CS440/ECE448 Spring 2019 Exam 2 Review

Be able to define the following terms and answer basic questions about them:

- **Bayesian inference**
  - Likelihood, prior, posterior
  - Maximum likelihood (ML), maximum a posteriori (MAP) inference
  - Naïve Bayes
  - Parameter learning
  - Laplace smoothing

- **Machine learning**
  - Naïve Bayes as a linear classifier
  - The perceptron learning rule; multi-class perceptrons
  - Training, development test, evaluation test, generalization, overfitting
  - Different types of supervision (supervised, unsupervised, self-supervised, etc.)

- **Bayesian networks**
  - Structure and parameters
  - Conditional independence assumptions
  - Calculating joint and conditional probabilities
  - The sum-product algorithm for trees: you can do this
  - The junction tree/join tree algorithm for almost-tree graphs: you know the name of this algorithm
  - Complexity of inference: the SAT problem as a Bayes net
  - Parameter estimation for fully observable problems: you can do this
  - Parameter estimation for partially observable problems: you know the name of the algorithm that is used
  - Hidden Markov models: definition, prediction, filtering

- **Markov decision processes and reinforcement learning**
  - Markov assumption, transition model, policy
  - Bellman equation
  - Value iteration, policy iteration
  - Model-based vs. model-free approaches
  - Exploration vs. exploitation
  - Q-learning: TD and SARSA
  - Deep Q-learning: Q(s,a) as a function, the advantage function A(s,a)
  - Convergence; the epsilon-greedy algorithm
  - Stability of learning; the experience replay buffer
  - Policy learning: softmax function, actor-critic network
  - Imitation learning: policy imitation

- **Deep learning**
  - Neural net forward propagation
  - Training criteria: sum-squared error and cross-entropy
  - Derivative of the cross-entropy w.r.t. softmax inputs
  - Back-propagation

    - Convolutional neural net: define convolution, channels, ReLU, max pooling, layer normalization

- **Natural language processing and Speech processing**
    - Hybrid HMM-DNN: you should know how to divide the softmax by the prior in order to turn it into a pseudo-likelihood for use in an HMM.
    - Recurrent neural net: you should know what it is (previous state is input to next state), and you should know what a sequence-to-sequence model is (time step calculated separately for input and output sequences). Details like LSTM, CTC, and attention are NOT required.

Sample exam questions

Problem 1

A friend who works in a big city owns two cars, one small and one large. Three-quarters of the time he drives the small car to work, and one-quarter of the time he drives the large car. If he takes the small car, he usually has little trouble parking, and so is at work on time with probability 0.9. If he takes the large car, he is at work on time with probability 0.6. Given that he was on time on a particular morning, what is the probability that he drove the small car?

Problem 2

We have a bag of three biased coins, *a*, *b*, and *c*, with probabilities of coming up heads of 20%, 60%, and 80%, respectively. One coin is drawn randomly from the bag (with equal likelihood of drawing each of the three coins), and then the coin is flipped three times to generate the outcomes X_1, X_2, and X_3.

a. Draw the Bayesian network corresponding to this setup and define the necessary conditional probability tables (CPTs).

b. Calculate which coin was most likely to have been drawn from the bag if the observed flips come out heads twice and tails once.

Problem 3

Consider the data points in the table below representing a set of seven patients with up to three different symptoms. We want to use the Naive Bayes assumption to diagnose whether a person has the flu based on the symptoms.

Sore Throat	Stomachache	Fever	Flu
No	No	No	No
No	No	Yes	Yes
No	Yes	No	No
Yes	No	No	No
Yes	No	Yes	Yes
Yes	Yes	No	Yes
Yes	Yes	Yes	No

a. Show the structure of the network and the conditional probability tables.

b. If a person has stomachache and fever, but no sore throat, what is the probability of him or her having the flu (according to your learned naive Bayes classifier)?

Problem 4

Consider training a perceptron for the same binary classification problem. Treat every "yes" as +1, treat every "no" as -1. Start with the weight vector equal to [0,0,0] and the bias equal to 0. Break ties in favor of the decision "no flu."

Sore Throat	Stomachache	Fever	Flu
No	No	No	No
No	No	Yes	Yes
No	Yes	No	No
Yes	No	No	No
Yes	No	Yes	Yes
Yes	Yes	No	Yes
Yes	Yes	Yes	No

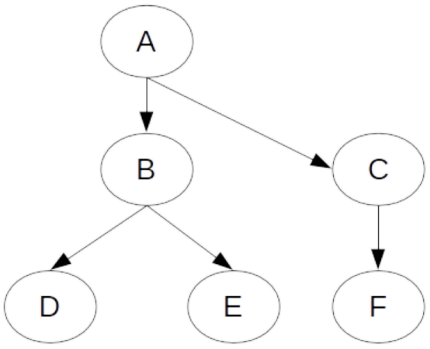
Assume that examples are presented one after the other, in the order shown in the table, and that the weight vector and bias are adapted after every example, with a learning rate of 1.0. Provide a table with seven rows; in the nth row, show the weight vector and bias after presentation of the nth training example.

Problem 5

Consider a Naïve Bayes classifier with 100 feature dimensions. The label Y is binary with $P(Y=0) = P(Y=1) = 0.5$. All features are binary, and have the same conditional probabilities: $P(X_i=1|Y=0) = a$ and $P(X_i=1|Y=1) = b$ for $i=1, ..., 100$. Given an item X with alternating feature values ($X_1=1, X_2=0, X_3=1, ..., X_{100}=0$), compute $P(Y=1|X)$.

Problem 6

Consider the Bayesian network with the following structure and conditional probability tables (all variables are binary):



$P(A) = 0.8$

$P(B \mid A) = 0.5,\ P(B \mid \neg A) = 0.2$

$P(C \mid A) = 0.8,\ P(C \mid \neg A) = 0.6$

$P(D \mid B) = 0.5,\ P(D \mid \neg B) = 0.6$

$P(E \mid B) = 0.8,\ P(E \mid \neg B) = 0.4$

$P(F \mid C) = 0.2,\ P(F \mid \neg C) = 0.01$

a. Is this a polytree?

b. Are D and E independent? Are they conditionally independent given B?

c. If you did not know the Bayesian network, how many numbers would you need to represent the full joint probability table?

d. If the variables were ternary instead of binary, how many values would you need to represent the full joint probability table and the conditional probability tables, respectively?

e. Write down the expression for the joint probability of all the variables in the network.

f. Find $P(A = 0, B = 1, C = 1, D = 0)$.

g. Find $P(B \mid A = 1, D = 0)$.

h. Using the sum-product algorithm, it is possible to compute the joint probability table P(C,E) using a series of sum-product operations in which the total computational complexity of each operation is never more than 8 (i.e., there are never more than 8 entries in the probability table after a product operation, and a sum operation never involves more than 8 additions). Write a series of sum-product operations that satisfies this constraint and results in the correct answer for the table P(C,E).

Problem 7

Two astronomers in different parts of the world make measurements M_1 and M_2 of the number of stars N in some small region of the sky, using their telescopes. Normally, there is a small probability *e* of error by up to one star in each direction (and if there is such an error, it is equally likely to be +1 or -1). Each telescope can also (with a much smaller probability *f*) be badly out of focus (events F_1 and F_2), in which case the scientist will undercount by three or more stars (or if N is less than 3, fail to detect any stars at all).

a. Draw a network for this problem and show the conditional probability tables.

b. Write out the conditional distributions for $P(M_1 \mid N)$ for the case where $N \in \{1,2,3\}$ and $M_1 \in \{0,1,2,3,4\}$. Each entry in the conditional distribution table should be expressed as a function of the parameters *e* and/or *f*.

Problem 8

In micro-blackjack, you repeatedly draw a card (with replacement) that is equally likely to be a 2, 3, or 4. You can either *Draw* or *Stop* if the total score of the cards you have drawn is less than 6. Otherwise, you must *Stop*. When you *Stop*, your utility is equal to your total score (up to 5), or zero if you get a total of 6 or higher. When you *Draw*, you receive no utility. There is no discount ($\gamma = 1$).

a. What are the states and the actions for this MDP?

b. What is the transition function and the reward function for this MDP?

c. Give the optimal policy for this MDP.

d. What is the smallest number of rounds of value iteration after which estimated utility of each state in this MDP will converge to its true utility (if value iteration will never converge exactly, state so).

Problem 9

In K-Means clustering, a dataset gets partitioned into "k" clusters where the algorithm tries to cluster similar data entries. Is this a type of supervised, unsupervised, semi-supervised, or active learning? Why?

Problem 10

We want to implement a classifier that takes two input values, where each value is either 0, 1 or 2, and outputs a 1 if at least one of the two inputs has value 2; otherwise it outputs a 0. Can this function be learned by a linear classifier? If so, construct a linear classifier that does it; if not, why not.

Problem 11

You are a Hollywood producer. You have a script in your hand and you want to make a movie. Before starting, however, you want to predict if the film you want to make will rake in huge profits, or utterly fail at the box office. You hire two critics A and B to read the script and rate it on a scale of 1 to 5 (assume only integer scores). Each critic reads it independently and announces their verdict. Of course, the critics might be biased and/or not perfect, therefore you may not be able to simply average their scores. Instead, you decide to use a perceptron to classify your data. The features and labels of your perceptron are defined as follows:

FEATURES: There are three features: a constant bias, and the two reviewer scores. Thus $f_0 = 1$ (a constant bias), f_1 = score given by reviewer A, and f_2 = score given by reviewer B.

LABELS: The label is Y=+1 if the movie returns a profit, Y=-1 otherwise.

a. Suppose that you are given the following five training examples, as shown in Table 1. The initial weights are $w_0 = -1, w_1 = 0, w_2 = 0$. Suppose you train using the examples in Table 1 with a learning rate of $\alpha = 1$. The perceptron is trained sequentially: each row in the table is classified, then the perceptron weights are either updated, or not updated, depending on the classification result. After this process has been performed for one row of the table, the updated weights are then used to classify the next row of the table, and so on. After learning has gone through the table once, what are the weights?

Movie Name	A	B	Profit
Pellet Power	1	1	No
Ghosts!	3	2	Yes
Pac is bac	4	5	No
Not a Pizza	3	4	Yes
Endless Maze	2	3	Yes

b. Instead of Table 1, suppose instead that you want to learn a perceptron that will always output $\hat{Y} = +1$ when the total of the two reviewer scores is more than 8, and $\hat{Y} = -1$ otherwise. Is this possible? If so, what are the weights w_0, w_1, and w_2 that will make this possible?

c. Instead of either Table 1 or part (b), suppose you want to learn a perceptron that will always output $\hat{Y} = +1$ when the two reviewers agree (when their scores are exactly the same), and will output $\hat{Y} = -1$ otherwise. Is this possible? If so, what are the weights w_0, w_1 and w_2 that will make this possible?

Problem 12

Explain the advantages of using convolutional neural networks for images (as opposed to fully connected networks).

Problem 13

When we apply the Q-learning algorithm to learn the state-action value function, one big problem in practice may be that the state space of the problem is continuous and high-dimensional. Discuss at least two possible methods to address this.

Problem 14

Suppose you have an MDP with N possible states, and with M possible actions. Specify, in terms of M and N, the space complexity TD learning and SARSA learning.

Problem 15

In an actor-critic deep learning paradigm, what does the actor compute? What does the critic compute? How are these two types of information combined to estimate the value of a state?

Problem 16

During the time that she lived in the White House, Malia Obama owned a dog named Bo. The front door of the White House had a dog door, so that Bo could come and go as he pleased. Every Sunday, a Secret Service agent made sure that the dog door was locked. Each weekday, on her way to school, Malia checked the dog door. If it was locked, she unlocked it with probability 1/3. If it was unlocked, she locked it with probability 1/4. On days when the dog door was unlocked, Bo escaped the house with probability 3/4, and went wandering about unleashed on the White House lawn. On days when the dog door was locked, the only way for Bo to escape was by begging to go out for a walk, and then breaking his leash; this happened with probability 1/10.

a. What was the probability that Bo was found unleashed on the White House lawn on any given Sunday?
b. Given that Bo escaped on a Monday, what was the probability that the dog door was open on that day?
c. A new janitor was hired at the White House. He sometimes locked the dog door, and sometimes unlocked it. He also sometimes let Bo escape by accident; at other times, he captured Bo and brought him back inside. As a result of these changes, Malia discovered that her old model of Bo's behavior was out of date, and had to be re-estimated. She observed the following data:

Day	Dog Door	Bo
Sunday	Locked	Outside
Monday	Unlocked	Inside
Tuesday	Unlocked	Outside
Wednesday	Locked	Inside
Thursday	Locked	Inside
Friday	Unlocked	Outside
Saturday	Locked	Inside

Define L_t=1 to be the event "dog door locked on day t," and define O_t=1 to be the event "Bo outside on day t." Estimate the conditional probability tables $P(L_{t+1}|L_t)$ and $P(O_t|L_t)$ from the table above.

Problem 17

The j'th softmax output, A_j, is computed from the j'th softmax input, Z_j, in two steps: (1) first, Z_j is exponentiated, (2) second, $\exp(Z_j)$ is divided by the summation of $\exp(Z_k)$ over all possible values of k. Each of these two steps is necessary so that the output, A_j, will satisfy all of the axioms of probability.

a. Which one of the three axioms of probability does A_j satisfy only because it is proportional to $\exp(Z_j)$? State the axiom, in words, or give an equation.
b. Which two of the three axioms of probability does A_j satisfy only because it is proportional to $1/\sum_k \exp(Z_k)$? State the axiom, in words, or give an equation.

Problem 18

The softmax function is given by

$$A_j = \frac{\exp Z_j}{\sum_k \exp Z_k}$$

Find $d \ln A_5 / dZ_2$. Express your answer in terms of only A_2; it should contain no other variables.

Problem 19

In an HMM-DNN hybrid, the DNN softmax outputs compute $P(Q_t|E_t)$, where Q_t is the HMM state variable, and E_t is the observed evidence at time t. Before using this probability in the HMM, we first need to divide it by $P(Q_t)$. Why?