CS B551, Fall 2016, Probability practice problems (2)

1. This first problem is a continuation of the spam problem from practice problem set 1. Assuming conditional independence between words (as discussed in class), what is the probability that an e-mail is spam if it contains only the two words `offer` and `program`? We could use Bayes' Law directly to get the answer, but let's use Variable Elimination for practice.

   We can use Bayes' Law directly, but let's use Variable Elimination for practice. Let's introduce some random variables S which equals 0 or 1 depending on whether the email is spam, and variables P, O, G, R which are 0 or 1 depending on whether the words pharmacy, offer, program, and research (respectively) are in the email. The graphical model in this case consists of a root node S with four children corresponding to P, O, G, and R. We want to compute $P(S = 1|P = 0, O = 1, G = 1, R = 0)$, which we can do by computing $P(S = 1, P = 0, O = 1, G = 1, R = 0)$ and $P(P = 0, O = 1, G = 1, R = 0)$,

   $$P(S = 1|P = 0, O = 1, G = 1, R = 0) = \frac{P(S = 1, P = 0, O = 1, G = 1, R = 0)}{P(P = 0, O = 1, G = 1, R = 0)}.$$

   The numerator is easy to compute by just looking at the factorization of the Bayes Net:

   $$P(S = 1)P(P = 0|S = 1)P(O = 1|S = 1)P(G = 1|S = 1)P(R = 0|S = 1),$$

   which is $(0.5)(0.75)(0.35)(0.015)(0.975) \approx 0.00191953125$. To compute the denominator, we just need to eliminate S:

   $$\tau_1 = \sum_{s \in \{0,1\}} P(S = s)P(P = 0|S = s)P(O = 1|S = s)P(G = 1|S = s)P(R = 0|S = s),$$

   which is $(0.5)(0.99)(0.025)(0.1)(0.8)+(0.5)(0.75)(0.35)(0.015)(0.975) \approx 0.00290953125$. So the final conditional probability of spam is about 0.66.

2. Recall that if two events A and B are independent, their joint probability factors as a product of the individual probabilities, $P(A, B) = P(A)P(B)$. If they are not independent, we say that they are *positively correlated* if $P(A, B) > P(A)P(B)$, and we say they are *negatively correlated* if $P(A, B) < P(A)P(B)$.

   In estimating the probability that a message is spam, we have assumed that the word occurrences are independent events, allowing the probabilities to factor as a product over words,

   $$P(w_1, ..., w_n|S) = \prod_{i=1}^{n} P(w_i|S),$$
   $$P(w_1, ..., w_n|\overline{S}) = \prod_{i=1}^{n} p(w_i|\overline{S}).$$

(a) Suppose a message contains a group of words that is in reality positively correlated in spam messages, but not in non-spam messages (i.e., could be independent or negatively correlated). Does the independence assumption bias towards making it more or less likely that the message is considered spam?

Suppose there are two words, w_1 and w_2, that are positively correlated in the spam messages. From Baye's law we have,

$$\begin{aligned} P(S|w_1, w_2) &= \frac{P(w_1, w_2|S)P(S)}{P(w_1, w_2)} \\ &= \frac{P(w_1, w_2|S)P(S)}{P(w_1, w_2|S)P(S) + P(w_1, w_2|\bar{S})P(\bar{S})} \end{aligned} \tag{1}$$

If we assume that the words are independent, we would approximate this as,

$$P(S|w_1, w_2) \approx \frac{P(w_1|S)P(w_2|S)P(S)}{P(w_1|S)P(w_2|S)P(S) + P(w_1|\bar{S})P(w_2|\bar{S})P(\bar{S})}. \tag{2}$$

Since the words are positively correlated in the spam messages but not in the non-spam messages, we know that $P(w_1, w_2|S) > P(w_1|S)P(w_2|S)$ and $P(w_1, w_2|\bar{S}) = P(w_1|\bar{S})P(w_2|\bar{S})$. Thus the numerator of equation (**??**) is smaller than the numberator of the exact probability in equation (**??**). The denominator is also smaller but by a lesser factor, so the approximation makes the estimated probability of spam *lower* than the actual probability.

(b) Suppose a message contains a group of words that is in reality negatively correlated in spam messages, but not in non-spam messages (i.e., could be independent or positively correlated). Does the independence assumption bias towards making it more or less likely that the message is considered spam?

By the same logic as (a), the independence assumption makes the estimated probability of spam *higher* than the actual probability.

3. What is the smallest group of people such that the probability of at least two people having the same birthday is at least 0.8? Assume there are 365 days per year and that birthdays are uniformly distributed.

As we saw in class, in a group of N people the probability of at least one common birthday is,

$$P(N) = 1 - \frac{365 \times 364 \times ... \times (365 - N + 1)}{365^N}.$$

We can use a "guess and check" strategy to find the answer, by checking $P(N)$ for some reasonable values of N: $P(25) \approx 0.56$, $P(30) \approx 0.706$, $P(34) \approx 0.795$, and $P(35) \approx 0.814$. So $N = 35$ is the smallest group for which the probability is at least 0.8.

Many calculators can't compute a number as large as 365^{35} (it's a number with almost 100 decimal digits — that's way more than the number of atoms in the universe). You can still do the computation, but you have to be a bit clever: instead of computing the numerator and then dividing by the denominator, instead alternate between the two, so that the intermediate results are never very large. For example, to compute the probability for $N = 35$, you could do the following sequence of operations on your calculator:

`365 / 365 * 364 / 365 * 363 / 365 * ...  * 336 / 365.`

4. A die is loaded so that the probability of a face coming up is proportional to the number on that face. Let X be a random variable specifying the number that comes up when the die is rolled. What is the expected value of X, $E[X]$?

   Let $P(i)$ denote the probability of rolling the number i, for $i \in \{1, 2, 3, 4, 5, 6\}$. The sum of the probabilities has to equal one, and we know that $P(i)$ is proportional to i. Let x be the unknown proportionality factor. Then we have,

$$\begin{aligned} 1 &= P(1) + P(2) + P(3) + P(4) + P(5) + P(6) \\ &= x + 2x + 3x + 4x + 5x + 6x \\ &= 21x, \end{aligned}$$

   so $x = \frac{1}{21}$ and $P(i) = \frac{1}{21}i$. Then we can compute the expected value,

$$\begin{aligned} E[X] &= \frac{1}{21}(1) + \frac{2}{21}(2) + \frac{3}{21}(3) + \frac{4}{21}(4) + \frac{5}{21}(5) + \frac{6}{21}(6) \\ &= \frac{13}{3} \end{aligned}$$

5. A student has three ways of commuting to her class:

— by bus, for which the fare is \$1.50. The bus runs late 25% of the time.

— by car, in which case she must pay \$10.00 for parking. With probability 10% she encounters a traffic jam along the route and is late to class.

— by bicycle, which is free and the probability of being late is only 5%.

Assume that she takes the bus 50% of the time, drives 25% of the time and rides her bike 25% of the time.

(a) One day the student is late to class. What is the probability that she took the bus that day?

Let B, C, and Y denote the events that she arrived on campus via bus, car, and bicycle, respectively. Let L denote that she is late and $\bar{L}$ denote that she is on-time. We want compute the probability that she took the bus given that she is late, $P(B|L)$. Using Bayes' Law,

$$\begin{aligned} P(B|L) &= \frac{P(L|B)P(B)}{P(L)} \\ &= \frac{P(L|B)P(B)}{P(L|B)P(B) + P(L|C)P(C) + P(L|Y)P(Y)} \\ &= \frac{(0.25)(0.5)}{(0.25)(0.5) + (0.1)(0.25) + (0.05)(0.25)} \\ &\approx 0.7692 \end{aligned}$$

(b) Another day the student is on time. What is the expected value of the cost of her commute that day?

Let X denote the cost of her commute. Using the definition of expected value,

$$E[X] = P(B|\bar{L})(1.5) + P(C|\bar{L})(10) + P(Y|\bar{L})(0).$$

We can use Bayes' law to compute $P(B|\bar{L})$ and $P(C|\bar{L})$,

$$\begin{aligned} P(B|\bar{L}) &= \frac{P(\bar{L}|B)P(B)}{P(\bar{L}L)} \\ &= \frac{P(\bar{L}|B)P(B)}{P(\bar{L}|B)P(B) + P(\bar{L}|C)P(C) + P(\bar{L}|Y)P(Y)} \\ &= \frac{(0.75)(0.5)}{(0.75)(0.5) + (0.9)(0.25) + (0.95)(0.25)} \\ &\approx 0.4478. \end{aligned}$$

$$\begin{aligned} P(C|\bar{L}) &= \frac{P(\bar{L}|C)P(C)}{P(\bar{L}L)} \\ &= \frac{P(\bar{L}|C)P(C)}{P(\bar{L}|B)P(B)+P(\bar{L}|C)P(C)+P(\bar{L}|Y)P(Y)} \\ &= \frac{(0.9)(0.25)}{(0.75)(0.5)+(0.9)(0.25)+(0.95)(0.25)} \\ &\approx 0.2687. \end{aligned}$$

So using the above formula for expected value, we have,

$$E[X] = P(B|\bar{L})(1.5)+P(C|\bar{L})(10)+P(Y|\bar{L})(0) \approx 0.4478(1.5)+0.268(10) = \$3.35.$$

6. A small airline provides daily service from Bloomington to Toronto using a 10-seat aircraft. On any given day, the probability that a given passenger will not show up for the flight is 10% (independent of the other passengers).

   (a) If the airline sells 10 tickets for the flight, what is the probability that all ten passengers show up?

   $$(0.9)^{10} \approx 0.3487$$

   (b) If the airline sells 12 tickets for the flight, what is the expected number of passengers that show up? (Hint: If X and Y are random variables, then $E[X+Y] = E[X]+E[Y]$).

   Let X denote the number of passengers that show up. Note that $X = X_1 + X_2 + ... + X_{12}$, where X_i is 1 if the i-th passenger shows up and 0 otherwise. The expected value of X_i for any i is,

   $$E[X_i] = (1)(0.9) + (0.0)(0.1) = 0.9,$$

   so $E[X]$ is equal to 10.8,

   $$E[X] = E[X_1] + E[X_2] + ... + E[X_{12}] = 12 * 0.9 = 10.8.$$

(c) Suppose that the price of a ticket is \$100. The airline gets to keep the fare of a passenger who doesn't show up, but they have to refund the ticket price and pay an extra \$200 fine for every passenger who is denied boarding because the plane is full. What is the expected net income (ticket sales minus fines and refunds) from the flight if the airline sells 12 tickets?

Let X be a random variable denoting the number of passengers who show up for the flight, and let Y be the net income. The expected value of Y is,

$$\begin{aligned} E[Y] &= P(X=0)(1200) + P(X=1)(1200) + P(X=2)(1200) + ... + P(X=10)(1200) \\ &\quad + P(X=11)(1200-300) + P(X=12)(1200-600) \\ &= 1200(1 - P(X=11) - P(X=12)) + 900P(X=11) + 600P(X=12) \\ &= 1200(1 - P(X=11) - P(X=12)) + 900P(X=11) + 600P(X=12) \end{aligned}$$

Note that X is binomially distributed, with p=0.9, so $P(X=11) = \binom{12}{11}(0.9)^1 1(0.1)^1 \approx 0.3765$ and $P(X=12) = \binom{12}{12}(0.9)^1 2(0.1)^0 \approx 0.2824$. Substituting these values into the equation above, we have $E[Y] = \$917.61$.

(d) Assuming the same conditions as in (c), how many tickets should the airline sell in order to maximize their expected net income?

If it sells 10 tickets, the expected income is \$1000, and we found in (c) that for 12 tickets the expected income is \$917.61. For 11 tickets, the expected income is,

$$\begin{aligned} E[Y] &= 1100(1 - P(X=11)) + 900P(X=11) \\ &= 1100(1 - P(X=11)) + 900P(X=11) \\ &\approx \$1037.24, \end{aligned}$$

so the optimal number of tickets is 11.

7. DNA molecules are made up of pairs of complementary nucleic acids. If any one of the two nucleotides is knocked out, the base can still be repaired. However, if both elements of the pair are destroyed, then the base is lost. A particular gene of interest contains n base pairs. When the segment is bombarded by radiation, exactly m nucleotides are destroyed at random. Find the expected number of lost base pairs.

Let X be the number of base pairs destroyed, and let X_i be 1 if the i-th pair is destroyed and 0 otherwise. Then we have,

$$X = X_1 + X_2 + ... + X_n,$$

and using linearity of expectations we can write,

$$E[X] = E[X_1] + E[X_2] + ... + E[X_n].$$

Using the definition of expectaction, we know that the expected value of X_i is,

$$\begin{aligned} E[X_i] &= P(i\text{-th base pair destroyed})(1) + P(i\text{-th base not destroyed pair})(0) \\ &= P(i\text{-th base pair destroyed}) \end{aligned}$$

All that remains is to determine the probability that a given base pair is destroyed. A pair is destroyed if radiation strikes both bases in the pair. There are m bases destroyed by radiation, and $2n$ bases total, so there are $\binom{2n}{m}$ possible combinations of destroyed. Of these outcomes, exactly $\binom{2n-2}{m-2}$ of them involve both bases in the i-th pair being destroyed, for any given value of i. Thus we have,

$$\begin{aligned} P(i\text{-th base pair destroyed}) &= \frac{\binom{2n-2}{m-2}}{\binom{2n}{m}} \\ &= \frac{(2n-2)!(2n-m)!m!}{(m-2)!(2n-2-m+2)!(2n)!} \\ &= \frac{m(m-1)}{2n(2n-1)} \end{aligned}$$

Now since $E[X] = E[X_1] + E[X_2] + ... + E[X_n] = (n)E[X_i]$, we have that the expected number of destroyed base pairs is,

$$\begin{aligned} E[X] &= (n)\frac{m(m-1)}{2n(2n-1)} \\ &= \frac{m(m-1)}{2(2n-1)}. \end{aligned}$$