Predictive Analytics World / Deep Learning World Exercises – Bandits

1. Create a 3-armed bandit environment. Let the user enter each action manually. Output each reward received. Set each reward to either 1.0 or 0.0. Fix the probabilities of non-zero reward to be [0.2, 0.5, 0.7] for the three possible actions.

2. Connect a uniform random agent to the environment. Output the mean reward received over 1000 steps.

3. Modify the agent to record the reward received for each action, and use the epsilon-greedy algorithm to balance exploration with exploitation. Find the value of epsilon that earns the most reward over 1000 total steps.

4. Modify the agent to use Thompson Sampling instead of epsilon-greedy. Compare the new mean reward received to epsilon-greedy's performance.

5. Which is the best definition of an optimal policy?

a. The highest sum of rewards that can possibly be received.
b. A set of rules for choosing actions that earns no less reward than any other policy.
c. A strategy for selecting the best value of the epsilon parameter.

6. Which of the following are potential problems with the epsilon-greedy algorithm?

a. When epsilon is near 0.5, exploration and exploitation can interfere with each other.
b. When epsilon is too low, it can take a long time to learn an optimal policy.
c. A fixed value of epsilon guarantees that a certain fraction of reward will always be lost.