Research on demography data reconstruction

Joseph Wang

Abstract A recommender system may explicitly solicit user demographics through user registration. However, online users are not always willing to provide such information due to privacy concern. On the other hand, user interactions such as ratings in recommender systems may provide an alternative way to infer demographic information. For example, a Netflix user who likes romance comedy and child-friendly movies may indicate that she is a mom. Existing attempts include the famous de-anonymization of Netflix Prize dataset that link private Netflix rating data with public databases such as IMDB to partially infer some user identities. Other attempts suggest that it is possible to infer user gender with as high as 80% accuracy given sufficient user ratings in recommender systems.[1]

Decision Tree Approach

Specifically, for predicting both ratings and demographics such as age, we adopt least mean square error (L_2) as the loss function. In such cases, the user profile has a closed-form solution:

$$u_L = \left(\lambda_r \sum_{i \in L(w)} \sum_{(i,j) \in O} v_j v_j^\top + \lambda_s \sum_{i \in L(w) \cap S} \theta\theta^\top + \lambda_u I\right)^{-1} \\ \lambda_r \sum_{i \in L(w)} \sum_{(i,j) \in O} r_{ij} v_j + \lambda_s \sum_{i \in L(w) \cap S} y_i \theta + \lambda_u u_p\Big) \tag{1}$$

The profiles u_D and u_U for the other two children can be computed in a similar way. In summary, we iterate over possible items and select the best one for single-item split at each node. While for multi-item split, we alternatively optimize using techniques until convergence. After the current node is constructed, we recursively construct its child nodes in a similar way.[2]

Example Li:2017

In cold-start scenario, the system queries the user's rating on several selected items and constructs a rough user profile, which is then used to predict ratings for other

1. Li 2017.
2. Sun, Li, and Zha 2017.

items and at the same time to infer demographics. assume that the demographic label y such as age or gender follows some distribution

$$y_i \in p\left(y_i \mid \theta^\top u_i\right) \tag{2}$$

where θ is the regressor for continuous label prediction or the classifier for discrete label prediction. Given observed ratings $O = \left\{(i, j) \mid r_{ij} \text{ is observed }\right\}$ and demographic information $S = \{i \mid y_i \text{ is observed }\}$, where $i = 1, 2, ..., m$ and $j = 1, 2, ..., n$, our goal is to learn the function T , item profile v_j for each item j, and the regressor θ to minimize the negative log posterior of the model, which is equivalent to the following objective:

$$\begin{aligned} \min_{T,V,\theta} \lambda_r \sum_{(i,j)\in O} \ell_r\left(r_{ij}, T\left(x_i\right)^\top v_j\right) + \lambda_s \sum_{i\in S} \ell_s\left(y_i, T\left(x_i\right)^\top \theta\right) \\ + \lambda_v \|V\|^2 + \lambda_\theta \|\theta\|^2 \end{aligned} \tag{3}$$

References

Li, Changbin. 2017. A Study on User Demographic Inference Via Ratings in Recommender Systems. *LSU Master's Theses,* no. 4466, 1–51. Available at <https://digitalcommons.lsu.edu/gradschool_theses/4466>.

Sun, Mingxuan, Changbin Li, and Hongyuan Zha. 2017. Inferring Private Demographics of New Users in Recommender Systems. Paper presented at the Proceedings of the 20th ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems, 237–244. MSWiM '17. Miami, Florida, USA: Association for Computing Machinery. ISBN: 9781450351621. https://doi.org/10.1145/3127540.3127566. Available at <https://doi.org/10.1145/3127540.3127566>.