CS440/ECE448 Section Q Fall 2017 Final Review

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

- **Probability**
  - Random variables, axioms of probability
  - Joint, marginal, conditional probability distributions
  - Product rule, chain rule
  - Independence and conditional independence
  - Bayes rule

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

- **Bayesian networks**
  - Structure and parameters
  - Conditional independence assumptions
  - Calculating joint and conditional probabilities
  - Inference (you should be able to do the math)
  - Hidden Markov models (definition, types of inference problems)
  - Complexity of inference (worst-case, special classes of networks for which efficient inference is possible)
  - Parameter learning

- **Machine learning**
  - Training, testing, generalization, overfitting, regularization
  - Different types of supervision (supervised, unsupervised, self-supervised, etc.)
  - Nearest neighbor classifiers
  - Support vector machines (incl. kernel support vector machines)
  - Perceptrons (incl. perceptron learning algorithm)
  - Neural networks (incl. training procedure)
  - Deep convolutional neural networks

- **Markov decision processes and reinforcement learning**
  - Markov assumption, transition model, policy
  - Bellman equation
  - Value iteration, policy iteration
  - Model-based vs. model-free reinforcement learning
  - Exploration vs. exploitation
  - TD Q-learning: Bellman equation for Q values, update rule
  - SARSA
  - Deep reinforcement learning: deep Q learning, actor-critic, imitation learning

- **Societal impacts of AI**
  - Privacy, bias and fairness, safety, effects on employment

Sample exam questions

1. Use the axioms of probability to prove that $P(\neg A) = 1 - P(A)$.

2. Let A and B be independent binary random variables with $P(A = 1) = 0.1$, $P(B = 1) = 0.4$. Let C denote the event that at least one of them is 1, and let D denote the event that exactly one of them is 1.

   a. What is $P(C = 1)$?
   b. What is $P(D = 1)$?
   c. What is $P(D = 1 \mid A = 1)$?
   d. Are A and D independent? Why?

3. Consider the following joint probability distribution:

   $P(A = \text{true}, B = \text{true}) = 0.12$
   $P(A = \text{true}, B = \text{false}) = 0.18$
   $P(A = \text{false}, B = \text{true}) = 0.28$
   $P(A = \text{false}, B = \text{false}) = 0.42$

   What are the marginal distributions of A and B? Are A and B independent and why?

4. 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?

5. 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.

6. 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

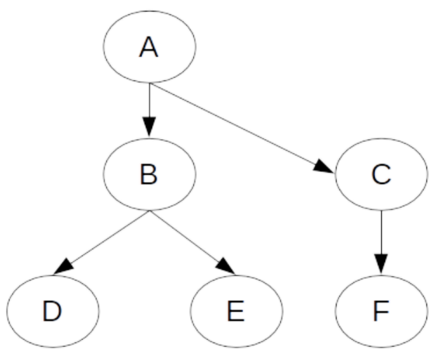
Yes	No	Yes	Yes
Yes	Yes	No	Yes
Yes	Yes	Yes	No

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)?

7. 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 $P(Y=1|X)$.

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



$P(A) = 0.8$

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

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

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

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

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

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

9. 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*.

10. 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 (number of consecutive games) 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).

11. 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?

12. 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 Perceptron? If so, construct a Perceptron that does it; if not, why not.

13. 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?

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

15. 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.