Machine Learning Engineer Nanodegree Capstone Proposal

Jonathan Leo Qi Xiang

12 August 2018

Domain Background

This project was a past competition titled “Expedia Hotel Recommendations” [1] from Kaggle. Hotel booking can be daunting and tedious task as there are hundreds of hotels to choose from at each destination. Through this competition, Expedia aims to provide personalized hotel recommendation to its users by predicting the likelihood the user will stay at different hotel groups. Some interesting references that I will be referencing include Vik Paruchuri from Dataquest [2] who has an interesting take on using non machine learning methods and from Aditi from USCD [3] who recommended a few good machine learning algorithms for this project. I am interested in this project because I personally would like to experience better hotel recommendations and improve my hotel booking experience.

Problem Statement

Expedia has in-house algorithms to form hotel clusters, where similar hotels for a search (based on historical price, customer star ratings, geographical locations relative to city centre etc) are grouped together. The goal here is to predict the booking outcome (hotel cluster) for a user event, based on their search and other attributes associated with that user event provided in the logs of customer behaviour. I would model this as a multi-class classification problem.

Datasets and Inputs

The customer behaviour dataset that Expedia provided includes information like what customers searched for, how they interacted with search results (click/book), whether the search result was a travel package. There is a total of 23 features. Table 1 below shows the schema of the datasets.

Column name	**Description**	**Data Type**
date_time	Timestamp	string
site_name	ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk, Expedia.co.jp, ...)	int
posa_continent	ID of continent associated with site_name	int
user_location_country	The ID of the country the customer is located	int
user_location_region	The ID of the region the customer is located	int
user_location_city	The ID of the city the customer is located	int
orig_destination_distance	Physical distance between a hotel and a customer at the time of search. A null means the distance could not be calculated	double
user_id	ID of user	int
is_mobile	1 when a user connected from a mobile device, 0 otherwise	tinyint

is_package	1 if the click/booking was generated as a part of a package (i.e. combined with a flight), 0 otherwise	int
channel	ID of a marketing channel	int
srch_ci	Checkin date	string
srch_co	Checkout date	string
srch_adults_cnt	The number of adults specified in the hotel room	int
srch_children_cnt	The number of (extra occupancy) children specified in the hotel room	int
srch_rm_cnt	The number of hotel rooms specified in the search	int
srch_destination_id	ID of the destination where the hotel search was performed	int
srch_destination_type_id	Type of destination	int
hotel_continent	Hotel continent	int
hotel_country	Hotel country	int
hotel_market	Hotel market	int
is_booking	1 if a booking, 0 if a click	tinyint
cnt	Number of similar events in the context of the same user session	bigint
hotel_cluster	ID of a hotel cluster	int

Table 1: Schema of the train and test datasets

The dataset is split into the training data from 2013 to 2014 and test data from 2015. Training data includes all the users in the logs, including both click events and booking events. Test data only includes booking events. However, the test dataset did not have the hotel cluster column which is not useful to me as I cannot evaluate my models with the test dataset. I will instead sample about 7% of the training data as my test dataset.

There is also an additional destination dataset which consists of 149 features extracted from hotel reviews text.

Solution Statement

Since this is a multi-class classification problem. I will use supervised machine learning algorithms to predict the hotel clusters. Models such as K nearest neighbour, Random Forest, Stochastic Gradient Descend and Neural Network can be trained on the training dataset with user search features and hotel cluster labels. I will then predict the hotel cluster in the test dataset using the trained models and use metrics such as accuracy, precision, recall or mean average precision @5 to determine the best model.

Benchmark Model

The benchmark model that I will be using is a naïve model that always predict the hotel cluster to be the cluster that appears most frequently in the training dataset. The hotel cluster that appears most frequently can be determined by plotting a frequency histogram of the hotel clusters in the training dataset.

Evaluation Metrics

I will be using the Mean Average Precision @ 5 (MAP@5) as the evaluation metric. The same metric was used for the Kaggle competition. The formula [4] is as follows:

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

where |U| is the number of user events, P(k) is the precision at cutoff k, n is the number of predicted hotel clusters. It basically means we are making 5 predictions per user event. For each prediction, calculate the precision at that position and sum all 5 precisions, then take the average across all user events.

Project Design

This project will be broken down into several categories:

Data Exploration

- Investigate correlation between hotel cluster and other features by creating a correlation matrix heat map.
- Find out the hotel cluster distribution by plotting a frequency histogram to check if the clusters are evenly distributed. The cluster with the highest frequency will also be used as the prediction for the benchmark naïve model.

Data Pre-processing

- Sample new test dataset, with only users that has made a booking, from training dataset given.
- Drop rows with missing data in both training and test dataset.
- Add new features to training dataset by reducing the dimension of the destination dataset using PCA.

Train Models & make predictions

- Models that I am considering includes K-nearest neighbour, Random Forest, Stochastic Gradient Descend and Neural Network.
- Models will be trained using the training dataset and be used to make hotel cluster prediction for the test dataset.

Evaluate Models

- Models will be evaluated using MAP@5 based on the predictions that are made with the test dataset. The model with the highest MAP@5 will be the best model.

Fine Tune Best Model

- The parameters of the best model will be fine tuned using the grid search method.

Acknowledgment

[1] Kaggle Expedia Hotel Recommendation Challenge

https://www.kaggle.com/c/expedia-hotel-recommendations

[2] Vik Paruchuri. How to get into the top 15 of a Kaggle competition using Python

https://www.dataquest.io/blog/kaggle-tutorial/

[3] A.Mavalankar, A.Gupta, C. Gandotra and R. Misra. Hotel Recommendation System

https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a038.pdf

[4] Kaggle Expedia Hotel Recommendation Challenge - Evaluation Metric

https://www.kaggle.com/c/expedia-hotel-recommendations#evaluation