Jenish Maharjan **Week 13**

Paper: *Maximum Entropy Inverse Reinforcement Learning*

Summary:

Ziebart et al, in their paper present a probabilistic approach of imitation learning based on the principle of maximum entropy. The goal in imitation learning problems is to predict the behavior and decision an agent would choose. Research has shown that looking at imitation learning as solutions to Markov Decision problems reduces learning to the problem of recovering a utility function that makes the behavior induced by a policy that is close to the optimal. Their method is based on this very idea of framing imitation learning problems as solutions to Markov Decision Problems. In the paper they provide a background and motivation for their approach, their algorithm along with its comparative evaluation and application areas.

The idea for looking at imitation learning problems to structure the space of learned policies to be solutions of search, planning or MDPs in general allows is powerful. This reduces the problem of imitation learning to recovering a reward function that causes the demonstrated behavior where the search algorithm works to "stitch-together" sequences of decisions that optimize the reward function. In the paper, the authors point out the limitation of previous algorithms about feature counts and inverse reinforcement learning (IRL). Both IRL and the matching of feature counts are ambiguous as each policy can be optimal for many reward functions and many policies may lead to the same feature counts. The ambiguity of suboptimality is unresolved. In their approach, they implement the principle of maximum entropy to resolve ambiguities in choosing distributions which leads to the distribution over behaviors constrained to match feature expectations while it is no more fixed to any specific path than required by the constraint.

The authors they investigate the problem of predicting driving behavior and route recommendation using their maximum entropy approach on IRL. For this, they model the structure for the road network surrounding Pittsburgh, as a deterministic MDP with over 300,000 states and 900,000 actions. Assumptions that the drivers executing the plans are trying to reach some goal while optimizing time, safety, stress, fuel costs and other factors is made for this task. The destination is represented within the MDS as an absorbing state without additional costs and each trip has a different destination and slightly different corresponding MDPs. The data was collected using GPS trace data from 25 cab taxi drivers over a 12-week period. Out of the collected data, 30% was discarded and 20% of remaining trips were used for training and 80% was used for testing set. The Maximum Entropy IRL model (MaxEnt) was applied to the task of learning taxi drivers' collective utility function for different features describing paths in the road network. The task is to maximize the probability of demonstrated paths within a smaller fixed class of reasonably good paths rather than the class of all possible paths. The algorithm is efficient for both classes but the reduction to smaller fixed class provided a significant speed up.

The approach was compared with two IRL models Maximum Margin Planning (MMP) and an action-based distribution model (Action) for evaluation of effectiveness. Each of the three models were evaluated for the ability to model paths in the testing set after training on the path's origin and destination using three metrics – first compared the model's most likely path estimate with the actual demonstrated path, second was the percentage of testing paths that matched at least 90% with the model's predicted path and the final metric measured the average log probability of paths in the train set under the given model. The results show that MaxEnt outperformed the other models in all three metrics with significant improvement. The paper also talks about other how their approach opens up possibilities for driver prediction. MaxEnt model could be used for tasks like driver behavior prediction by learning a probability distribution over driver preferences, destinations and routes. The paper successfully presents a novel approach to inverse reinforcement and imitation learning that resolves ambiguities present in previous works.

Strengths:

- The authors describe the concepts of Maximum Entropy IRL in detail, with a full section dedicated to it, which is useful for people who are new to it.
- They have a related works section where they mention other works in the realm of IRL models.

Weaknesses:

- Numbering the sections would have been helpful.
- The section explaining the concepts of Maximum Entropy IRL felt mathy to me.

Confusions:

- What is the difference between a deterministic and a non-deterministic MDP?
- How does the Expected edge frequency calculation algorithm work?

Discussion Questions:

- A walk-through on the concepts presented in the Maximum Entropy IRL section would be really helpful.
- What are other possible application areas of Maximum Entropy IRL?
- What is the current state of research with regards to IRL?