Machine Learning
Self-perception in dating
Carine Candel, Deniz Ovalioglu and Isabel Walter

We are using a dataset called the Speed Dating Experiment from Columbia Business School. They acquired data from participants in speed dating events, where the participants dated for four minutes and asked if they would like to meet again at the end. Participants also rated themselves and the partner they met on six attributes: Attractiveness, Sincerity, Intelligence, Fun, Ambition, and Shared Interests. The original experiment was designed to discover if race and gender preferences matter in dating. Therefore, there is a significant number of attributes included in the original data, that are not relevant to our experiment so we excluded it and made our own adapted dataset.

Our dataset consists of 8378 instances and each instance includes the following attributes:

1. How person 1 would rate him- or herself according to the six attributes above, on a scale from 1-10. Here we have a problem.
2. How person 2 rated person 1 according to the six attributes on a scale from 1-10.
3. Whether person 2 wants to meet person 1 again.

With this data we want to find out whether the difference between how you perceive yourself and how someone else perceives you has an influence on whether someone wants to meet you again.

To answer this question we would calculate the squared difference (to avoid negative differences) between attribute 1 and 2. These differences are our instances and we divide them in a training set, a cross validation set and a test set. Attribute 3 represents the labels of our data. That is, it shows whether with a certain difference person 2 wanted to meet person 1 again.

First we want to do supervised learning by implementing a decision tree classifier, because our data is categorical. Also, there is some missing data and decision trees handle missing values well. Naturally, we want to train our classifier with the training set and then use the CV to properly prune the tree. Perhaps we also want to include a random forest classifier.

Then we want to do unsupervised learning by implementing k-means clustering. We would use k = 2 because person 2 either wants to meet person 1 again or not. Then we will check whether the two clusters have a predominant label to see whether there actually is a correlation between the difference between attribute 1 & 2 and attribute 3.

Till now we acquired our data and adapted the amount of attributes in the dataset to include only the significant attributes for our experiment.