Learning Set Functions with Limited Complementarity

Hanrui Zhang

Duke University

Set Functions

- f: 2[n] → [0, ∞) maps subsets of {1, …, n} to real numbers
- **marginal:** $f(S \mid T) = f(S \cup T) - f(T)$
- **normalization:** $f(\varnothing) = 0$
- **monotonicity:** $f(S \mid T) \geq 0$
- **1-Lipschitz continuity:** $f(S \mid T) \leq |S \setminus T|$

"Probably Mostly Approximately Correct"

Given **samples of f*(S)**, the learner **outputs f** such that:

- with probability $1 - \delta$ the learner succeeds, in which case,
- for a random set S, w.p. $1 - \varepsilon$, f(S) approximates f*(S), i.e.,
- $f(S) \leq f^*(S) \leq \alpha \cdot f(S)$

PMAC-learning generalizes PAC-learning!

(Limited) Complementarity

complement each other!

Supermodular Sets ([2]): T is supermodular w.r.t. f, iff ∃ v, S, s.t., $\forall T' \subseteq T$, $f(v \mid S \cup T') < f(v \mid S \cup T)$

"T is supermodular if there is environment S and item v, such that if everyone has S, T provides more marginal for v than any of T's proper subsets."

Supermodular Width ([2]): SMW(f) := max{ |T| | T is a supermodular set}

"The supermodular width of f is the cardinality of the largest supermodular set w.r.t. f."

[2]: f is submodular iff SMW(f) = 0.

"SMW generalizes submodularity."

Our Results

[1]: Restricted to product distributions, there is an algorithm that PMAC-learns submodular functions with approximation factor $\alpha = O(\log(1/\varepsilon))$ and $\ell = O(n^2 \log(1/\delta))$ samples.

— What if there is complementarity?

— We show:

Restricted to product distributions, there is an algorithm that PMAC-learns SMW-d functions with approximation factor $\alpha = O((d+1)^2 \log(1/\varepsilon))$ and $\ell = O(n^2 \log(1/\delta))$ samples.

— What happens with non-product distributions?

— [1]: PMAC-learning is hard even with submodular functions.

Concentration of Set Functions

— Is learnability surprising?

— No! Actually:

$\Pr[|f(X) - \mathrm{Med}(f(X))| \geq t \cdot (\mathrm{Med}(f(X))^{1/2}] \leq 4 \exp(-t^2 / 4(d+1)^2)$

"Under a product distributions, the value of f on a random set X is tightly concentrated around its median."

- Pf: Talagrand's or concentration of self-bounding functions
- Implies concentration around the mean
- Directly implies an algorithm: approximate f* by the empirical mean
- Need to handle 0's of f* separately

References

1. Maria-Florina Balcan and Nicholas Harvey. Learning submodular functions.
2. Wei Chen, Shang-Hua Teng and Hanrui Zhang. Capturing Complementarity in Set Functions by Going Beyond Submodularity/Subadditivity.