# Рударење на масивни податоци – Домашна бр. 2

Симона Ристовска 221003

## 1. Објаснување на алгоритамот

### *ALS*

ALS алгоритамот работи на тој начин што на почеток конструира матрица R со димензионалност **uxm**, каде **u** е бројот на корисници, а **m** е бројот на филмови во нашиот пример.

$$R = \begin{bmatrix} 5 & 3 & ? \\ 4 & ? & 2 \\ ? & 1 & 3 \end{bmatrix}$$

### *Rank*

Според овој пример, првата редица кореспондира со тоа кои оцени првиот корисник ги дал за филмовите (колоните). Доколку имаме прашалник на одредено поле, тоа значи дека корисникот не дал оценка за тој филм. ALS креира матрици U со димензионалност **uxl** и P со димензионалност **mxl**, каде l е бројот на латентни фактори (**als.rank**).

### *MaxIter*

Најпрвин матриците U и P се иницијализираат со рандом вредности. Потоа за матрицата U се наоѓа матрица P така што разликата помеѓу R и $UP^T$ да е најмала (разликата само помеѓу оние позиции кои се достапни, а не и тие кои се празни во R). Значи пробуваме со $UP^T$ да ја апроксимираме R. Потоа за фиксна вредност на новопронајдената P се пронаоѓа U, и така **maxIter** пати. На крај од матрицата $UP^T$ ги земаме позициите кои фалат во матрицата R и тоа ни се проценките направени од ALS алгоритамот.

### *regParam*

**regParam** претставува регуларизациски параметар кој ја контролира L2 регуларизацијата и ги пенализира големите вредности во матриците U и P.

## 2. Крос-валидација и евалуација

За избор на најдобри параметри на ALS моделот ја користев **CrossValidation** техниката. Податочниот сет за тренирање го дели на повеќе подмножества (fold-ови), во мојот сличај на **3** дела, при што за секоја комбинација од хиперпараметри моделот се тренира на **2** делчиња и евалуира на преостанатото делче (ова се прави за сите комбинации од делчиња за тренирање и тестирање).

За една комбинација на хиперпараметри ја имаме 3 пати искалкулирано **MSE** метриката, па финалната MSE метрика за оваа комбинација од хиперпараметри е средна вредност од трите претходни евалуации. И моделот со најдобра MSE метрика се зима за најдобар модел и потоа го тренирам на целото тренирачки множество се со цел да направам предикции врз тестирачкото множество.

Избраниот модел потоа го евалуирав на тест подмножеството со повеќе метрики, вклучувајќи **MSE** и **MAE .** Добиените резултати покажуваат дека моделот има добра способност за апроксимација на реалните оценки, при што RMSE и MAE укажуваат на релативно мала просечна грешка. Доколку го интерпретираме MAE (Mean absolute error) можеме да кажеме дека предикцијата за рејтинг на моделот во просек се разликува од реалниот рејтинг за 0.73, што не е многу.

### *Бонус*

Покрај ALS алгоритамот, испробав два други алгоритми, конкретно **KNN** и **NMF**, со цел да се направи споредба на нивните перформанси врз истиот податочен сет, исто така користејќи алгоритам за најдобар избор на хиперпараметри (**GridSearchCV**).

- Резултатите покажуваат дека ALS моделот има најдобри перформанси, со најниска вредност на **MSE** (0.84) и **MAE** (0.73), што укажува на помала просечна грешка во предвидувањето на оценките.
- **KNN** моделот покажува послаби резултати, со повисоки вредности на MSE (0.96) и MAE (0.77). **NMF** моделот постигнува резултати подобри од KNN, но сепак полоши од ALS, со MSE (0.88) и MAE (0.74).
- Според овие метрики, може да се заклучи дека ALS е најсоодветниот модел, бидејќи обезбедува најмала грешка и најдобра апроксимација на рејтинзите.