Match prediction enhancement for dating applications

Project proposal for the MALIS course @EURECOM

- Marco Cavenati
- Mayank Narang
- Ilaria Pilo

Motivation

With the pandemic, the usage of dating applications has seen a rapid increase. However, users often struggle to find like-minded people with whom to build a durable and successful relationship.

To overcome these limitations, we aim to develop a machine learning model that allows the system to propose higher quality matches between users in less time.

Our model will be trained starting from data collected during a speed dating experiment. In such an experiment, participants were posed different questions about themselves and their ideal partner, with the goal of studying dating behaviour. We believe that the same questions could be easily replicable in any dating application during the user's registration process, and used as input to predict whether a given match would be successful or not.

Methodology and experiments

Our task is a binary classification problem, with the goal of predicting the outcome of a match (successful or not successful).

The data will be taken from Speed Dating | Kaggle. Such a dataset is unbalanced—7k "No match" records against 1.4k "match" records—and it contains mainly categorical attributes.

First of all, we will preprocess our data, removing uninteresting features (or, more specifically, features we cannot have information about during the user's registration process).

After splitting the dataset into three subsets for training, validation and testing, we will briefly analyze our data (how it is distributed, if features are correlated). Then, according to these results, we will train different models based on machine learning techniques such as K-Nearest Neighbours, Logistic Regression, Random Forest and Support Vector Machine. We will also analyze the effect of feature dimensionality reduction techniques (e.g., PCA).

We will evaluate our models on the validation test using normalized DCF (being an unbalanced dataset, accuracy could be misleading).

Finally, we will select the most promising model, and we will apply it to the (unknown) test set.