Report: Recommender Systems Analysis (Part 2)

6. Advanced Ranking & Hybrid Systems Analysis

Moving beyond point-wise prediction, we explored ranking-oriented models, hybrid architectures, and neural recommenders to address the limitations of classical collaborative filtering.

The Power of Pairwise Optimization (BPR)

Our implementation of Bayesian Personalized Ranking (BPR-MF) demonstrated the fundamental difference between predicting a rating (MSE) and optimizing a ranking (Pairwise Loss).

- **Objective Alignment:** By explicitly training the model to score a known positive item higher than a sampled negative item, BPR-MF achieved an NDCG@10 of 0.0656. This outperforms our best classical MF model (FunkSVD) and even slightly beats our best overall classical model (Item-Item CF, NDCG@10 = 0.0625). The comparison with FunkSVD is particularly relevant as both are latent factor models, highlighting the superiority of pairwise ranking loss over point-wise MSE for top-K retrieval tasks.
- **Catalog Discovery:** The most significant advantage of BPR was its catalog coverage (45.8%). Unlike Item-Item CF, which collapsed into a narrow set of popular items, BPR learned to rank a much broader spectrum of the catalog, effectively halving the popularity bias.
- **The Sampling Trade-off:** We found that the negative sampling ratio is critical. Sampling 5 negatives per positive interaction provided the optimal balance of contrast for the model to learn. Too few samples (1:1) led to poor convergence and low coverage, while too many (20:1) distracted the model, slightly degrading accuracy.

Hybrid Architectures: Bridging the Gap

To address the cold-start problem inherent in pure collaborative models, we designed a Hybrid Recommender combining BPR-MF (collaborative signal) with an Enhanced Content-Based model (TF-IDF genres + popularity).

- **The Synergy:** The Hybrid model (Weighted Blending, $\alpha = 0.8$) outperformed both base models. The Content-Based signal acted as a crucial fallback for niche or cold items where BPR lacked sufficient interaction data.

- **Who Benefits?** Our segmented analysis revealed a stark contrast:
    - **Cold Users ($\leq$ 30 ratings):** The Hybrid approach was the clear winner, significantly boosting accuracy by leveraging content features when collaborative history was sparse.
    - **Warm Users ($\geq$ 100 ratings):** The pure BPR model actually performed better. For users with rich histories, the collaborative signal is strong enough; injecting generic content features only added noise and diluted the personalized recommendations.

Deep Learning: Capacity vs. Practicality

We implemented two neural architectures: Neural Collaborative Filtering (NeuMF) and a Two-Tower model.

- **Representational Limits:** Interestingly, both NeuMF and the Two-Tower model performed slightly worse than the linear BPR-MF baseline on this dataset. Neural networks are data-hungry. In a relatively dense dataset like MovieLens 1M, where interactions are strictly structural (IDs), a heavily regularized linear dot product often generalizes better than a highly parameterized MLP prone to overfitting.
- **The Two-Tower Advantage:** The Two-Tower model outperformed NeuMF because it allowed us to inject TF-IDF content features directly into the item representation tower.
- **Production Reality:** While NeuMF requires a computationally prohibitive forward pass for every user-item pair at inference time, the Two-Tower architecture allows for precomputing and caching item embeddings. This reduces inference to a fast Approximate Nearest Neighbor (ANN) search, making it the only viable deep learning architecture for web-scale retrieval.

7. Online Evaluation & System-Level Synthesis

Offline metrics are proxies. To understand how these models would perform in a live environment, we simulated an online evaluation using Multi-Armed Bandits.

The Exploration-Exploitation Dilemma

We deployed ϵ-greedy, UCB1, and Thompson Sampling bandits to dynamically route traffic between four static policies (Popularity, BPR-MF, Hybrid, Random).

- **Thompson Sampling:** This Bayesian approach was the clear winner, achieving an average reward almost identical to the theoretical maximum. It quickly learned to ignore the poor models and confidently committed the vast majority of its traffic to the optimal Hybrid policy.
- **The Cost of Hard-Coded Exploration:** The ϵ-greedy (0.3) bandit performed poorly because it permanently wasted 30% of its traffic pulling sub-optimal arms. It also suffered from early noise,

mistakenly getting "stuck" exploiting BPR-MF instead of the true winner.

- **UCB1 Failure:** UCB1 struggled to converge. Because our simulated rewards were small fractions (around 0.04), the mathematical "uncertainty bonus" in the UCB1 formula was disproportionately large, causing it to over-explore and fail to separate the signal.

Final Deployment Choice & Iteration Strategy

If deploying this system to production today, we would implement the **Two-Tower Architecture** serving as the primary candidate generator, feeding into a lightweight **Candidate Reranking Hybrid**.

1. **Why Two-Tower?** It provides the best balance of representational flexibility (combining collaborative IDs with content features) and inference speed (via cached embeddings and ANN search).
2. **Why Reranking?** Blending scores across the entire catalog is too slow. Generating 100 candidates via the Two-Tower model and reranking them with a business-logic aware model (e.g., boosting recent items or demoting already-seen genres) is the industry standard.

Post-Deployment Iteration:

1. **A/B Testing:** We would immediately launch an A/B test comparing the new Two-Tower/Hybrid pipeline against the existing Item-Item CF baseline. The primary metric would be **Click-Through Rate (CTR)** on the top-5 recommendations, with **Watch Time** as a guardrail metric to ensure we aren't just optimizing for clickbait.
2. **Monitoring Failure Modes:** We must actively monitor the "Filter Bubble" effect. If the Hybrid model begins to narrow a user's genre exposure too aggressively, we would need to introduce an explicit diversity boost (or redundancy penalty) in the reranking stage. We also need to monitor the performance of the Two-Tower model on newly added items (Cold-Start) to ensure the content features are providing sufficient signal before collaborative interactions accumulate.