# COMP 652: Machine Learning - Assignment 4

## Posted Wednesday, April 8, 2015
## Due Tuesday April 14, 2015
## No penalties until Wednesday April 22, 2015

**This assignment is optional. If you complete it, it will be averaged as part of your homework grade. Otherwise, your homework grade will be the average of the assignments completed already.**

1. [25 points] **Markov Decision Processes**

   Jack has a car dealership and is looking for a way to maximizes his profits. Every week, Jack orders a stock of cars, at the cost of d dollars per car. These cars get delivered instantly. The new cars get added to his inventory. Then during the week, he sells some random number of cars, $k$, at a price of $c$ each. Jack also incurs a cost of $u$ for every unsold car that he has to keep in inventory. Formulate this problem as a Markov Decision Process. What are the states and actions? What are the rewards? What are the transition probabilities? Describe the long-term return.

2. [40 points] **Bandit problems**

   In this problem, you will compare two bandit algorithms. First, write a little simulation of a two-armed bandit, with the two arms having Bernoulli rewards. One arm has probability of success 0.5, while the other has probability of success 0.6. For all your experiments, you should perform 10 independent runs consisting of 100 trials each of the algorithms and present average results over these runs. You should plot 3 graphs, measuring the average regret as a function of the number of trials, the average reward obtained for the trial, and an indicator showing whether the optimal action is estimated to be best at that trial or not.

   The algorithms to compare are:

   (a) $\epsilon$-greedy exploration with $\epsilon = 0.1$

   (b) $\epsilon$-greedy exploration with $\epsilon = 0.01$

   (c) The UCB algorithm (also known as UCB1)

   Please write a paragraph explaining the results that you see.

3. [35 points] **Advantage functions**

   We discussed in class the notion of a state-value function, $V^\pi$ and of a state-action value function, $Q^\pi$. Another quantity which can sometimes be useful is the advantage function:

$$A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$$

(a) [5 points] Can the advantage function be positive for all actions? Justify your answer

(b) [10 points] Explain how the advantage function can be used to decide if an action should be eliminated from the choices made by a policy $\pi$

(c) [10 points] How will the advantage function of the optimal policy look like?

(d) [10 points] Derive a Bellman equation for advantages and based on it, a learning rule for advantages