, 0)$, and $z = (1, 0)$. Then

     $$\begin{aligned} 4 = (0 - 2)^2 + (0 - 0)^2 = L_2^2\big((0, 0), (2, 0)\big) &\not\leq L_2^2\big((0, 0), (1, 0)\big) + L_2^2\big((1, 0), (2, 0)\big) \\ &= (0 - 1)^2 + (0 - 0)^2 + (1 - 2)^2 + (0 - 0)^2 \\ &= 1^2 + 0^2 + 1^2 + 0^2 = 2 \end{aligned}$$

Ex 3

Let two feature vectors contain the following attributes:

- person's height (between 5 and 6 feet)
- person's weight (between 90 and 260 lbs)
- person's annual income (between 10,000 and 1 Million dollars)

How would you calculate the distance between the two vectors? What kind of transformations would you apply, and which distance would you choose?

Solution

Use max-min standardization and take the logarithm of the income.

1. Use max-min standardization:

$$\frac{X_1 - \min(X_1)}{\max(X_1) - \min(X_1)} = \frac{X_1 - 5}{6 - 5} = X_1 - 5$$

2. Use max-min standardization:

$$\frac{X_2 - \min(X_2)}{\max(X_2) - \min(X_2)} = \frac{X_2 - 90}{260 - 90} = \frac{X_2 - 90}{170}$$

3. First calculate the base-10 logarithm of the income, then apply max-min:

$$\frac{\log_{10}(X_3) - \min\big(\log_{10}(X_3)\big)}{\max\big(\log_{10}(X_3)\big) - \min\big(\log_{10}(X_3)\big)} = \frac{\log_{10}(X_3) - 4}{6 - 4} = \frac{\log_{10}(X_3) - 4}{2}$$

   Why do we compute the logarithm of the income component?
   The idea is to emphasize the "diminishing marginal utility" of transforming income into human capabilities. This means that the concave logarithmic transformation makes clearer the notion that an increase of income by $10,000$ for somebody whose annual income is $50,000$ has a much greater impact on the standard of living than the same $10,000$ increase for a person whose income is $500,000$.

After standardization any distance metric can be used (e.g. Minkowski distance), while Mahalanobis distance is an even better choice since it also takes into consideration the covariance structure of the data (since height and weight is probably highly correlated).

Ex 4

Consider the following three documents:

- d_1: "ant bee"
- d_2: "dog bee hog ant"
- d_3: "cat gnu dog eel fox"

These documents can be represented by 8-dimensional vectors in a so-called *document-term matrix*, which describes the frequency of terms that occur in a collection of documents:

	ant	bee	cat	dog	eel	fox	gnu	hog
d_1	1	1	0	0	0	0	0	0
d_2	1	1	0	1	0	0	0	1
d_3	0	0	1	1	1	1	1	0

1. Calculate simple matching coefficient (SMC) and Jaccard-coefficient of d_1 and d_2.
2. Determine distances derived from these coefficients. Which is the better coefficient to handle the problem? Why?

Note.: Of course, this approach cannot distinguish "John is quicker than Mary" and "Mary is quicker than John" documents.

Solution

$$\mathrm{SMC}(d_1, d_2) = \frac{M_{11} + M_{00}}{M_{11} + M_{10} + M_{01} + M_{00}} = \frac{6}{8} \Longrightarrow \mathrm{SMC}_{\mathrm{dist}}(d_1, d_2) = 1 - \frac{6}{8} = \frac{1}{4}$$

$$\mathrm{Jacc}(d_1, d_2) = \frac{M_{11}}{M_{11} + M_{01} + M_{10}} = \frac{2}{4} \Longrightarrow \mathrm{Jacc}_{\mathrm{dist}}(d_1, d_2) = 1 - \frac{2}{4} = \frac{1}{2}$$

The Jaccard index may be more reasonable since zeros and ones have asymmetric meaning and the fact that none of the documents contain a certain word does not make them more similar while the fact that a word appears in both of the document is more important regarding calculating the similarity.

Ex 5

Let $tf_{i,j}$ denote the entry in the i-th row and j-th column of the document-term matrix from the previous task. For instance, $tf_{1,1} = 1$, where row = d_1 and column = ant. Consider the following tf–idf transformation.

- Let m be the number of documents.
- Let df_j denote the *document frequency*, i.e. the number of non-zero elements in the j-th column (number of documents containing the j-th word). E.g. $df_1 = 2$.
- Let idf_j denote the *inverse document frequency*, which is defined as follows:

$$idf_j = \log\left(\frac{m}{df_j}\right)$$

With these notations the tf–idf tranformation is defined as follows:

$$tf\text{–}idf_{i,j} = tf_{i,j} \cdot idf_j = tf_{i,j} \cdot \log\left(\frac{m}{df_j}\right)$$

The tf–idf abbreviation stands for term frequency – inverse document frequency. What is the impact of this transformation? In case of a concrete, real document-term matrix, what could be the purpose of this transformation?

Solution

The transformation is intended to reflect how important a word is to a document in a collection or corpus. The tf–idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general (e.g. articles, auxiliary verbs). Multiplying the term frequencies by the inverse document frequency factor is incorporated which diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.

Ex 6

The following table's rows correspond to customers (A, B, C), and the columns correspond to products $(a, b, \ldots, h)$. The table contains 1 if a given costumer bought the given item, 0 otherwise. Determine the Jaccard similarity and the Cosine similarity of A and B.

	a	b	c	d	e	f	g	h
A	1	1	0	1	1	0	1	1
B	0	1	1	1	1	1	1	0
C	1	0	1	1	0	1	1	1

Solution

$$\text{Jacc}(A, B) = \frac{4}{8} = \frac{1}{2} \qquad \cos(A, B) = \frac{1+1+1+1}{\sqrt{6}\sqrt{6}} = \frac{4}{6}$$

Ex 7

Assuming that the cost of compression/stretching is 0, determine the DTW distance of the following time series (let the inner distance function be the absolute distance)! Find optimal alignment between two time series (the warping path)!

$$t_1 = (3, 2, 5, 7, 8, 9), \qquad t_2 = (2, 3, 2, 3, 6, 8)$$

Solution

	2	**3**	**2**	**3**	**6**	**8**
3	1	1	2	2	5	10
2	1	2	1	2	6	11
5	4	3	4	3	3	6
7	9	7	8	7	4	4
8	15	12	13	12	6	4
9	22	18	19	18	9	5