Learning Nearest Neighbor Graphs from Noisy Distance Samples

Blake Mason, Ardhendu Tripathy, & Robert Nowak
Motivation

Wish to learn ‘most similar’ or ‘closest’ items to a given from noisy measurements
Motivation

Wish to learn ‘most similar’ or ‘closest’ items to a given from noisy measurements

amazon.com/discover
Motivation

Wish to learn ‘most similar’ or ‘closest’ items to a given from noisy measurements
Motivation

Wish to learn ‘most similar’ or ‘closest’ items to a given from noisy measurements.

We don’t know the given a priori. We want to answer ‘closest’ queries for any item quickly!
The Nearest Neighbor Graph Problem

$\mathcal{X} = \{x_1, \ldots, x_n\}$ is a set of $n$ points with unknown distance function $d(\cdot, \cdot)$. In few queries to a noisy distance oracle $Q(\cdot, \cdot)$, learn

$$x_{j^*} := \arg\min_{x \in \mathcal{X} \backslash \{x_j\}} d(x_j, x) \quad \forall j \in [n],$$

that are all correct with probability at least $1 - \delta$.  

Sharma et al. (2015)
Preliminaries and Notation

• $\mathcal{X} := \{x_i\}_{i=1}^{n}$ and $x_j^* := \min_{x \in \mathcal{X} \setminus x_j} d(x_j, x)$

• $Q(i,j)$ yields a realization of $d(x_i, x_j) + \eta$ where $\eta$ is a 1-sub-Gaussian random variable

• $\Delta_{i,j} := d(x_i, x_j) - d(x_i, x_i^*)$
Outline of ANNTri

- Iterate over \( \{x_1, \ldots, x_n\} \) in order and find \( x_{j^*} \) correctly w.p. \( 1 - \delta/n \) in the \( j^{th} \) round.
Elimination via the triangle inequality

\[ d_{i,j} \leq d_{i,l} + d_{j,l} \text{ and } d_{j,k} \geq |d_{k,l'} - d_{j,l'}|, \]

so \[ d_{i,l} + d_{j,l} < |d_{k,l'} - d_{j,l'}| \ \Rightarrow \ x_k \neq x_{j*} \]
Triangle Inequality Bounds
Theoretical Results

Theorem 1. In the good event when all the conf. intervals are valid, which occurs with probability \( \delta \), simplified ANOVA can be computed on the graph by making

\[
O \left( \sum_{j=1}^{n} \sum_{k=1}^{n} A_{j,k} + \sum_{k<j} 1_{[A_{j,k}] (0, \infty]} - \sum_{k,j} H_{k,j} \right)
\]

- Worst case complexity is always \( O(n^2) \)
- In general, order matters
Theoretical Results

• Often, we can do better:

\[ \{ x_k : d_{i,k} < 6C_{\delta/n}(1) + 2d_{i,j} \} \subseteq C_m \quad \forall i, j \in C_m \]
Theoretical Results

- An example of separation:
Theoretical Results

**Theorem 2:** For hierarchical datasets of $\nu$ clusters, ANNTRi learns the correct nearest neighbor graph in

$$O(n \log(n) \Delta^{-2})$$

samples where

$$\Delta^{-2} := \frac{1}{n\nu} \sum_{i=1}^{n/\nu} \sum_{j,k \in C_i} \log(n^2/(\delta \Delta_{j,k})) \Delta_{j,k}^{-2}$$

is the average number of samples between nearby points and is due to the noise.
Experimental Results

- Simulated data

- 100 points in $\mathbb{R}^2$
- 10 clusters of 10 points
- Euclidean distance
- Gaussian noise, $\sigma^2 = 0.1$
Experimental Results

- Compare against Random sampling
- Test effect of triangle inequality
Experimental Results

- The metric is (2d) Euclidean
- We can compare against (distance) matrix completion
- With a distance matrix, the graph can be computed easily
Experimental Results

• What shoes are most similar?
Experimental Results

- What shoes are most similar?
- 85 images from UTZappos50K dataset
Experimental Results

- What shoes are most similar?
- 85 images from UTZappos50K dataset
Experimental Results

Click on the two most similar shoes
Experimental Results

• What shoes are most similar?
• 85 images from UTZappos50K dataset
• Human judgements collected by Heim et al., (2015).

\[ d_{i,j} := \mathbb{E}_{k \sim \text{Unit}(\mathcal{X} \setminus \{i,j\})} \mathbb{E}[\mathbf{1}_{E_{i,k}^j} | k] , \]

the probability that \( i, j \) are not chosen when queried with random \( k \).
Experimental Results

- What shoes are most similar?
- 85 images from UTZappos50K dataset
Main takeways for ANNTri

1. ANNTri finds the nearest neighbor graph for general metrics using the triangle inequality

2. Only requires access to noisy oracle

3. In favorable settings, requires $O(n \log(n) \Delta^{-2})$ queries versus $O(n^2 \Delta^{-2})$ needed by brute force!