Weeks 10–11 Summary: Recommender Systems and Matrix Factorization

Z3003 Data Mining: Scaling to Large Data (Part 2 · Dr. Mahendran V)

1. Content-Based vs. Collaborative Filtering

Content-based recommenders build TF-IDF or attribute profiles per item, whereas collaborative filtering uses observed user-item interactions (R_{ui}) directly. Centering each user's ratings by subtracting user mean $\bar{r}_u$ turns cosine similarity into Pearson correlation, neutralizing harsh-rater bias.

2. Biased Latent-Factor Matrix Factorization (SVD)

To avoid imputing missing entries as zero, fit user factors $p_u \in \mathbb{R}^k$ and item factors $q_i \in \mathbb{R}^k$ strictly over observed pairs $(u, i) \in \Omega$:

$$\min_{p,q,b} \sum_{(u,i)\in\Omega} \big(r_{ui} - (\mu + b_u + b_i + p_u^\top q_i)\big)^2 + \lambda\big(\|p_u\|^2 + \|q_i\|^2 + b_u^2 + b_i^2\big).$$

Stochastic gradient descent (SGD) and alternating least squares (ALS) scale linearly in $|\Omega|$.