Games and Computation Homework #5: Exploration, Exploitation, & Monte Carlo Methods

Answer these questions within the HW #5 Moodle quiz:

Real-Life Exploration-Exploitation Tradeoff

There are many real-life examples of the exploration-exploitation tradeoff. If I try a new restaurant menu item, I may regret having chosen it compared to my known favorite, yet by only choosing what I've enjoyed most, I may never discover a menu item I might enjoy even more.

Describe a real-life exploration-exploitation trade-off where you have felt a tension between choosing the best known versus an unknown in your experience:

Epsilon-Greedy 3-Armed Bandit Selection

Let epsilon = 0.2. After selecting arms 0-2 once each, epsilon-greedy selection will choose a random arm from arms 0-2 with probability 0.2. Otherwise, the arm with the best average payout will be chosen. Arm 0 has been pulled 3 times with a total payout of 5.894. Arm 1 has been pulled 1 time with a total payout of 1.992. Arm 2 has been pulled 2 times with a total payout of 4.072. Given this information, what is the probability that arm 2 will be pulled next?

a) 0.100
b) 0.800
c) 0.067
d) 0.867

Epsilon-Greedy 4-Armed Bandit Selection

Let epsilon = 0.1. After selecting arms 0-3 once each, epsilon-greedy selection will choose a random arm from arms 0-3 with probability 0.1. Otherwise, the arm with the best average payout will be chosen. Arm 0 has been pulled 2 times with a total payout of 13.044. Arm 1 has been pulled 1 time with a total payout of 6.884. Arm 2 has been pulled 4 times with a total payout of 29.065. Arm 3 has been pulled 3 times with a total payout of 19.628. Given this information, what is the probability that arm 1 will be pulled next?

a) 0.033
b) 0.025
c) 0.900
d) 0.925

Epsilon-Greedy 5-Armed Bandit Selection

Let epsilon = 0.5. After selecting arms 0-4 once each, epsilon-greedy selection will choose a random arm from arms 0-4 with probability 0.5. Otherwise, the arm with the best average payout will be chosen. Arm 0 has been pulled 3 times with a total payout of 5.066. Arm 1 has been pulled 5 times with a total payout of 8.084. Arm 2 has been pulled 2 times with a total payout of 4.792. Arm 3 has been pulled 1 time with a total payout of 1.822. Arm 4 has been pulled 4 times with a total payout of 6.872. Given this information, what is the probability that arm 1 will be pulled next?

a) 0.600
b) 0.500
c) 0.125
d) 0.100

UCB1 3-Armed Bandit Selection 1

Let c = 5. After selecting arms 0-2 once each, UCB1 will select the arm that maximizes v_i/n_i + c*sqrt(ln(t)/n_i), where v_i is the total payout for arm i, n_i is the number of times arm i has been selected, t = the total number of times all arms have been selected, sqrt is the square root function, and ln is the natural log function. Arm 0 has been pulled 100 times with a total payout of 564.750. Arm 1 has been pulled 1 time with a total payout of -3.978. Arm 2 has been pulled 10 times with a total payout of 34.905. Given this information, which arm will UCB1 select next?

- ☐ Arm 0
- ☐ Arm 1
- ☐ Arm 2

UCB1 3-Armed Bandit Selection 2

Let c = 4. After selecting arms 0-2 once each, UCB1 will select the arm that maximizes v_i/n_i + c*sqrt(ln(t)/n_i), where v_i is the total payout for arm i, n_i is the number of times arm i has been selected, t = the total number of times all arms have been selected, sqrt is the square root function, and ln is the natural log function. Arm 0 has been pulled 100 times with a total payout of 904.474. Arm 1 has been pulled 10 times with a total payout of 70.402. Arm 2 has been pulled 1 time with a total payout of 0.962. Given this information, which arm will UCB1 select next? {

- ☐ Arm 0
- ☐ Arm 1
- ☐ Arm 2

UCB1 3-Armed Bandit Selection 3

Let c = 3. After selecting arms 0-2 once each, UCB1 will select the arm that maximizes v_i/n_i + c*sqrt(ln(t)/n_i), where v_i is the total payout for arm i, n_i is the number of times arm i has been selected, t = the total number of times all arms have been selected, sqrt is the square root function, and ln is the natural log function. Arm 0 has been pulled 1 time with a total payout of 3.863. Arm 1 has been pulled 100 times with a total payout of 976.246. Arm 2 has been pulled 10 times with a total payout of 81.142. Given this information, which arm will UCB1 select next?

- ☐ Arm 0
- ☐ Arm 1
- ☐ Arm 2

Monte Carlo Tree Search Expansion 1

Initially containing only a root node for the current state, Monte Carlo Tree Search (MCTS) then iterates through 4 stages: Selection (within the tree), Expansion (growing the tree), Simulation (beyond the tree), and Backpropagation (back up through the tree). Imagine MCTS on the initial state of a western Mancala game with 6 pits per side and 4 pieces per pit. Assuming that expansion grows a single node per iteration when a selection first leads beyond existing nodes, how many nodes will the tree contain after 5 iterations?

_____

Monte Carlo Tree Search Expansion 2

Initially containing only a root node for the current state, Monte Carlo Tree Search (MCTS) then iterates through 4 stages: Selection (within the tree), Expansion (growing the tree), Simulation (beyond the tree), and Backpropagation (back up through the tree). Imagine MCTS on the initial state of a western Mancala game with 6 pits per side and 4 pieces per pit. Assuming UCB1 selection and expansion that grows a single node per iteration when selection first leads beyond existing nodes, what is the maximum depth of the tree grown after 5 iterations? (The root node is at depth 0.)

_____