Network Influence Seed Optimization

Lecture Note of Big Data Algorithm and Design, Prof: Xiaotie Deng, 2017 Fall

By Hao Zhou

1 Submodular Function

1.1 Definition

If V is a finite set, a **Submodular Function** is a set function $f : 2^V \to \mathbb{R}$, where 2^V denotes the power set of V , which satisfies one of the following equivalent conditions:

- For every $X, Y \subseteq V$ with $X \subseteq Y$ and every $x \in V \setminus Y$ we have that $f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y)$.
- For every $S, T \subseteq V$ we have that $f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$.
- For every $X \subseteq V$ and $x_1, x_2 \in V \backslash X$ we have that $f(X \cup \{x_1\}) + f(X \cup \{x_2\}) \geq f(X \cup \{x_1, x_2\}) + f(X)$.

1.2 Monotone Submodular Function

f is a **Monotone Submodular Function** if f satisfies both the following conditions:

- f is a submodular function.
- $f(S \cup \{v\}) \geq f(S)$.

1.3 Submodular Function Optimization

Let $f : 2^V \to \mathbb{R}$ is a nonegative monotone submodular function, the **Submodular Function Optimization** is

$$\begin{aligned} \max \quad & f(S) \\ \text{s.t.} \quad & S \subseteq V \\ & |S| \leq k \end{aligned}$$

1.3.1 Hardness

Theorem 1.1. *Submodular Function Optimization is NP-hard.*

1.3.2 Greedy Algorithm

1. Define $\Delta_x f(A) = f(A \cup \{x\}) - f(A)$. A greedy algorithm is as follows:

Algorithm 1: Greedy Algorithm

```
S_0 ← ∅;
for i ← 0 to k do
    choose v ∈ V to maximize Δ_v f(S_{i−1});
    S_i ← S_{i−1} ∪ {v};
Output S_k;
```

2. Approximation Ratio: $(1 - e^{-1})$

Proof. Suppose $x_1, x_2, \ldots, x_k$ are obtained by Alg. 1. Denote $S_i = \{x_1, \ldots, x_i\}$ and let $S_* = \{u_1, u_2, \ldots, u_k\}$ be the optimal solution. We can get

$$\begin{aligned} f(S^*) &\leq f(S_i \cup S^*) \\ &= f(S_i) + \Delta_{u_1} f(S_i) + \Delta_{u_2} f(S_i \cup \{u_1\}) + \ldots + \Delta_{u_k} f(S_i \cup \{u_1, u_2, \ldots, u_{k-1}\}) \\ &\leq f(S_i) + \Delta_{u_1} f(S_i) + \Delta_{u_2} f(S_i) + \ldots + \Delta_{u_k} f(S_i) \qquad (Submodular!) \\ &\leq f(S_i) + k(f(S_{i+1}) - f(S_i)) \qquad (Greedy!) \end{aligned}$$

Denote $a_i = f(S^*) - f(S_i)$. Then

$$a_i \leq k(a_i - a_{i+1}) \Rightarrow a_{i+1} \leq (1 - \frac{1}{k})a_i \leq e^{-1/k} a_i$$

Hence, $a_k \leq e^{-1} a_0 = e^{-1} f(S^*)$, i.e.,

$$f(S_k) \geq (1 - e^{-1}) f(S^*)$$

□

2 Influence Maximization

2.1 Two Models

2.1.1 Independent Cascade Model

Given a directed graph $G = (V, E; p, A_0)$, where

- $A_0 \subseteq V$ is an initial activated seed set.
- $p : E \to Q+$ where $p(i, j)$ is the independent state transition probability $P(j = 1 | i = 1)$ of influence of i on j.
- If i is activated, j will be influence to become active with probability $p(i, j)$.

2.1.2 Linear Threshold Model

Given a directed graph $G = (V, E; \theta, b, A_0)$, where

- $A_0 \subseteq V$ is an initial activated seed set.
- $\theta : V \to Q+$ where $\theta_v \sim U[0, 1]$ is generated randomly for every node $v \in V$.
- A given set of weight $b_{v,w}$ such that $\sum_{w \in \Gamma(v)} b_{w,v} \leq 1$ is given where $\Gamma(v)$ is the neighbors of v.
- A node v becomes active if

$$\sum_{w:\ active\ neighbor\ of\ v} b_{w,v} \geq \theta_v$$

.

2.2 Influence Maximization Problem

2.2.1 Definition

Given a directed graph $G = (V, E)$, a diffusion model m and an integer $k \leq |V|$, find a set S which satisfies $S \subseteq V$ and $|S| = k$, such that the expected influence spread $\sigma_m(S)$ is maximum.

2.2.2 Hardness

Theorem 2.1. *Influence Maximization Problem with Independent Cascade Model is NP-hard.*

Proof. Consider the reduction from Set Cover Problem.
Set Cover Problem: Given a collection of subsets $S_1, \ldots, S_m$ of a ground set $U = \{u_1, u_2, \ldots, u_n\}$ and integer $k \geq 0$, find k ones of the subsets whose union is equal to U.
For each set cover problem instance, construct the bipartite graph as follows:

1. There are m left nodes, named $S_1, S_2, \ldots, S_m$;
2. There are n right nodes, named $u_1, u_2, \ldots, u_n$;
3. There is an edge (S_i, u_j) if and only if the subset S_i contains the element u_j.

Let $p(i, j) = 1, \forall (i, j) \in E$ and A_0 be the m left nodes. Then, Set Cover Problem has a solution $\Leftrightarrow$ $\exists k$ seeds influence $n + k$ nodes in this bipartite graph. □

2.2.3 Submodularity for Independent Cascade Model

Consider a sample space consisting of all subgraph of input digraph G. For each sample X, let $\sigma^X(A)$ denote the number of active nodes for seed set A. Then

$$\sigma(A) = \sum_X Pr[x] * \sigma^X(A)$$

Notice $\sigma^X(A)$ is monotone submodular since $\sigma^X(A) = |\{u \mid u \text{ can be reached form } A \text{ via a directed path in } X\}|$. Hence, $\sigma(A)$ is also monotone submodular.