COMP-424: Artificial intelligence

Homework 4

Due on *myCourses* Wednesday, Apr 10, 9:00pm.
Head TA (for grading inquiries): Jade Yu <lei.yu@mail.mcgill.ca>

Exceptional late policy: Because of the overlap with the project submission deadline, you may hand in this assignment without penalty up until Saturday, April 13, 9:00pm. No submissions will be accepted past this second deadline.

General instructions.

- This is an individual assignment. You can discuss solutions with your classmates, but should only exchange information orally, or else if in writing through the discussion board on *myCourses*. All other forms of written exchange are prohibited.
- Unless otherwise mentioned, the only sources you should need to answer these questions are your course notes, the textbook, and the links provided. Any other source used should be acknowledged with proper referencing style in your submitted solution.
- Submit a single pdf document containing all your pages of your written solution on your McGill's *myCourses* account. You can scan-in hand-written pages. If necessary, learn how to combine many pdf files into one.

Question 1: HMMs for Part-of-Speech (POS) Tagging

In computational linguistics, **part-of-speech tagging** (or **POS tagging**) is the process of marking up words in texts as corresponding to their "part of speech identifications". A simplest form of this task is to identify each word as noun, verb, adjective, adverb, etc. Here we demonstrate the power of HMM by applying it on POS tagging.
You need to design an HMM to make predictions regarding the POS identifications of words in an English sentence. The **observed states** are the words themselves in the given sequence, while the **hidden states** would be the POS tags for the words. The **transition probabilities** would be somewhat like P(verb | noun) that is, what is the probability of the current word having a tag of verb given that the previous tag was a noun. **Emission probabilities** would be like P(that | noun) or P(good | adjective), which denote the probability that we observed the word, say "good", given that the tag is an adjective.
Now suppose you have the following training corpus, where each word is annotated with its POS tag:

*That/**(conjunction)** that/**(noun)** is/**(verb)**, is/**(verb)**.*
*That/**(conjunction)** that/**(noun)** is/**(verb)** not/**(noun)**, is/**(verb)** not/**(verb)**.*
*Is/**(verb)** that/**(verb)** it/**(noun)**?*
*Is/**(verb)** it/**(noun)** that/**(noun)** good/**(adjective)**?*

For each of these following tasks, write down your calculations, or provide the code that you wrote to compute the answer.

(a) Ignoring capitalization and punctuations, there should be a tag set of size 4 and a lexicon of size 5. Write out the corresponding sets of hidden states and observed states. Then define the HMM of this process by giving: 1) the initial probability vector; 2) the transition probability matrix; 3) the emission probability matrix.
All these probabilities should be learned from the training data. To avoid very sparse empirical distributions, we'll apply a technique called **add-one smoothing**: for each entry in a k-class multinomial distribution with N trials, the smoothed version of MLE would be:

$$p_i = \frac{N_i+1}{N+k}$$

where N_i is the number of observations from class *i*.

(b) What is the probability of the observed sentence "Not that good"?

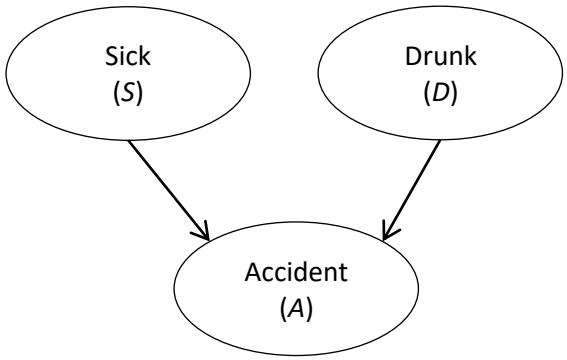
(c) What is the probability of that, if there's another word after "good" in the previous sequence "Not that good", the 4th word has POS tag as a noun?

(d) What is the most likely sequence of POS tags for the sentence in (b)?

S (F)	S (T)
0.75	0.25

D (F)	D (T)
0.85	0.15

Question 2: Utility



Consider the Bayes Net shown here, with all Bernoulli variables, which models driving accidents. Having an accident has a utility of -100 if the car was insured, but has a utility of -500 if the car was not insured (or the insurance claim is denied). Having insurance when there is no accident has a utility of -20, and not having insurance and an accident has a utility of 0.

Use the principle of Maximum Expected Utility and Value of Information to answer the following questions. For parts a)-c) assume the insurance company pays for 100% of the cases.

S	D	A (T)
F	F	0.05
F	T	0.1
T	F	0.1
T	T	0.75

a) Given no information on whether the driver will be driving in a sick or drunk state, how much should they pay to insure the car?

b) How much should they pay for insurance if you know for certain that they will drive both sick *(S=True)* and **not drunk** *(D=False)*?

c) How much should they pay for insurance if you know that they will drive drunk *(D=True)*?

d) A company is offering cheaper insurance but has a reputation of rejecting 10% of insurance claims. How much should they charge for this insurance, to make it competitive with the insurance offered by the more reliable company? (*Hint*: Set the cost of the new insurance to have the same MEU as the other insurance.)

Question 3: Jade's Diet

Here we show how reinforcement learning could be applied to encourage our TA Jade to eat a healthier diet. Suppose Jade's daily lunch choices consist of six foods, with each of them assigned a relative "health score" reflecting its healthfulness:

Food (State) Name	**Health Score (c_i)**
Salad	50
Yogurt	40
Hamburger	30
Burrito	20
Pizza	10
Poutine	0

We could therefore construct an MDP with each food choice as a state. At each state s_i, our TA could take actions to switch into any food state that is one level high or one level below s_i, or he could also stay in s_i. (For instance, at state "Yogurt", Jade could select both "Salad" or "Hamburger", or stay at "Yogurt".) The transitions for all actions have a Pr=0.8 of success (in which case our TA goes to the new state) and Pr=0.2 of failure (in which case our TA stays in current state). The reward of moving into state s_j from any state s_i would be the health scores c_j of food s_j. Assume a discount factor of $\gamma=0.9$. When updating policies, if necessary, break ties by choosing the food state with highest health score.

a) Describe the space of all possible policies for this MDP. How many are there?
b) Assume initially, Jade acts on a fairly unhealthy policy π^0: at state s_i, he would move into the food (state) that is one level below s_i in the health score form above: for instance, $\pi^0(burrito) = \text{"choose pizza tomorrow"}$ (except in state "Poutine", where Jade would stay on his favorite). What is the initial value function $V^0(s)$ for each state in this case?
c) Given the initial estimate, V^0, if you run one iteration of policy improvement, what will be the new choice for Jade at each food state?
d) What is the optimal value function at each state for this domain?
e) Is the optimal value function unique? Explain.
f) What is the optimal policy at each state for this domain?
g) Is the optimal policy unique? Explain.
h) If we divide all the health scores by a scaling constant of 10, would this change influence the optimal values and optimal policy? Explain.

Question 4: Bandits

Consider the following 6-armed bandit problem. The initial value estimates of the arms are given by Q = {1, 2, 2, 1, 0, 3}, and the actions are represented by A = {1, 2, 3, 4, 5, 6}. Suppose we observe that each lever is played in turn: (from lever 1 to lever 6, and then start from lever 1 again):

$$A_t = ((t-1) mod\ 6)\ + 1 \qquad (1)$$

We also observe that the rewards R_t seem to fit the following function:

$$R_t = \ 2\cos\left[\frac{\pi}{6}(t-1)\right] \qquad (2)$$

So, the first two action-reward pairs are $A_1 = 1,\ R_1 = 2$, and $A_2 = 2,\ R_2 = \sqrt{3}$.

a) Show the estimated Q values from t=1 to t=12 of the trajectory using the average of the observed rewards, where available. Do not consider the initial estimates as samples.

b) It turns out the player was following an ϵ-greedy strategy, which just happened to coincide with the scheme described above in (1) for the first 12 time steps. For each time step t from 1 to 12, report whether it can be concluded with certainty that a random action was selected.

c) Suppose now we continue to visit the levers iteratively as in (1), and that the observed rewards continue to fit the pattern established by (2). Is there a limiting expected reward $Q^*(a)$ for each action $a \in A$ as t approaches infinity? Justify your answer.