CPSC 422,
Homework 4

General Instructions:

- Working in Teams: You may work with at most one other person. That person must also be enrolled in CPSC 422 this term. If you are working with a partner, the two of you must submit only one assignment, listing both of your names. Keep in mind that when you work in pairs, each of you must understand the work that you submit.
- Note: You are permitted to derive your solutions by implementing the respective algorithms, even when it is not explicitly required in the question. You are NOT permitted to take existing implementations of the algorithms and use those for your solutions. Such existing algorithms include those available from friends, CD, or off the web. You are also NOT permitted to provide your implementation to anyone else outside your team.
- Important: It is best if you type your assignment. Handwritten work will be marked only if clearly legible, and will still have to be submitted via handin.
- Make sure to include your name and student ID (and the names and student IDs of the two team members if you are working with someone else) and the assignment number, at the top of the assignment. Ensure that each question and sub-question is appropriately labeled and clearly identifiable.
- The assignment is due Tu., Apr. 13, midnight.

Question 1 [35 points] Bayesian Learning

Fig.1 reproduces the candy bag example we used in the lectures on Bayesian learning. The data in the example (drawing 10 lime candies in a row), can be seen as having been generated by distribution h_0 in the figure.

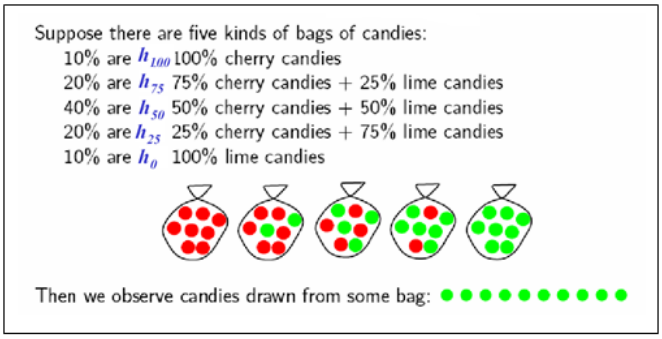


Figure 1: The hypothesis set for the candy example.

a **[20 points]** For this question, you need to generate a dataset of length 100 from

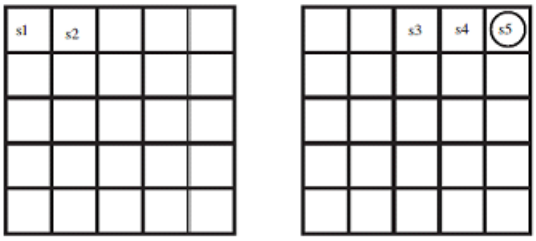
- h_{50} if your are doing the homework by yourself
- h_{25} both of h_{50} and h_{25} if you are doing this in group

Then, using full Bayesian learning on the above dataset(s), plot the corresponding graphs for $P(h_i|d_1...d_m)$, with $m = 1...N$, for all five hypothesis (as we have seen in class) and $P(d_{m+1} = \text{lime}|d_1...d_m)$.

b **[15 points]** Using the dataset(s) generated in question (a) above, plot $P(d_{m+1} = \text{lime}|h_{MAP})$ and $P(d_{m+1} = \text{lime}|h_{ML})$. Comment on your results.

Question 2 [20 points] Reinforcement learning

Consider a 5x5 grid world as shown in the diagram below, where the agent can move up, right, down or left. A prize can appear at one of the corners, generating a positive reward. Monsters can (stochastically) attack at certain locations, generating a negative reward. Suppose that the agent steps through the state space in the order of steps given in the diagram below, (i.e., going from s1 to s2 to s3 to s4 to s5), each time doing a right action. You can assume that this is the



first time the robot has visited any of these states. All Q-values are initialized to zero. Assume that the discount is 0.9.

a **[10 points]** Suppose the agent received a reward of -10 entering state s3 and received a reward of +10 on entering the state s5, and no other rewards. What Q-values are updated during Q-learning based on this experience? Explain what values they get assigned. You should assume that $\alpha_k = \frac{1}{k}$ (note that for Q-learning with varying α_k, there needs to be a different count k for each state-action pair).

b **[10 points]** Suppose that, at some later time in the same run of Q-learning, the robot revisits the same states: s1 to s2 to s3 to s4 to s5, and has not visited any of these states in between (i.e, this is the second time visiting any of these states). Suppose this time, the agent only receives a reward of +10 on entering the state s5. You should assume that $\alpha_k = \frac{1}{k}$. What Q-values have their values changed? What are their new values? Explain the answers.

Question 3 (20 points) Reinforcement learning

Consider four different ways to derive the value of α_k from k in Q-learning (note that for Q-learning with varying α_k, there needs to be a different count k for each state-action pair).

(i) Let $\alpha_k = \frac{1}{k}$

(ii) Let $\alpha_k = \frac{10}{9+k}$

(iii) Let $\alpha_k = 0.1$

(iv) Let $\alpha_k = 0.1$ for the first 10,000 steps, $\alpha_k = 0.01$ for the next 10,000 steps, $\alpha_k = 0.001$ for the next 10,000 steps, $\alpha_k = 0.0001$ for the next 10,000 steps, etc.

In this question, we compare these four methods theoretically and experimentally.

a **[8 points]** Which of these will converge to the true Q-value in theory?

b **[8 points]** We will now evaluate how these methods perform in practice. Method i), iii), and iv) can be evaluated using the Q-learning applet at
http://www.cs.ubc.ca/spider/poole/demos/rl/q.html,

method ii) is implemented in the applet at:
http://www.cs.ubc.ca/spider/poole/demos/rl/q10.html.

For the simple 10x10 grid world in these applets, execute up to 100,000 steps with each method of setting α_k and comment on the resulting Q-values as compared to the true Q-values (which you computed in assignment 2 using the value iteration applet at
http://www.cs.ubc.ca/spider/poole/demos/mdp/vi.html).

c **[4 points]** Which methods can adapt when the environment adapts slowly?

Question 4 [10 points]: Reinforcement learning

In this question, we will study the tradeoff between exploration and exploitation in reinforcement learning, using the Q-learning applet at
http://www.cs.ubc.ca/spider/poole/demos/rl/q.html.

In order to get reliable results in this noisy domain, perform a long run of 5 million steps for each strategy. For each method, evaluate the following two performance metrics:

(1) The total reward received after 5 million steps.

(2) The additional reward collected by an additional 5 million greedy steps (i.e. with parameter greedy exploit set to 100

Intuitively, the first metric measures how well the method exploits and the second metric measures how good the learned Q-values are, i.e. how well the method has explored. Initializing Q-values to zero (initial value set to 0), evaluate the following values for the greedy exploit parameter: 0, 80%, and 100%. Explain your findings.

Question 5 [10 points]

[Note that this question is worth marks, so dont forget to do it.]

(a) (3 points) For each question in this assignment, say how long you spent on it.

(b) (3 points) Rate each question in this assignment by how much you learnt doing it, on a scale from 1 (very little) to 5 (a whole lot)

(c) (4 points) For each question in this assignment, what did you learn?