Predicting Expedia Hotel Cluster Groupings with User Search Queries

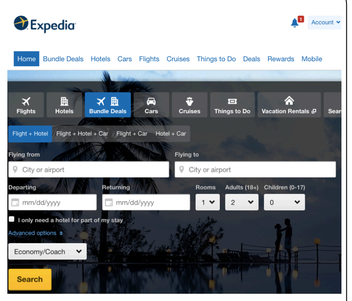
Jarrod Cingel and Liezl Puzon

Stanford University

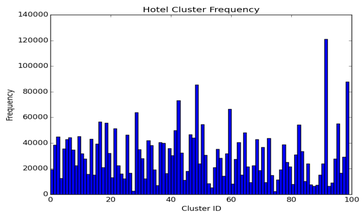
Project Objective

- Provide smarter suggestions for Expedia users by predicting what group of hotels a user will book a hotel from based on certain search criteria

Data



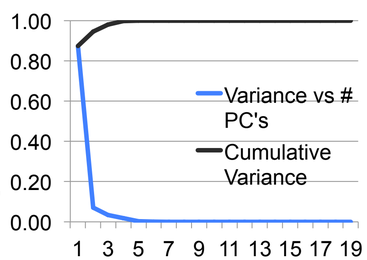
- **Source: Kaggle.com**
  - Input: user queries on Expedia.com
  - Output: Hotel cluster booking
  - Hidden test data on Kaggle
- **Statistics**
  - Training data
    - N = 3,000,693
  - Test Data
    - N = 2,528,243
  - 100 hotel clusters
  - No strong individual correlations between 19 features and hotel_cluster



- **Preprocessing**
  - Normalization [0, 1] for each feature
  - Standardization (N~(0,1))

Feature Selection

- **PCA**
  - 5 PC's account for 99% of the variance
- **Wrapper method** to select 5 features with **forward search**
  - Average k-fold cross validation score as evaluation function



Methodology

Model Selection and Combination

- **Generalize SVM** Prediction to compute an ordered list of 5 clusters
  - n(n-1)/2 one vs. one classifiers
  - Each classifier gets a "vote"
  - Sort the cluster numbers in decreasing order by number of votes, and choose 5
- **Decision Tree**: 5 with highest probability
- Unique **proportional ensembling** of of both classifiers
  - Multipliers for each classifier normalized and made proportional to generalization error
    - Higher weight to the model with **lowest generalization error**
  - For each item in each of the 2 lists, calculate score by position
- Use 5 highest-scoring items in sorted order

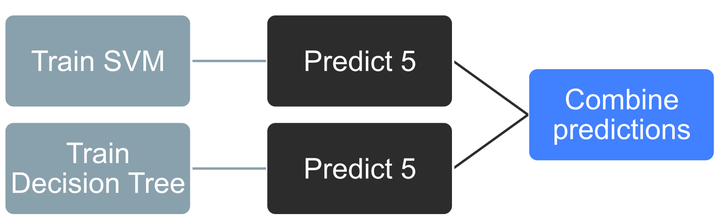
$$c_i = \frac{1 - \mathrm{err}_i}{\sum_{j=1}^{2}(1 - \mathrm{err}_j)} \qquad \mathrm{score}_i := \mathrm{score}_i + c_i(5 - n_i)$$

Scoring

- MAP@5 (**Mean Average Precision @ 5**) score to evaluate list of 5 predictions for hidden test data

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

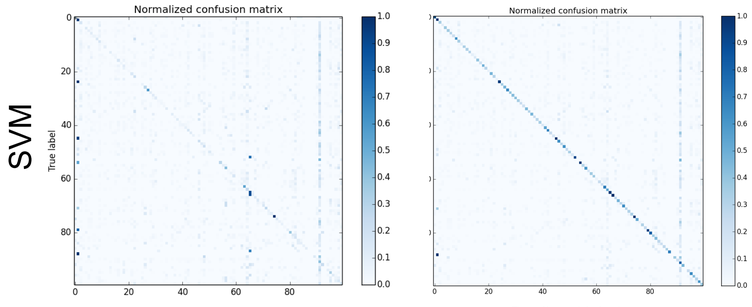
|U| = number of user events
P(k) = precision at cutoff k
n = number of predicted hotel clusters.



Results

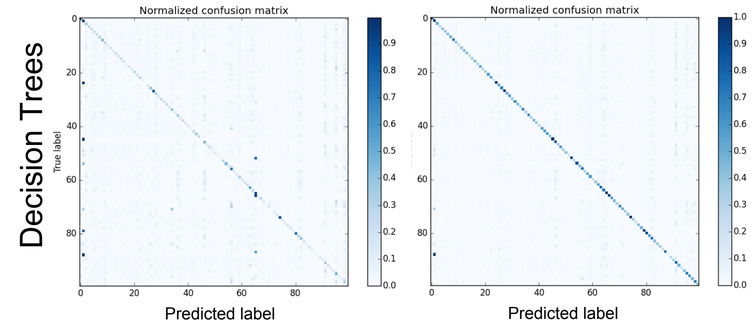
Cross Validation on Local Training Set

First guess only | All five guesses



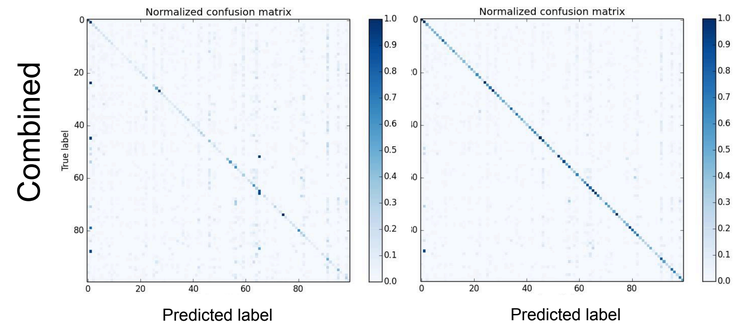
Precision
Mean: 0.620
Range:
0.198 (id 71)
0.999 (id 24)

Recall
Mean: 0.525
Range:
0.048 (id 88)
1.000 (id 74)



Precision
Mean: 0.620
Range:
0.067 (id 91)
1.000 (id 24)

Recall
Mean: 0.525
Range:
0.029 (id 14)
1.000 (id 24)

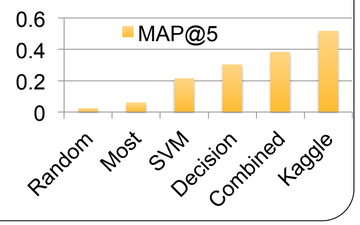


Precision
Mean: 0.670
Range:
0.000 (id 24)
1.000 (id 35)

Recall
Mean: 0.533
Range:
0.000 (id 24)
1.000 (id 27)

MAP@5 on Hidden Set

Optimal model combination has greater precision than the individual SVM, decision tree models

0.38315	8	Thu, 02 Jun 2016 20:00:57

MAP@5 Score