Machine Learning

Recommendation System:
Collaborative Filtering and Content -based

Recommender System

- Like many machine learning techniques, a recommender system makes prediction based on users' historical behaviors. Specifically, it's to predict user preference for a set of items based on past experience. To build a recommender system, the most two popular approaches are Content-based and Collaborative Filtering.
- **Content-based** approach requires a good amount of information of items' own features, rather than using users' interactions and feedbacks. For example, it can be movie attributes such as genre, year, director, actor etc., or textual content of articles that can be extracted by applying Natural Language Processing.

towardsdatascience.com

Recommender System

- **Collaborative Filtering**, on the other hand, doesn't need anything else except users' historical preference on a set of items. Because it's based on historical data, the core assumption here is that the users who have agreed in the past tend to also agree in the future. In terms of user preference, it usually expressed by two categories.
- **Explicit Rating**, is a rate given by a user to an item on a sliding scale, like 5 stars for Titanic. This is the most direct feedback from users to show how much they like an item.
- **Implicit Rating**, suggests users preference indirectly, such as page views, clicks, purchase records, whether or not listen to a music track, and so on. In this article, I will take a close look at collaborative filtering that is a traditional and powerful tool for recommender systems.

Recommender System

- Item recommendation
- Rating Prediction System
- U: set of users
- I: set of items
- Probability p: U X I -> R
- Learn p from data
- Use to predict the utility value each item to each user
- Content based
  - Based on content similarity
- Collaborative filtering
  - Types: User-based nearest neighbour, Item-based nearest neighbour

Neighborhood formation phase

Let the record (or profile) of the target user be **u** (represented as a vector), and the record of another user be **v** (**v** $\in T$).

The similarity between the target user, **u**, and a neighbor, **v**, can be calculated using the **Pearson's correlation coefficient**:

$$sim(\mathbf{u},\mathbf{v}) = \frac{\sum_{i \in C}(r_{\mathbf{u},i} - \bar{r}_{\mathbf{u}})(r_{\mathbf{v},i} - \bar{r}_{\mathbf{v}})}{\sqrt{\sum_{i \in C}(r_{\mathbf{u},i} - \bar{r}_{\mathbf{u}})^2}\sqrt{\sum_{i \in C}(r_{\mathbf{v},i} - \bar{r}_{\mathbf{v}})^2}},$$

Recommendation Phase

Use the following formula to compute the rating prediction of item *i* for target user **u**

$$p(\mathbf{u},i) = \bar{r}_{\mathbf{u}} + \frac{\sum_{\mathbf{v}\in V} sim(\mathbf{u},\mathbf{v}) \times (r_{\mathbf{v},i} - \bar{r}_{\mathbf{v}})}{\sum_{\mathbf{v}\in V} |sim(\mathbf{u},\mathbf{v})|}$$

where V is the set of k similar users, $r_{\mathbf{v},i}$ is the rating of user **v** given to item *i*,