Mid-Report(2)

Session-based Purchase Prediction and Explanation

Team GAZUA

2011147064 JuHyuk Lee

2012147562 Inho Choi

2014154009 YooJin Kim

**Contents**

1. Subject
   1. Session-based Purchase Prediction and Explanation
   2. Necessity
2. Previous Research and limitation
   1. Previous Research
   2. Limitation & Our development
3. Current Progress

A. Data Analysis

B. Entropy representation for buyer decision process

C. Session representation for GAN

D. Training ACGAN

1. Future Plan
   1. Dealing with mode collapsing
   2. Prediction and explanation
2. Reference
3. **Subject**
   1. **Session-based Purchase Prediction and Explanation**

Purchase prediction is about predicting user’s product purchase probability, to know whether a user will purchase the product or not. In this project, we’ll use session-based user data to predict the user’s product purchase by developing a neural model.

Also, using neural network model could be hard to understand since it is a black-box model. People wouldn’t trust the output from simple black-box model without reasonable explanation about why the model is reliable. According to this problem, we’ll use an explanation model LIME(Local Interpretable Model-agonistic Explanation) to explain the predictions of our classifier.

* 1. **Necessity**

Recommending a worthy item to a user is important for sellers to gain high profit. Since online market becomes popular to users in these days, lots of sellers use online market for their marketing. Not for only off-line marketing, now also for on-line marketing product recommendation is necessary. To recommend a worthy item for a user in eCommerce market, seller should know which product user is interested in, and wants to purchase finally. For this reason, purchase prediction now becomes important in eCommerce market.

Since session data is a reasonable data to get in online, session-based purchase prediction will be a good solution. And with some reasonable explanation, it could be a reliable solution for eCommerce marketers to manage customers.

1. **Previous Research and limitation**
   1. **Previous Research**
      1. Purchase prediction with user data

In paper about predicting online purchase conversion for retargeting, researchers used logined user data to predict the user purchase model. Given an e-commerce site, each customer has a unique cookie ID. Researchers use this unique cookie ID and its page visit (click or hit) and product purchase.

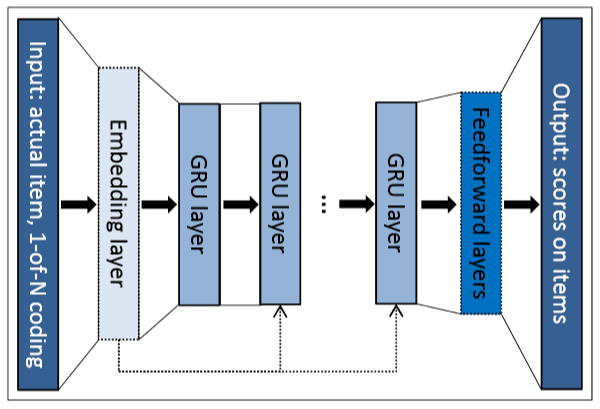
Using this data, researchers study about the conversion rate, which means the rate of conversion from window-shoppers to a real buyer. For the result of their experiments, researchers proposed a joint modeling of customer-level and product-level conversion patterns based on the buying decision process, to get the reliable conversion rate.

* + 1. Session-based Recommendation with RNN

In paper about session-based recommendation with RNN, researchers used short session-based data to build a recommender system with RNN. Since RNN is used for sequential data modeling, their basic idea was trying to modeling the whole session as it is a sequential data.

Researchers used GRU-based RNN, which means elaborating model of an RNN unit dealing with vanishing gradient problem which is showed in Figure 1. Also they used ranking loss function to modify classic RNN, and use ranking method to get top items that the user might be interested in.

For the result of these experiments they got a recommendation system for eCommerce marketers to predict user’s purchase and recommend some reasonable product for the user.



<Figure 1>

* 1. **Limitation & Our development**
     1. Limitation

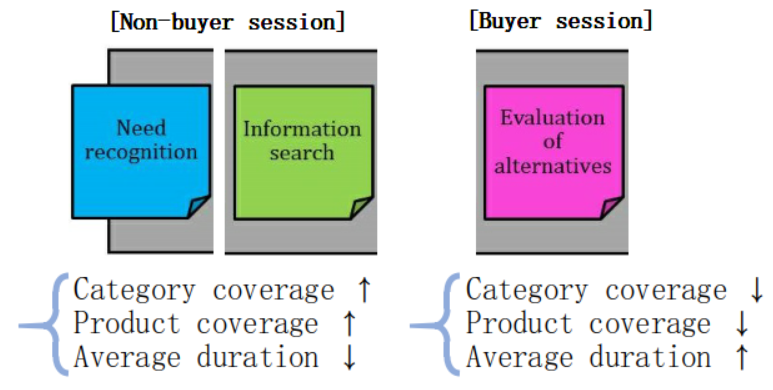
The previous studies have two major big limitations. First, the study about purchase prediction with user data used only logined data of users to purchase their prediction. In their research, used data was with user cookie ID data. Since people usually visit online shop without login, we thought that there must be more reasonable data that related with all users who does not logined.

The study about session-based recommendation with RNN also has a serious problem. In this research, researchers used RNN model to make prediction. However, since RNN is a black-box model, it is hard to understand or trust the output from the model. It needs more reliable explanation model for the output. So we try to find out some more explainable model by refer some explanation method, LIME (Local Interpretable Model-agnostic Explanations).

* + 1. Our development
       1. Our research will find a explainable model for predicting user’s purchase.
       2. To develop a prediction model, we’ll use session-based user data which can be applied to more broad area than research using login user data.
       3. Since the dataset is imbalanced, we’ll use ACGAN to oversample the minor class (buyer session) of the dataset.
       4. We expect that our new prediction model could be applied to eCommerce customer management.

1. **Current Progress**
   1. **Data Analysis**

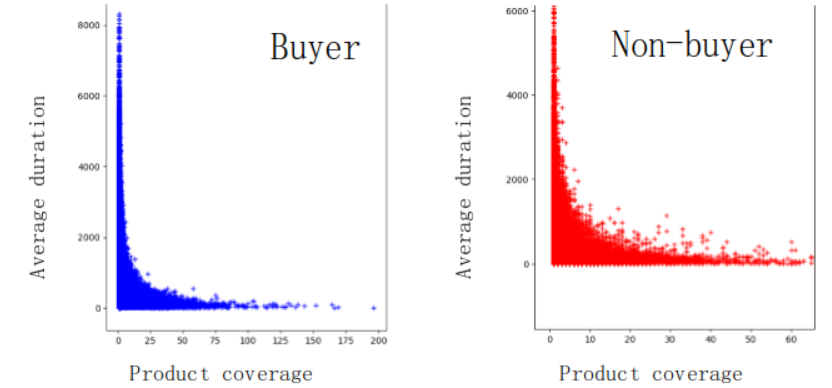
We had some assumptions about category coverage, product coverage, average duration. First, Category coverage of Non-buyer session would be bigger than Buyer session. Second, Product coverage of Non-buyer session would be bigger than Buyer session. Third, Average duration of Non-buyer session would be shorter than Buyer session.



In order to check if our assumptions are correct or not, we did data analysis by next procedure.

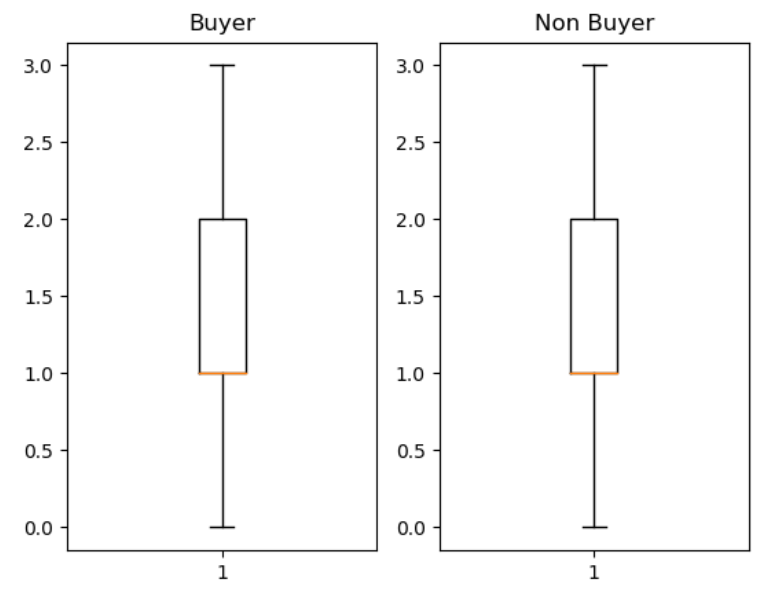
Our data is composed of buyer session(509,696) and non-buyer session(8,740,033). We sampled same number(509,696) of non-buyer session with buyer session. After that, we visualized following four graphs and analyzed them.

**Product coverage - Average duration**



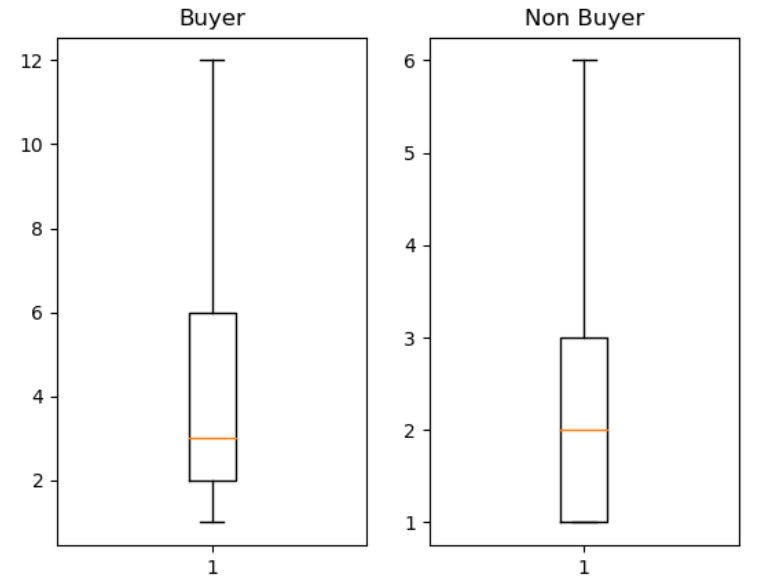
⇒ Product coverage and Average duration are inversely proportional.

**Category coverage**



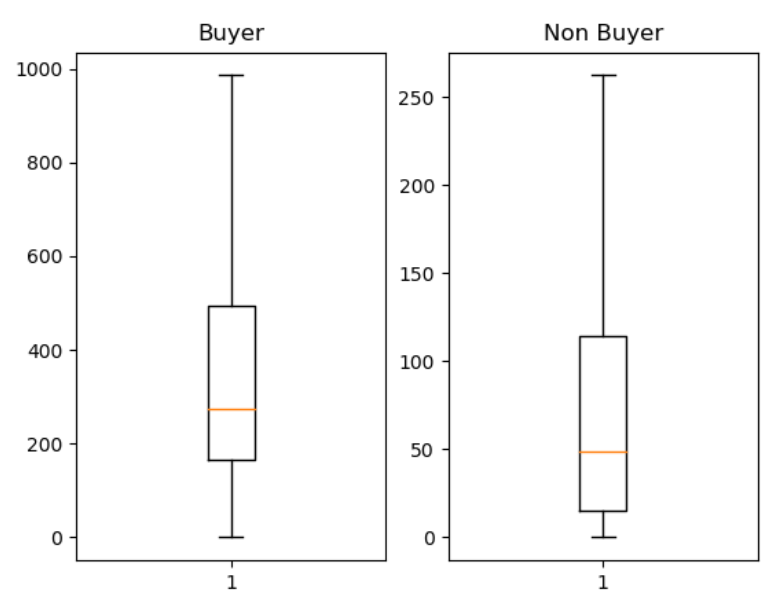
⇒ Category coverage is not an important factor because each graph shows similar coverage.

**Product coverage**



⇒ Product coverage of buyer session is bigger than it of non-buyer session.

**Average duration**

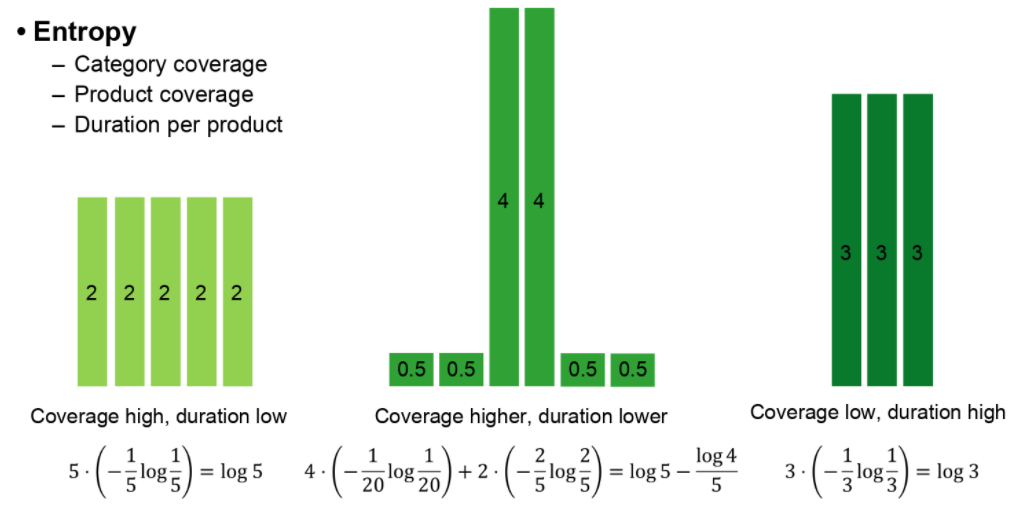


⇒ Average duration of buyer session is much bigger than that of non-buyer session.

As a result, we can figure out that our second and third assumptions are correct but the category coverage is not an important factor. Also, we get the relation between product coverage and average duration.

* 1. **Entropy representation for buyer decision process**

We also analyzed Entropy of session while change the value of Product coverage and Average duration. Entropy is a good way to measure the uncertainty of specific state. The bigger the entropy, uncertainty of that state is bigger. Thus, we can extract whether phase of the session is in the information search or alternative evaluation.



* 1. **Session representation for GAN**

To use Generative Adversarial Neural Network, we need a session representation for our GAN model’s input. Each session is recorded as feature below.

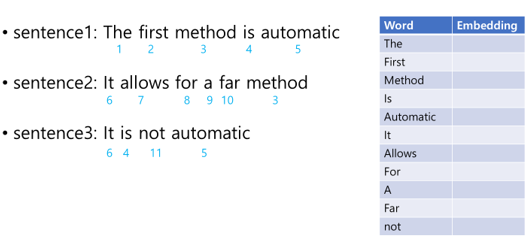


There could be many session data per one session ID, because of timestamp or different items. To transform this session data as GAN model’s input, we get some features from this data.

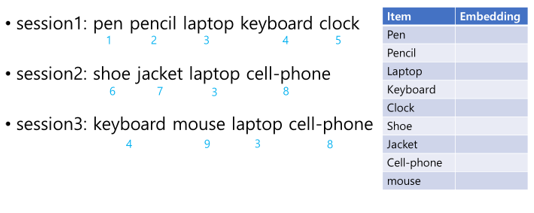
|  |
| --- |
| **Feature from Session Data** |
| (Product, Category) in one session |
| Real category |
| Time slot |
| Average duration of item |
| Average duration of category |
| Total duration |
| Item entropy |
| Category entropy |
| Item count |
| Average item count per category |

* + 1. (Product, Category) in one session

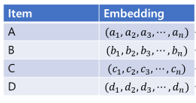
To find out the relationship between items, we used item-to-vector method. This method is from word-to-vector method, which use skip gram. Word2Vec method learns some words that are close with each other in one sentence have close distance. In the example below, for example, words like ‘The’ and ‘first’ will have close distance in skip gram.



From this previous research, we adopted this skip gram method and applied it to session and item distance. In the example below, for example, items in one session have close distance if they are close to each other in each session.



Each item will be embedded by skip gram and the result will be like table below. After get this table, we calculated component wise average for each component.



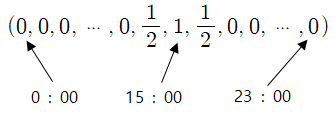
Component wise average

* + 1. Real category

Since there are 12 categories, we can express category by 13-dimension vector. For example, if the item belongs to second, third, eighth, and tenth category, it’s representation will be (0,1,1,0,0,0,0,1,0,1,0,0).

* + 1. Time slot

Session will be represented by time slot. For example, if the session is appeared in 15:00, it’s representation will be like feature below.



* + 1. Average duration of item

Average duration of items can be calculated by item count/total duration.

* + 1. Average duration of category

Average duration of category can be calculated by item count/total duration.

* + 1. Total duration

Total duration can be calculated by timestamps.

* + 1. Item entropy

After calculating total duration and duration of each item, entropy can be calculated by feature below. means total duration, and means duration of item.

* + 1. Category entropy

After calculating total duration and duration of each category, entropy can be calculated by feature below. means total duration, and means duration of category.

* + 1. Item count

We can get item count value by counting distinct item id.

* + 1. Average item count per category

This can be calculated by item count / # of category.

After getting these features, we produce GAN input. The item embedding becomes 128-dimension vector, category is a 13-dimension vector and time slot is a 24-dimension vector. For single values like average of item, category duration, total duration, item count, average item count per category, we produce them into 4-dimension vector with log, sqrt, raw (do nothing), and square. This process is to make feature’s influence more diversely. For item and category entropy, we did not do log exceptionally because much of them was smaller than 1, which can make log value to be negative infinity. After expand dimension, we concatenated all features so total input becomes 191-dimension vector.

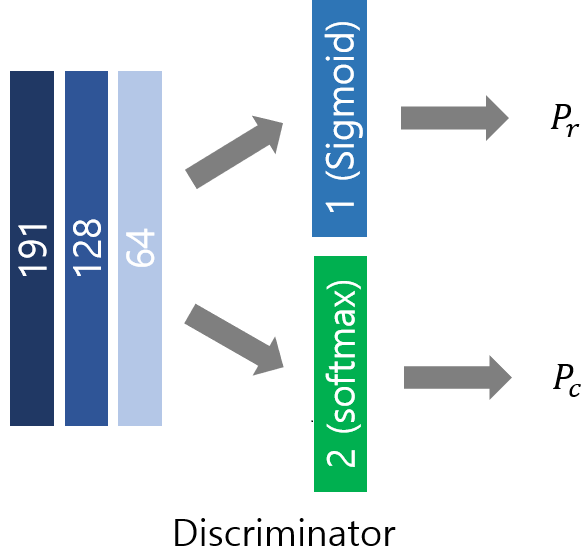
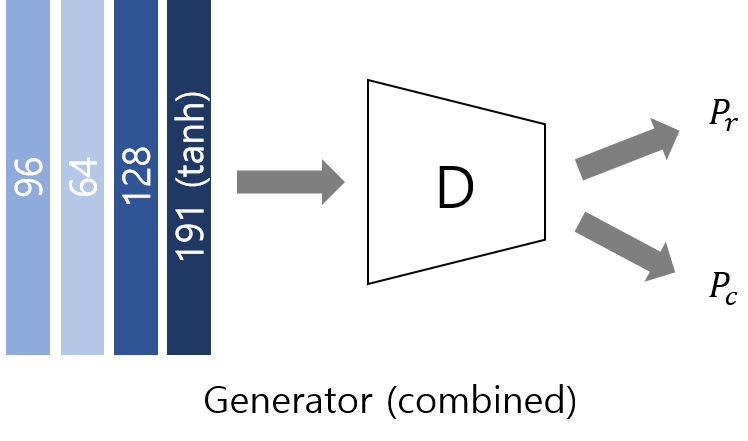
Finally, since generator’s output should pass through the activation function, , all vector’s component should be between -1 and 1. For this reason, we do min-max normalizing to input.

**D. Training ACGAN**

* + 1. ACGAN model

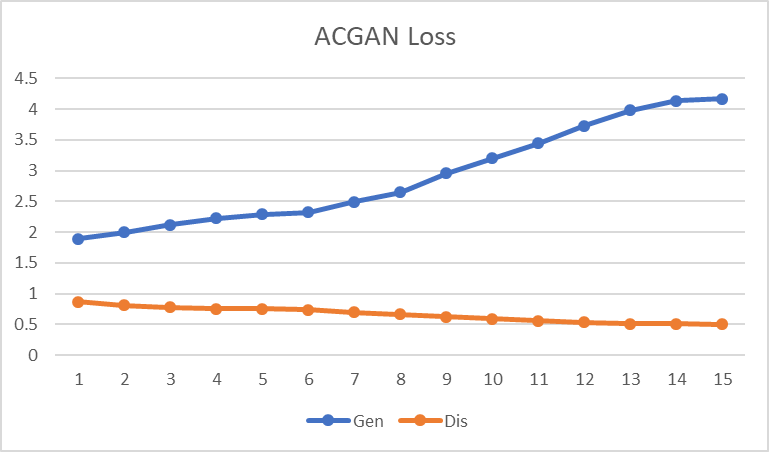
We used adam(adam\_lr, adam\_beta\_1) optimizer with hyperparameters adam\_lr = 0.0002, adam\_beta\_1 = 0.6. And, the size of latent vector was 96.

For training method, we used mini-batch and the batch-size was 128. During the training, we repeated discriminator and generator once at a time alternatively. We set the epoch to 15. We extracted 50,000 fake buyer and nonbuyer session for each epoch.

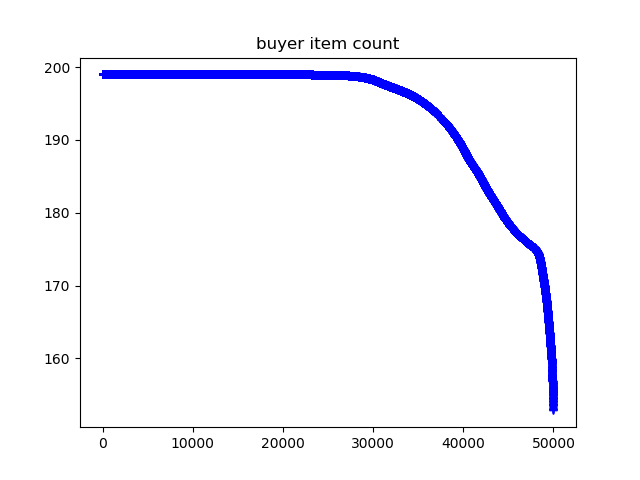
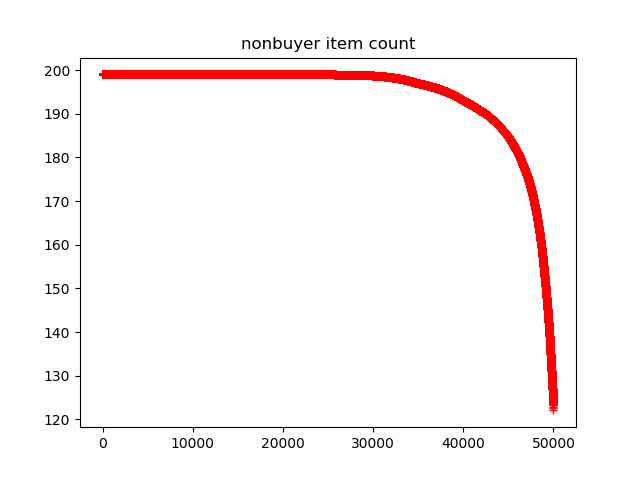


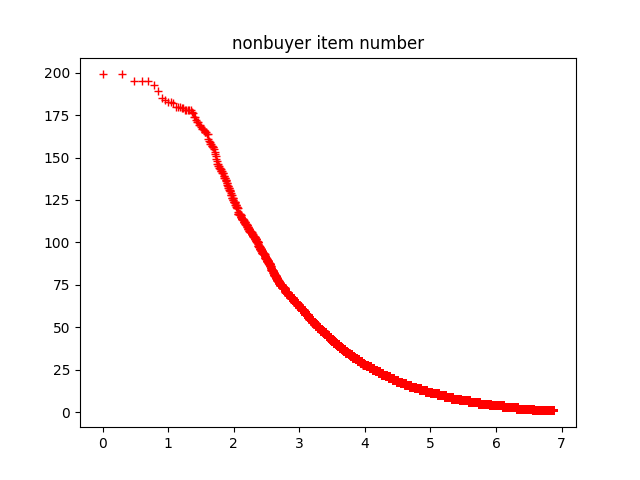
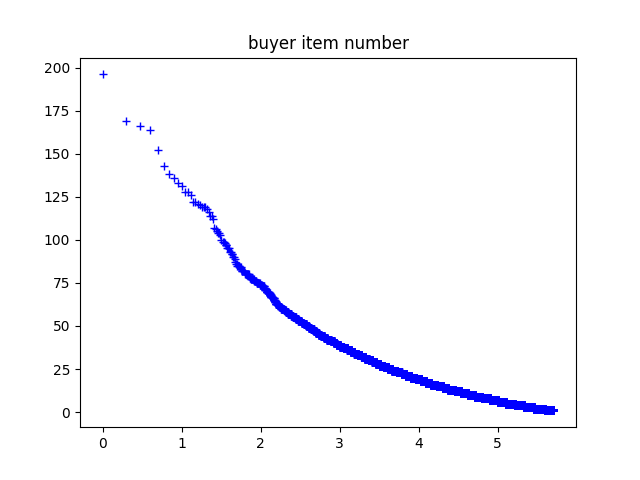
* + 1. Result

The figure shows that generator loss is increasing. However, generator loss and discriminator loss should converge between 0 and 1. This means there is some problem to our model.



To analyze accurately, we compared feature graphs. For example, item number (item count) feature of real session and fake session is totally different. Above figures are item count graph of real buyer session and real nonbuyer session. X axis is session number (0 to number of sessions) and y axis is item count. We took log to x axis to make graph visual. Above figure illustrates that maximum value is 200 and very few sessions have value close to maximum.

However, by observing figures below which are item count graph of fake buyer session and fake nonbuyer session, it is easy to see that so many sessions have maximum value of item count. This problem is well-known and called mode collapsing.



1. **Future plan**
   1. **Dealing with mode collapsing**
      1. **Add distance loss**

Mode collapsing means generator can generate only small portion of real data. Thus, we can overcome mode collapsing by adding distance loss between real and fake data. There are two methods called feature matching and mini-batch discrimination.

* + 1. **Historical averaging**

In general, deep learning model forgets previous training information after many training epochs and this would lead to mode collapsing problem. However, by using historical averaging method, the model can remember before training epoch.

* + 1. **WGAN**

WGAN loss is better than binary cross entropy loss because WGAN loss is proposed to dealing with the problem of JS divergence distance. Thus we can expect that WGAN can solve mode collapsing problem.

* 1. **Prediction and Explanation**

After generating fake buyer session, we are going to combine generated data and original data to train neural model to predict whether sessions in the test set is buyer or non-buyer. Then, we are going to explain the prediction with buyer decision process.

1. **References**
2. "Why Should I Trust You?" Explaining the Predictions of Any Classifier (<http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf>)
3. Session-based Recommendations with Recurrent Neural Networks (<https://arxiv.org/pdf/1511.06939.pdf>)
4. 1시간만에 GAN(Generative Adversarial Network) 완전 정복하기 (<https://www.youtube.com/watch?v=odpjk7_tGY0>)
5. Conditional Image Synthesis With Auxiliary Classifier GANs (<https://arxiv.org/pdf/1610.09585.pdf>)
6. eCommerceGAN : A Generative Adversarial Network for E-commerce (<https://arxiv.org/abs/1801.03244>)
7. Predicting Online Purchase Conversion for Retargeting (<https://dl.acm.org/citation.cfm?id=3018715>)
8. Predicting User Purchase in E-commerce by Comprehensive Feature Engineering and Decision Boundary Focused Under-Sampling (<https://dl.acm.org/citation.cfm?id=2813517>)
9. RecSys Challenge 2015 (<http://2015.recsyschallenge.com/challenge.html>)
10. Item2Vec: Neural Item Embedding for Collaborative Filtering (<https://arxiv.org/vc/arxiv/papers/1603/1603.04259v2.pdf>)
11. Deep Neural Networks for YouTube Recommendations ([https://static.googleusercontent.com/media/research.google.com/ko//pubs/archive/45530.pdf](https://static.googleusercontent.com/media/research.google.com/ko/pubs/archive/45530.pdf))