[PPML] Movie recommender system

Rodion “rodde” Efremov a.k.a “Machine Funkeehs”

February 17, 2015

1 Introduction

This document right here is the report on a movie recommender system developed on behalf of the project course **Project in Practical Machine Learning**, taught at Department of Computer Science, University of Helsinki, spring term.

2 Choice of data

The **movielens** group provides three different packages:

- a small package with 100000 ratings,
- a medium size package with one million ratings,
- a large package with ten million ratings.

We, Machine Funkeehs, began developing the machine learning code for the smallest of the packages as we were afraid that our approach would not scale well at both stages: database and processing. Now the smallest package provides along the ratings, naturally, 943 users and 1682 movies.

3 Choice of machine learning technique

If we ask ourselves, how do we recommend movies to a user, one answer might be: make the new user U_0 rate some movies, find, say k, other users $U_1, \dots, U_k$ that “act like U”, and recommend some movies that U_i tend to like. That is exactly what we did. We defined extended Jaccard-coefficient between two users U, U'

$$f = \frac{f_{11} - \sigma}{f_{01} + f_{10} + f_{11}},$$

where f_{10} is the amount of movies that only U has seen, f_{01} is the amount of movies that only U' has seen, $f_{11} = |M|$ is the amount of movies that both U

and U' have seen, and

$$\sigma = \sum_{m \in M} \frac{|r_U(m) - r_{U'}(m)|}{5},$$

where M is the set of movies that both U and U' have seen, and $r_X(m)$ gives the rating for movie $m \in M$ given by the user X.

What comes to ratings' scores, they are assumed to be integers within the range $[1, 5]$. Now it is easy to see that σ attempts to penalize those movies in M that have drastically different scores, and not to penalize those movies at all that has same scores. After defining the similarity measure, the rest is just running k-nearest neighbor algorithm, choose k "closest" users, see what they tend to like and recommend that to our new user U.