Perception Prediction

by Sven Nilsen, 2021

In this paper I discuss the idea of Perception Prediction.

An agent using Perception Prediction outputs actions and uses the history of actions to predict its input. With other words, the environment can be thought of as a non-deterministic function:

environment : action → input

The Perception Prediction agent does not know how it performs the actions.
However, it can observe the action it makes and use the history to predict input.

Perception Prediction can be thought of as a non-determinstic function:

perception_prediction : [action] → input

Based on previous actions, the agent predicts the next input. This function can be sampled to estimate the probability distribution of inputs.

Now, let us assume that the agent has a goal:

goal : input → utility

To maximize the utility of the goal, the agent can perform a sequence of actions which produces an input such that the computed utility is maximum.

This means that Perception Prediction in principle can be used to construct an Instrumental Rational[1] agent that optimises a single goal. However, it can also be used to construct a Higher Order Instrumental Rational[2] agent, where the goal is uncertain.

In perfect information environments[3], the Perception Prediction agent can predict the input perfectly. This is done by simulating the environment and replaying the effects of a sequence of actions, while also computing the input.

next_environment : environment × action → environment

input_state : environment → input

The next input is a function of the current environment and the action that the agent performs.

next_input(e : environment, a : action) = input_state(next_environment(e, a))

This is a slightly different interpretation of the environment as a deterministic function.
It means, the input state is computed deterministically from a state of the environment.

In practice, it is not possible to predict input perfectly, even for perfect information environment. However, it is useful as a theoretical tool to reason about agents who can behave rationally in principle, even their goal is not known.

References:

[1] “Instrumental Rationality”
Stanford Encyclopedia of Philosophy
https://plato.stanford.edu/entries/rationality-instrumental/

[2] “Zen Rationality”
Sven Nilsen, 2018
https://github.com/advancedresearch/path_semantics/blob/master/papers-wip/zen-rationality.pdf

[3] “Perfect information”
Wikipedia
https://en.wikipedia.org/wiki/Perfect_information