Tech Review: Netflix: Multi-Armed Bandit for Personalization

Author: Sreyashi Das

For my tech review, I will be covering Netflix's Multi-Armed Bandit technique for recommending movies to users. Netflix has a very large catalog and there is very little time, usually around 90 secs before a member may lose interest and abandon the session. The goal of the Netflix homepage is to provide relevant and personalised content to each of its members. So, Netflix condenses the large catalog into a few choices that users would like to play. With the company growing globally and number of members the scale becomes larger and larger on top of adding new content everyday. Recommendations form the core component of the Netflix homepage. Netflix homepage consists of a variety of algorithms defined for various objectives such as Top Picks, Trending Now, Because you Watched, New Releases, Search and Adhoc Personalization. In Ad Hoc Personalization, the right image is presented for the right title for a given member.

It is very hard to predict what resonates with the user before a title launches on the service. Most recommender systems use Collaborative Filtering or Content-based methods to predict new items of interest for users. Collaborative filtering systems work by collecting user feedback in the form of ratings for items in a given domain and exploit similarities or differences among profiles of several users in determining how to recommend an item. Testing has shown that the predicted ratings aren't actually super useful, while what you're actually playing is. Rather than focusing exclusively on ratings and rating predictions Netflix now depends on a more complex ecosystem of algorithms. Netflix keeps track of what content a user has watched, searched for, or rated, as well as the time, date and device. User interactions such as browsing or scrolling behaviour is also tracked. All that data is fed into several algorithms, each optimized for a different purpose. In a broad sense, most of Netflix's algorithms are based on the assumption that similar viewing patterns represent similar user tastes.

The goal of the billboard (the large banner on the Netflix homepage) is to recommend a single relevant title to each of its members and to respond quickly to member feedback. Traditional approaches for recommendation such as collaborative filtering have been widely applied in various industrial settings. One of the most popular algorithms is the Matrix Factorization where the idea is to use the wisdom of the crowd. These algorithms face several practical challenges when applied to a practical setting. For example, scarce feedback, dynamic catalog whereby new titles are constantly getting added to the service, and the member base is changing and growing. There is a time-sensitive component that the traditional batch approaches are not meant to respond quickly to member feedback such as content popularity changes over time and member interest evolves over time. A new set of techniques called multi-armed bandit techniques which are basically online learning techniques have become increasingly popular in these practical settings because they are more robust in handling challenges mentioned above. The idea of a multi-armed bandit is quite old and is inspired from the idea of gambling in a casino. There is a gambler and there are a number of slot machines with unknown reward distribution. The gambler has multiple arms and the goal of the gambler is to maximise his earnings or to maximise his reward. The key question for the gambler is which machine he should play to maximise his reward. There is a learner which interacts with the environment by selecting an action and the environment responds back to the learner by presenting a reward. For each round, the learner chooses an action from a set of actions and presents it to the environment and the environment generates a real-valued reward and presents it back to the learner. The goal of the learner is to maximise the cumulative reward or to minimize the cumulative regret. The regret is the difference in the reward when the action was selected by the learner compared to the optimal action that could have been selected by the learner in hindsight. At the heart of these bandit algorithms is the Exploration-Exploit tradeoff. The trade off is whether you should recommend the optimal title given the evidence so far which is also called Exploit or whether you should recommend lesser known titles to gather feedback which is called as Explore. Exploration allows us to gather information that helps us to determine what would have been an overall best action. The simplest exploration strategy is Naive exploration where noise is added to the greedy policy. Another famous strategy is optimism in the face of uncertainty which is to prefer actions about which you have little information. In this way you are going to explore more around that action and as a result gather more feedback. The extensions of Multi-Armed bandit handling context/ features are associated with the action, user or the user-action pair and are called contextual bandits and are very widely used in practice.

Now, let us focus on a very naive multi-armed bandit strategy which is called the epsilon greedy approach. The advantage of this approach is that it provides unbiased data which can be then used for training. When a user comes in, the recommendation is generated online and at the same time information is logged about why this recommendation was made and the member feedback that further enriches the training data set and updates the model. In the epsilon greedy case, a coin is flipped to decide to explore or exploit. In the case of explore, you randomly sample from the candidate pool to pick a title. In the exploit case, you apply a model to pick the most optimal recommendation. In this apprach with probability $\boldsymbol{\epsilon}$ you will select a random title. The rest of the time you will pull the title based on the current knowledge of the titles. ε is typically rather low (0.1), so mostly you play greedy, but with some exploration to find the best machine. The rewards will be low in the beginning but in the long run you will see significant optimal selection of movie titles.

Netflix has grown to be a leader in the streaming video on demand industry due to various recommendation algorithms working in collaboration. As a developer at Netflix and a long-time user of the product myself, I was intrigued to know more about the reason why a particular title shows up on the billboard when I login to Netflix. There have been countless times when I watched the movie on my billboard. Understanding the implementation of the recommendation system and why it shows that perfect title on the billboard has been a very interesting journey and I love the product even more.

Sources

- https://en.wikipedia.org/wiki/Collaborative_filtering
- https://info.dataengconf.com/hubfs/SF%2018%20-%20Slides/DS/A%20Multi-Armed%20Bandit%20Framework%20for%20Recommendations%20at%20Netflix.pdf
- https://netflixtechblog.com/ml-platform-meetup-infra-for-contextual-bandits-and-reinforcement-learning-4a90305948ef
- https://www.wired.com/2013/08/qq-netflix-algorithm/
- https://en.wikipedia.org/wiki/Recommender_system