Categorical Feature Compression via Submodular Optimization

Mohammad Hossein Bateni, Lin Chen, Hossein Esfandiari, Thomas Fu, Vahab Mirrokni, and Afshin Rostamizadeh

Pacific Ballroom #142
Why Vocabulary Compression?
Why Vocabulary Compression?

Embedding layer

Huge!

Video ID: ~7 billion values
99.9% of neural net
How to Compress Vocabulary?
How to Compress Vocabulary

Group similar feature values into one.

Good compression preserves *most information of labels.*

U.S. -> U.S./Canada
Canada
China
Japan -> Chn/Jpn/Kor
Korea

Supervised
Problem Formulation
Problem Formulation

<table>
<thead>
<tr>
<th>User ID</th>
<th>Feature</th>
<th>Compressed feature</th>
<th>Favorite fruit (label)</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1843</td>
<td>China</td>
<td>China/Japan/Korea</td>
<td></td>
</tr>
<tr>
<td>#429</td>
<td>Japan</td>
<td>China/Japan/Korea</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#9077</td>
<td>Brazil</td>
<td>Brazil/Argentina</td>
<td></td>
</tr>
</tbody>
</table>

Max $I(f(X); C)$

s.t. $f(X)$ can take at most $m$ values

Random variable $X \in \{\text{Afghanistan, Albania, ..., Zimbabwe}\}$

Compressed feature $f(X) \in \{\text{China/Japan/Korea, Brazil/Argentina, U.S./Canada}\}$

Random variable $C \in \{\text{pear, apple, ..., mango}\}$
Our Results
Our Results

There is a \textit{quasi-linear} ($O(n \log n)$) algorithm that achieves 63\% $f(\text{OPT})$ if label is \textit{binary}.

- Design a new submodular function after re-parametrization

There is a $\log(n)$-round distributed algorithm that achieves 63\% $f(\text{OPT})$ with $O(n/k)$ space per machine.

- $k$ is \# of machines
Reparametrization for Submodularity

- Sort feature values $x$ according to $P(X=x|C=0)$.
- A problem of placing separators
- $I(f(X); C)$ is a function of the set of separators.
Experiment Results

The graphs illustrate the average MI loss (log-scale) and log-loss for different vocabulary sizes, comparing submodular, divisive (MI), and divisive (U) strategies. The y-axis values range from $1.00E-09$ to $1.00E-04$ for both MI and log-loss graphs. The x-axis represents the vocabulary size in thousands, ranging from 25 to 150.
Pacific Ballroom #142

See you this evening