Gower Distance

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 our choice of an appropriate distance measure fell on the Gower distance measure to calculate an overall *similarity* between mixed data points. The gower similarity distance distinguishes between three types of variables:

Binary variables, where two variables of value 0 are *not* considered a match [CITE]. 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 comparible 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

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

where $s_{xy,k}$ denotes a score for the similarity of variables of index k between the data points x and y, that's definition is dependent 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, where the value 1 means, when variables are comparable and 0 when not. This could happen, if some values are not defined. Within this work, the dataset is complete, therefore, the sum of $\delta_{xy,k}$ for all $0 <= k <= v$ equals to v, the number of considered variables in each data point. Thus, the similarity coefficient can be interpreted as the average value of all similarity scores.

With respect to the variable type, $s_{xy,k}$ is defined by

- **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

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

  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.

TODO: sources, citations