Adaptive Submodularity for Active Classification

Sergey Karayev

December 2013
CS 270 Project

Table of Contents

Problem

- Have:
  - Hypothesis $H \in \mathcal{H}$, set of possible observations $\mathcal{X} = \{X_1, \ldots, X_F\}$.
  - Each X_i has cost c_i and depends on H in some way.
- Goal: select subset $\mathcal{X}_S$ to maximize reward $f(\mathcal{X}_S)$.
  - e.g. loss: $f(\mathcal{X}_S) = -\ell(h^*, \operatorname{argmax}_h P(h \mid \mathcal{X}_S))$.
  - e.g. entropy: $f(\mathcal{X}_S) = \sum_x P(h \mid \mathcal{X}_S) \log P(h \mid \mathcal{X}_S)$
- Variants:
  - *Budgeted maximization*: Maximize objective without exceeding cost budget.
  - *Minimum-cost cover*: Minimize cost while achieving some minimum objective value.

Active classification

- Furthermore, want to do this *adaptively*: selecting and observing values in sequence (motivation to follow).
- Lots of recent work in this setting: POMDPs, cascades, trees for spam detection, object recognition, sentence parsing.
- Different heuristic methods, but no performance guarantees.
- Here, we synthesize a few recent papers that may construe a promising avenue for theoretical analysis of the active classification problem.

Offline problem formulation

Take the problem of selecting the whole subset before observing anything; assume uniform costs.

Budgeted maximization

$$\underset{\mathcal{X}_S \subseteq \mathcal{X}}{\text{argmax}}\ f(\mathcal{X}_S)$$
$$s.t\ |\mathcal{X}_S| \leq N$$

Min-cost cover

$$\underset{\mathcal{X}_S \subseteq \mathcal{X}}{\text{argmin}}\ |\mathcal{X}_S|$$
$$s.t\ f(\mathcal{X}_S) \geq Q$$

- For most interesting f, this is NP-hard.
- Nemhauser et al. (1978) showed that if f is *monotone submodular*, then a simple greedy algorithm is near-optimal with these bounds:

$$f(\mathcal{X}_{S_N}) \geq (1 - \frac{1}{e})\, f(\mathcal{X}^*_{S_N}) \qquad |\mathcal{X}_{S_Q}| \leq |\mathcal{X}^*_S|\,(1 + \log \max_X f(X))$$

Monotone submodular functions

- Intuitively, submodularity is the property of diminishing returns. For proving near-optimality of the greedy method, we additionally require monotonicity.
- Define the *gain* of adding another item to a set as

$$\Delta_f(X \mid \mathcal{X}_A) = f(\mathcal{X}_A \cup \{X\}) - f(\mathcal{X}_A)$$

- For all X, the submodularity requirement must hold. "Adding an element to a smaller set helps more than adding the same element to a larger set."

$$\mathcal{X}_A \subseteq \mathcal{X}_B \implies \Delta_f(X \mid \mathcal{X}_A) \geq \Delta_f(X \mid \mathcal{X}_B)$$

- For all X, the monotonicity requirement must hold. "Adding an element can never hurt."

$$\Delta_f(X \mid \mathcal{X}_A) \geq 0$$

Greedy algorithm

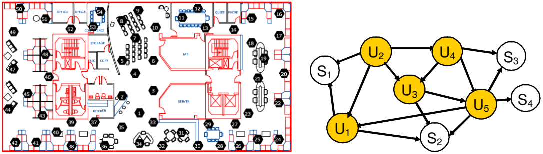
Input: Budget N; observations $\mathcal{X}$; objective f.
Output: Set $\mathcal{X}_S \subseteq \mathcal{X}$ of size N
begin
  $\mathcal{X}_S \leftarrow \emptyset$;
  for $i = 1$ **to** k **do**
    foreach $X \in \mathcal{X} \setminus \mathcal{X}_S$ **do** compute $\Delta_f(X \mid \mathcal{X}_S)$;
    Select $X^* \in \underset{X}{\operatorname{argmax}} \frac{\Delta_f(X|\mathcal{X}_S)}{c(X)}$;
    Set $\mathcal{X}_S \leftarrow \mathcal{X}_S \cup \{X^*\}$;
  end
end

- Costs have been omitted in the exposition, but are handled by the algorithm if the cost function is modular:

$$c(\mathcal{X}_S) = \sum_{X \in \mathcal{X}_S} c(X)$$

.

Example

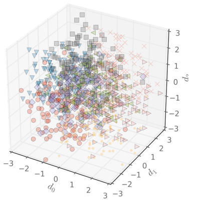


(a) *Sensor network deployment* (b) *Graphical model*

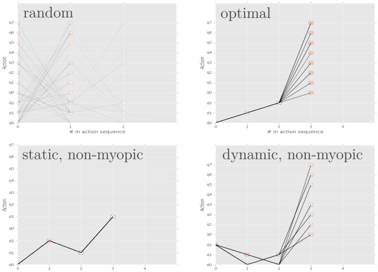
- Sensor network: want minimal set of observations (Sensors) to maximize utility over hidden variables (Fire): f(P(Fire | Sensors)).
- If the observation model satisfies certain conditional independence requirements, then Information Gain (reduction in entropy) is submodular [Krause and Guestrin (2005)].
- Classification models like Naive Bayes are very similar.

Adaptivity motivation

- With the greedy algorithm and monotone submodular objective, we can find a near-optimal subset.
- But what if we can observe the results of our actions upon execution?
- In this case, offline submodularity also can be exponentially far from optimal [Hollinger et al. (2011)].



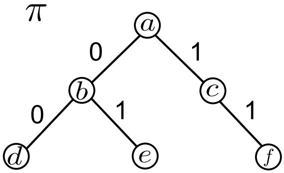
Feature	Number	Cost
d_i: sign of dimension i	D	1
q_o: label of datapoint, if in quadrant o	2^D	10



Adaptive submodularity

- Introduced by Golovin and Krause (2011)[1].
- All F potential observations X_i are now random variables taking on values in $\mathcal{O}$, and have *realization* $\mathbf{x}$.
- Selected subset $\mathcal{X}_S \subseteq \mathcal{X}$ forms a *partial realization* $\mathbf{x}_S$.
- The objective function $f : 2^F \times |\mathcal{O}|^F \rightarrow \mathcal{R}$ takes a realization as well as the selected subset.
- The *policy* π maps partial realizations to new observations, and can be thought of as a decision tree.
- $F(\pi) = \mathbb{E}[f(\pi(\mathbf{x}), \mathbf{x})] = \sum_{\mathbf{x}} P(\mathbf{x}) f(\pi(\mathbf{x}), \mathbf{x})$

$$\begin{aligned}
&\pi(\emptyset) = a \\
&\pi(\{(a,0)\}) = b \\
&\pi(\{(a,1)\}) = c \\
&\pi(\{(a,0),(b,0)\}) = d \\
&\pi(\{(a,0),(b,1)\}) = e \\
&\pi(\{(a,1),(c,1)\}) = f
\end{aligned}$$



[1] JAIR Best Paper Prize 2013

Adaptive submodular functions

- We transfer the guarantees of greedy optimization with submodular functions to the adaptive setting by taking expectation over realizations, conditioned on the current partial realization.
- Define the *conditional expected gain* of adding another item to a set as

$$\Delta_f(X \mid \mathbf{x}_A) = \mathbb{E}\left[f(\mathcal{X}_A \cup \{X\}, \mathbf{x}) - f(\mathcal{X}_A, \mathbf{x}) \mid \mathbf{x}_A\right]$$

- For all X, the function must be adaptive submodular.

$$\mathbf{x}_A \preceq \mathbf{x}_B \implies \Delta_f(X \mid \mathbf{x}_A) \geq \Delta_f(X \mid \mathbf{x}_B)$$

- For all X, the function must be adaptive monotone.

$$\Delta_f(X \mid \mathbf{x}_A) \geq 0$$

Greedy policy

Input: Budget N; observation set $\mathcal{X}$; $p(\mathbf{x})$; objective f.
Output: Set $\mathcal{X}_S \subseteq \mathcal{X}$ of size N
begin
  $\mathcal{X}_S \leftarrow \emptyset$; $\mathbf{x_S} \leftarrow \emptyset$;
  for $i = 1$ **to** k **do**
    foreach $X \in \mathcal{X} \setminus \mathcal{X}_S$ **do** compute $\Delta_f(X \mid \mathbf{x}_S)$;
    Select $X^* \in \underset{X}{\operatorname{argmax}} \frac{\Delta_f(X|\mathbf{x}_S)}{c(X)}$;
    Set $\mathcal{X}_S \leftarrow \mathcal{X}_S \cup \{X^*\}$;
    Observe value $\mathbf{x}^*$ and update $\mathbf{x}_S \leftarrow \mathbf{x}_S \cup \mathbf{x}^*$;
  end
end

Submodularity $\rightarrow$ Adaptive submodularity

Remarkably, optimality bounds stay the same!

Submodularity

- Works with sets.
- $\Delta_f(X \mid \mathcal{X}_A) = f(\mathcal{X}_A \cup \{X\}) - f(\mathcal{X}_A)$
- $\Delta_f(X \mid \mathcal{X}_A) \geq \Delta_f(X \mid \mathcal{X}_B)$ if $\mathcal{X}_A \subseteq \mathcal{X}_B$
- $\Delta_f(X \mid \mathcal{X}_A) \geq 0$
- $\max_{\mathcal{X}_S} f(\mathcal{X}_S)$ s.t. $|\mathcal{X}_S| \leq N$
- Budgeted maximization: greedy is within $(1 - 1/e)$ of OPT.
- Min-cost cover: greedy subset size is less than $1 + \mathit{log} \max_X f(X)$ of OPT subset size.

Adaptive submodularity

- Works with policies.
- $\Delta(X \mid \mathbf{x}_A) = \mathbb{E}\left[f(\mathcal{X}_A \cup \{X\}, \mathbf{x}) - f(\mathcal{X}_A, \mathbf{x}) \mid \mathbf{x}_A\right]$
- $\Delta_f(X \mid \mathbf{x}_A) \geq \Delta_f(X \mid \mathbf{x}_B)$ if $\mathbf{x}_A \preceq \mathbf{x}_B$
- $\Delta_f(X \mid \mathbf{x}_A) \geq 0$
- $\max_{\pi} F(\pi)$ s.t. $|\pi| \leq N$
- Budgeted maximization: greedy policy is within $(1 - 1/e)$ of OPT policy.
- Min-cost cover: greedy policy size is less than $1 + \mathit{log} \max_X f(X, \mathbf{x})$ of OPT policy size.

Benefit: Lazy evaluation

- A neat algorithmic speed-up follows from the submodularity of the objective: **the expected utility of an observation cannot increase**. This means not all observations need to be evaluated on every iteration.

	Expected Utility 1	Expected Utility 2
action 1	0.85	
action 2	0.72	
action 3	0.52	
action 4	0.63	

	Expected Utili
action 1	0.85
action 2	0.72
action 3	0.52
action 4	0.63

Formulating the objective

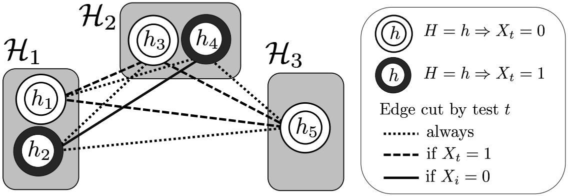
- So far, things look promising!
- Unfortunately, infogain is only adaptive submodular in the noiseless setting (in Naive Bayes, only one outcome $\mathbf{x}_i$ given h).
- If there is noise, need to find an objective that both optimizes what we want and is adaptive submodular.
- Have two previously explored options:
  - *Hypothesis Pruning* [Golovin et al. (2010)]
  - *Equivalence Class Edge Cutting* [Javdani et al. (2012)]
- Both have proofs only for the Min-Cost Cover setting.

Hypothesis pruning

- Motivation: efficiently figure out where an object is in 3D space so that a robotic arm can grasp it.
- Idea: downweight hypotheses with a non-normalized Gaussian, which effectively removes a portion of the hypothesis space. By construction, the probability mass of the posterior distribution can only become smaller with each step.
- Proof: lengthy (in report). Tricky to prove that the *expected* utility does not increase.

Equivalence Class Edge Cutting

- Motivation: request most informative instance to be labeled by an oracle.
- Idea: expand the space of realizations to account for all possible noise settings. A single hypothesis can now be true in multiple states. Find the test that removes the most hypotheses.
- Proof: lengthy (in report).



Future Work

- Empirically evaluate the two adaptive submodular objectives against my current POMDP approach and on my datasets.
    - Adaptive submodularity of a function can depend on the distribution $p(\mathbf{x})$.
- Modular[2] costs are fine. How about submodular costs?
- Analyze the adaptivity gap of these methods via data-dependent bounds.

[2] $c(\mathcal{X}_S) = \sum_{X \in \mathcal{X}_S} c(X)$

References

Golovin, D. and Krause, A. (2011). Adaptive submodularity: A new approach to active learning and stochastic optimization. *Journal of Artificial Intelligence Research*.

Golovin, D., Krause, A., and Ray, D. (2010). Near-optimal bayesian active learning with noisy observations. *CoRR*, abs/1010.3091.

Hollinger, G. A., Mitra, U., and Sukhatme, G. S. (2011). Active classification: Theory and application to underwater inspection. *CoRR*, abs/1106.5829.

Javdani, S., Klingensmith, M., Bagnell, D., Pollard, N. S., and Srinivasa, S. S. (2012). Efficient touch based localization through submodularity. *CoRR*, abs/1208.6067.

Krause, A. and Guestrin, C. (2005). Near-optimal nonmyopic value of information in graphical models. In *UAI*.

Nemhauser, G. L., Wolsey, L. A., and Fisher., M. L. (1978). An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294.