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 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

Scoring

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

\[
\text{MAP@5} = \frac{1}{|U|} \sum_{k=1}^{5} \min(U, n_k) \frac{P(k)}{P(k) + n_k}
\]

\(|U| = \text{number of user events}
\]

\(P(k) = \text{precision at cutoff } k
\]

\(n_k = \text{number of predicted hotel clusters.}

Results

Cross Validation on Local Training Set

- First guess only
- All five guesses

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

Precision

<table>
<thead>
<tr>
<th>Mean</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.620</td>
<td>0.198 (id 71) 0.999 (id 24)</td>
</tr>
</tbody>
</table>

Recall

<table>
<thead>
<tr>
<th>Mean</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.525</td>
<td>0.045 (id 88) 1.000 (id 74)</td>
</tr>
</tbody>
</table>

MAP@5 on Hidden Set

- Precision

<table>
<thead>
<tr>
<th>Mean</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.670</td>
<td>0.000 (id 24) 1.000 (id 35)</td>
</tr>
</tbody>
</table>

- Recall

<table>
<thead>
<tr>
<th>Mean</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.533</td>
<td>0.000 (id 24) 1.000 (id 27)</td>
</tr>
</tbody>
</table>