Data Science and Big Data

Summer Term 2019

3 - Exact Nearest Neighbor Search

submission deadline: May 22, 2019 before the lecture

solutions will be discussed: May 22, 2019

Tutor: Tobias Elvermann (t.elvermann@gmail.com)
Grading: Rajkumar Ramamurthy (ramamurt@cs.uni-bonn.de)

1 Exact Nearest Neighbor Search in High Dimensions [15 points]

Run an experiment comparing the runtime of exact Nearest Neighbor Search

1. using kd-trees
2. using the brute force algorithm

on datasets of 10000 random points and 1000 query points for dimensions $d < 100$. Plot and comment the runtime behavior of initialization and querying runtimes for the different approaches.

Hint: You are not required to implement the methods yourself. For python, e.g., the `scipy` library contains a kd-tree implementation and an effective method for pairwise distance computation in the module `scipy.spatial`.

2 Radius of the Inscribed d-Ball [15 points]

Give your answer (with proof) to the question on slide 13 of Lecture 2019-05-08, i.e. is there a finite value x with

$$\lim_{d \to \infty} r_d = x \ ?$$

3 Volume of d-Balls and d-Cubes [15 points]

Let $\text{ball}(d, r)$ be a d-dimensional ball with radius r centered at the origin and let $\text{cube}(d, a)$ be a d-dimensional cube centered at the origin with sides of length a.

1. (5 points) What fraction of the volume of $\text{ball}(d, r)$ lies in the ϵ-*shell* $\text{ball}(d, r) \setminus \text{ball}(d, r - \epsilon)$ for some $\epsilon \in (0, r)$ (cf. Slide 19 of Lecture 2019-05-08)?

2. (5 points) What is the probability that a point x chosen uniformly at random from $\text{cube}(d, a)$ lies in the ϵ-*surface* $\text{cube}(d, a) \setminus \text{cube}(d, a - 2\epsilon)$ (cf. Slide 20 of Lecture 2019-05-08)?

3. (2.5 points) Let $\epsilon = 0.01$ and $r = 1$. For which d is more than 90% of the volume of $\text{ball}(d, r)$ in its ϵ-surface?

4. (2.5 points) Let $\epsilon = 0.005$ and $a = 2$. For which d is the probability that a random point from $\text{ball}(d, a)$ lies in its ϵ-surface at least 90%?

4 ϵ-NNS to (ϵ, r)-PLEB [15 points]

Let $P \subset \mathbb{R}^2$ be

$$P = \{(0, 0), (0, 2), (0, 4), (2, 0), (4, 0)\}$$

and let $q = (4, 3)$. Using the algorithm on slide 33 of Lecture 2019–05–08, compute an ϵ-NNS of q in P for $\epsilon = 0.2$. What is the smallest r for which the algorithm outputs a $p \in P$, which p is it?

Hint: Feel free to implement a basic variant of the algorithm on slide 34 that uses an exact oracle for (ϵ, r)-PLEB.