Hands-on sessions on Interactive Learning

Mohamed CHETOUANI

mohamed.chetouani@sorbonne-universite.fr

TP : Introduction to generative models of a demonstrator's actions

Objectives :
— Develop basic interactive learning algorithms : human behavior / demonstrations
— Compare performances of different strategies

Description :
Human decisions are not random and could be modeled using Bayesian models for example. There have been several attempts in such directions. The softmax policy (Boltzmann model) is usually considered as robust model for human decision prediction. This approach allows quantifying the likelihood that a human will select any particular option $o \in O$. If each option o has an underlying reward $R(o)$, the Boltzman model computes the desirability of an option as :

$$P(o) = \frac{e^{R(o)}}{\sum_{i \in O} e^{R(i)}} \tag{1}$$

The Boltzman model has been employed to derive policies to simulate human behaviours by considering environments modeled as a Markov Decision Process (MDP) defined by the tuple $< S, P, R, \gamma >$. with S a set of states S ; A a set of actions ; a transition function that maps states and actions to distributions over next states,$T_i : S \times A \rightarrow P(S)$; a reward function that maps state/action/next-state transitions to scalar rewards, R ; and a discount factor $\gamma \in [0, 1)$ that captures a preference for earlier rewards. The Bellman equation is used to compute the optimal agent behaviour.

$$\pi(a_t|s_t) = \frac{e^{Q^*(s_t,a_t)/\tau}}{\sum_{a' \in A(s_t)} e^{Q^*(s_t,a')/\tau}} \tag{2}$$

With $Q^*(s_t, a_t)$ the optimal policy, τ a temperature parameter.

The Boltzman model is known as noisily-rational behaviour. The aim of this practical is to generate such noisily-rational behaviours in order to produce demonstrations. More details of how such approaches are employed are availaible in [1, 2, 3] (no need to read them but there are good examples of how the models are employed to generate rational behaviours).

Environment :
The environment is a traditional gridworld (figure 1). The gridworld has one goal state that is worth 10 points and three other types of tiles (orange, purple, cyan). Each type of tile can each be either "safe" (0 points) or "dangerous" (-2 points).

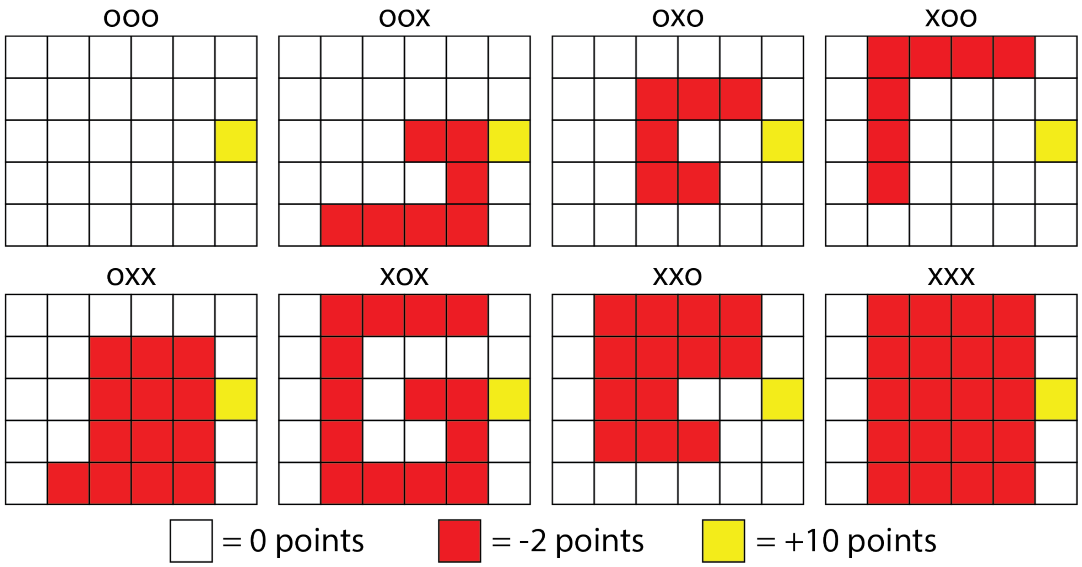

FIGURE 1 – Gridworld environments defined in [1]

Tasks :

1. Develop a gridworld environment (figure 1). You could use this environment : `https://github.com/MattChanTK/gym-maze`
2. Develop the Boltzman model to generate noisily-rational behaviors. Note : we could consider different costs for the directions.
3. Simulate the noisily-rational behaviors for environments : "oox", "oxo", "xoo", "xoo". Note : consider several simulations
4. Explain the role of the parameter τ in the noisily-rational behaviors generation.
5. (Optional) Explain why the approach proposed in [3] is relevant for human demonstrator's actions generation

Références

[1] Ho, M., Littman, M., Cushman, F., & Austerweil, J.L. (2018). Effectively Learning from Pedagogical Demonstrations. Cognitive Science.

[2] Milli, S., & Dragan, A.D. (2019). Literal or Pedagogic Human ? Analyzing Human Model Misspecification in Objective Learning. UAI.

[3] Bobu, A., Scobee, D. R. R., Fisac, J. F., Sastry, S. S., & Dragan, A. D. (2020). LESS is more : Rethinking probabilistic models of human behavior. In HRI 2020 - Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (pp. 429-437).