A Preference-Based Bandit Framework for Personalized Recommendation

Maryam Tavakol[1] and **Ulf Brefeld**[2]

Abstract. We present contextual bandits for personalized recommendation scenarios where user preferences are available. The model is a composite of a personalized model and a one-serves-all component where the latter resembles the mainstream recommendation. The derivation is consequentially carried out using Fenchel-Legrende conjugates and thus applicable in many different learning tasks. We present a unified framework that allows for quickly adapting our contextual bandits to different applications.

1 Introduction

Recommender systems are designed to serve user needs by extracting relevant content from a large amount of available information. User needs are generally characterized by individual interests of a user. However, considering many users at once gives rise to joint interests (e.g., topselling items) and needs that could be captured by a one-serves-all recommendation. Therefore, personalized recommendation aims to derive the individual preferences as well as their collective aggregate over all the users.

Traditional recommender systems focus on the recommendation problem from different perspectives. There are non-personalized approaches that focus on short-term goals and inferring topics of user sessions [7, 6], while collaborative filtering methods, on the other hand, aim to capture long-term preferences of users [4, 3]. The collaborative approach computes probably interesting items to a user by focusing on interests of similar peers. Whenever user preferences are available, it is convenient to directly learn user profiles from the partial order of items. Since preferences are often contradictive across serveral users, such an approach naturally extends to personalized recommendations with a dedicated model for every user. To also provide recommendations for new users with only little historical traffic, the individual models can be used as offsets to a one-serves-all (or average) model.

In this paper, we present a unified contextual bandit framework for personalized recommendation. The underlying scheme models the preferences between items which consists of an average part and an individual mnodel to compute the expected reward. We propose to leverage ideas from [5] to model our preference-based approach as a contextual bandit that is augmented by an individual offset. All derivations are carried out in dual space and using Fenchel-Legendre conjugates of the loss functions which renders our approach for a wide range of loss functions. In the next section, we derive a generalized model for personalized preference-based recommendation in dual space.

2 Linear Bandits in Dual Space

In this paper, we focus on sequential recommender systems for m users, $U = \{u_1, u_2, ..., u_m\}$, and n items, $A = \{a_1, a_2, ..., a_n\}$. Every item a_i is characterized by a set of attributes given by a feature vector $\boldsymbol{z}_i \in \mathbb{R}^k$. At each time step t, the goal of the system is to recommend items to the current user that are more likely to be clicked. We show how to derive the general optimization framework for linear bandits in dual space considering both the average and personalized models.

The proposed model is defined by a single bandit which learns the preferences between items for all the users. Assume that $\mathbf{z}_i$ and $\mathbf{z}_k$ belong to the items a_i and a_k, respectively, thus, we assign $\mathbf{z}_{i\succ k} := \mathbf{z}_i - \mathbf{z}_k$ to show the preference of item a_i over a_k. The payoff is therefore determined as a linear function of the preference,

$$\mathbb{E}[r_{t,i\succ k}|u_t = u_j] = \boldsymbol{\theta}^\top \mathbf{z}_{i\succ k} + \boldsymbol{\beta}_t^\top \mathbf{z}_{i\succ k} + b_{ik},$$

where $\boldsymbol{\theta}$ is the weight vector for the average model, while $\boldsymbol{\beta}_t = \boldsymbol{\beta}_j$ is the individual parameter for user j. $r_{t,i\succ k}$ shows the reward obtained by choosing item a_i over a_k at time t. For simplicity, we augment the feature vectors by a constant term (e.g., $\mathbf{z}_{i\succ k,0} = 1$) and move b_{ik} accordingly into the $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$.

Let $a_t = a_i$ be the selected item to recommend at time t, the problem in [5] is then relaxed to a simple one-armed bandit which learns the preferences between contexts, i.e., item features,

$$h_{\boldsymbol{\theta},\boldsymbol{\beta}_j}(\boldsymbol{z}_{i\succ l}; \exists l) = (\boldsymbol{\theta} + \boldsymbol{\beta}_j)^\top \boldsymbol{z}_{i\succ l},$$

where hypothesis h predicts the expected payoff for the specific user u_j on item a_i. Moreover, we substitute $\boldsymbol{z}_{i\succ l}$ by $\boldsymbol{z}_t$ as the context at time t. Given an appropriate loss function $V(., r_t)$, the regularized optimization problem can be stated as

$$\inf_{\substack{\boldsymbol{\theta}\\ \boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_m}} \quad \frac{1}{T}\sum_{t=1}^{T} V([\boldsymbol{\theta} + \boldsymbol{\beta}_t]^\top \boldsymbol{z}_t, r_t) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 + \frac{\hat{\mu}}{2}\sum_j \|\boldsymbol{\beta}_j\|^2. \tag{1}$$

Let $C = \frac{1}{\lambda T}$ and $\mu = \frac{\hat{\mu}}{\lambda}$, by incorporating y_t as shorthand for the predicted payoff we have

$$\inf_{\substack{\boldsymbol{\theta},\boldsymbol{y}\\ \boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_m}} \quad C\sum_{t=1}^{T} V(y_t, r_t) + \frac{1}{2}\|\boldsymbol{\theta}\|^2 + \frac{\mu}{2}\sum_j \|\boldsymbol{\beta}_j\|^2$$
$$s.t. \quad \forall t: \quad (\boldsymbol{\theta} + \boldsymbol{\beta}_t)^\top \boldsymbol{z}_t = y_t,$$

The equivalent unconstrained problem is derived by incorporating

[1] Leuphana University of Lüneburg, email: tavakol@leuphana.de

[2] Leuphana University of Lüneburg, email: brefeld@leuphana.de

Lagrange multipliers, $\boldsymbol{\alpha} \in \mathbb{R}^T$,

$$\begin{aligned}\sup_{\boldsymbol{\alpha}} \quad \inf_{\substack{\boldsymbol{\theta},\boldsymbol{y}\\ \boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_m}} \quad & C\sum_{t=1}^{T} V(y_t, r_t) + \frac{1}{2}\|\boldsymbol{\theta}\|^2 + \frac{\mu}{2}\sum_j \|\boldsymbol{\beta}_j\|^2 \\ & - \sum_{t=1}^{T} \alpha_t([\boldsymbol{\theta} + \boldsymbol{\beta}_t]^\top \boldsymbol{z}_t - y_t).\end{aligned}$$

Setting the partial derivatives w.r.t. $\boldsymbol{\theta}$ to zero, leads to $\boldsymbol{\theta} = \sum_{t=1}^{T} \alpha_t \boldsymbol{z}_t = Z^\top \boldsymbol{\alpha}$, where $Z \in \mathbb{R}^{T\times k}$ is the design matrix given by the training data. The derivatives w.r.t. $\boldsymbol{\beta}_j$ gives

$$\boldsymbol{\beta}_j = \frac{1}{\mu}\sum_{\substack{t\\ \boldsymbol{\beta}_t=\boldsymbol{\beta}_j}} \alpha_t \mathbf{z}_t = \frac{1}{\mu}\sum_t \phi_{jt}\alpha_t \mathbf{z}_t = \frac{1}{\mu}(Z \circ \boldsymbol{\phi_j})^\top \boldsymbol{\alpha},$$

where $\boldsymbol{\phi}_j \in \mathbb{R}^{T\times 1}$ is a binary vector which is 1 when $\boldsymbol{\beta}_t = \boldsymbol{\beta}_j$, and zero otherwise, and $\circ(.,.)$ stands for element-wise product. Substituting the optimality conditions into the optimization function yields

$$\begin{aligned}\sup_{\boldsymbol{\alpha}} \quad \inf_{\boldsymbol{y}} \quad & C\sum_{t=1}^{T} [V(y_t, r_t) + \frac{1}{C}\alpha_t y_t] - \frac{1}{2}\boldsymbol{\alpha}^\top Z Z^\top \boldsymbol{\alpha} \\ & - \frac{1}{2\mu}\sum_j \boldsymbol{\alpha}^\top (Z \circ \boldsymbol{\phi_j})(Z \circ \boldsymbol{\phi_j})^\top \boldsymbol{\alpha}.\end{aligned}$$

Recall that the Fenchel-Legendre conjugate of a function g is defined as $g^*(\boldsymbol{u}) = \sup_{\boldsymbol{x}} \boldsymbol{u}^\top \boldsymbol{x} - g(\boldsymbol{x})$ [2]. Thus, by moving the infimum inside the summation and given the dual loss

$$V^*(-\frac{\alpha_t}{C}, r_t) = \sup_{y_t} -\frac{\alpha_t}{C} y_t - V(y_t, r_t),$$

the generalized optimization problem in dual space reduces to

$$\begin{aligned}\sup_{\boldsymbol{\alpha}} \quad & -C\sum_{t=1}^{T} V^*(-\frac{\alpha_t}{C}, r_t) - \frac{1}{2}\boldsymbol{\alpha}^\top Z Z^\top \boldsymbol{\alpha} \\ & - \frac{1}{2\mu}\sum_j \boldsymbol{\alpha}^\top (Z \circ \boldsymbol{\phi_j})(Z \circ \boldsymbol{\phi_j})^\top \boldsymbol{\alpha}. \qquad (2)\end{aligned}$$

2.1 Upper Confidence Bound

The challenge in bandit-based approaches is to balance exploration and exploitation to minimize the regret. Auer [1] demonstrates that confidence bounds provide useful means to balance the two oppositional strategies. The idea is to use the predicted reward together with its confidence interval to reflect the uncertainty of the model given the actual context.

In our contextual bandit, the expected payoff is approximated by a linear model with an (arbitrary) loss function. The uncertainty U of the obtained value for each arm is therefore proportional to the variance σ^2 of the expected payoff, $U = c\sigma$, where σ^2 is estimated from training points in neighboring contexts as well as the model parameters. The uncertainty is added as an upper bound to the prediction to produce a confidence bound for selection strategy across the arms.

2.2 Optimization

Equation (2) can be optimized with standard techniques such as gradient-based approaches. The unconstrained problem needs to be maximized w.r.t. the dual parameters $\boldsymbol{\alpha}$ and is given by

$$\begin{aligned}\sup_{\boldsymbol{\alpha}} \quad & -C\mathbb{I}^\top V^*(-\frac{\boldsymbol{\alpha}}{C}, \boldsymbol{r}) - \frac{1}{2}\boldsymbol{\alpha}^\top Z Z^\top \boldsymbol{\alpha} \\ & - \frac{1}{2\mu}\sum_j \boldsymbol{\alpha}^\top (Z \circ \boldsymbol{\phi_j})(Z \circ \boldsymbol{\phi_j})^\top \boldsymbol{\alpha}.\end{aligned}$$

The gradient wrt $\boldsymbol{\alpha}$ is obtained by setting the derivative to zero.

$$-C\frac{\partial V^*(-\frac{\boldsymbol{\alpha}}{C}, \boldsymbol{r})}{\partial \boldsymbol{\alpha}} - (ZZ^\top - \frac{1}{\mu}[\sum_j (Z \circ \boldsymbol{\phi_j})(Z \circ \boldsymbol{\phi_j})^\top])\boldsymbol{\alpha} = 0$$

The actual form of the gradient depends on the dual loss V^*. Note that instantiations often give rise to more efficient optimization techniques than the general form in Equation (2) allows. Nevertheless, the sketched gradient-based approach will always work in case a general optimiser is needed, e.g., in cases where several loss functions should be tried out. Once the optimal $\boldsymbol{\alpha}^{opt}$ has been found, it can be used to compute the primal parameters. Alternatively, a kernel $K_Z = \phi_Z(Z, Z)$ could be deployed in the dual representation to allow for non-linear and convoluted feature space.

2.2.1 *Instantiation: Squared Loss*

We present the optimization algorithm for the special case of squared loss. The dual of squared loss is given by

$$V^*(-\frac{\alpha_t}{C}, r_t) = \frac{1}{2C^2}\alpha_t^2 - \frac{1}{C}\alpha_t r_t,$$

which leads to the following objective,

$$\begin{aligned}\max_{\boldsymbol{\alpha}} \quad & -\frac{1}{2C}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} + \boldsymbol{r}^\top \boldsymbol{\alpha} \\ & - \frac{1}{2}\boldsymbol{\alpha}^\top [ZZ^\top + \frac{1}{\mu}(\sum_i \boldsymbol{\phi}_i \otimes \boldsymbol{\phi}_i^\top) \circ ZZ^\top]\boldsymbol{\alpha},\end{aligned}$$

where $\otimes$ denotes the vector outer product. Rephrasing the problem as a minimization task and setting $P = \frac{1}{C}\mathbb{I} + ZZ^\top + \frac{1}{\mu}(\sum_i \boldsymbol{\phi}_i \otimes \boldsymbol{\phi}_i^\top) \circ ZZ^\top$, and $\mathbf{q} = -\mathbf{r}$, the task becomes a standard quadratic optimization problem,

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\boldsymbol{\alpha}^\top P \boldsymbol{\alpha} + \mathbf{q}^\top \boldsymbol{\alpha}.$$

The confidence bound for the linear bandit setting with squared loss is given by

$$U = c\sqrt{\boldsymbol{z}_t^\top (Z^\top Z + \lambda I)^{-1} \boldsymbol{z}_t}.$$

REFERENCES

[1] Peter Auer, 'Using confidence bounds for exploitation-exploration trade-offs', *Journal of Machine Learning Research*, **3**, 397–422, (2003).

[2] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.

[3] Yifan Hu, Yehuda Koren, and Chris Volinsky, 'Collaborative filtering for implicit feedback datasets', in *Proceedings of the 8th IEEE International Conference on Data Mining*, pp. 263–272. IEEE, (2008).

[4] Yehuda Koren, Robert Bell, and Chris Volinsky, 'Matrix factorization techniques for recommender systems', *Computer*, **8**, 30–37, (2009).

[5] L. Li, W. Chu, J. Langford, and R. E. Schapire, 'A contextual-bandit approach to personalized news article recommendation', in *Proceedings of the International World Wide Web Conference*, (2010).

[6] Maryam Tavakol and Ulf Brefeld, 'Factored mdps for detecting topics of user sessions', in *Proceedings of the 8th ACM Conference on Recommender Systems*, pp. 33–40. ACM, (2014).

[7] Chong Wang and David M Blei, 'Collaborative topic modeling for recommending scientific articles', in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 448–456. ACM, (2011).