Cover Trees

Mike Izbicki

November 22, 2016

1 Introduction

Definition 1. A metric space is a set $\mathcal{X}$ and a distance function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

2 Implementation

A cover tree satisfies the following properties.

Invariant 1. Every node p has an associated integer $\texttt{level}(p)$. For all nodes $q \in \texttt{children}(p)$, $\texttt{level}(q) < \texttt{level}(p)$.

Invariant 2. Every node p has an associated real number $\texttt{covdist}(p) = 2^{\texttt{level}(p)}$. For all nodes $q \in \texttt{children}(p)$, $d(p, q) \leq \texttt{covdist}(p)$.

Invariant 3. For all nodes $q_1, q_2 \in \texttt{children}(p)$, $d(q_1, q_2) \leq \texttt{covdist}(p)$.

3 Analysis

Definition 2. A ball is defined as

$$B_{\mathcal{X}}(x, \delta) = \{y : y \in \mathcal{X}, d(x, y) \leq \delta\}. \tag{1}$$

Definition 3. A δ-packing of a set $\mathcal{X}$ with respect to a distance d is a set $\{x_1, x_2, ..., x_M\} \subseteq \mathcal{X}$ such that $d(x_i, x_j) > \delta$ for all distinct $i, j \in [M]$. The δ-packing number $M_\delta(\mathcal{X})$ is the cardinality of the largest δ-packing.

Definition 4. The double-packing number of a set $\mathcal{X}$ is defined as

$$\text{dpnum}(\mathcal{X}) = \max_{x \in \mathcal{X}, \delta \in \mathbb{R}^+} M_\delta(B(x, 2\delta)). \tag{2}$$

The double-packing dimension is defined to be the base 2 logarithm of the double-packing number. That is,

$$\text{dpdim}(\mathcal{X}) = \lg \text{dpnum}(\mathcal{X}). \tag{3}$$

Lemma 1. Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be two sets satisfying $\mathcal{X}_1 \subseteq \mathcal{X}_2$, and let d be a metric over both sets. Then $\text{dpnum}(\mathcal{X}_1) \leq \text{dpnum}(\mathcal{X}_2)$.

Proof. Let x be a point in $\mathcal{X}_1$ and $\delta \in \mathbb{R}^+$. Then any valid δ-packing of $B_{\mathcal{X}_1}(x, 2\delta)$ is also a valid δ-packing of $B_{\mathcal{X}_2}(x, 2\delta)$. $\square$

Lemma 2. Let $\mathcal{X}_1 \subset \mathcal{X}_2$, and x be a point in $\mathcal{X}_2$ but not in $\mathcal{X}_1$. Then,

$$\text{dpnum}(\mathcal{X}_1 \cup \{x\}) \leq \text{dpnum}(\mathcal{X}_1) + 1. \tag{4}$$

Proof. Let p be a maximal δ-packing of $\mathcal{X}$. Assume for contradiction that there exists a δ-packing p' of $\mathcal{X} \cup \{x\}$ such that $|p'| > |p| + 1$. If $x \notin p'$, then p' is a δ-packing of $\mathcal{X}$. But $|p'| > |p|$, which violates the assumption that p is maximal. If $x \in p$, then the set $p' - \{x\}$ is a packing of $\mathcal{X}$. $\square$

Lemma 3. Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be metric spaces with the same distance function d. Then,

$$\text{dpdim}(\mathcal{X}_1 \cup \mathcal{X}_2) \leq \text{dpdim}(\mathcal{X}_1) + \text{dpdim}(\mathcal{X}_2). \tag{5}$$

Definition 5. The radius of a dataset is defined as

$$r(\mathcal{X}) = \max_{x_1, x_2 \in \mathcal{X}} d(x_1, x_2), \tag{6}$$

the dispersion of a dataset is defined as

$$d(\mathcal{X}) = \min_{x_1, x_2 \in \mathcal{X} : x_1 \neq x_2} d(x_1, x_2), \tag{7}$$

and the condition number of a dataset is their ratio

$$\kappa(\mathcal{X}) = \frac{r(\mathcal{X})}{d(\mathcal{X})}. \tag{8}$$

Lemma 4. The depth of a cover tree is bounded by the log of the condition number.

Definition 6. The doubling dimension of a metric space $(\mathcal{X}, d)$

$$c = \lg \max_{x \in \mathcal{X}} \frac{\mu B(x, 2\delta)}{\mu B(x, \delta)} \tag{9}$$

Theorem 1. Insertion takes time $O(c^{12} \log n)$.

Example 1. Consider a data set of m points in $\mathbb{R}$. Let $x_0 = 1$, and $x_t = x_{t-1}/2$. Then there is a valid cover tree over this data set with height m.