Where will a new Airbnb user book their first travel experience...

Mohamed-Laid HEDIA
2018280581
Department of Computer Science and Technology
Tsinghua university
Beijing, China
xinli18@mails.tsinghua.edu.cn

Abstract

This document provides a report for the final Machine Learning course project. It addresses the chosen problem and the data set used, and discuss the different techniques used to handle this problem.

1 Introduction

The cold start problem is a very diffucult challenge on recommandation systems. Actually, the most famous recommandation algorithms use the relation user/item (e.g: Collaborative Filtering [1]) to know which item will interest the user. Or, the new users are a challenging problem for these kind of systems. This problem which intend to resolve Airbnb, a privately held global company that operates an online marketplace and hospitality service which is accessible via its websites and mobile apps. This has been in the form of a challenge on the kaggle platform [2]

1.1 Data

Airbnb provides around 213,000 sample of users and there are 11 countries as destinations(France, Germany...). These destinations are the targets that we have to predict: this may be also "no destination found" which means the user haven't booked. So the goal of this challenge is to predict in which destination will book a new user. A great amount of information is provided for each user: the signup mehod and date, the device used, the signup-flow... etc Other interesting informations are provided: the sessions of some users; In fact, the challenge provides the actions that some users have realized on the website/mobile application before make a reservation and these actions are sorted chronologically.

The percentage of the users whose have session informations is 34% (Figure 1) and they are around 73 000 which is a significant number of samples.

1.2 Published solutions

As the competition is dated from 2015, many solutions are published on internet. Most of these models are built using XGBoost. This kind of models is powerful on challenges and especially when there is a huge number of features per row. So, the accuracy by using these techniques on this competition was usually around 0.88 which is a good accuracy. The idea here is to try something different

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

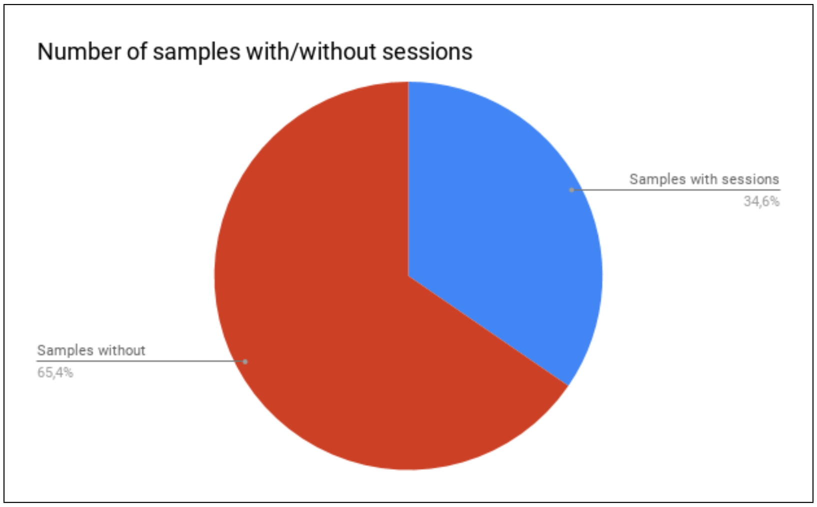

Figure 1: Number of rows with/without session informations

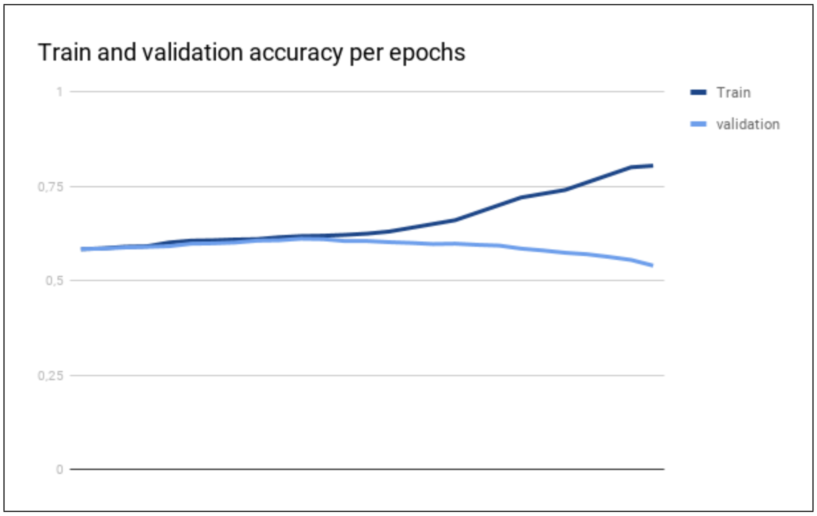

Figure 2: The accuracy of train and validation set during the training

2 Approaches

2.1 Getting started

To getting started, first, I tried the fully-connected neural network. Using the techniques to avoid overfit problem (regularization), this model has diffuculty to converge to an optimal accuracy and it stucks in an accuracy of 0.60 despite the great number of epochs. However, whthout regularization the model continue to converge but only on the train data and not on the validation set: there is an overfit.

This overfit (Figure 2) is due mainly by the huge number of features. In fact, as mentioned above, the challenge provides a big number of information for each user and the majority of this features are in form of categories: about 300 different actions, 50 different browsers and many others. So, by encoding these features in one hot vectors the number of features by user will be around 1000.

2.2 Use sessions informations and Long-Short Term Memory network

The usage of recurrent neural network here make this approach different. In fact, Long-Short term memory is a kind of recurrent neural network that is able to learn through time steps. The idea here

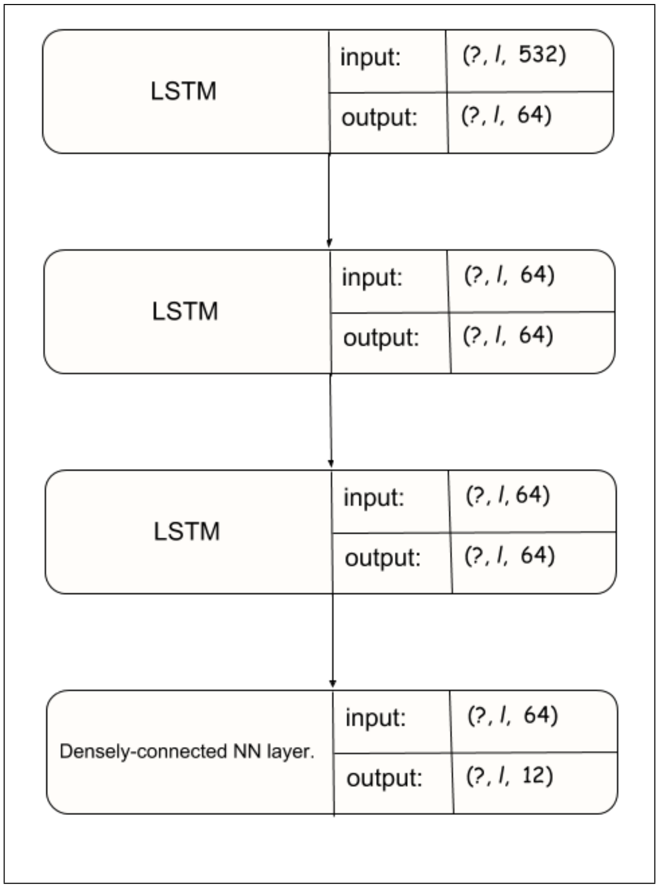

Figure 3: Network architecture

to use the sessions of each user as time step and then try to predict the destination. First to define a time step x_i for an user j: x_i will contain informations about the user (age, gender...) and of course informations about the action like the type of the action, spent time, the device used.. the time here is normalized as the following equation

$$t_i = \frac{(x_i - \mu)}{\sigma} \tag{1}$$

With μ and σ being respectively the mean and the variance of the spent time vector on each action.

To define the size of the window:

$$l = \max_j n_j \tag{2}$$

with n_j being the number of action x_i for the user j in the data set. In other words, the size of the window l will be the maximum number of actions for an unique user. So, for each user, if the number of actions is less than l than the window will be filled by zero vectors.

Network architecture The architecture of the network is 3 stacked LSTM networks followed by a densely-connected neural network layer. This architecture will take as input a window of size l. This window will contain the actions realized by a user to make a reservation (or not). In the other side, and thanks to the densely connected layer, the output of the network will be a vector of size 12 which is the prediction of the destination for user j. See Figure 3

3 results and discussions

As we can see in Table 1 the accuracy of this model is also 0.60 but this time after just 5 epochs and trained only on 22 000 users to avoid memory problems. Due to the big number of dimensions

Table 1: Accuracy score

	Train	Validation
Accuracy	0.6035	0.6085

((2238, 532) for each window) and to the complexity of the network used, one epoch takes more than 3 hours and requires a lot of memory. So the next step will be how to reduce these dimensions: Make another definition for a window ? Feature selection ?

Resolve this challenge differently was not an obvious thing. In fact, this idea was be the result of finding how to use session informations as features in the fully-connected model. And by doing some research I found many approaches to handle a goal recognition problem: the goal recognition means here predict in which country intend to book a user by regarding the session informations

Some papers suggest to handle this kind of problems by using LSTM [3] but the architecture used in my work is different. The results got in this architecture can be explained by the big number of features passed to the network and the definition of a window.

Also, the data has some missed values(age, gender...) and the pre process wasn't an easy task.

As conclusion, The approach of the Long-Short term memory network and using the sessions of users to predict the first destination of the new users got the results as the fully-connected Layers model. However, this approach resolve the problem differently and can be performed to solve other problems using sessions informations.

4 Other trials after the presentation

I tried other things after the presentation. First I started by reducing the number of features by applying some features techniques selctions. This came to the same results as the other techniques. Secondly, I tried other network architectures, by using architectures more complex and more suitable for multi class classification, but the same thing the results are usually around 0.60. After doing some researches in the competition forum, it seems be very diffucult to get good results just using neural networks. Some solutions use neural networks but combined with other techniques (random forest, XGBoost...) and then they do a kind of regressions using the predictions of each models (assigning a weight for each model).

5 Conclusion

Thanks to this Project I learned a lot of things: features engineering, feature selection, goal recognition and many others. It was a great experience to work in a classification problem but as we can see this problem wasn't an obvious thing. I tried to handle this problem using different techniques and thanks to that I could discover interesting techniques those I didn't know before.

References

[1] R. Zhang, Q. Liu, Chun-Gui, J. Wei and Huiyi-Ma, "Collaborative Filtering for Recommender Systems," 2014 Second International Conference on Advanced Cloud and Big Data, Huangshan, 2014, pp. 301-308. [2] `https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings`

[3] Leonardo Amado, Ramon F Pereira, Joao P Aires, Mauricio Magnaguagno, Roger Granada, and Felipe Meneguzzi. Goal recognition in latent space. In Proceedings of the 2018 International Joint Conference on Neural Networks, IJCNN'18. IEEE, 2018