Planned solution of recommender systems' analysis based on different methods

Faculdade de Ciências e Tecnologias - Universidade Nova de Lisboa (DI - Web Search Course)

Carlos Quendera (c.quendera@campus.fct.unl.pt)
David Pais (dm.pais@campus.fct.unl.pt)
Rebekka Görge (r.gorge@campus.fct.unl.pt)

1 THE PLANNED SOLUTION

In this document we want to present our planed solution for the project on the analysis of recommender systems using some implicit methods. After making an introduction on recommender systems, implicit methods, and the data set used, we are going to approach the chosen click through-rate algorithms. We decided to analyze "deep factorization machines" and "product based neural networks" in detail (if we have time, other methods will be analysed*). For this, we will base our analysis on chapters 16.9 and 16.10 of the book "Dive into deep learning" and also on some papers about the different methods for click-trough rate predictions and their implementations [1]. The base for our project is the criteo data set [2]. Next, the planed table of contents is visible.

1.1 Table of contents

1.2 Recommender systems

We start with an introduction on the general subject of recommender systems. We want to give an overview over the use of recommender systems and the different types of methods that can be used.

1.3 Implicit methods

Here, we introduce the subject of implicit methods. Implicit feedback consists of actions as clicks, including items viewed/bought, but not explicitly rated by users, which may indicate what the user likes. We decided to analyse implicit methods due to being a bigger challenge in terms of modeling the user behaviour and interests, in order to make useful recommendations.

[1] https://paperswithcode.com/sota/click-through-rate-prediction-on-criteo
[2] https://www.kaggle.com/leonerd/criteo-small

1.4 The criteo data set

In this chapter we will describe the criteo data set and the preprocessing made. The criteo data set is a well known benchmark data set for click-through rate predictions. It contains 45 million data samples and each sample has 13 numerical features, 26 categorical features, and a label (1 if the user clicked the advertisement, 0 otherwise). We are going to cutout the data set, using a smaller version of it (with 1 million data samples), since the original one it is not computable in useful time with the computational resources we have. Furthermore, we will split it into a train and test data sets.

1.5 Click-through-rate prediction

In the field of implicit methods we want to focus on click-through rate predictions (CTR). Interaction data indicates in a basic way the preferences and interests of users. Click-through rate is a metric that measures the number of clicks on a link in ratio to the total number of users on the page. A click trough rate prediction predicts the likehood that a link on a website will be clicked. In order to optimize the results obtained, these algorithms use hot encoding of categorical features.

1.5.1 Deep Factorization Machines. The first method we want to focus on are deep factorization machines. For the deep factorization machine deep learning and factorization machines are combined. This is useful as factorization machines model features into a linear paradigm and this is insufficient for real world data. Again the online advertising data is used.

1.5.2 Product based neural networks. The second method we want to analyse in this paper are product based neural networks (PNN) These PNNs have a layer to learn a distributed representation of the categorical data, a product layer to capture interactive patterns between interfield categories, and further fully connected layers to explore high-order feature interactions.

1.6 Comparison of the different methods

This chapter presents a comparison between the used methods and their results on the criteo data set. In order to compare the results obtained from each algorithm we will use several evaluation metrics: AUC (area under ROC curve), Logloss (cross entropy), and others.