Portfolio 2

Anita Kurm: 201608652

The delta learning rule

Explain the delta learning rule with reference to the Rescorla-Wagner model of classical conditioning.

The delta learning rule is what decision-making agents in Rescorla-Wagner model use to **update their beliefs** (i.e. **learn**) based on weighted prediction error from prior experience.

$$Q_{t+1}^k = Q_t^k + \alpha(r_t - Q_t^k)$$

There Q_{t+1}^k stands for utility value attributed to option *k* for future trial, which is calculated based on the current utility value of option k summarised with the prediction error term (the difference between given on this trial reward r_t and the expected value Q_{t+1}^k) weighted by learning rate parameter α. The learning rate is a number between 0 and 1 and determines how much the prediction error affects the future utility value, which is necessary for adequate value updating (the smaller learning rate makes for slower value adjustment and therefore smaller chance to 'overshoot'). This framework for understanding learning as a prediction error driven process originated in Rescorla & Wagner's research on classical conditioning (1972), where they aimed to explain associative learning in terms of prediction errors, and saliency of conditioned stimuli and reinforcers.

The softmax function

Explain the softmax function with reference to the Luce choice rule as a model of choice.

The softmax function is used to calculate probability of making a certain choice based on expected utility values and explorative tendencies of an agent. It has the following properties: 1) takes into account expected value of an option; 2) chooses the most valuable option most often; 3) leaves a certain room for exploration via the inverse heat parameter β (consistency with expected utility). Softmax function for choosing option k with probability p_t^k looks like the following equation:

$$p_t^k = \frac{\exp(\beta Q_t^k)}{\sum_{i=1}^{K} \exp(\beta Q_t^i)}$$

Softmax function can be interpreted as follows: p_t^k (probability of choosing option k on trial t) is equal to the exponentiated product of Q_t^k (utility value of option k on trial t) multiplied by 'exploration parameter' β (inverse heat parameter) – all divided by the sum of all such exponentiated products for all options. The parameter β can range from 0 to ∞, where 0 stands for random choice (lots of exploration) and as β approaches ∞, the choice of highest expected value becomes inevitable (no exploration). This framework of thinking of choices in probabilistic terms was derived from Luce Choice Axiom (LCA), which was proposed by Luce in 1950s and summarised by Pleskac in 2012. Softmax is used to calculate the probability of a choice, which is influenced by preferences of an agent (utility values assigned to different options) and relates to the context of that choice (utility values of all other available options).

Delta rule and softmax in Q-learning: bandit tasks

Explain how the delta rule and the softmax function work together in the Q-learning algorithm as a model of sequential choice on bandit tasks.

Together, delta rule and the softmax function compose a Q-learning model, where an agent can learn utility values of different options based on rewards, and can use relative utility values of all options to make probabilistic choices. While the delta rule is used to **update** utility values Q based on reward, resulting prediction error and a certain learning rate, the softmax function is used to **make a probabilistic choice** based on learned utility values of all options and agent's consistency parameter. An example of a task that Q-learning model describes is a bandit task, where every trial an agent chooses one of the two bandit machines. From the very start of the task, both machines have a certain utility value in eyes of an agent, but in reality have different probabilities of yielding rewards and rewards of different size. Throughout the task, every trial the agent picks a machine, either gets or does not get a reward, and updates the utility value for the picked machine. Given utility values of both machines, the agent uses the softmax function to evaluate probabilities of choosing either of the machines. Depending on the inverse heat parameter in the softmax function, the agent will choose a machine with a higher utility value more or less frequently.

Q-learning vs Choice Kernel

The choice kernel (CK) model presented in Wilson and Collins (2019) imitates an agent, who keeps track of previous choices and tends to repeat them. In this model, we still use the delta rule and softmax function, but instead of evaluating utility value of an option (Q_t^k in the Q-learning model), the agent evaluates the choice kernel of the option (CK_t^k in this model), which reflects how often this option has

been chosen before. Moreover, in calculation of the prediction error, instead of rewards the model simply takes value of either 0 or 1 corresponding to whether the options was chosen on last trial. Then choice kernels of all options is taken into account, as well as the inverse heat parameter to make a probabilistic choice out of options presented.

To evaluate the Q-learning in relation to the CK model, we need to conduct parameter recovery and model recovery.

Parameter recovery

For both models, parameter recovery was conducted on 75 different parameter combinations, for which each time the task was running for 100 trials. Results presented in Figures 1a-2b show that in both models parameter recovery was successful for β, but not reliable for α. This suggests that neither Q-learning model, nor CK model performed at an acceptable level.

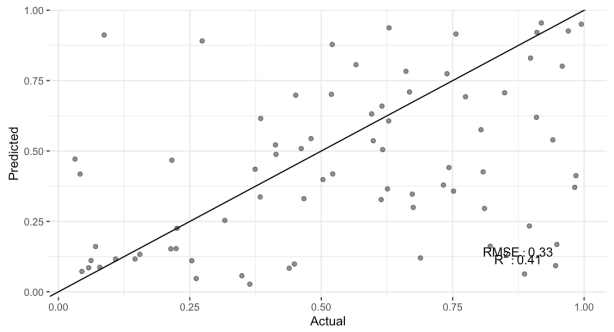

Figure 1a. Parameter α recovery for Q-learning model

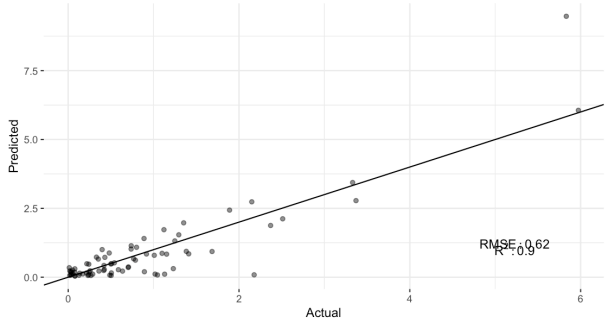

Figure 1b. Parameter β recovery for Q-learning model

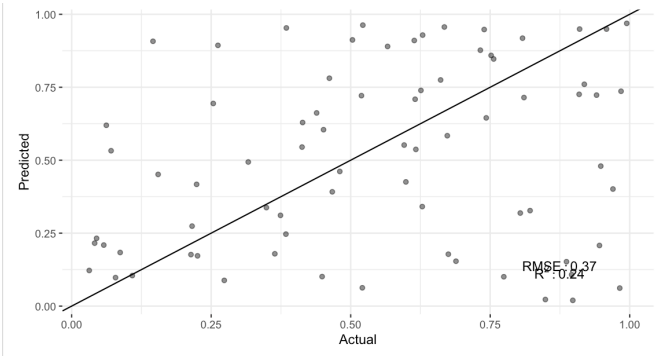

Figure 2a. Parameter α recovery for CK model

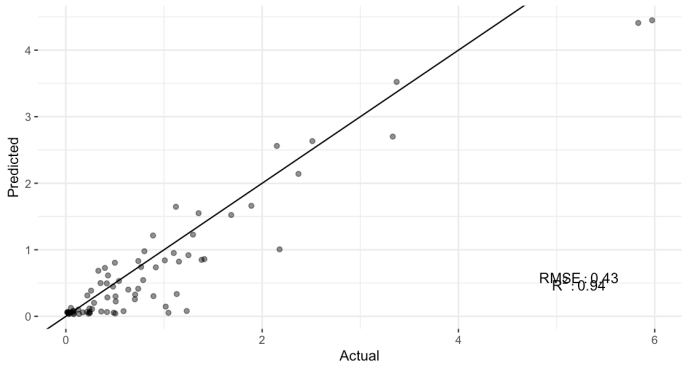

Figure 2b. Parameter β recovery for CK model

Model recovery

To further evaluate quality of the Q-learning model, a model recovery process has been conducted to make sure that it is best at describing its own data. The results presented in Figure 3 suggest that both Q-learning model and CK model were best at explaining their own data, which is a good result.

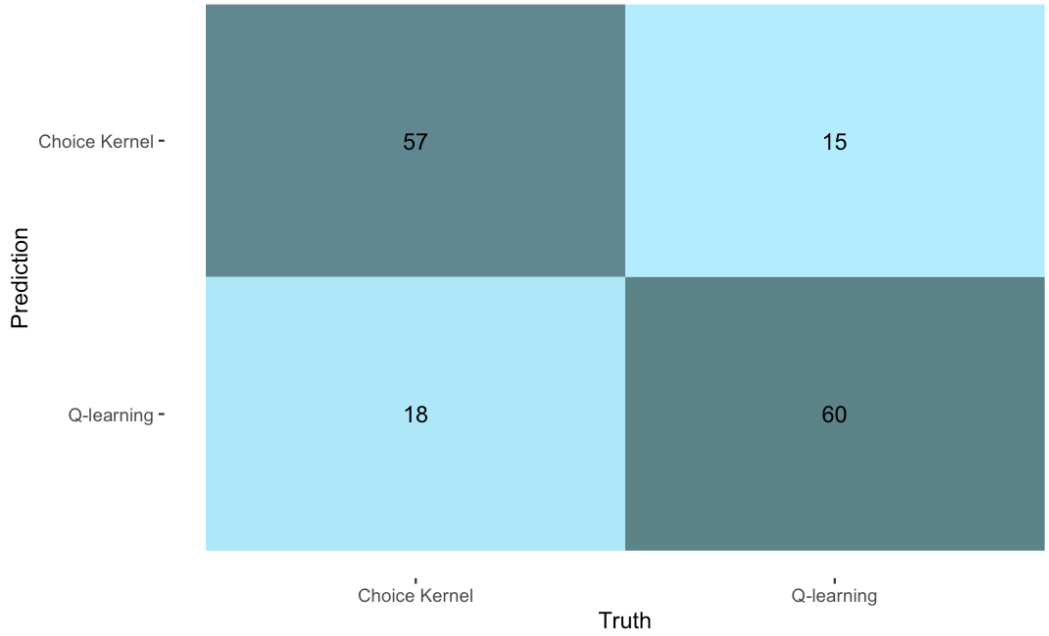

Figure 3. Confusion matrix of 75 simulations. 'Truth' represents the model that data was generated with and 'Prediction' represents which model had the smallest DIC on that data.

References:

Rescorla, R.A., & Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black & W.F. Prokasy (Eds.) *Classical Conditioning II*. pp. 64–99. Appleton-Century-Crofts.

Pleskac, T. J. (2012). Decision and Choice: Luce's Choice Axiom. In J. D. Wright (Ed). *International Encyclopedia of the Social & Behavioral Sciences* (Second Edition).

Wilson and Collins (2019) Ten simple rules for the computational modeling of behavioral data