Statistical Topological Data Analysis

Chaitanya Ahuja[1]
Bhuwan Dhingra[1]

[1]Language Technologies Institute
Carnegie Mellon University

Contents

Topological Data Analysis

- **Topological Data Analysis (TDA)** refers to data analysis methods which study properties such as shape, topology and connectedness of the data.
- This includes:
  - Clustering (particularly Density Based Clustering)
  - Density Modes and Ridge Estimation
  - Manifold Learning / Dimension Reduction
  - Persistent Homology
- TDA is useful as a visualization tool and for summarizing high-dimensional datasets.

This Project

- We review recent work [1] on performing statistical inference for *Density Trees*—a particular class of hierarchical clustering methods.
- Outline:
  - Definitions and Tree Topology
  - Constructing confidence sets via bootstrap
  - Pruning trees to remove insignificant features
- As an application, we generate density trees to visualize distribution of words in documents

Density Trees

Suppose the data lies in $\mathcal{X} \subset \mathbb{R}^d$. Given a density function $f : \mathcal{X} \to [0, \infty)$,

- Let $T_f(\lambda)$ denote the connected components of the upper level set $\{x : f(x) > \lambda\}$. These are the high density clusters at level λ.
- The density tree is the collection of all such clusters: $\{T_f\} = T_f = \cup_\lambda T_f(\lambda)$.
- This is a tree by construction, i.e. if $A, B \in \{T_f\}$, then either $A \subset B$, or $B \subset A$ or $A \cap B = \phi$.

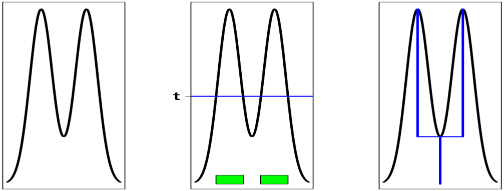


Figure: Obtained from [2]

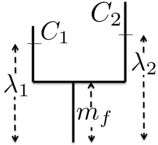
Estimated Tree

In general we have an iid sample from the true density $X_1, X_2, \ldots, X_N \sim p$. The **Estimated Tree** $T_{\hat{p}_h}$ is the tree constructed from the Kernel Density Estimate:

$$\hat{p}_h(x) = \frac{1}{nh^d} \sum_{i=1}^{N} K(\frac{\|x - X_i\|}{h})$$

$$T_{\hat{p}_h}(\lambda) = \{x : \hat{p}_h(x) > \lambda\}$$

Tree Topology



- Given a tree $\{T_f\}$, we can define the **tree distance function** between elements of the tree:

$$d_{T_f}(C_1, C_2) = \lambda_1 + \lambda_2 - 2m_f(C_1, C_2) \qquad C_1, C_2 \in \{T_f\}$$

- It can be shown that d_{T_f} is a metric on $\{T_f\}$, and hence induces a metric topology on it.

Lemma

If the true unknown density p is a morse function, then $\exists$ a constant $h_0 > 0$, such that $\forall h$ s.t. $0 < h \leq h_0$, the true cluster tree, T_p and the estimated tree $T_{\hat{p}_h}$ have the same metric topology above.

Hence we do not need to let the KDE bandwidth $h \to 0$. This leads to a dimension-independent rate of convergence for the bootstrap confidence set.

Confidence Sets via Bootstrap

- To construct confidence sets, we first need a metric to measure the "closeness" of two trees. The l_∞ metric is defined as,

$$d_\infty(T_p, T_q) = \sup_{x \in \chi} |p(x) - q(x)| = \|p - q\|_\infty$$

- The confidence set is defined as $C_\alpha = \{T : d_\infty(T, T_{\widehat{p}_h}) \leq t_\alpha\}$ for T_{p_h}.
- t_α can be obtained by the bootstrap:

$$\begin{aligned} \hat{F}(s) &= \frac{1}{B} \sum_{i=1}^{B} \mathbb{I}(d_\infty(\tilde{T}^i_{p_h}, T_{\widehat{p}_h}) < s) \\ \hat{t_\alpha} &= \hat{F}^{-1}(1 - \alpha) \end{aligned}$$

Where $\{\tilde{T}^1_{p_h}, \ldots, \tilde{T}^B_{p_h}\}$ are the estimated trees for the bootstrap samples $\{\tilde{X}^1_1, \ldots, \tilde{X}^1_n\}, \ldots, \{\tilde{X}^B_1, \ldots, \tilde{X}^B_n\}$.

Convergence Rate

Theorem

Under regularity conditions on the kernel, the constructed confidence interval is asymptotically valid and satisfies,

$$\mathbb{P}\left(T_p \in \hat{C}_\alpha\right) = 1 - \alpha + O\left(\frac{\log^7 n}{nh^d}\right)^{\frac{1}{6}} \tag{1}$$

where $\hat{C}_\alpha = \{T : d_\infty(T, T_{\hat{p}_h}) \leq \hat{t}_\alpha\}$

From the Lemma presented previously, we can fix h to a small constant, to obtain a dimension-independent rate of $O\left(\frac{\log^7 n}{n}\right)^{\frac{1}{6}}$.

Notions of Tree Simplicity

- The confidence set $\hat{C}_\alpha$, contains infinitely many trees—including very complex ones obtained by small perturbations of the density estimate.
- We would like to obtain "simple" trees by removing statistically insignificant features.
- A notion of simplicity is given by the following partial ordering:

Definition

For any $f, g : \mathcal{X} \to [0, \infty)$ and their trees T_f, T_g we say $T_f \preceq T_g$ if $\exists$ a map $\Phi : \{T_f\} \to \{T_g\}$ which preserves set inclusion relationships, i.e. for any $C_1, C_2 \in \{T_f\}$ we have $C_1 \subset C_2$ iff $\Phi(C_1) \subset \Phi(C_2)$.

- This partial ordering matches intuitive notions of simplicity, for e.g. if T_f is obtained by removing edges from T_g, then $T_f \preceq T_g$.

Pruning Rules

Following two strategies are suggested to prune the empirical tree $T_{\hat{p}_h}$:

1. **Pruning leaves:** Remove all leaves of the tree with length less than $2\hat{t}_\alpha$.
2. **Pruning leaves and internal branches:** Remove all leaves and internal branches of the tree with *cumulative length* less than $2\hat{t}_\alpha$.

It can be shown that the tree obtained after pruning from either of these two strategies,

- Is simpler than $T_{\hat{p}_h}$.
- Is generated from a valid density function.
- And the density function lies in the constructed confidence set.

Visualization of Word Embeddings

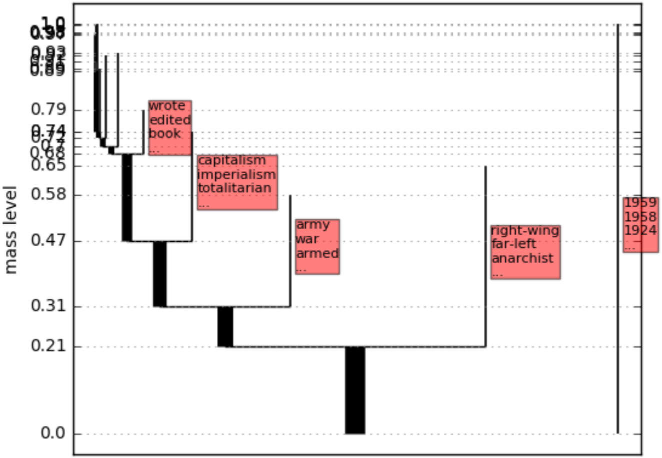


Figure: Cluster tree for Wikipedia Page on **Noam Chomsky**

Visualization of Word Embeddings

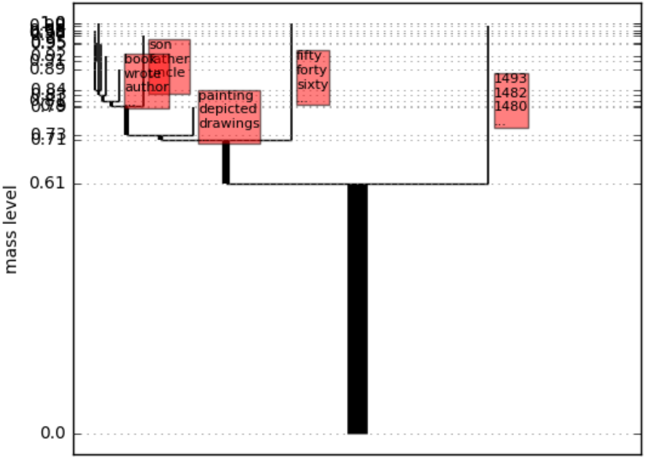


Figure: Cluster tree for Wikipedia Page on **Leonardo da Vinci**

References I

Jisu Kim et al. “Statistical Inference for Cluster Trees”. In: *arXiv preprint arXiv:1605.06416* (2016).

Larry Wasserman. “Topological Data Analysis”. In: *arXiv preprint arXiv:1609.08227* (2016).