Item2vec in the Expedia Hotel Recommendations

Domain Background

I currently work on the team responsible for recommendations at OLX as a software engineer. In this capstone project, I decided to work on a recommender system based on the item2vec paper [1] and apply it in the "Expedia Hotel Recommendations Kaggle competition" [2].

Item2vec is a recent implementation of the ideas from word2vec for recommendations. It represents multiple features of an item as embeddings, combine those embeddings in a multi-dimensional vectorial space and items closer to it other in this space have higher similarity.

The "Expedia Hotel Recommendations" was launched in 2016 and its goal was to predict the hotel cluster more likely to be purchased from a user click stream. Using user data and latent item data for recommendations are classified as hybrid recommenders, implementing item2vec in this context will be challenging so I don't expect to achieve state-of-the-art results.

Even though my team builds and maintains a few recommender systems, I never build one myself, so this experience will be very rewarding to me.

Problem Statement

This project is defined as: applying item2vec in the Kaggle competition "Expedia Hotel Recommendations".

By leveraging a Kaggle competition I will be able to work on a common dataset and benchmark my solution against its leaderboard. I don't expect to beat the winner of the competition with my approach, my goal, however, is to have a score that would at least put me in the top 20%.

Datasets and Inputs

The dataset is provided by Expedia for this Kaggle competition. It consists of the following files:

train.csv and test.csv - logs of customer behavior, and its action regarding a hotel offer.
destinations.csv - hotel search latent attributes (mainly reviews)

Solution Statement

My proposal is to create recommender system based on item2vec using PyTorch to predict the hotel cluster more likely to be purchased analysing a log stream. The dataset consists of nearly 38M log events and 62k destinations latent data.
I will try to leverage Sagemaker for training and predicting if I see that it will simplify the development of my model.

Benchmark Model

The Leaderboard from this Kaggle competition serves as a benchmark. I will submit my model and compare my Mean Average Precision @ 5 score against the current Leaderboard to evaluate the performance of my model.

Evaluation Metrics

Every Kaggle competition defines a metric the competitors are optimizing for. In this competition the metric is: Mean Average Precision @ 5 (MAP@5). This metric is defined as:

$$MAP@5 = \frac{1}{|U|}\sum_{u=1}^{|U|}\sum_{k=1}^{min(5,n)} P(k)$$

where |U| is the number of user events, P(k) is the precision at cutoff k, n is the number of predicted hotel clusters.

Project Design

I will follow a standard machine learning workflow:

- Data exploration and data cleaning
- Model development and feature engineering
- Model training and tuning
- Evaluation

Tools I intend to use are:

- Python 3.7
- PyTorch
- Amazon Sagemaker

In this project I will provide all my auxiliary jupyter notebook used for data exploration and model development and the code of the model itself.

[1] https://arxiv.org/abs/1603.04259
[2] https://www.kaggle.com/c/expedia-hotel-recommendations/data