MAP554 – Networks: PC 4

28 September 2016

Viral marketing: submodular functions and greedy algorithms

Consider the following SIR epidemics model. A graph G on n nodes is given together with an array of infection probabilities p_{ij} for each $i, j \in [n]$. p_{ij} is the probability that i succeeds to infect j if it has been infected. Given a budget $k < n$ of nodes that one can infect initially, the objective of this problem is to determine a subset $S \subset [n]$ of size k such that the expected number of nodes infected in the SIR epidemics is maximized.

1.1 (Algorithmic hardness) The so-called set cover algorithmic problem is defined as follows. Given a collection $C_1, \ldots, C_m$ of subsets of $[N]$, and a budget $k < m$, determine whether there is a sub-collection $C_{i(1)}, \ldots, C_{i(k)}$ of subsets whose union is exactly $[N]$. This problem is known to be NP-hard.

Show that even when the probabilities p_{ij} are restricted to equal either 0 or 1, it is NP-hard to find a subset of $[n]$ of size k such that the corresponding SIR epidemics has the largest possible size.

1.2 A function F defined on subsets of a base set $[n]$ and taking real values is called submodular if for all $A, B \subset [n]$ one has

$$F(A \cup B) + F(A \cap B) \leq F(A) + F(B).$$

Show that the function F defined by letting $F(A)$ be the expected number of eventually infected nodes when $A \subset [n]$ is the set of initially infected nodes is submodular.

Hint: consider first the case where the p_{ij} equal 0 or 1, and show that in that case $M(A \cup B) = M(A) \cup M(B)$ for all $A, B \subset [n]$, where $M(A)$ denotes the set of eventually infected nodes with initial set A.

1.3 Consider a submodular function $F : 2^{[n]} \to \mathbb{R}_+$ taking non-negative values, that is also non-decreasing, i.e. $A \subset B \Rightarrow F(A) \leq F(B)$, and submodular.

Consider a **greedy** selection $v_1, \ldots, v_k$ defined by

$$v_i \in \text{argmax}_{v \in [n] \setminus \{v_1, \ldots, v_{i-1}\}} F(\{v_1, \ldots, v_{i-1}, v\}).$$

Let then $C_i = \{v_1, \ldots, v_i\}$, $i \in [k]$ and $C_0 = \emptyset$. Show that for any set $C = \{w_1, \ldots, w_k\} \subset [n]$ of size k, and all $i \in \{1, \ldots, k-1\}$, one has:

$$F(C_{i+1}) - F(C_i) \geq \frac{1}{k} \left[F(C) - F(C_i) \right]. \tag{1}$$

Hint: Introduce for each $j = 0, \ldots, k$ the set $D_j := C_i \cup \{w_1, \ldots, w_j\}$ and reason about the sum $\sum_{j=1}^{k} F(D_j) - F(D_{j-1})$.

1.4 Deduce from (1) that for any subset C of size k, one has

$$F(C) - F(C_k) \leq \left(1 - \frac{1}{k}\right)^k [F(C) - F(\emptyset)]$$

1.5 Deduce that a greedy selection algorithm yields a solution C_k such that the corresponding performance $F(C_k)$ is at least $(1-1/e) \approx 0.632$ times that of the optimal solution $\max_{|C|=k} F(C)$. This result applies to the present viral marketing scenario but also has a large variety of other applications.