Data mining: Syllabus

Edo liberty

Data Mining is concerned with efficiently extracting statistics, patterns, structures, or meanings from raw data. This task becomes hard when the amount of data is large, which is often the case in modern data sets. This course will survey modern algorithms, concepts, and data structures common to data mining in large data sets. We will try to cover, among other topics: data sampling, finding frequent item sets, counting in data streams, ranking and sorting, approximating large matrix operations, dimensionality reduction and efficient searching in high dimensions. We will also discuss modern cluster architectures and computational models.

I recommend that students be familiar with probability theory, combinatorics, linear algebra, basic complexity theory, and traditional data structures, at least on at introductory level. The class will attempt to be self contained nonetheless.

Class topics

1. Introduction
   (a) Overview of data mining
   (b) Very brief recap of probability, expectation, linearity of expectation, useful inequalities
   (c) Using the "birthday pardox" to probe huge data sets
2. Introduction part 2 and Sampling
   (a) Recap of hash functions and tries
   (b) Bloom filters
   (c) Chernoff's inequality
   (d) I.i.d sampling
   (e) Fisher-Yates and online reservoir sampling
3. Frequent item sets, and frequent items
   (a) The "Apriory" algorithm
   (b) Dynamic Item set Counting (DIC algorithm)

(c) Simple algorithm for common items in streams

(d) Finding frequent items in data streams (Charikar-Chen-Colton)

4. Frequency Moments and counting distinct elements in streams

   (a) Size approximation and transitive closures (Cohen)

   (b) Approximate counting (Flajolat)

   (c) Frequency moments (Alon-Matias-Szegedy)

5. Ranking and sorting

   (a) Graph Markov models and the graph Laplacian

   (b) PageRank

   (c) HITS

6. 'Bag of words' models

   (a) Principal Component analyses

   (b) Latent Semantic Indexing

   (c) Sums of independent random matrices (Ahlswede-Winter)

   (d) Matrix rank-k approximation using matrix sampling (Rudelson-Vershynin)

7. Distance preserving dimensionality reduction.

   (a) Random Projections (Johnson-Lindenstrauss)

   (b) Fast random projections (Ailon-Chazele-Liberty)

8. Final Project, choosing an interesting method and data

9. Indexing and searching

   (a) The structure of a search index

   (b) Searching with ranks

10. Nearest neighbor search

    (a) Space Partition approaches (KD trees)

    (b) Random hyperplane space partition

    (c) Searching the hypercube

    (d) Locality sensitive hashing

11. The partial match problem

    (a) Hardness result

    (b) Partial match algorithm (Charikar-Indyk-Panigrahy)

12. Final Project class presentations