Course Name - Recommender Systems
Professor Name - Prof. Mamata Jenamani
Department Name - Industrial and Systems Engineering
Institute Name - Indian Institute of Technology Kharagpur
Week - 04
Lecture – 20

Lecture 20: Other advanced models

Hello everyone. Today is going to be the last lecture of this module. This module is on collaborative filtering, model based collaborative filtering. In fact, for the entire collaborative filtering series which we did last week and this week this is the last lecture. So, here in the context of model based collaborative filtering we are going to learn about few new ideas. Starting ahead here specifically we will talk about two advanced models.

One is how to integrate the concept of time in collaborative filtering, model based collaborative filtering and then we will see how to integrate neighborhood based models along with this ah latent factors ah latent factor models. Now starting with we will now try extending the baseline predictor that we saw considering the effect of time. Now this temporal effect can it has been observed by the researchers that is temporal effects can significantly improve the accuracy. Now decomposing rating into distinct terms allow us to treat different temporal effects separately.

So, now temporal effects on all the biases and user preferences can vary and on item characteristics they are mostly static in lecture because over the time the item features will not change much. However, there can be little ah there can be little deviations as well. For example, if there is a new movie and it is getting released then probably if it is a good one probably many people will rate it and ratings will be very high. And at a later time the number of ratings will decrease and if let us say the situation is different probably the nature of the rating will also vary. So, however, the variation is not going to be much.

$$b_{ui} = \mu + b_u(t_{ui}) + b_i(t_{ui})$$

$$b_i(t) = b_i + b_{i,\mathrm{Bin}(t)}$$

So, with this idea now let us look at this time aware factor models. So, this time sensitive baseline model can be written in this manner. This is exactly same, but here we have 2 functions that time we were having how many variables number same as that of number of users and number of item I mean the users and the items. Now, this B u will be treating as a function of how the rating is changes over the time how the rating and B i is also a function of how the rating changes in in in over the time. So, these 2 are some real valued function.

Now the question is what are the choices for these functions? So, as we discussed this item bias is not going to change very fast ok. So, typically this is one example actually I am following the recommender system handbook for this in fact, for the last lecture and this one. If you read the chapters corresponding to this, you will see that not only the ones that I am covering there are many models. So, when we say there are many models based on what these models are developed. This basically these models are developed based on the observations that various researchers made with respect to the data sets and the typical domain knowledge that the person the researcher is having.

Now why I am telling this? Discovery of these functions also depends on these observations. Now if somebody is using this linear model that does not mean that model always has to be linear. We are going to of course, limit ourselves to this linear model, but there can be spline functions there can be other functions and not only it is limited to the functions that is there in the book. If you wish you can develop your own function, but you should have sufficient intuition to justify why you are taking that model ok. So, now here in this we are splitting this function into 2 part.

Once one part is a stationary part that is b i and then the other part we are taking bins over the time. So, this part is going to vary if let us say we are observing making a model using the total time horizon of 1 month depending on our nature of the data we can may be taking it 1 week 1 week 4 bins we can make or if the variation is too much and number of observations are much then what you can do? You can take maybe 1 day 1 day 1 day or if your observation period is very large maybe 1 year then maybe 1 month is a good choice. So, it depends. So, how you decide this number of how many bins you are supposed to use it depends on you. So, but however, you are supposed to keep one thing in mind like you have you definitely have a desire to have finer very fine resolution over this.

So, that your bin size it is small, but in that small bin size if you do not have enough rating it is not going to be of much use. So, therefore, you also have to have enough rating for bin. So, larger ratings are good. So, experimentally you can decide what is a good bin size a good number of bins with respect to your kind of data set. Now modeling the user bias is not that straightforward because of two reasons.

First the volatility associated with the users is much more. The temporal effect let us say today user is not in a good mood probably he will is going to give bad ratings. Moreover, there are situations in which one account is used by multiple users. So, therefore, using this concept of isolated bins probably is not going to be appropriate. So, here we are considering two things.

One there can be a gradual drift in the behavior of the user. This can happen when you are talking about a single user who is behaving different at different point of time. But it may so happen that at particular point in time sometimes your rating is going some kind of change in behavior which can be considered as a spike some instantaneous change. And that particularly is going to happen when you have multi user accounts which often is the case. If one Netflix account is taken by your home probably many people will be watching movies from that.

$$\mathrm{dev}_u(t) = \mathrm{sign}(t - t_u) \cdot |t - t_u|^{\beta}$$

$$b_u \ \ (t) = b_u + \alpha_u \cdot \mathrm{dev}_u(t)$$

So, everybody's choice is different and their rating behavior is also going to be different. So, to capture this concept of gradual drift gradual concept drift of user bias we may consider this function. So, here let us say T u is this is the you know the number of days between the dates T u T and T u this is the T is the date. So, this is the difference. And we can put certain beta to give certain weightage to this and the sign the sign function we have already seen if this is more ok.

This is higher then this is going to be negative if this is higher this is going to be positive if both are same this is going to be 0. So, the how you rate depends on the difference between these two times. So, we now add a new parameter and this new parameter is influenced by this difference in time. So, how do we determine this new parameter alpha? This can be determined experimentally through cross validation ok. So, with this idea now we model this not only with this, but we add one additional term.

Now, what is this additional term? Now, besides gradual concept drift in many applications there are sudden drifts which can be called as spikes on a single day on a single session and so on. So, this may so happen with respect to a single person as well if his behavior is different in that particular day or if there is a group effect people are somehow using one account then they can also create some kind of sudden change. So, to address such short lived effects we assign a single parameter per user and day and we try absorbing the day specific variability. So, this parameter is this one ok. So, as I told you this are flexible it is not that you have to go by this model, but for the purpose of this course we are limiting ourselves to this, but depending on the kind of data that you are using this model is going to change.

Now, considering both this is now new user bias. So, this function together is your function of time and this is the function for the what for the item bias. So, this is so these two functions now we can integrate in this model instead of. So, original model was mu plus b u t b i t. So, now we have two functions instead of that.

So, now as usual we try minimizing the regularized error. So, how many regularized errors? How many parameters? All the b u's all the b i's this is m this is n this is m this is this depends on n into number of bins this is n into interval that I have considered how many intervals you have considered. So, this is the model. Now, as usual what will you do with respect to all the variables? You will try taking the partial derivative set it to 0 and derive some kind of update equations. So, after deriving the update equations you will be using stochastic gradient descent to do it to learn them from the data.

Now we are going to look at something called global neighborhood model. In fact, while talking about neighborhood based methods we also saw that how regression concepts can be brought in and instead of using the similarity values which are determined heuristically we we were determining those values from the regression model. So, now in a global neighborhood model this concept is extended. So, you do not rely on these heuristics based item item similarity and why it is called global neighborhood because it mostly considers the item related ah item relationships and why item we cannot call it similarity they are equivalent to similarity, but they are item item relationships. So, these relationships are stable that is why we say this item item based neighborhoods are global neighborhoods ok.

$$\hat{r}_{ui} = b_{ui} + \sum_{j \in \mathrm{R}(u)} (r_{uj} - b_{uj}) w_{ij}$$

$$\hat{r}_{ui} = b_{ui} + \sum_{j \in \mathrm{R}(u)} \left[(r_{uj} - b_{uj}) w_{ij} + c_{ij} \right]$$

$$\hat{r}_{ui} = b_{ui} + \sum_{j \in \mathrm{R}(u)} (r_{uj} - b_{uj}) w_{ij} + \sum_{j \in \mathrm{N}(u)} c_{ij}$$

Now, in the new model we find a solution for the values these relationships that we try to establish and try minimizing the error. So, this is the idea this global optimization avoids user specific weights in favor of global item item weights. It may consider user item bias together or separately the weights from some item j to i we can denote by w i j additionally it may consider implicit feedback from rating matrix or from certain additional source and how to get it from the rating matrix last class we saw. Now, all this parameter it tries learning from the data and the parameter update equations are supposed to be developed as we have discussed in the earlier lecture and we can use stochastic gradient descent with respect to those equations the concept of stochastic gradient descent and slowly read slowly update the values. Now, this is what we have been talking rating at a particular point can be certain bias related to both user and item of course, we can separate them as we can see in the next slide and for all the items which are rated by the single user this actual rating minus this value which is supposed to be predicted into this weight.

So, you give weight to strengthen this rating. So, along with this we can add something which can represent some kind of implicit feedback implicit feedback and this implicit feedback if it comes from the rating matrix it can be part of this and if it comes from some other source where this is no longer the rating matrix, but for this the users feedback is available then this is from this is implicit feedback from rating matrix and another model this can be implicit feedback from other source. So, now with this we move ahead. So, these are basically 3 different models this in this model the concept of only the weight is there for each for the item and in this case with respect to rest of the items and in this case along with that weight we consider one implicit feedback which is available which is derived from this rating matrix and in the this is a third model. So, this is the first model, this is second model, this is third model.

The model

$$b_{ui} = \mu + b_u + \alpha_u \cdot \mathrm{dev}_u(t_{ui}) + b_{u,t_{ui}} + b_i + b_{i,\mathrm{Bin}(t_{ui})}$$

The minimization problem for regularized squared error

$$\min \sum_{(u,i)\in\mathcal{K}} (r_{ui} - \mu - b_u - \alpha_u \mathrm{dev}_u(t_{ui}) - b_{u,t_{ui}} - b_i - b_{i,\mathrm{Bin}(t_{ui})})^2 + \lambda_7 (b_u^2 + \alpha_u^2 + b_{u,t_{ui}}^2 + b_i^2 + b_{i,\mathrm{Bin}(t_{ui})}^2)\,.$$

In fact, you can make any model depending as I have told you any model of your choice, but this can always help you in guiding what was the best practice followed ever ok. So, moving ahead this global neighborhood model can can be seen in many different ways here instead of taking only one bias like our baseline predictor we took mu plus user bias plus item bias and these 3 models that we saw can now be extended and in this particular here we were trying to take the average considering only the L 2 norm, but here we can even consider this L 2 norm this is with respect to all the neighbors this is with respect to nearest k neighbors. So, k with respect to only k nearest neighbor we can do this this one as well now look here this is this additional k is there. We can now think of integrating this user and item factors in this model as well we can think of integrating user and item factor with this. So, how do we with neighborhood model? So, this part comes from the neighborhood model now in this we are limiting ourselves only to items.

$$\hat{r}_{ui} = \mu + b_u + b_i + \sum_{j \in \mathrm{R}(u)} [(r_{uj} - b_{uj})w_{ij} + c_{ij}]$$

$$\hat{r}_{ui} = \mu + b_u + b_i + |\mathrm{R}(u)|^{-\frac{1}{2}} \sum_{j \in \mathrm{R}(u)} [(r_{uj} - b_{uj})w_{ij} + c_{ij}]$$

$$\hat{r}_{ui} = \mu + b_u + b_i + |\mathrm{R}^k(i;u)|^{-\frac{1}{2}} \sum_{j \in \mathrm{R}^k(i;u)} [(r_{uj} - b_{uj})w_{ij} + c_{ij}]$$

So, with respect to item this is the item rating this is the this detail is coming from the neighborhood. So, additionally we can find 2 parameters x j and y j. So, now instead of w we are considering this latent factor and x j which means w is no more one value we have latent factor for all the latent factor that we are considering maybe k number of such latent factors. So, also the situation over here where this w i j is defined by this and this implicit feedback is decided by this. So, considering both the weight and implicit feedback with respect to only item we can come up with this model.

$$\hat{r}_{ui} = \mu + b_u + b_i + |\mathrm{R}(u)|^{-\frac{1}{2}} \sum_{j \in \mathrm{R}(u)} [(r_{uj} - b_{uj})q_i^T x_j + q_i^T y_j]$$

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T \left(|\mathrm{R}(u)|^{-\frac{1}{2}} \sum_{j \in \mathrm{R}(u)} (r_{uj} - b_{uj})x_j + y_j \right)$$

Now, such model you can think of deriving for the users from considering the user similarities or user relationships and we can even combine these items with users to make a factorized neighborhood model with user and item additional parameters for users and items. So, these are again the references and all these contents are taken from this book recommender system handbook. These are my conclusions the temporal effects can significantly improve can significantly improve the accuracy of the latent factor model as has been observed by many researchers. Such models must capture the drifts and spike in the data neighborhood models can be designed considering the latent factors as well then the choice of the model depends on the data in hand and the related domain. Thank you.