Methods of classification and dimensionality reduction - Report 1

March 29, 2022

1. Statement of the problem

We have the data containing information how users rate some movies. Our task is to create a recommender system, so having only some data we want to predict all ratings.

1.1. **Description of the data.** The data contains information about around 100000 ratings - around 600 users rated around 9000 movies. Our columns are: userId, movieId and rating. We keep this data in two-dimensional matrix such that ... If the user i haven't rated the movie j we leave $Z[i, j]$ empty.

1.2. **Quality of the system.**

2. Description of methods

In this problem, we try different approaches.

First 3 methods, so SVD1, SVD2 and NMF get a $n \times d$ dimensional matrix Z and approximate it by a different matrix $\tilde{Z}$. Since we want somehow $\tilde{Z}$ to maintain only "the most important" information from Z, then the rank of Z is to be much smaller than rank of Z. Precisely, we want to find matrix $\tilde{Z}_r$ of rank r ($r < rank(Z)$ and r is a parameter), so that $\|Z - \tilde{Z}_r\|$ is small.

Because these methods are given a whole matrix, so they need the missing data to be imputed before performing. The way we impute the data we describe later in the report in ??section??

SVD1. Using SVD decomposition $Z = U\Lambda^{\frac{1}{2}}V^T$ we construct $\tilde{Z}$ as

$$\tilde{Z}_r = U_r \Lambda_r^{\frac{1}{2}} V_r^T,$$

where Λ_r contains r biggest eigenvalues of Z and U_r, V_r contains only columns corresponding to those eigenvalues.

SVD2. SVD2 is an iterative method using SVD1, so we perform SVD1 on matrix Z, then on the result of first SVD1 and so on. Since $\tilde{Z}$ can have different values than actual in elements that we actually now (so these from training set), so we in every step we have to make a correction and impute the real values there. We stop when...

NMF. This time we approximate Z with $\tilde{Z}_r = W_r H_r$, where W_r and H_r are matrices with non-negative elements (W_r has r columns and H_r has r rows). Precisely, we look for such W_r and H_r that $\|Z - W_r H_r\|^2$ is the smallest, where $\|A\|^2 = \sum_{i,j} A_{ij}^2$.

SGD.

3. Our data

Description.

Performing methods.

Choosing parameters.

4. Results