SC4/SM8 Advanced Topics in Statistical Machine Learning
Collaborative Filtering

Dino Sejdinovic
Department of Statistics
Oxford

Slides and other materials available at:
http://www.stats.ox.ac.uk/~sejdinov/atsml/

Ratings and Recommendations

movie \ user	Alice	Bob	Chuck	Dan	Eve
Happy Gilmore	?	2	5	1	4
Click	1	?	4	?	?
Ex Machina	?	4	?	?	2
Blade Runner	5	?	1	?	?
The Matrix	5	5	?	?	4

- Data: a **partially observed** matrix $\mathbf{Y} \in \mathbb{R}^{n_1 \times n_2}$ where $y_{i,j}$ is the rating (e.g. between 1 and 5) of item i by user j.
- Most entries will be **missing/unknown** since most users will not have rated most movies.
- **Exposures**: $e_{i,j} = 1$ if the user j has rated movie i and $e_{i,j} = 0$ otherwise.

Features of items and users
Content based recommendation

- each item i has a **feature vector** $\phi_i = [\phi_{i1}, \ldots, \phi_{ik}]^\top \in \mathbb{R}^k$
- If ϕ_i observed, simply solve one linear model per user:

$$\min_{\psi_j} \sum_{i\,:\,e_{i,j}=1} (y_{i,j} - \phi_i^\top \psi_j)^2 + \lambda_\psi \|\psi_j\|_2^2, \quad j = 1, \ldots, n_2.$$

- ψ_j is the corresponding vector of coefficients in the linear model corresponding to user j - can be treated as a feature (**preference**) vector of user j.
- If ψ_j observed, but ϕ_i is hidden, solve one linear model per item:

$$\min_{\phi_i} \sum_{j\,:\,e_{i,j}=1} (y_{i,j} - \phi_i^\top \psi_j)^2 + \lambda_\phi \|\phi_i\|_2^2, \quad i = 1, \ldots, n_1.$$

Alternating linear regressions

- Assume neither features are observed.
- Formulate recommendations solely based on the ratings matrix: alternating regression model.
- Often simply use stochastic gradient descent (SGD) updates: as soon as new rating becomes available:

$$\begin{aligned}\phi_i \leftarrow &\quad (1-\epsilon_t\lambda_\phi)\phi_i + \epsilon_t\psi_j(y_{ij} - \phi_i^\top \psi_j),\\ \psi_j \leftarrow &\quad (1-\epsilon_t\lambda_\psi)\psi_j + \epsilon_t\phi_i(y_{ij} - \phi_i^\top \psi_j).\end{aligned}$$

- **Collaborative**: predictions for each user can potentially depend on ratings of all other users.
- Potentially results in features/preferences which do not have a readily interpretable meaning.

Probabilistic Matrix Factorization

Introduced in [Salakhutdinov and Mnih, 2007], the generative model corresponding to CF can be described as follows:

- For each movie $i = 1, \dots, n_1$, sample independently the latent vector of features $\phi_i \sim \mathcal{N}(0, \sigma_\phi^2 I_k)$ from a k-dimensional normal distribution,
- For each user $j = 1, \dots, n_2$, sample independently the latent vector of preferences $\psi_j \sim \mathcal{N}(0, \sigma_\psi^2 I_k)$ from a k-dimensional normal distribution,
- For each movie-user pair (i, j), sample $e_{i,j} \sim Bernoulli(p)$ independently and if $e_{i,j} = 1$, sample $y_{i,j} | \phi_i, \psi_j \sim \mathcal{N}(\phi_i^\top \psi_j, \sigma_y^2)$.

Beyond Gaussian "ratings" likelihood

Binary ratings:

- Logistic link: $p(y_{i,j}|\phi_i, \psi_j) \sim \sigma(y_{i,j}\phi_i^\top \psi_j)$
- Probit link: $p(y_{i,j}|\phi_i, \psi_j) \sim \Phi(y_{i,j}\phi_i^\top \psi_j)$

"Count" ratings:

- Poisson link: $y_{i,j} \sim \text{Poisson}(\exp(\phi_i^\top \psi_j))$