Semi-Supervised Inverse Reinforcement Learning

Michal Valko, Mohammad Ghavamzadeh, and Alessandro Lazaric
@inria.fr

Problem: Apprenticeship Learning

- Inverse reinforcement learning: expert trajectories → policy (via reward)
- Problem: expert trajectories are expensive to get or not available
- Solution: learn also from unlabeled trajectories and use the structure in the feature counts

Semi-Supervised Inverse Reinforcement Learning

- If we assume that the reward is linear in feature counts, \( R^*(s) = w^* \cdot \phi(s) \), then:
  \[
  \mathbb{E}_{s_t \sim D}[V^*(s_0)] = \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t R(s_t) | \pi \right] = w \cdot \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t \phi(s_t) | \pi \right] = w \cdot \mu(\pi).
  \]
- IRL of Abbeel and Ng [1] is based on matching the feature counts of the expert performer:
  \[
  \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t R(s_t) | \pi \right] - \mathbb{E}_\hat{\pi} \left[ \sum_{t=0}^{\infty} \gamma^t R(s_t) | \hat{\pi} \right] = \|w^\mu(\hat{\pi}) - w^\mu(\pi)\| \leq \|w\|_2 \|\mu(\hat{\pi}) - \mu(\pi)\|_2 \leq \varepsilon
  \]
- Semi-supervised learning (SSL) makes distributional assumptions such as compactness (gap, unsupervised) or smoothness (manifold). We choose to use the gap assumption and the related semi-supervised support vector machines (SVMs).
- Semi-supervised SVMs use besides the standard hinge loss \( \hat{V}(f, x, y) = \max(1 - y f(x), 0) \), also the hat loss \( \hat{V}(f, x) = \max(1 - |f(x)|, 0) \) on unlabeled data [2] to compute max-margin decision boundary \( \hat{f} \) that avoids dense regions of data:
  \[
  \hat{f} = \min_f \left\{ \sum_{i \in L} V(f, x_i, y_i) + \gamma \|f\|^2 + \gamma_u \sum_{i \in U} \hat{V}(f, x_i) \right\}
  \]
- In semi-supervised IRL (SSIRL) we penalize the decision boundary \( w \) that crosses the empirical feature counts from unlabeled trajectories:
  \[
  \min_w \left\{ \max \left\{ 1 - w^\mu(\pi), 0 \right\} + \gamma \|w\|_2 + \gamma_u \sum_{j < i} \max \left\{ 1 + w^\mu(j, \pi), 0 \right\} + \gamma_u \sum_{i \in U} \max \left\{ 1 - |w^\mu_u|, 0 \right\} \right\}
  \]

Discussion

- Contributions:
  - first IRL method to take advantage of the unlabeled trajectories
  - assuming clustered feature counts can learn a better performing policy
- Disadvantages:
  - similar to [1] only outputs a mixture policy
  - stopping criterion is needed, because the method converges to IRL [1]
- Future directions:
  - enhance other inverse RL methods (MaxEnt IRL, MMP) with unlabeled trajectories
  - investigate manifold assumption for inverse RL

References