Embedding-Based Multinomial Logit Model for Modeling Object Similarity

Daan Noordenbos

August 23, 2023

1. Introduction

Similarity learning is a machine learning technique that focuses on training models to quantify the similarity between objects. This technique is commonly used in various applications, such as recommendation systems, image retrieval, text analysis, and more. In the context of similarity learning, we consider a set of objects denoted by $A = \{x_1, \ldots, x_n\}$ and data describing the similarities between these objects. The goal is to construct a function $\mathrm{Dis} : A \times A \to \mathbb{R}$ that effectively captures the relatedness among the objects. Specifically, when comparing objects x_j and x_k to x_i, the function should adhere to the condition $\mathrm{Dis}(x_i, x_j) < \mathrm{Dis}(x_i, x_k)$ if x_i is more similar to x_j than to x_k.

A significant challenge in similarity learning is acquiring data that accurately represents the similarities between objects. This difficulty arises from the fact that quantifying similarity consistently is often problematic, due to the fact that the desired property of Dis remains invariant under a monotonic transformation. Consequently, it can be advantageous to use data regarding relative similarity, as this eliminates the subjectivity introduced by scales inherent in quantitative similarity. However, the literature concerning lightweight similarity models for relative similarity data is sparse. This paper addresses the gap by presenting a novel lightweight method for constructing Dis from relative similarity data. This method involves a Euclidean embedding of the set A coupled with a non-linear multinomial logit model for determining the parameters of the embedding.

2. Model

Formally, we consider the following problem: Given a set $A = \{x_1, \ldots, x_n\}$ and a dataset comprising K observations, where each observation takes the form of $(x_{i,k}, A'_k, x^*_k)$, with $x_{i,k} \in A$, $A'_k \subseteq A \setminus x_{i,k}$, and $x^*_k \in A'_k$ denoting the object in A'_k that is most similar to $x_{i,k}$. We aim to construct a function $\mathrm{Dis} : A \times A \to \mathbb{R}$ that maximizes

$$\frac{1}{K} \sum_{k=1}^{K} \mathbb{1}\{\mathrm{Dis}(x_{i,k}, x^*_k) \leq \mathrm{Dis}(x_{i,k}, x_j)\ \forall x_j \in A'_k\}\,. \tag{1}$$

In this paper, we propose to embed A into a Euclidean space and use the Euclidean norm to measure similarity. Specifically, each x_i is mapped to a point $\boldsymbol{p}_{x_i} \in \mathbb{R}^m$, and we define $\mathrm{Dis}(x_i, x_j) = \|\boldsymbol{p}_{x_j} - \boldsymbol{p}_{x_i}\|$. However, since Equation 1 is non-differentiable, gradient-based optimization methods cannot be applied. To circumvent this, we assume that there is uncertainty regarding the similarity between objects by assuming that $\mathrm{Dis}(x_i, x_j) = \|\boldsymbol{p}_{x_j} - \boldsymbol{p}_{x_i}\| + \varepsilon_{x_i,x_j}$. Now to find candidates for $\boldsymbol{p}_{x_1}, \ldots, \boldsymbol{p}_{x_n}$ we can, instead of maximizing Equation 1, maximize the likelihood of observing the data. Moreover, if we assume that the ε_{x_i,x_j}'s are i.i.d. and follow a Gumbel distribution we can use the multinomial logit model. The likelihood of each observation is then as follows

$$\begin{aligned}
\mathbb{P}((x_{i,k}, A'_k, x^*_k)) &= \mathbb{P}(\text{out of } A'_k\ x^*_k \text{ is most like } x_{i,k}) \\
&= \mathbb{P}(\mathrm{Dis}(x_{i,k}, x^*_k) \leq \mathrm{Dis}(x_{i,k}, x_j)\ \forall x_j \in A'_k) \\
&= \mathbb{P}(-\mathrm{Dis}(x_{i,k}, x^*_k) > -\mathrm{Dis}(x_{i,k}, x_j)\ \forall x_j \in A'_k \setminus x^*_k) \\
&= \frac{\exp(-\|\boldsymbol{p}_{x^*_k} - \boldsymbol{p}_{x_{i,k}}\|)}{\sum\limits_{x_j \in A'_k} \exp(-\|\boldsymbol{p}_{x_j} - \boldsymbol{p}_{x_{i,k}}\|)}.
\end{aligned}$$

With these probabilities the log-likelihood of the problem can be constructed.

$$l(\boldsymbol{p}_{x_1}, \ldots, \boldsymbol{p}_{x_n}; (x_{i,1}, A'_1, x^*_1), \ldots, (x_{i,K}, A'_K, x^*_K)) = \sum_{k=1}^{K} \log \left(\frac{\exp(-\|\boldsymbol{p}_{x^*_k} - \boldsymbol{p}_{x_{i,k}}\|)}{\sum\limits_{x_j \in A'_k} \exp(-\|\boldsymbol{p}_{x_j} - \boldsymbol{p}_{x_{i,k}}\|)} \right).$$

By maximizing the log-likelihood estimates for $\boldsymbol{p}_{x_1}, \ldots, \boldsymbol{p}_{x_n}$ can be determined. The details of this maximization process and its efficiency are discussed in the Appendix. Using $\boldsymbol{p}_{x_1}, \ldots, \boldsymbol{p}_{x_n}$, we define our proposal of Dis to maximize Equation 1. Specifically, we define

$$\mathrm{Dis}(x_i, x_j) = \frac{\|\boldsymbol{p}_{x_j} - \boldsymbol{p}_{x_i}\|}{\max\limits_{x_v, x_w \in A} \|\boldsymbol{p}_{x_v} - \boldsymbol{p}_{x_w}\|},$$

ensuring that Dis maps to the interval $[0, 1]$.

2.1 Scaling

A problem with the model is that sometimes does not have a maximizer. This occurs when there is no stochasticity in the dataset. To elaborate, if there exist $\boldsymbol{p}_{x_1}, \ldots, \boldsymbol{p}_{x_n}$ such that $\|\boldsymbol{p}_{x_{i,k}} - \boldsymbol{p}_{x^*_k}\| < \|\boldsymbol{p}_{x_{i,k}} - \boldsymbol{p}_{x_{j,k}}\|$ for all $x_j \in A'_k$ and k, then no stochasticity is present in the dataset. As a consequence, we have that the maximum of the log-likelihood is zero, and that it has no maximizers. The latter follows from the fact that for $t > 1$ we have that

$$l(t\boldsymbol{p}_{x_1}, \ldots, t\boldsymbol{p}_{x_n}; (x_{i,1}, A'_1, x^*_1), \ldots, (x_{i,K}, A'_K, x^*_K)) > l(\boldsymbol{p}_{x_1}, \ldots, \boldsymbol{p}_{x_n}; (x_{i,1}, A'_1, x^*_1), \ldots, (x_{i,K}, A'_K, x^*_K)).$$

Therefore it can be the case that the log-likelihood is not maximal, but the points $\boldsymbol{p}_{x_1}, \ldots, \boldsymbol{p}_{x_n}$ already have the right structure. To deal with this we use two stopping criteria, one regarding the change in the log-likelihood and another pertaining to the structure of the points. Specifically, if after scaling the points based on their structure they change little we have found candidates for $\boldsymbol{p}_{x_1}, \ldots, \boldsymbol{p}_{x_n}$. We propose two methods to do this scaling. The first method scales using the maximum distance between points. The stopping criteria we then get is

$$\frac{1}{n} \sum_{x_i \in A} \frac{\|\boldsymbol{v}_{\boldsymbol{x_i}}\|}{\max\limits_{x_j, x_k \in A} \|\boldsymbol{p}_{\boldsymbol{x_j}} - \boldsymbol{p}_{\boldsymbol{x_k}}\|} < \varepsilon, \text{ with } \boldsymbol{v}_{x_i} = \frac{\partial}{\partial \boldsymbol{p}_{x_i}} l(\boldsymbol{p}_{x_1}, \ldots, \boldsymbol{p}_{x_n}; (x_{i,1}, A'_1, x^*_1), \ldots, (x_{i,K}, A'_K, x^*_K)).$$

The down side of this method is that the maximum distance between points needs to be determined. This scales quadratically with the number of points and is therefore not preferable. The second method remedies this by scaling based on the average distance to the center of the points, where the center is defined as the point for which the total squared distance to the other points is minimal. The center is therefore the weighted sum of the points $\boldsymbol{p}_{x_1}, \ldots, \boldsymbol{p}_{x_n}$, because

$$\operatorname*{argmin}_{\boldsymbol{t} \in \mathbb{R}^m} \sum_{x_i \in A} \|\boldsymbol{p}_{x_i} - \boldsymbol{t}\|^2 \Rightarrow \boldsymbol{t} = \frac{1}{n} \sum_{x_i \in A} \boldsymbol{p}_{x_i}.$$

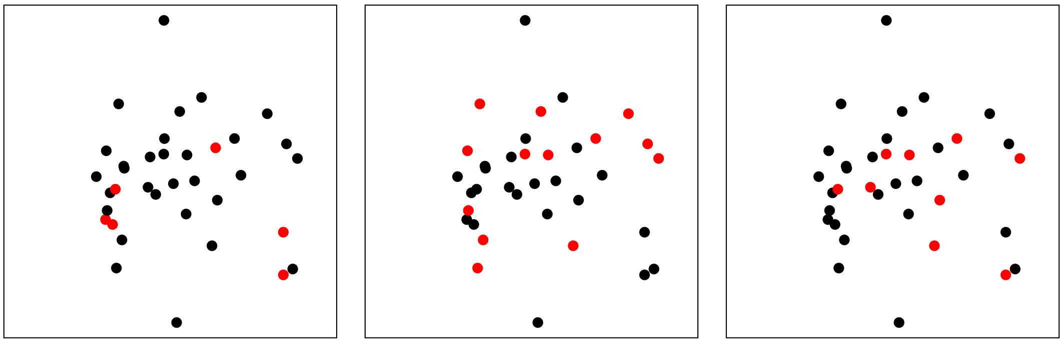
Now, because $\sum_{x_i \in A} \boldsymbol{v}_{\boldsymbol{x_i}} = \boldsymbol{0}$ (see Appendix), we have that $\frac{1}{n} \sum_{x_i \in A} \boldsymbol{p}_{x_i}$ is invariant, so we only need to center the points once. After which we get the stopping criteria

$$\frac{\sum_{x_i \in A} \|\boldsymbol{v}_{\boldsymbol{x_i}}\|}{\sum_{x_i \in A} \|\boldsymbol{p}_{\boldsymbol{x_i}}\|} < \varepsilon.$$

This criteria will trigger when no or little stochasticity is present in the data, as in those cases the log-likelihood will converge at a glacial pace.

3. Case Study

For chess players, the ability to quickly locate games featuring positions similar to the ones they are studying would be useful. As they can learn from the moves made in these games. However, to systematically identify such games, a similarity measure for chess positions is required. Such a measure does not yet exist, but a component which would probably go into it, would regard the balance of minor pieces, as this is something (strong) chess players consider when assessing how similar positions are. There are 36 distinct ways to balance minor pieces, with an example being two knights versus two bishops. To assess the similarity between various balances of minor pieces, we employed the model described above. We enlisted a player of master-level to compare the similarities among these balances, resulting in 228 observations. Utilizing this dataset, we conducted a two-dimensional embedding of the 36 balances. This embedding effectively predicts 91% of the data and yields a pseudo-R^2 of 0.71.



a) Opposite colored bishops

b) Bishop pair

c) Three pieces

Figure 1: Characteristic clusters

The embedding is visualized in Figure 1, which highlights three distinct groups of balances. These clusters reveal that certain groups exhibit higher similarities to each other than the average similarities among all balances.

Appendix

A.1 Gradient

To determine the gradient of the log-likelihood note that

$$\begin{aligned}
l(\boldsymbol{p}_{x_1},\ldots,\boldsymbol{p}_{x_n};(x_{i,1},A'_1,x^*_1),\ldots,(x_{i,K},A'_K,x^*_K)) &= \sum_{k=1}^{K}\log\left(\mathbb{P}((x_{i,k},A'_k,x^*_k))\right)\\
&= \sum_{k=1}^{K}\log\left(\frac{\exp(-\|\boldsymbol{p}_{x^*_k}-\boldsymbol{p}_{x_{i,k}}\|)}{\sum\limits_{x_j\in A'_k}\exp(-\|\boldsymbol{p}_{x_j}-\boldsymbol{p}_{x_{i,k}}\|)}\right)\\
&= \sum_{k=1}^{K}\left(-\|\boldsymbol{p}_{x^*_k}-\boldsymbol{p}_{x_{i,k}}\|-\log\left(\sum_{x_j\in A'_k}\exp(-\|\boldsymbol{p}_{x_j}-\boldsymbol{p}_{x_{i,k}}\|)\right)\right).
\end{aligned}$$

Differentiating each term in the sum yields:

$$\begin{aligned}
\frac{\partial\log\left(\mathbb{P}((x_{i,k},A'_k,x^*_k))\right)}{\partial\boldsymbol{p}_{x_j}} &= \boldsymbol{0},\text{ for } x_j\notin A'_k\cup x_{i,k}\\
\frac{\partial\log\left(\mathbb{P}((x_{i,k},A'_k,x^*_k))\right)}{\partial\boldsymbol{p}_{x_j}} &= \frac{\exp(-\|\boldsymbol{p}_{x_j}-\boldsymbol{p}_{x_{i,k}}\|)}{\sum\limits_{x_l\in A'_k}\exp(-\|\boldsymbol{p}_{x_l}-\boldsymbol{p}_{x_{i,k}}\|)}\frac{\boldsymbol{p}_{x_j}-\boldsymbol{p}_{x_{i,k}}}{\|\boldsymbol{p}_{x_j}-\boldsymbol{p}_{x_{i,k}}\|},\text{ for } x_j\in A'_k\setminus x^*_k\\
\frac{\partial\log\left(\mathbb{P}((x_{i,k},A'_k,x^*_k))\right)}{\partial\boldsymbol{p}_{x^*_k}} &= \frac{\exp(-\|\boldsymbol{p}_{x^*_k}-\boldsymbol{p}_{x_{i,k}}\|)}{\sum\limits_{x_l\in A'_k}\exp(-\|\boldsymbol{p}_{x_l}-\boldsymbol{p}_{x_{i,k}}\|)}\frac{\boldsymbol{p}_{x^*_k}-\boldsymbol{p}_{x_{i,k}}}{\|\boldsymbol{p}_{x^*_k}-\boldsymbol{p}_{x_{i,k}}\|}-\frac{\boldsymbol{p}_{x^*_k}-\boldsymbol{p}_{x_{i,k}}}{\|\boldsymbol{p}_{x^*_k}-\boldsymbol{p}_{x_{i,k}}\|}\\
\frac{\partial\log\left(\mathbb{P}((x_{i,k},A'_k,x^*_k))\right)}{\partial\boldsymbol{p}_{x_{i,k}}} &= -\sum_{x_j\in A'_k}\frac{\partial\log\left(\mathbb{P}((x_{i,k},A'_k,x^*_k))\right)}{\partial\boldsymbol{p}_{x_j}}.
\end{aligned}$$

Where we made use of $\frac{\mathrm{d}\|\boldsymbol{x}\|}{\mathrm{d}\boldsymbol{x}}=\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$. Moreover, observe that

$$\sum_{k=1}^{K}\sum_{x_j\in A}\frac{\partial\log\left(\mathbb{P}((x_{i,k},A'_k,x^*_k))\right)}{\partial\boldsymbol{p}_{x_j}}=\boldsymbol{0},$$

which has as neat consequence that the points to not drift.

A.2 Physical Analog

Using these partial derivatives the gradient of the log-likelihood can be determined. To then maximize the function we suggest to use gradient descent as we suspect that it has a strong tendency to move to the global maximum. This suspicion is founded on the fact that the above set equations corresponds to a physical system. In this system we have n particles in m dimensional space which repel and attract each other.

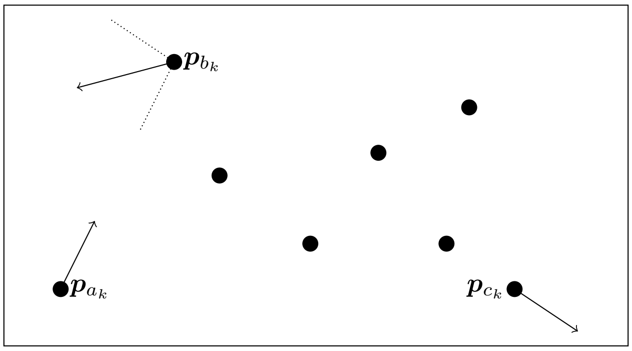


Figure 2: Effect of an observation on the gradient

This is made more clear in Figure 2 where the forces that an observation induce on the relevant set of particles are visualised. Like most physical systems, we suspect it will tend to an equilibrium. This equilibrium should correspond to a set of points which produce a nearly maximal log-likelihood.

A.3 C++ Code

Due to the lightweight nature of the model all computations can and are performed using a single short function. Below is the code for when A'_k contains two elements for all k. The proposed gradient descent is augmented with momentum term to find the equilibrium faster.

```cpp
void euclideanEmbedding(Eigen::Matrix<int, Eigen::Dynamic, Eigen::Dynamic>& data,
                        Eigen::Matrix<double, Eigen::Dynamic, Eigen::Dynamic>& pos,
                        int observations, int dimensions, double tol, int iterMax,
                        double delta, double alpha)
{
    Eigen::MatrixXd vel = Eigen::MatrixXd::Zero(pos.rows(), pos.cols());
    int x, y, z;
    double scale;
    Eigen::VectorXd YmX = Eigen::VectorXd::Zero(dimensions);
    Eigen::VectorXd ZmX = Eigen::VectorXd::Zero(dimensions);
    LL1 = logLikelihood(data, pos, observations, dimensions);

    for (int iter = 0; iter < iterMax; iter++)
    {
        for (int i = 0; i < observations; i++)
        {
            x = data(i, 0);
            y = data(i, 1);
            z = data(i, 2);

            ZmX = pos.col(z) - pos.col(x);
            YmX = pos.col(y) - pos.col(x);
            scale = 1 / (1 + exp(ZmX.norm() - YmX.norm()));
            vel.col(x) += scale * (YmX.normalized() - ZmX.normalized());
            vel.col(z) += scale * (ZmX.normalized());
            vel.col(y) -= scale * (YmX.normalized());
        }

        pos += delta * vel;
        vel *= alpha;
        /*if (stopping criteria)
        {
            ...
        }*/
    }

    std::cout << "Maximum iterations reached.\n";
}
```