$\mathcal{U}$-Bootstrap percolation : critical probability, exponential decay and applications, by Ivailo HARTARSKY

Leo Davy Martin Gjorgjevski Aleksandr Pak

ENS Lyon
M2 Advanced Mathematics

March 2022

Principal Component Analysis (PCA)

PCA

The goal is to find the linear subspace P_k that best fits the d-dimensional data $\{x_i\}_{i=1}^N$ in the LS sense, i.e., find an orthogonal family of k vectors $\{u_l\}_{l=1}^k$ that maximizes

$$\sum_{l=1}^{k}\sum_{i=1}^{N} |u_l^T x_i|^2.$$

A solution is the k-principal eigenvectors of the empirical autocorrelation matrix

$$\hat{R} = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T =: \frac{1}{N}\sum_{i=1}^{N} \Phi(x_i).$$

$\hat{R}$ is a *sketch* of our data (of dim d^2).

The sketch

$$\hat{R} = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T =: \frac{1}{N}\sum_{i=1}^{N} \Phi(x_i) \in \mathbb{R}^{d^2}$$

is a very compressed version of the data $\{x_i\}_{i=1}^N$, *but*, it still contains the geometry of the data.

CS inspired idea

Take m random measurements[a] of each sample and use the sketch defined by $\Phi(x) = \mathcal{M}(xx^T)$. Provided $m > kd$, the principal eigenvectors can be recovered.

[a] $\mathcal{M} : \mathbb{R}^{d\times d} \to \mathbb{R}^m$ satisfying RIP on matrices of rank at most $2k$.

k-means centroids

The problem

The goal is to recover k centroids $\{c_l\}_{l=1}^{k}$ from some data $\{x_i\}_{i=1}^{N}$ that minimize

$$\sum_{i=1}^{N} \min_{l} ||x_i - c_l||^2.$$

For $N >> 1$ traditional algorithms are not very efficient because they take the whole dataset at once...
But, $N >> 1$ allows to use the laws of large numbers and concentration. It is reasonable to consider that the data will accumulate on small portions of the space.

The binning map

Assume the centroids are spaced by at least ε and have a norm smaller than r, then cover $[-r, r]^d$ by, $N = (\frac{2r}{\varepsilon})^d$, d-dimensional cubes (*bins*). For each bin, count the average number of points that belong to it. This defines the binning map $\hat{p} \in \mathbb{R}_+^N$.

This gives us a sketch of the data, but in a large dimensional space.
But, if the model that generated the data is "structured", i.e., the data concentrates in a few centroids, then the problem is a *sparse* problem.

CS inspired idea

Use a Gaussian random matrix in $\mathbb{R}^{m \times N}$ and define the sketch as $\tilde{z} = A\hat{p}$ and solve

$$\tilde{p} = argmin_{p \in \Sigma_k^+} ||\tilde{z} - Ap||^2$$

Sketching estimates the underlying data-distribution

We can do k-means clustering with m measures using:

- $m \geq k \log N$ for Gaussian sampling
- $m \geq k \log(k)^3 \log N$ for DFT of $\hat{p}$

What if we consider the *continuous* Fourier Tranform (FT) ?
We only know the empirical distribution $\bar{p}_{\mathcal{X}} = \frac{1}{N} \sum_{i=1}^{N} \delta(x_i - x)$, so

$$FT(\bar{p})(\omega) = \int_{\mathbb{R}^d} \bar{p}_{\mathcal{X}}(x) e^{-i2\pi\langle w,x\rangle} dx = \frac{1}{N} \sum_{i=1}^{N} e^{-i2\pi\langle w,x\rangle} =: \bar{\Psi}_{\bar{p}_{\mathcal{X}}}(\omega)$$

$$\longrightarrow^{\mathbb{E}} \bar{\Psi}_{p^*}(\omega) \quad \text{The "true" characteristic function at } \omega.$$

CS inspired idea

If the true distribution p^* is "simple", then interpolating the characteristic function, using simple models, on "few" of its samples should give the true distribution.