Course Name - Recommender Systems
Professor Name - Prof. Mamata Jenamani
Department Name - Industrial and Systems Engineering
Institute Name - Indian Institute of Technology Kharagpur
Week - 07
Lecture - 31

Lecture 31: Introduction to evaluation ofrecommender system

Hello everyone. Today we are going to start a new module. So, in this lecture we are going to introduce the topic. To start with let us see what evaluation is. So, evaluation is a systematic determination and assessment of a subject's merit, worth, significance using some criteria or a set of standards which is established beforehand. So, in this line evaluation of recommender system aims to understand its success in terms of certain measurable criteria.

So, in this series of lectures we are going to look at such measurable criteria and we are going to see how to this matrix these measurable criteria which are also called matrix we will be conceptualizing this and we will try to utilize them with as many examples as possible. So, basically when we talk about evaluating the recommender system there are two perspectives. One is as a researcher or as a algorithm developer you would like to develop new and better algorithms. So, to show how the new algorithm works whether it is more accurate or fast whether some kind of other performance measure is being improved.

So, those things the researcher would like to do. But what the business people would like to do? So, these business people would like to have a different perspective altogether. So, what is that perspective? Whether the users were actually satisfied, whether they got really new items, whether they could get opportunity to explore diverse item, whether their preserve privacy was preserved, how much profit did you generate after introducing your system then could you actually make some effort to change the behavior of the user and in turn generated more profit. So, many of the things many of this topics in the very first lecture of this course we have discussed. So, mostly now we will be focusing on the research perspective from the point of view of the developer.

There are basically three approaches for evaluating recommender system. Offline experiment, user studies and online experiment. Offline user offline experiment on user studies and online experiment. In case of offline experiment, you have some kind of historical data, the data which is already collected and on that you try conducting various experiments. So, there is no interaction with the real user.

Whereas, in both these studies you need a real user. And in this first kind that is user study you have the real user but you have a controlled environment. So, which means you

are available as an evaluator you are available to be there with the subjects and guide them how to participate in the experiment. But in case of online experiment where once again you use real users you really do not directly interact with the users and users are also not aware that they are participating in the experiment. So, this is most unbiased and most trustworthy.

However, there are issues with all this along with these advantages that we are going to see shortly. So, whether it is online, offline or user study. Every time you have to compare quantities you have to establish some hypothesis and you have to see whether your hypothesis is correct. If you say algorithm A is more accurate than algorithm B, then we have to match the some kind of error that they are making. And how do we compare this mirror? Are they statistically same or they statistically different? See two quantities let us say 0.001 and 0.0012 are they same or they are different? After so many decimal points they are appearing are they significant? So, simply by saying looking at the numbers and saying that this is better does not ensure statistical significance of this. So, in this context we have to decide the hypothesis and we have to take our decision depending on whether this is hypothesis can be accepted or it can or it it cannot be accepted. So, various tests can be performed for this. Now, when you conduct this experiment you also have to have some variables.

So, some of these variables will remain fixed and with respect to the other variable you will be checking. For example, if you are comparing two algorithms it is not that for both algorithms you will be using two different. So, when we talk about the variable please do not confuse it with the features or something it is the experimental variable. So, suppose you have 5 datasets and you are going to test 10 algorithms. So, each algorithm it is not that each algorithm will be tested with different data set all the algorithms will be tested with all the datasets.

So, if you do not keep one of the things fixed let us say decide that I will be testing all the algorithms using this dataset. So, what is fixed now? My data size is fixed my algorithms are varying. So, similarly if I decide I will be finding out for this kind of data let us say sparse data what is the better algorithm what is the what is the best among all the 5 algorithms. So, what do I do? I have to now fix my dataset and check which algorithm works best. Which algorithm works is more generalizable.

So, what do I do? If that is my objective I keep the algorithm fixed I change the dataset. So, that is what is the next observation generalization power generalization beyond the experimental data. So, with experimental data you can generalize of course, taking 5, 10 whatever is your available data, but at the same time you can try for analytical proofs you can think beyond the experimental data wherever possible. But let me tell you because all these are machine learning algorithms analytically proving that some algorithm is better

than other will be little tough if not impossible. Now, we start our discussion on offline experiments.

So, offline experiments are the easiest of all because the dataset is already with you. It is pre collected it is historical data and most of the time you have some kind of benchmark datasets available. Your movie lens dataset etcetera are the some kind of standard datasets which are typically used by many developers who are trying different algorithms. Now, while conducting experiments it may so, happen that you have to model the user behavior. Why do you need to model user behavior? Because in the absence of real user how will you check how it is going to affect the rating behavior.

So, some of the part of the rating probably will hide and test with the rest. Now, this offline experiments mostly test accuracy hence they are very restrictive the other approaches like serendipity etcetera they cannot test. During this experiment the experimenter the algorithm developer generally has a tendency to create such a situation so that the accuracy improves. So, you can eliminate some sort of ratings some sort of users with very few ratings, some items on which rating is not available and so on. However, the benefit of offline experiment is both of both user studies as well as online experiment both of them are very expensive.

They usually require real user and modification of the existing system. So, therefore, the major purpose of this offline experiment is you try with various algorithms and you make sure that you select few algorithms which work best in the offline experiment setting and use them in online experiment environment. So, these offline experiments have a tendency to introduce systematic bias in the result due to inappropriate experimental setting. One of these I have already explained the experimenter may try pre filtering the data. So, if it pre filters the data looking at let us say some low count on certain rating drop one user and so on.

The algorithm which works best with that pre filter data may not actually work with the real data. Similarly, he may the experimenter may try randomization in choice. He understands the nature of the data and select certain algorithm. Because this data set is a test data set it may not actually represent the reality. So, which means you can introduce some bias in terms of randomness.

Then when you are talking about this offline data this users bias which we have seen we have discussed in at length at during our discussion on model based collaborative filtering. If you do not take care of this bias probably offline experiments are going to give you misleading results. There are issues with simulating user behavior as such. Now hiding some of the interactions and making the recommendation for this interaction is the process in which you simulate. Now, in this process you can use a fixed number of known items and the fixed number of hidden items per test user.

So, not only that you have to hide ratings sometimes you have to hide the items as well. Now, in case of time stamped data while selecting these samples you have to see that if suppose you are selecting for the period 2 then you should make sure that you are using period 1 data. So, prior data you should be using to predict the future ratings. Now next come to the user studies. As I told you user studies are conducted in a controlled environment controlled environment in the sense you have you have to have your own systems and take the users to use your algorithms sitting in your experimental environment that you have created.

So, during this process 2 things can happen. You can give them certain questionnaire which they will fill up and submit which can be assessed. Second thing that can happen is you have the now the opportunity to observe the users. Maybe you can track their eye movement, browsing activity, number of clicks to reach a particular item and so on. So, in during this questions that you ask you can ask the questions like how the what portion of the task they completed because you will be giving them some tasks based on that you will be judging the quality of the recommendation.

So, preparing that questionnaire is again tricky. Now, whether the results are accurate for users expectation, whether the appropriateness of the what is the appropriateness of the user interface, did you find the item in one click or you took multiple clicks. So, those things are can to be asked and there are certain aesthetics issues as well with the interface design which makes improves the convenience of the user. So, those things also you can ask. The real advantage of the user study is the subjects.

The subjects are the real humans you can directly communicate with them and you have the opportunity to observe these individuals and collect certain qualitative data. The major disadvantage is this is very costly. You have to set up the experimental process scenario, you have to pay the users and because of this process if it is not run properly you may introduce bias. Experiments are after all experiments they may fail. So, suppose they fail and you try utilizing the same subjects again they already know about the experiment.

So, they can introduce some bias. If you are not controlling the interaction among the subjects, they will be talking to each other and they will be getting influenced by each other. So, the answers the exact scenario that you are expecting to emerge you cannot get it. Moreover, you pay them and because you pay them because of the payment they may try to please you. Second thing that can happen because of the payment when there you are repeating an experiment they may find that first experiment failed they may try making the first experiment fail. So, that they can be paid twice or if they are not paid twice they may not be rightly giving the feedback.

So, maybe to know the validity of your questionnaire and your experimental setting you can conduct a set of pilot studies. And when you conduct the pilot studied studies you should be careful that you are getting a small sample of the people the variety of people who will be participating in the final experiment. You have to follow proper design of experimental procedure and questionnaire design is also a problem. Next is your online experiment. Here also you use real users, but in a not in a control environment in the you will be using you will be these real users will be actually participating in the actually be using the recommender system.

So, which means if you are all that them are using the recommender system they are using the same recommender system. No, what you have to do you have to create two versions of recommendation engine possibly using the algorithms that you are trying to trying to adopt. So, programming wise you need additional effort to do so. Again while collecting users activity data it is not that just like your of user studies you will be giving them a questionnaire and you expect that they will be actually using that questionnaire filling up that questionnaire. Here in a online setting if they you must have participated in the beta testing of many online software they ask you certain questions what do you do you try avoiding that.

Is it it most of the time we say that later we will just press the later button and we will go away. So, which means we try avoiding giving feedback. So, getting first of all you need you need to create you need to have a lot of programming effort to create two recommendation engine which will be tested parallely. You create additional effort for creating a questionnaire to get the data, but user itself avoids what are you going to do.

So, it is a very tricky situation. Now, the search engine that you are trying to introduce new search engine the user without users knowledge you are actually diverting them one group to search engine one second group to search engine two. So, what is happening let us say the new reintroduce search engine two sorry the recommendation engine two does not work well. So, the group of user which who are using that will feel that this website from this e-commerce website from which we are trying to get the trying to buy the items is providing irrelevant recommendation. So, it may discourage the user to come to the website again. So, these are the general goals of evaluation design.

Accuracy is something which is the which is mostly used in nature and in offline setting calculating this is it is computationally possible to calculate and get a numerical value for this. However, there are others like coverage confidence etcetera. Now, coming to this accuracy you can measure the accuracy of rating prediction, accuracy of usage prediction and you can get the accuracy of the rank in case of a top n recommendation. Now, there are other evaluation matrices about accuracy and other we are going to talk in detail, but there are other evaluation matrix as well. For example, coverage this coverage can be item space coverage or it can be user space coverage.

So, what is item space coverage the proportion of the item that the recommender system can recommend. What is user space coverage? It is the coverage which shows what proportion of the users the system could recommend the item with. Sales diversity how on equally the items are chosen given a recommender system. So, coverage can also address cold start problem the coverage and the performance of the system on the new item you can find out. Second is confidence, confidence in the recommender system can be defined as the systems trust in its recommendation or prediction.

The confidence in the predicted property typically also grow with the amount of data. The measurement of the measurement in terms of confidence interval can be done for this purpose. So, next metric is your trust. So, users trust in the system of recommendation is difficult to measure and specifically cannot measure in offline experiment. In the user studies or online experiment, we can ask questions and get this value.

Novelty is another method another evaluation metric it is about recommending items that the user did not know about. So, in user studies it can be asked directly it is difficult to measure in offline experiment so also the sensitivity. So, how surprising the successful recommendations are? It can be done by comparing the metadata that is the item details about the product. Next is your diversity. This is about how dissimilar the recommended products are.

And as we know while discussing about content based recommender system we knew that this novelty, serendipity and diversity cannot be taken care of by such systems because we are always trying to measure the similarity with the items which the user has already seen. The next metric is about measuring the utility. So, this utility is measured in terms of something which the recommender system is intended to maximize. For this purpose, you have to build a utility function. For example, if your aim is to maximizing the revenue you have to build a function accordingly and collecting the data you have to test it.

Next is your risk. So, it is measured in terms of variance of the expected utility. Few more evaluation metrics robustness it shows the stability of the recommendation in the presence of fake information. Now, what is this fake information? I have told several times earlier also. Recommendation systems are quite prone to attack. Attack in the sense in case of in case of rating some spurious users can come and give ratings to increase or decrease the average rating of the item.

Similarly, you can also give misleading information about the product features as a result the item is getting recommended. So, the stability of the recommender system in the presence of fake information can be assessed if it can be assessed this metric is called robustness. Then comes your privacy. When you are utilizing the users data there are many privacy issues because you can utilize users demographic data if it is available to

you, you can utilize their rating behavior. So, it should not happen that I am watching a particular kind of movie and somebody else should not know that I am watching such kind of movies such kind of situation should not arise.

So, your algorithm should be such that it should not invade into privacy of specific users. Next is your adaptivity. It can it is about to find out if you can cope with when the item collection changes rapidly ok. So, new items are very fast getting added ok. So, you should be able to predict with the new addition of items how the users interest is shifting.

And the last one is your scalability which shows the capability to scale up to a real large data setting. So, these are my references of course, I did not add a recommender system handbook here this was not particularly useful here rather recommender system handbook. That was the base of all this and I have somehow missed it. So, please read this from recommender system handbook.

So, these are my concluding remarks. The evaluation of recommender system aims to understand its success in terms of certain measurable criteria. Offline experiment user studies and online experiments are three approaches for evaluating recommender system. And most of the literature on performance evaluation are on offline experiments. Now, besides accuracy which is a very well discussed measure for performance there are other criteria like coverage, robustness, mobility, serendipity etcetera. So, those things should be care should be taken care of while doing evaluation. With this we conclude this lecture. Thank you.