Ultra Fast Medoid Identification via Correlated Sequential Halving

Tavor Z. Baharav, David N. Tse
tavorb,dntse@stanford.edu

Stanford University

Problem Formulation

\[ x_1, \ldots, x_n \in \mathbb{R}^d \quad i^* = \arg \min_{i \in [n]} \theta_i \quad \theta_i = \frac{1}{n} \sum_{j=1}^n d(x_i, x_j) \]

- High dimensional generalization of median
- Unlike mean, medoid is inside dataset

Computation \rightarrow Estimation: Bandits!

- Estimate \( \theta_i \) via random sampling
  \[ \mathbb{E}\{d(x_i, x_j)\} = \theta_i \quad J \sim \text{Unif}([n]) \]
- RAND: estimate each \( \theta_i \) to same degree of accuracy
  \[ \hat{\theta}_i = \frac{1}{|J|} \sum_{j \in J} d(x_i, x_j) \]
- Medoid Bandit (Med-dit): sample adaptively (UCB) [1]
  - Can we do better than UCB?

Intuition

- UCB ignores structure of problem: consider dist matrix D

```plaintext
\begin{array}{c}
\text{UCB} \\
\begin{array}{c}
\|d(x_i, x_j)\| \\
\vdots \\
\|d(x_i, x_j)\|
\end{array}
\end{array}
```

- Can overcome via correlating our sampling
  - Sample rows of D, \( D_{ij} = d(x_i, x_j) \)
- Need to prove \( \hat{\theta}_i < \hat{\theta}_j \)
  - Control \( \hat{\theta}_i - \hat{\theta}_j \) instead of \( \hat{\theta}_i \) \( \hat{\theta}_j \)

Simulation Results

<table>
<thead>
<tr>
<th>Dataset, Metric</th>
<th>n, d</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-Seq 20k, ( \ell_1 )</td>
<td>20k, 28k</td>
</tr>
<tr>
<td>Netflix 100k, cos</td>
<td>100k, 18k</td>
</tr>
<tr>
<td>MNIST Zeros, ( \ell_2 )</td>
<td>6424, 784</td>
</tr>
</tbody>
</table>

- Figures arranged top to bottom, left to right, following the table

Theorem Statement

- Notation: \( d(x_i, x_j) - d(x_i, x_j) \) is subgaussian
- Theorem: \( \text{corrSH} \) identifies the medoid within \( T \) distance computations with probability at least

\[
1 - 3 \log n \exp \left( \frac{T}{16 \sigma^2 \log n} \cdot \min_{i \in [n]} \Delta_i^2 \right)
\]

Summary

- Convert computational problem to statistical estimation
- Fast randomized algorithm for data science primitive
- Incorporating structure of the computational problem in this reduction can yield massive gains
- Similar approach can work for k-NN

References