UNIVERSITY OF ENGINEERING AND TECHNOLOGY
Department of Computer Science and Engineering

Probability and Random Variables
Fall 2018

Problem Set 2
Due: September 17, 2018

Note: These problems have been liberally copied from older versions of MIT's 6.041/6.431, Harvard's Stat 110, Al Drake's book and Gian Carlo Rota's notes. It is difficult to acknowledge each problem individually. There is no claim to originality in these problems, and the debt to all these resources is gratefully acknowledged. This applies to all problems in this course.

1. Spam filter is designed by looking at commonly occurring phrases in spam. Suppose that 80% of email is spam. In 10% of the spam emails, the phrase "free money" is used, whereas this phrase is only used in 1% of non-spam emails. A new email has just arrived, which does mention "free money". What is the probability that it is spam?

2. Joe is a fool with probability of 0.6, a thief with probability 0.7, and neither with probability 0.25. Determine the
   (a) probability that he is a fool or a thief but not both.
   (b) conditional probability that he is a thief, given that he is not a fool.

3. Fischer and Spassky play a sudden death chess match. Each games ends up with either a win by Fischer, this happens with probability p, a win for Spassky, this happens with probability q, or a draw, this happens with probability $1 - p - q$. The match continues until one of the players wins a game (and the match).
   (a) What is the probability that Fischer will win?
   (b) Given that the match lasted no more than 5 games, what is the probability that Fischer won the first game?
   (c) Given that the match lasted no more than 5 games, what is the probability that Fischer won the match?
   (d) Given that Fischer won the match, what is the probability that he won at or before the 5th game?

4. Die A has five olive faces and one lavender face; die$B$ has three faces of each of these colors. A fair coiri is flipped once. If it falls heads, the game coritinues by throwing die A alone; if it falls tails, die B alone is used to continue the game. However awful their face colors may be, it is known that both dice are fair.
   (a) Determine the probability that the nth throw of the die results in olive.
   (b) Determine the probability that both the nth and $(n + 1)$st throw of the die results in olive.
   (c) If olive readings result from all the first n throws, determine the conditional probability-of an olive outcome on the $(n + 1)$st toss. Interpret your result for large values of n.

5. Oscar has lost his dog in either forest A (with a priori probability 0.4) or in forest B (with i a priori probability 0.6). If the dog is alive and not found by the Nth day of the search, it will die that evening with probability$N/(N + 2)$.

   If the dog is in A (either dead or alive) and Oscar spends a day searching for it in A, the conditional probability that he will find thedog that day is 0.25. Sinlilarly, if the dog is in B and Oscar spends a

day looking for it there, he will find the dog that day with probability 0.15. The dog cannot go from one forest to the other. Oscar can search only in the daytime, and he can travel from one forest to the other only at night. All parts of this problem are to be worked separately.

(a) In which forest should Oscar look to maximize the probability he finds his dog on the first day of the search?

(b) Given that Oscar looked in A on the first day but didn't find his dog, what is the probability that the dog is in A?

(c) If Oscar flips a fair coin to determine where to look on the first day and finds the dog on the first day, what is the probability that he looked in A?

(d) Oscar has decided to look in A for the first two days. What is the a priori probability that he will find a live dog for the first timeon the second day?

(e) Oscar has decided to look in A for the first two days. Given the fact that he was unsuccessful on the first day, determine the probability that he does not find a dead dog on the second day.

(f) Oscar finally found his dog on the fourth day of the search. He looked in A for the first 3 days and in B on the fourth day. What is the probability he found his dog alive?

(g) Oscar finally found his dog late on the fourth day of the search. The only other thing we know is that he looked in A for 2 days and and in B for 2 days. What is the probability he found his dog alive?

6. For each of the following statements, indicate whether it is True or False, and provide a brief explanation.

(a) $P(A|B) = P \implies P(B|A^c) = P(B)$

(b) If 5 out of 10 independent fair coin tosses resulted in tais, the events "first toss was tails" and "10th toss was tails" are independent.

(c) If 10 out of 10 independent fair coin tosses resulted in tails, the events "first toss was tails" and "10th toss was tails" are independent.

(d) B and C are some events in Ω. If events $A_1, \ldots, A_n$ form a partition of the sample space, then

$$P(B|C) = \sum_{i=1}^{n} P(A_i|C)P(B|A_i)$$

7. We want to design a spam filter for email. As described in Question 1, a major strategy is to find phrases that are much more likely to appear in a spam email than in a non- spam email. In that question, we only consider one such phrase: "free money". More realistically, suppose that we have created a list of 100 words or phrases that are much more likely to be used in spam than in non-spam. Let W_j be the event that an email contains the jth word or phrase on the list. Let

$$p = P(\text{spam}), pj = P(W_j|\text{spam}), rj = P(W_j|\text{not spam}),$$

where "spam" is shorthand for the event that the email is spam. Assume that $W_1, \ldots, W_{100}$ are conditionally independent given$M$, and also condition- ally independent given M^c. A method for classifying emails (or other objects) based on this kind of assumption is called a naive Bayes classifier. (Here "naive" refers to the fact that the conditional independence is a strong assumption, not to Bayes being naive. The assumption may or may not be realistic, but naive Bayes classifiers sometimes work well in practice even if the assumption is not realistic.) Under this assumption we know, for example, that

$$P(W_1, W_2, W_3^c, W_4^c, \ldots, W_c^{100}|\text{spam}) = p_1p_2(1-p_3)(1-p_4)...(1-p_{100})$$

Without the naive Bayes assumption, there would be vastly more statistical and com- putational difficulties since we would need to consider $2^{100} \approx 10^{30}$ events of the form $A_1 \cap A_2 \ldots \cap A_{100}$ with each A_j equal to either W_j or W_j^c . A new email has just arrived, and it includes the 23rd, 64th, and 65th words or phrases on the list (but not the other 97). So we want to compute

$$P\left(\text{spam}|W_1^c, \ldots, W_c^{22}, W_{23}, W_c^{24}, \ldots, W_{63}^c, W_{64}, W_{65}, W_{66}^c, \ldots, W_{100}^c\right).$$

Note that we need to condition on all the evidence, not just the fact that $W_{23} \cap W_{64} \cap W_{65}$ occurred. Find the conditional probability that the new email is spam (in terms of p and the p_j and r_j).