Poison-1-Mean Experiments

May 15, 2022

Abstract

We conduct experiments from the study that was performed on poisoning a geometric data sets. In the research paper Algorithms for Poisoning Geometric Data Sets, we have an algorithm for poisoning 1 mean clustering. This paper will run analysis on iris data, diabetes data and random generated data. All of theses experiments are conducted in $\mathbb{R}^2$ Euclidean space

1 Poisoning 1-Mean Clustering

In the k-means clustering problem, the input consists of a set of points, $X = \{1, \ldots, x_n\} \subset [0,1]^d$, and the goal is to find a set of means, $M = \{\mu_1, \ldots, \mu_k\} \subset [0,1]^d$, such that $\text{cost}_k(X, M) = \sum_{x \in X} \min_{\mu \in M} \|x - \mu\|_2^2$ is minimised. We use $\text{cost}_k(X)$ to denote the optimal cost of k-means on X; that is $\text{cost}_k(X) = \inf_M \text{cost}_K(X, M)$. We refer to a set $\{\mu_1, \ldots, \mu_k\}$ that minimizes $\text{cost}(X)$ as *optimal means*. We can define a function $\mu : X \to M$ as $\mu(x) = \arg\min_{\mu \in M} \|\mu - x\|_2$ where the ties are broken arbitrarily. The function μ then defines a partition $\{X_1, \ldots, X_k\}$ of X by setting $X_i := \mu^{-1}(\mu_i)$.

The m-poisoning of k-means seeks a poison multiset $P = \{p_1, \ldots, p_m\} \subset [0,1]^d$ such that $\text{cost}_k(X \cup P)$ is maximized.

Algorithm Poison-1-Mean for 1-Mean Poisoning: The input consists of $n, m \in \mathbb{N}$, and a set of points $X = \{x_1, \ldots, x_n\} \subset [0,1]^d$, with $|X| = n$. The output is a poison $P = \{p_1, \ldots, p_m\} \subset [0,1]^d$.

Step 1. Let

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

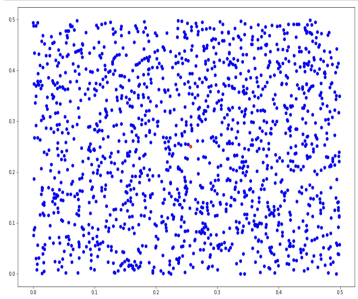
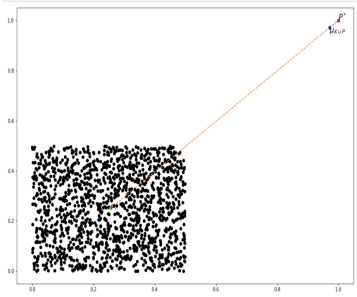
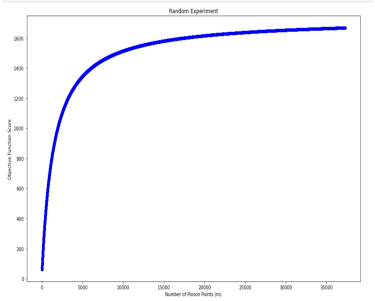
Step 2. Find $p^* = \arg\max_{p \in [0,1]^d} \|p - \mu\|_2$

Step 3. Let P_{Alg} be the multiset containing m copies of p^*. Return P_{Alg}.

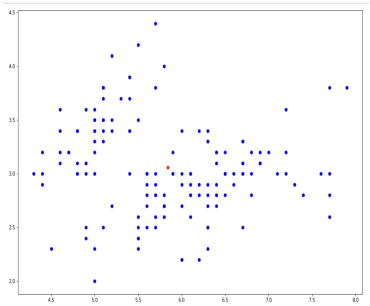
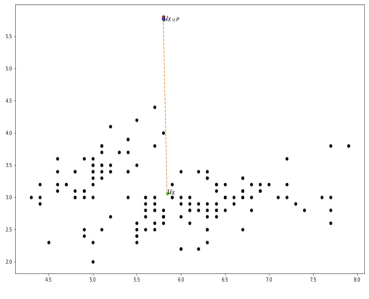
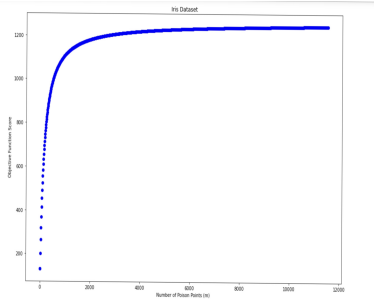
This concludes the description of the algorithm.

2 Random Data Experiment

The random data experiment consisted of a bounding box with $p_1 = (0, 0)$ and $p_2 = (1, 1)$

3 Iris Data Experiment

4 Diabetes Data Experiment