IT492: Recommendation Systems

Lecture - 07

Collaborative Filtering: Model-based Methods

Arpit Rana

14th Feb 2022

Baseline Estimates

Typical *CF* data exhibit systematic tendencies -

- for some users to give higher ratings than others, and
- for some items to receive higher ratings than others.

We account for these systematic tendencies within the ***baseline estimates***.

Baseline Estimates

A ***baseline estimate*** for an unknown rating r_{ui} is denoted by b_{ui} and accounts for the user and item effects:

$$b_{ui} = \mu + b_u + b_i$$

Here, μ is the overall average rating, b_u is the observed deviation in user *u's* ratings and b_i is the observed deviation in item *i's* raings.

Baseline Estimates

In order to estimate b_u and b_i one can solve the following least squares problem:

$$\min_{b_*} \sum_{(u,i)\in\mathcal{K}} (r_{ui} - \mu - b_u - b_i)^2 + \lambda_1\left(\sum_u b_u^2 + \sum_i b_i^2\right)$$

Here, $\boldsymbol{K} = \{(u, i) \mid r_{ui} \text{ is known}\}$.

MF with Baseline Estimates

A typical SVD model associates each user u with a user factors vector $p^{\cancel{u}}_{u} \in R^k$, and each item i with an item-factors vector $q^{\cancel{i}}_{i} \in R^k$.

The prediction is done by taking an inner product, i.e.,

$$\hat{r}_{ui} = \mu + b_i + b_u + q_i^T p_u$$

To learn the factor vectors (p_u and q_i), the system minimizes the regularized squared error on the set of known ratings:

$$\min_{p_*, q_*, b_*} \sum_{(u,i) \in \mathcal{K}} (r_{ui} - \mu - b_u - b_i - p_u^T q_i)^2 + \lambda_3(\|p_u\|^2 + \|q_i\|^2 + b_u^2 + b_i^2)$$

Neighborhood-based Models Revisited

In item-item CF, the predicted value of r_{ui} is taken as a weighted average of the ratings of neighboring items:

$$\hat{r}_{ui} = b_{ui} + \frac{\sum_{j \in N_i} w_{ij} \cdot (r_{uj} - b_{uj})}{\sum_{j \in N_i} w_{ij}}$$

We can see ratings are adjusted for user and item effects through the baseline estimates; and w_{ij} is a ***shrunk correlation coefficients***.

$$w_{ij} = \frac{n_{ij}}{n_{ij} + \beta} \cdot s_{ij}$$

Neighborhood-based Models Revisited

Note the dual use of the similarities in the previous formulation -

- identification of nearest neighbors, and
- as the interpolation weights

This raise some concerns -

- SImilarity weights do not take into account the relationship among neighbours.
- By definition, the interpolation weights sum to one, which may cause overfitting.

Neighborhood-based Models Revisited

This raise some concerns -

- SImilarity weights do not take into account the relationship among neighbours.
- By definition, the interpolation weights sum to one, which may cause overfitting.

To resolve this, we search for optimum interpolation weights without regard to values of the similarity measure.

$$\hat{r}_{ui} = b_{ui} + \sum_{j \in N_i} \theta_{ij}^{u} \cdot (r_{uj} - b_{uj})$$

Learning-based Methods

Learning-based methods that use neighborhood or similarity information can be divided in two categories:

- Factorization methods (e.g. MF), and
- **Neighborhood learning methods (e.g. SLIM).**

Neighborhood Learning Methods

- Standard neighborhood-based recommendation algorithms determine the neighborhood of users or items directly from the data, using some pre-defined similarity measure like PC.
- Recent developments have shown the advantage of learning the neighborhood automatically from the data, instead of using a pre-defined similarity measure.

Neighborhood Learning Methods

- A representative neighborhood-learning recommendation method is the Sparse Linear Model (SLIM) algorithm, developed by Ning et al. 2011.
- In SLIM, a new rating is predicted as a sparse aggregation of existing ratings in a user's profile,

$$\hat{r}_{ui} = \mathbf{r}_u \mathbf{w}_i^\top ,$$

- Here r_u is the uth row of the rating matrix R and w_i is a sparse row vector containing |I| aggregation coefficients.
- Essentially, the non-zero entries in w_i correspond to the neighbor items of an item i.

Neighborhood Learning Methods

- The neighborhood parameters are learned by minimizing the squared prediction error.

$$
\begin{aligned}
\underset{W}{\text{minimize}} \quad & \frac{1}{2}\|R - RW\|_F^2 + \frac{\beta}{2}\|W\|_F^2 + \lambda\|W\|_1 \\
\text{subject to} \quad & W \geq 0 \\
& \operatorname{diag}(W) = 0.
\end{aligned}
$$

- The non-negativity constraint on W imposes the relations between neighbor items to be positive.
- The constraint diag(W) = 0 is also added to the model to avoid trivial solutions (e.g., W corresponding to the identity matrix) and ensure that r_{ui} is not used to compute the predicted rating.

Next lecture -

Wrap-up
Collaborative Filtering