MADALGO
CENTER FOR MASSIVE DATA ALGORITHMICS

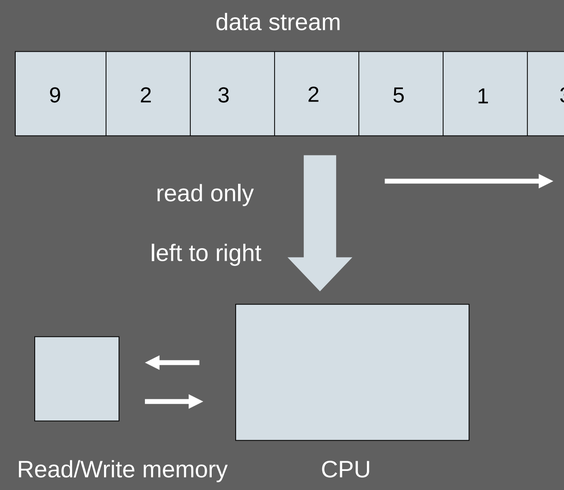

Kevin L. Chang
Max-Planck-Institute for Informatics

Streaming Algorithms for Clustering and Learning

Motivation

Data mining and machine learning techniques have become very important for automatic data analysis. Our goal is to design algorithms for performing these computations that are suitable for **Massive Data Sets**. The streaming model of computation models many of the important constraints imposed by computation on large data sets.

Streaming algorithms



Details of the model

- The large input is a **read-only array**. The array may only be accessed through a few sequential passes.
- Main memory is modeled as extra space that supports random access, reads and writes.

Important resources to minimize

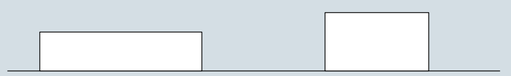
- Ideally, the number of passes should be small.
- Ideally, the amount of memory required should be small.

What types of problems we study, Part 1: Statistical problems

Suppose the data in your data stream consists of random numbers drawn from some probability distribution, but we don't know what the distribution is. Can we "learn" the distribution from the data?

Suppose your data look like this

We want to recover the distribution that generated the data



We focus on the special case where the data are known to be drawn from a mixture of distributions.

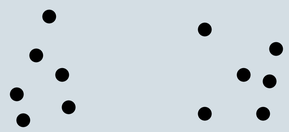
Mixtures of Distributions

- Here, there are k different probability distributions, but don't know what exactly they are..
- Each data point is drawn from one of these distributions, but you don't know which one.
- Can you reconstruct the k different distributions from the data?
- These are very popular models.

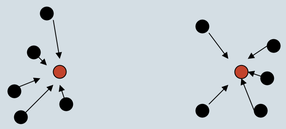
What types of problems we study, Part 2: Clustering

Suppose you have many data points, and want to find a small number of points that "represent" the large set of data points.

Input points



2 representative points and their "clusters"



Two natural applications

Document Clustering: points represent web pages of news web sites.

- Clustering is used to group news pages with similar topics.
- i.e. a cluster for all pages with European Central Bank news, a cluster for the US election, etc.

Intrusion detection: points represent internet connections to computers in your network.

- Clustering is used to group connections into different categories.
- i.e. clusters could represent "denial of service attack", "password guessing", "normal connection"

Highlights of the theoretical results

Trade-off between memory and passes

We focus on algorithms that solve statistical and clustering problems with a **strong trade-off** between the number of passes over the data stream required and the amount of memory required.

Example: if a 2 pass algorithm would require say 100 MB of memory, then a 4 pass algorithm may only require ~10 MB.

Optimality of trade-off

For many of our algorithms, we can mathematically prove that our trade-off is optimal.

So, it's **impossible** to achieve a significantly better trade-off!

Future directions

For the statistical problem, can we design streaming algorithms for learning distributions that are more general?

One approach to accomplish this may be **mathematical programming** over data streams.

A linear program is an example of a mathematical program:

Find numbers x, y that

Maximize $50 x + 7 y$

Subject to $100x + 20 y \leq 10$

$x, y \geq 0$

The distribution learning problem in more general contexts can be cast as a mathematical program.