Personalized Expedia Hotel Searches – 1st place

ICDM 2013 – Dallas, 8 December 2013

Author: Owen Zhang

Compiled and presented by Adam Woznica, PhD

Agenda

- Preprocessing / Feature Engineering
- Models
- Remarks / Observations

Agenda

- Preprocessing / Feature Engineering
- Models
- Remarks / Observations

Preprocessing Steps

- Missing value imputation
  - Imputed with a negative value
- Bounding numerical variables (e.g. price)
- Down sampling negative instances
  - Faster learning

Five groups of features

- All original features
- Numerical features averaged over
  - srch_id
  - prop_id
  - destination_id
- Composite features
- EXP features
- Estimated position

Composite features

Feature name	Description
price_diff_from_recent	Difference between hotel price and recent price
price_order	order of the price within same srch_id
…	…

EXP Features: categorical features converted into numerical features

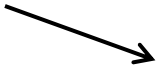
- Each factor F replaced with an average of the target variable related with F, excluding the current observation

W(x, y) – weighted average of x and y

Cat. feature	Target
A	1
A	1
A	0
C	0
C	0

Factor A	Factor C
0.5	0
0.5	0
1	0
0	0
0	0

Factor A	Factor C
W(0.5, 0.4)	W(0, 0.4)
W(0.5, 0.4)	W(0, 0.4)
W(1, 0.4)	W(0, 0.4)
W(0, 0.4)	W(0, 0.4)
W(0, 0.4)	W(0, 0.4)



0.4: overall average of the target

Estimated position

- EXP feature of position based on prop_id/dest_id/target_month
- Position of the same hotel in same destination in the previous and next search
- Average of the two above

Agenda

- Preprocessing / Feature Engineering
- Models
- Remarks / Observations

Ensemble of Gradient Boosting Machines (GBM)

- R GBM implementation (NDCG loss function)
- Two types of models
  - without EXP features (A)
    - 5000 elementary trees
    - 30 hours to train
  - with EXP features (B)
    - 2500 elementary trees
    - 20 hours to train

26 GBM models

Model Type	EXP feature included	Problem fix	# Instances Trained
A1	N	N	8
A2	N	Y	2
B1	Y	N	12
B2	Y	Y	4

$$\text{Final score} = \frac{\sum A1 + \sum A2}{10} + 2 * \frac{\sum B1 + \sum B2}{16}$$

Agenda

- Preprocessing / Feature Engineering
- Models
- Remarks / Observations

Remarks / Observations

- Most important features:
  - Position
  - Price
  - Location desirability (ver. 2)
- Random impressions are not fully random
- Down sampling negative instances improves training time *and* predictive performance
- Ideas:
  - Release user id

Thank you