Gower Distance

TODO: citations
When clustering data with algorithms such as *k-means* or *k-medoids*, the distance between datapoints is usually calculated by measures such as the Euclidean or Manhatten norm. These distance measures come with constraints though, since they are only defined for numerical variables. In the scope of this work, we are facing data points with mixed variable types, therefore we investigated the **Gower distance measure** as an appropriate measure to calculate an overall *similarity* between mixed data points. The Gower distance measure distinguishes between three types of variables:

Binary variables, where two variables of value 0 are *not* considered a match [**gower1971general**]. In the scope of this work, the only considerable candidate for a binary variable would have been *gender*. Though, during our research process, we concluded that the similarity measure for binary variables was not a suitable choice, as the distance does not match 0 values. Hence, we consider all variables, that are not explicitly numerical, as categorical variables.
Categorical variables form a set of unordered values and are comparable to ENUMs in programming languages.
Numerical variables hold ordered numerical values that support arithmetic operations.

Calculating the distances

Given two data points x and y, that each form a tuple of v variables of arbitrary type, the similarity coefficient between the two points is given by

$$S_{xy} = \sum_{k=1}^{v} s_{xy,k} / \sum_{k=1}^{v} \delta_{xy,k} \tag{1}$$

where $s_{xy,k}$ denotes a *score for the similarity* of the two variables at position k in the data points x and y. Note, that the definition of the score dependends on the type of the variable, as defined below. In the divisor, $\delta_{xy,k}$ basically represents the possibility of comparing the two variables at index k, such that it evaluates to 1, when the two variables are comparable and to 0 when not. For example, variables are not comparable, if values are undefined in the data points or the variable types do not match at index k, among other reasons. Within this work, the dataset is complete, therefore, $\sum_{k=1}^{v} \delta_{xy,k} = v$, the number of considered variables in each data point, for all x, y in the data set. Thus, the similarity coefficient in (1) can be interpreted as the average value of all similarity scores.

With respect to the variable type, the similarity score $s_{xy,k}$ is defined as follows:

- **Binary:** The score for binary variables is basically the result of an logical AND operation. As pointed above, 0 values are not considered a match and even further, not considered to be comparable. Hence, the values result as in the table

i	1	1	0	0
j	1	0	1	0
$s_{xy,k}$	1	0	0	0
$\delta_{xy,k}$	1	1	1	0

- **Categorical:** The similarity score of categorical variables is 1, if the variables are completely identical in x and y and 0, if they differ.

- **Numerical:** for numerical variables, the similarity score is calculated by

$$s_{xy,k} = 1 - \frac{|x_k - y_k|}{range(k)}$$

where $range(k)$ is the total range of values, that the numerical variable at index k can accept. This can be a global range of acceptable values for variable k or chosen on the basis of the dataset.

References

[1] Gower, John C, *A general coefficient of similarity and some of its properties*, Biometrics, JSTOR pp. 857-871, 1971.