DAT440 / DIT471 Advanced topics in machine learning - Solutions Re-exam

Course Responsible and Examiner: Morteza Haghir Chehreghani

August 15, 2022, 2:00 pm

1. Multiple choice questions
   (a) ii
   (b) i, iv
   (c) ii
   (d) ii
   (e) i
2. Successive Elimination
   (a) All arms are still active during the current phase, so all arms will be played. Arms are deactivated in the end of the phase.
   (b)

$$
\begin{aligned}
\mathrm{UCB}_t(1) &\approx \frac{112}{224} + \sqrt{\frac{2\cdot 7}{224}} = \frac{3}{4}\\
\mathrm{LCB}_t(1) &\approx \frac{112}{224} - \sqrt{\frac{2\cdot 7}{224}} = \frac{1}{4}\\
\mathrm{UCB}_t(2) &\approx \frac{56}{224} + \sqrt{\frac{2\cdot 7}{224}} = \frac{1}{2}\\
\mathrm{LCB}_t(2) &\approx \frac{56}{224} - \sqrt{\frac{2\cdot 7}{224}} = 0\\
\mathrm{UCB}_t(3) &\approx \frac{168}{224} + \sqrt{\frac{2\cdot 7}{224}} = 1\\
\mathrm{LCB}_t(3) &\approx \frac{168}{224} - \sqrt{\frac{2\cdot 7}{224}} = \frac{1}{2}\\
\mathrm{UCB}_t(4) &\approx \frac{84}{224} + \sqrt{\frac{2\cdot 7}{224}} = \frac{5}{8}\\
\mathrm{LCB}_t(4) &\approx \frac{84}{224} - \sqrt{\frac{2\cdot 7}{224}} = \frac{1}{8}
\end{aligned}
$$

   (c) All 4 arms are played in the current phase. Since $\mathrm{UCB}_{t+4}(2) < 0.5 < \mathrm{LCB}_{t+4}(3)$, arm 2 will be deactivated.
   (d) When the confidence intervals of the arms with highest and lowest mean estimates don't overlap, the arms with the lowest mean estimates have a low probability of being the best arm and thus may be discarded.

3. Thompson Sampling
   (a) Thompson Sampling samples arms to play according to their posterior probability of being optimal. This means that an arm is more likely to be sampled if the posterior distribution over the expected reward has high mean (exploitation) or high variance (exploration).

   (b) Since all of the posterior distributions have the same variance, the arm of the one with the highest mean has the highest posterior probability of being the best arm, i.e., the second arm, with posterior mean 0.8. However, due to variance in the samplings, this choice is not guaranteed as another arm just by chance might yield a larger sample.

   (c) The posterior distribution of the arm with the largest estimated mean reward needs to be concentrated (with a narrow variance) such that it does not have an overlap with the posterior distributions of the other arms. The other arms may have wide and overlapping distributions with each other.

   (d) Since the samples are drawn i.i.d. from the same posterior distributions and only one is selected, the expected regret is the same as if the first sample is selected (as in standard Thompson Sampling).

4. Policy improvement

   See Chapter 4.2 in RL course literature by Sutton & Barto.

5. Deep Q-network
   (a) It is off-policy since the behavior policy and target policy are different. The behavior policy is given by ϵ-greedy($\hat{q}$), while the target policy is given by greedy($\tilde{q}$).

   (b) Mainly line 17 and 18 of the algorithm since this is where the loss is computed and used to update $\hat{q}$.

   (c) The part of the algorithm where the experience replay memory is used. Using experience replay have the following advantages: (1) More data efficient learning since it allows each stored experience to be used for many updates; (2) Reduces variance of updates since successive updates are not correlated with each other; (3) More stable training since experience replay removes the dependence of successive experiences from the current weights.

   (d) Duplicate network is used since otherwise the target depends on the current action-value function estimate. Then, when using a parameterization of the action value, the target is a function of the same parameters that are being updated. This can lead to oscillation and/or divergence.

6. REINFORCE
   (a) It is on-policy since the algorithm evaluates/improves the same policy that is used to observe experience.

   (b) Actor-Critic (AC) learns both a policy and value function estimation, while REINFORCE only learns a policy. AC replaces the full return of REINFORCE with the one-step return. AC uses a baseline while this variant of REINFORCE does not. Both uses policy gradients and stochastic gradient ascent to update the policy.

   (c) It is a Monte-Carlo algorithm, since a full episode is used to estimate the return.

   (d) See Chapter 13.3 in RL course literature by Sutton & Barto.