Imitation Learning by Coaching

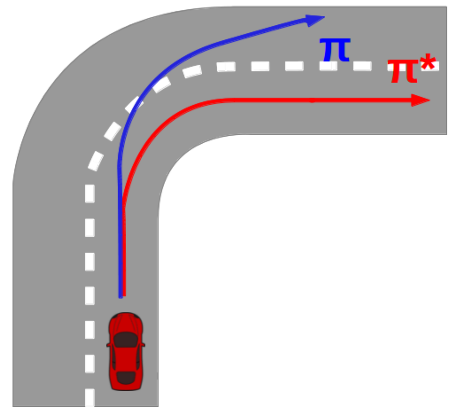
He He[1], Hal Daumé III[1] and Jason Eisner[2]

[1]University of Maryland, College Park [2]Johns Hopkins University

Introduction

- Imitation learning by classification
- DAgger [1]: iterative policy training via a reduction to online learning
- Coaching (new): update towards easy-to-learn intermediate actions when the oracle is too good to imitate
- Experiments on test-time cost-sensitive dynamic feature selection

Imitation Learning by Classification



- Markov Decision Process
  - state $s \in S$, action $a \in A$, policy $\pi : S \to A$
  - d_π: average distribution of states over T steps
  - immediate loss $L(s, a) \in [0, 1]$, expected loss $J(\pi) = T\mathbb{E}_{s\sim d_\pi}[L(s, \pi(s))]$
- Oracle action: $\pi^*(s) = \arg\min_{a\in A} C(s, a)$

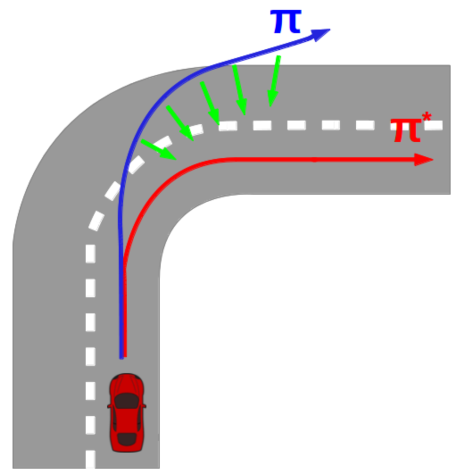
  $C(s, a)$: oracle's measure of the quality of a in s
- Goal: minimize the task loss $J(\pi) \to$ minimize a local regret ℓ
- Policy as a multiclass classifier: $\hat{\pi} = \arg\min_{\pi\in\Pi} \mathbb{E}_{s\sim d_{\pi^*}}[\ell(\pi(s), \pi^*(s))]$

Dataset Aggregation (DAgger)

Problems with the classification approach

- Different distributions of states at training and test time
- Learner may go to states never visited by oracle
- Quadratic loss: Let $\mathbb{E}_{s\sim d_{\pi^*}}[\ell(\pi(s), \pi^*(s))] = \epsilon$, then $J(\pi) \le J(\pi^*) + T^2\epsilon$

Iterative policy training



- Execute the most recently trained policy
- Retrain classifier on all states *ever* encountered
  - Teaches learner how to recover from past mistakes
- Supervised action at states $s = \pi^*(s)$

Theoretical guarantee

- $Q_t^{\pi'}(s, \pi)$: t-step loss of executing π in the initial state and then running π'
- Test-time surrogate loss: $\mathbb{E}_{s\sim d_\pi}[\ell(\pi(s), \pi^*(s))] = \epsilon$
- General case: If $Q_{T-t+1}^{\pi^*}(s, \pi) - Q_{T-t+1}^{\pi^*}(s, \pi^*) \le u$ for all actions a, $t \in \{1, 2, \dots, T\}$, then $J(\pi) \le J(\pi^*) + uT\epsilon$.
- DAgger:

  N iterations, $\pi_1, \pi_2, \dots, \pi_N$ denoted by $\pi_{1:N}$

  Error of best policy in hindsight: $\epsilon_N = \min_{\pi\in\Pi} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{s\sim d_{\pi_i}}[\ell(\pi(s), \pi^*(s))]$

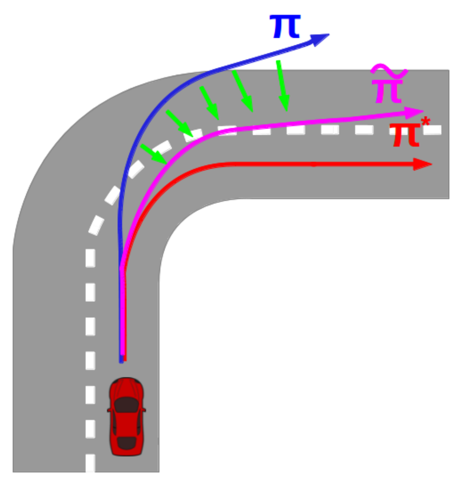
  If N is $O(uT \log T)$ and $Q_{T-t+1}^{\pi^*}(s, \pi) - Q_{T-t+1}^{\pi^*}(s, \pi^*) \le u$, there exists a policy $\pi \in \pi_{1:N}$ s.t. $J(\pi) \le J(\pi^*) + uT\epsilon_N + O(1)$.

Coaching

A too-good-to-learn oracle

- Policy space far from the learning policy space – limited learning ability
- Information not inferable from the state – limited learning resources
- Large training error in each iteration and large ϵ_N

Coach



- Easy-to-learn actions: scored high by the learner's current policy
- Good actions: low task loss
- Hope action: not much worse than the oracle action but easier to achieve

  λ: specifying how close the coach is to the oracle

  $\tilde{\pi}_i(s) = \arg\max_{a\in A} \lambda \cdot \text{score}_{\pi_i}(s, a) - C(s, a)$

DAgger by coaching

Initialize $\mathcal{D} \leftarrow \emptyset$, $\pi_1 \leftarrow \pi^*$
for $i = 1$ **to** N **do**
  Sample T-step trajectories using π_i
  Collect coaching dataset $\mathcal{D}_i = \{(s_{\pi_i}, \tilde{\pi}_i(s))\}$
  Aggregate datasets $\mathcal{D} \leftarrow \mathcal{D} \bigcup \mathcal{D}_i$ and train policy π_{i+1} on $\mathcal{D}$
end for
Return best π_i evaluated on validation set

Theory

Reduction to online learning

- Treat trajectories collected in each iteration as one online-learning example
- Choose the best policy so far: *Follow-The-Leader*
- No-regret online learning algorithm:
  $\frac{1}{N}\sum_{i=1}^{N} \ell_i(\pi_i) - \min_{\pi\in\Pi} \frac{1}{N}\sum_{i=1}^{N} \ell_i(\pi) \le \gamma_N$ and $\lim_{N\to\infty} \gamma_N = 0$
- DAgger theoretical guarantee holds under any no-regret algorithm

Coaching guarantee

- Test-time loss: $\tilde{\ell}_i(\pi) = \mathbb{E}_{s\sim d_{\pi_i}}[\ell(\pi(s), \tilde{\pi}_i(s))]$
- Error of best policy in hindsight w.r.t. hope actions: $\tilde{\epsilon}_N = \frac{1}{N}\min_{\pi\in\Pi} \sum_{i=1}^{N} \tilde{\ell}_i(\pi)$
- Linear policy:

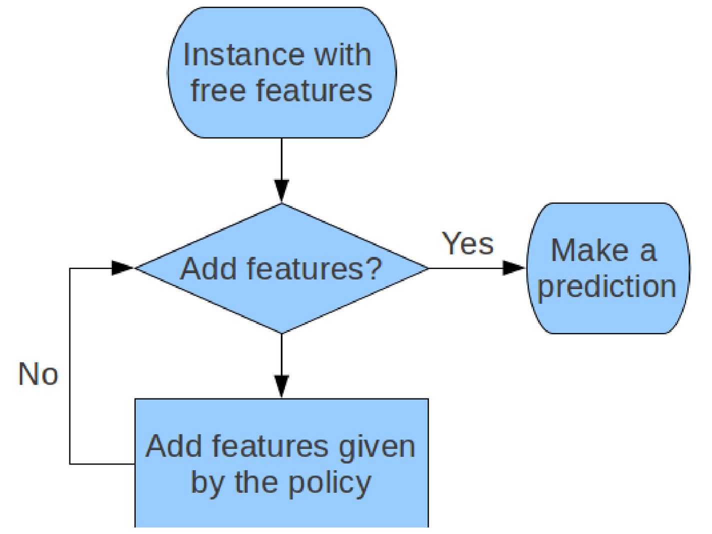
  predicted action $\hat{a}_{\pi,s} = \arg\max_{a\in A} \mathbf{w}^T\phi(s, a)$

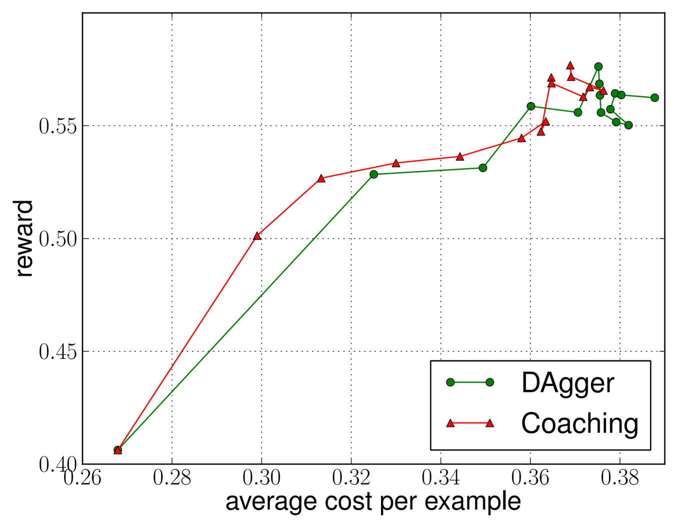
  hope action $\tilde{a}_{\pi,s} = \arg\max_{a\in A} \lambda\mathbf{w}^T\phi(s, a) - L(s, a)$
- For DAgger with coaching, if N is $O(uT \log T)$ and $Q_{T-t+1}^{\pi^*}(s, \pi) - Q_{T-t+1}^{\pi^*}(s, \pi^*) \le u$, there exists a policy $\pi \in \pi_{1:N}$ s.t. $J(\pi) \le J(\pi^*) + uT\tilde{\epsilon}_N + O(1)$.

Experimental Results

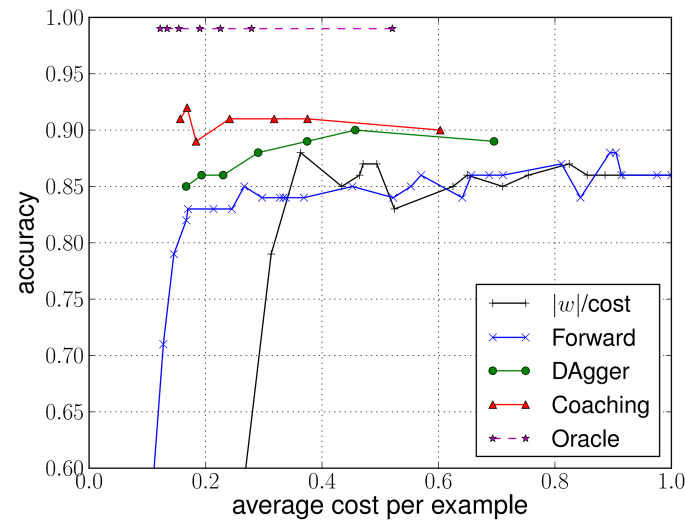
Dynamic feature selection

- Instance-specific feature selection at test time
- User-specified accuracy-cost trade-off
- state s_t: selected features and their values
- action a_t: features to add and *stop* (i.e. make a prediction with obtained features)
- immediate loss:
  $L(s, a) = \alpha \cdot \text{cost}(s) - \text{margin}(a)$

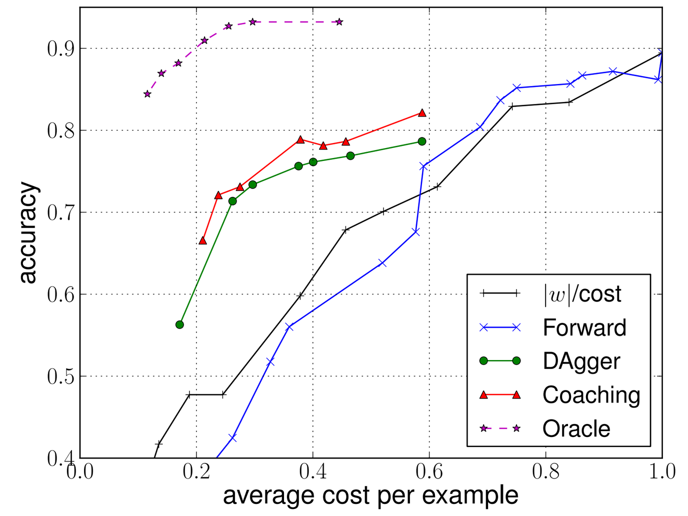




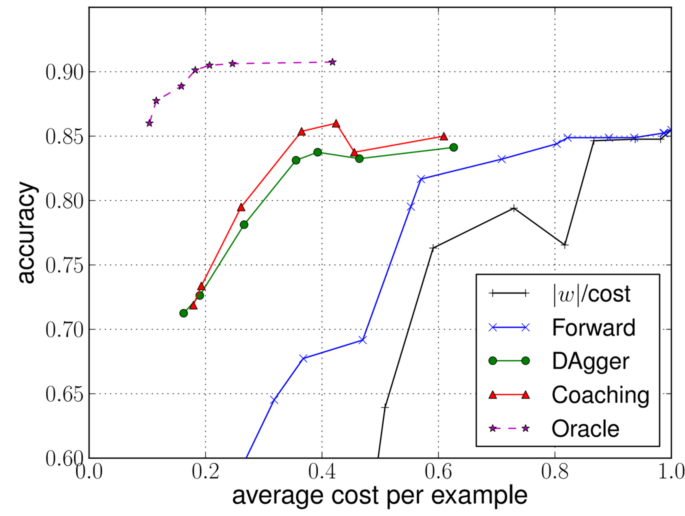
(a) Reward of DAgger and Coaching.



(b) Radar dataset.



(c) Digit dataset.



(d) Segmentation dataset.

Conclusion and Future Work

- Coaching: target at easier goals first and gradually approach the oracle
- Application in natural language processing and computer vision
- Relate to regularized methods in online convex optimization

Reference

[1] Stéphane. Ross, Geoffrey J. Gordon, and J. Andrew. Bagnell.
A reduction of imitation learning and structured prediction to no-regret online learning.
In *AISTATS*, 2011.