Lecture 6: Match Data with Algorithms

Xiaotie Deng

AIMS Lab
Department of Computer Science
Shanghai Jiaotong University

October 16, 2017

Algorithm and Data

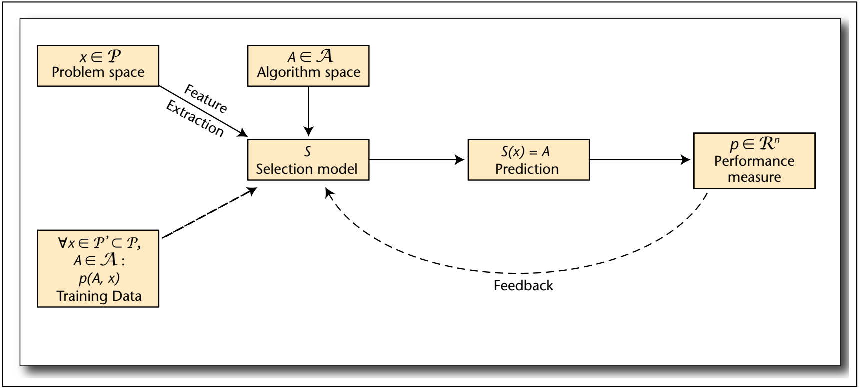
Asymptotic Worst Case Complexity of Algorithm on Data

- Given a problem $\mathcal{P}$.
  - Each $p \in \mathcal{P}$ is characterized by its input data.
- Design an algorithm for all problem instance in $\mathcal{P}$.
  - Return a correct output of the problem for each input instance.
- Time(/space/communication) complexity of algorithm A.
  - $t(A, n) = \max\{t(A, x) : \text{time to return output } \forall x : |x| = n\}$
- Complexity of the problem.
  - The best algorithm $\min\{t(A, n)|A \in \mathcal{A}\}$.

Algorithm on Restricted Data

- Data input with fixed parameter.
  - MOOC assignment: https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-046j-design-and-analysis-of-algorithms-spring-2015/lecture-videos/lecture-18-complexity-fixed-parameter-algorithms/
- Data input following a given distribution.
  - Impagliazzo, Russell. "A personal view of average-case complexity." Structure in Complexity Theory Conference, 1995., Proceedings of Tenth Annual IEEE. IEEE, 1995.
- Fixed data with random noise.

Algorithm Selection



Five Worlds of Impagliazzo

- Algorithmica: $P = NP$, verification equivalent to solving a problem.
- Heuristica: $NP \neq P$ but some tractable on the average for some distribution.
- Pessiland: Hard average problems exist but no one way function.
- Minicrypt: One way function exists but PKC is impossible.
- Cryptomania: The existence of PKC (Public Key Cryptography).

Five Worlds in Big Data

- Propose Some in your assignment

Progressive Data Sequence

An Ensemble of Distribution and Class *AvgP*

- An ensemble of distribution is a sequence of distribution μ_n, $n \in Z$, on the set of positive integers with bit size n.
- A function $T : Z^+ \rightarrow Z^+$ is a polynomial on average with respect to μ_n, $n \in Z^+$, if there is an $\epsilon > 0$ such that the expectation of $T(i)^\epsilon$ when i is chosen according to μ_n is $O(n)$.
- A problem f on μ_n is in *AvgP* if there is an algorithm to compute f whose running time is polynomial on average with respect to μ_n.

Polynomial Time Benign Algorithm Scheme

- Algorithm $A(x, \delta)$ computes f with benign fault: Output f or '?' and output is the correct function value if not '?';
- runs in polynomial in $|x|$ and $1/\delta$;
- $\forall \delta : 1 > \delta > 0$, and $\forall n \in Z^+$: $Prob_{x \in_{\mu_n} Z^+}(A(x, \delta) =' ?') \leq \delta$.

Heuristic Polynomial Time Algorithm $\mathcal{HP}$

- For x randomly chosen according to $\{\mu_n : n \in Z^+\}$, and $\forall\delta > 0$, there is a deterministic polynomial time algorithm $A(x, \delta)$ that computes $f(x)$ correctly except an error of upto δ.
- $\mathcal{HPP}$: probability version.
- $\mathcal{HPP}/poly$: non-uniform algorithm version.

Data of Fixed Parameters

Vertex Cover Set of Fixed Constant Size k

- Idea: There are $\binom{n}{k}$ possible such subsets
- Algorithm: Go over all ($\binom{n}{k}$ loops) and check (times m).
- Total time: $O(n^{k+1})$, a polynomial where k is a constant.

FPT: Reduced time to $O(f(k)) * n^{O(1)}$

- Idea: There is a node selected in an edge.
- Algorithm:
  - Go over all edges one by one,
  - binary step: choose u or v if none of nodes u and v is already chosen.
- Total time: $O(2^k + n + m)$
  - no more than depth k, total binary steps bounded by 2^k.

Kernelization in Fixed Parameter Complexity

A preprocessing stage to reduce the input to a smaller input, called a "kernel", that is easier to solve.

- Idea: Remove all vertices of degree $k+1$.
  - The remaining graph has maximum degree k.
  - The size of kernel, the resulted graph, is no more than k^2+k since its vertex cover set has no more than k vertices and each connects to no more than k other vertices
- Algorithm:
  - Remove each such vertex one by one
  - Work all choices of the remaining graph
- Total time: $O(f(k)+n+m)$

Data of Fixed Distribution

Order Statistics

- IID random varialbes: $X_1, X_2, \cdots, X_n$.
- Order Statistics: $X_{i-1,n} < X_{i,n}$, $i = 1, 2, \cdots, n$, with $X_{0,n} = 0$.
- For exponential distribution: $Pr[X_{i,n} > t] =???$

The Case for Exponential Distributions

- CDF (cumulative distribution function)
  - $Pr[X_{1,n} > t] = Pr[X_i > t, i = 1, \cdots, n] = e^{-nt}$
  - $Pr[X_{n,n} \leq t] = Pr[X_i < t, i = 1, \cdots, n] = (1 - e^{-t})^n$

Joint Distributions of Order Statistics

- Lemma: $f_{X_{1,n},X_{2,n},\cdots,X_{n,n}}(t_1 < t_2 < \cdots < t_n) = n! * \Pi_{i=1}^{n} f_X(t_i)$
- Proof: by symmetry, LHS$= n! * f_{X_1,X_2,\cdots,X_n}(t_1, t_2, \cdots, t_n)$
  - where $t_1 < t_2 < \cdots < t_n$.
- Note: Condition $t_1 < t_2 < \cdots < t_n$ is important and useful.
- Corollary:

Generating Order Statistics of Exponential Distributions

- Define $Y_1 = X_{1,n}$, $Y_2 = X_{2,n} - X_{1,n}$, $\cdots$, $Y_n = X_{n,n} - X_{n-1,n}$.
- Then Y is linear in X: $Y = AX$ where A is lower triangle with diagonal terms all 1.
- $f_{X_{1,n}, X_{2,n}, \cdots, X_{n,n}}$

Assignment II (last part)

1. For a graph $G = (V, E)$, design a polynomial time algorithm to find a clique, i.e., a subset of vertices which has an edge between each other, of size constant k.