COSI 101A PS2
Zexi Josh Jin
Feb 14, 2017

13.13 Consider two medical tests, A and B, for a virus. Test A is 95% effective at recognizing the virus when it is present, but has a 10% false positive rate (indicating that the virus is present, when it is not). Test B is 90% effective at recognizing the virus, but has a 5% false positive rate. The two tests use independent methods of identifying the virus. The virus is carried by 1% of all people. Say that a person is tested for the virus using only one of the tests, and that test comes back positive for carrying the virus. Which test returning positive is more indicative of someone really carrying the virus? Justify your answer mathematically.

Answers:

Let A be the event that test A returns positive, ¬A be the events that test A returns negative.
Let B be the event that test B returns positive, ¬B be the events that test B returns negative.
Let V be the event that virus is presented on a patient, and ¬V be the events that virus is not presented on a patient.
From the description, we have:

$P(A|V) = 0.95, P(A|\neg V) = 0.10, P(B|V) = 0.90, P(B|\neg V) = 0.05$

Also, we have $P(V) = 0.01, P(\neg V) = 1 - P(V) = 0.99.$

We want to find $P(V|A)$ and $P(V|B)$

For $P(V|A)$:

$$P(V|A) = \frac{P(V \wedge A)}{P(A)} = \frac{P(A|V)P(V)}{P(A|V)P(V)+P(A|\neg V)P(\neg V)} = \frac{(0.95)(0.01)}{(0.95)(0.01)+(0.10)(0.99)} = 0.0876$$

For $P(V|B)$:

$$P(V|B) = \frac{P(V \wedge B)}{P(B)} = \frac{P(B|V)P(V)}{P(B|V)P(V)+P(B|\neg V)P(\neg V)} = \frac{(0.90)(0.01)}{(0.90)(0.01)+(0.05)(0.99)} = 0.154$$

Test B returning positive is more indicative.

13.22 Text categorization is the task of assigning a given document to one of a fixed set of categories on the basis of the text it contains. Naive Bayes models are often used for this task. In these models, the query variable is the document category, and the "effect" variables are the presence or absence of each word in the language; the assumption is that words occur independently in documents, with frequencies determined by the document category.

a. Explain precisely how such a model can be constructed, given as "training data" a set of documents that have been assigned to categories.

b. Explain precisely how to categorize a new document.

c. Is the conditional independence assumption reasonable? Discuss.

Answers:

Part a:

First, using the training set, we first determine which document belongs to which category. For example, if there is a category a, we say this document belongs to category a.

From that, we can construct the conditional probability of P(word | category), which is the frequency of the appearance of word given a category. For example, P(word = b | category = a) is the frequency of word b to appear in category a. This is obtained from the training set. We do this for all words appeared in the training set.

Also, for each document, we can basically express it as document = $\prod_i word_i$, where $\prod_i word_i$ means all words in a document.

Part b:

For each new document, we can basically go through every word in the document and assign the word to a previously assigned category.

That is, we want P(category=a | document) = P(category=a | $\prod_i word_i$). Then, by applying the Naïve Bayes theorem and Bayes' Rule using what we get from the training data in part a, we calculate the probability of which category the new documents are mostly likely to belongs to:

$$\mathrm{P}(\text{category} = \text{a} \mid \text{document})$$
$$= P(\text{category} = \text{a} \mid \prod_i word_i)$$

$$= \frac{P(\prod_i word_i | \text{ category=a}) P(category=a)}{P(\prod_i word_i)}$$

By Naïve Bayes theorem, we get:

$$P(\text{category} = \text{a} \mid \text{document}) = \frac{\prod_i P(word_i \mid \text{category=a}) P(category=a)}{P(\prod_i word_i)}$$

Thus, by solving the equation above, we get the probability given a document if it belongs to category a. After that, we decide if this probability is large enough to classify the new documents.

Part c:

No, it is not reasonable, because different words are dependent. On a smaller scale, there are word combinations. For example, “spider web”, “web like structure”, and “web browser” are word combinations all containing “web” but all containing different meanings. On a bigger scale, basically all sentences have the structure “nouns + verbs + something else”. That means there are correlations in all words in every sentence.

COSI 101A PS2
Zexi Josh Jin
Feb 14, 2017

14.7 The Markov blanket of a variable is defined on page 517. Prove that a variable is independent of all other variables in the network, given its Markov blanket and derive Equation (14.12) (page 538).

Answer:

We start with the probability of one x_i given all other unknowns. Assuming there is a total of n unknowns and the unknown variable we are trying to show independence is x_i. From the Bayes Rules, we can derive a basic equation. In another word:

$$P(x_i|x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = \frac{P(x_1,\dots,x_{i-1},x_i,x_{i+1},\dots,x_n)}{P(x_1,\dots,x_{i-1},x_{i+1},\dots,x_n)} \dots \#1$$

By Bayes Rules, we know that $P(x) = \sum_{y_i} P(x, y_i)$. That is, if we add all the possibilities of y_i in P(x,y) for a determined x and all y's, we can get the possibility for P(x). Thus, we can rewrite #1 as:

$$\frac{P(x_1,\dots,x_{i-1},x_i,x_{i+1},\dots,x_n)}{P(x_1,\dots,x_{i-1},x_{i+1},\dots,x_n)} = \frac{P(x_1,\dots,x_{i-1},x_i,x_{i+1},\dots,x_n)}{\sum_{x_i} P(x_1,\dots,x_{i-1},x_i,x_{i+1},\dots,x_n)} \dots \#2$$

In the lecture note, we know that $P(x_1, \dots, x_n) = \prod_i P(x_i|parent(x_i))$. Apply this rule to #2:

$$\frac{P(x_1,\dots,x_{i-1},x_i,x_{i+1},\dots,x_n)}{\sum_{x_i} P(x_1,\dots,x_{i-1},x_i,x_{i+1},\dots,x_n)} = \frac{\prod_j P(x_j|parent(x_j))}{\sum_{x_i} \prod_j P(x_j|parent(x_j))} \dots \#3$$

To extract the uncommon part in #3 and delete the repeated part in the denominator and nominator, we focus on x_i's parents, x_i's children, and x_i's children's parents. Observe that the nodes we need to pay attentions to are the Markov's blanket.

In detail, we focus on x_i's parents, x_i's children, and x_i's children's parents because the only differences in the denominator and nominator in #3 is x_i. by $P(x_1, \dots, x_n) = \prod_i P(x_i|parent(x_i))$, we get two kinds of probability that can be related to x_i.

Case 1: $P(x_i|parent(x_i))$.

Case 2: $P(y_k|parent(y_k))$ $given$ $(\forall y_k \ | y_k \ is \ child \ of \ x_i)$.

From here, we reduce #3 by canceling out the terms that is not one of these two cases:

$$\frac{\prod_j P(x_j|parent(x_j))}{\sum_{x_i} \prod_j P(x_j|parent(x_j))} = \frac{P(x_i|parent(x_i)) \prod_{y_k \ given \ (\forall y_k \ | y_k \ is \ child \ of \ x_i)} P(y_k|parent(y_k))}{\sum_{x_i} P(x_i|parent(x_i)) \prod_{y_k \ given \ (\forall y_k \ | y_k \ is \ child \ of \ x_i)} P(y_k|parent(y_k))} \dots \#4$$

We find in the book the following equation:

$$P(x'_i \mid mb(X_i)) = \alpha \, P(x'_i \mid parents(X_i)) \times \prod_{Y_j \in Children(X_i)} P(y_j \mid parents(Y_j)) \ . \ (14.12)$$

In the question, we have x's Markov blanket given, and we can write is as a constant. Also, since in the nominator, we are counting for all possible values of x_i. we have sum of $x_i = 1$. Then, the value of the denominator is 1 times the value of the Markov Blanket, which is a constant of Markov blanket given. Let α equals to 1 over the constant, then we can rewrite #4 as:

#4 = $\alpha P(x_i|parent(x_i)) \prod_{y_k \, given \, (\forall y_k \, | y_k \, is \, child \, of \, x_i)} P(y_k|parent(y_k))$ … #5

Despite the slightly different wording between #5 and (14.12), clearly #5 and (14.12) are equivalent.

In another word, by bringing back to the most basic formats, we get:

$P(x_i|x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i|mb(x_i))$, which proves desired result.

14.12 Two astronomers in different parts of the world make measurements M1 and M2 of the number of stars N in some small region of the sky, using their telescopes. Normally, there is a small possibility e of error by up to one star in each direction. Each telescope can also (with a much smaller probability f) be badly out of focus (events F1 and F2), 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). Consider the three networks shown in Figure 14.22.

a. Which of these Bayesian networks are correct (but not necessarily efficient) representations of the preceding information?

b. Which is the best network? Explain.

c. Write out a conditional distribution for P(M1 | N), for the case where N ∈ {1, 2, 3} and M1 ∈ {0, 1, 2, 3, 4}. Each entry in the conditional distribution should be expressed as a function of the parameters e and/or f.

d. Suppose M1 = 1 and M2 = 3. What are the possible numbers of stars if you assume no prior constraint on the values of N?

e. What is the most likely number of stars, given these observations? Explain how to compute this, or if it is not possible to compute, explain what additional information is needed and how it would affect the result.

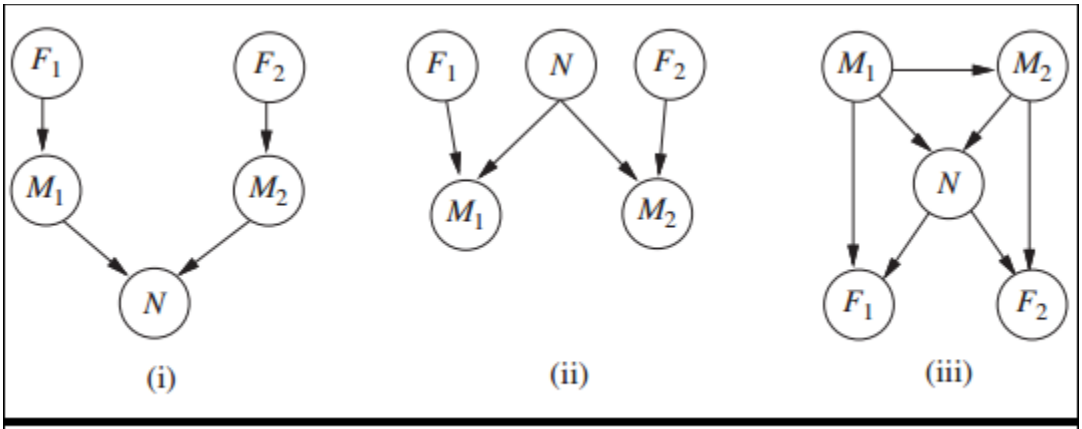

Figure 14.22 Three possible networks for the telescope problem.

Answers:

Part a:

(ii) and (iii) are correct if we ignore efficiency.

For (i), M1 and M2 cannot be independent of each other. It should be at least somewhat related.

For (ii), the network makes sense. F1, F2, and N can be independent of each other. M1 and M2 are not independent of each unless given N.

For (iii), it makes sense but it is not very efficient. There are simply too many connections.

Part b:

Like I mentioned in part a, it should be (ii). (iii) is simply too complicated. With M1 somehow linked to M2, what is exactly the connection is not very clear. (ii) is much easier to understand.

Part c:

Conditional Distribution Table		M1				
		0	1	2	3	4
N	1	$\frac{e}{2}+f-\left(\frac{e}{2}\right)f$	$(1-e)(1-f)$	$\frac{e(1-f)}{2}$	0	0
	2	f	$\frac{e(1-f)}{2}$	$(1-e)(1-f)$	$\frac{e(1-f)}{2}$	0
	3	f	0	$\frac{e(1-f)}{2}$	$(1-e)(1-f)$	$\frac{e(1-f)}{2}$

Part d:

We construct a conditional distribution table for all possible Ns.

	N					
	1	2	3	4	5	6
M1=1	$(1-e)(1-f)$	$\frac{e(1-f)}{2}$	0	f	f	f
M2=3	0	$\frac{e(1-f)}{2}$	$(1-e)(1-f)$	$\frac{e(1-f)}{2}$	0	f

N cannot be 1, 3, or 5.

The probability for N = 2 is $(\frac{e(1-f)}{2})^2$

The probability for N = 4 is $\frac{ef(1-f)}{2}$

The probability for any N greater or equal to 6 is f^2.

Part e:

The most likely number of stars is 2. As I mentioned in part d the probability for N = 2 is $(\frac{e(1-f)}{2})^2$, where $\frac{e(1-f)}{2}$ should be larger than $\frac{ef(1-f)}{2}$ and f^2, given that f is a much smaller possibility than e.

14.14 Consider the Bayes net shown in Figure 14.23.

a. Which of the following are asserted by the network structure?

(i) P(B,I,M) = P(B)P(I)P(M).

(ii) P(J | G) = P(J | G, I).

(iii) P(M | G, B, I) = P(M | G, B, I, J).

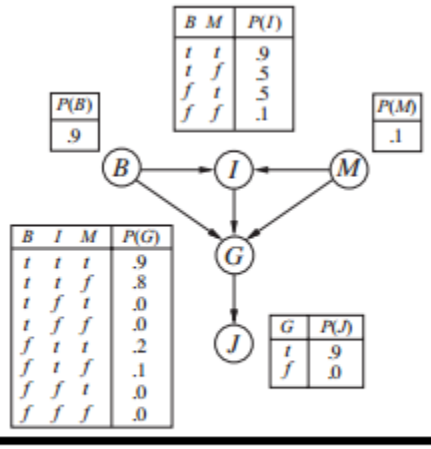

Figure 14.23 A simple Bayes net with Boolean variables $B = BrokeElectionLaw$, $I = Indicted$, $M = PoliticallyMotivatedProsecutor$, $G = FoundGuilty$, $J = Jailed$.

b. Calculate the value of P(b, i,¬m, g, j).

c. Calculate the probability that someone goes to jail given that they broke the law, have been indicted, and face a politically motivated prosecutor.

d. A context-specific independence (see page 542) allows a variable to be independent of some of its parents given certain values of others. In addition to the usual conditional independences given by the graph structure, what context-specific independences exist in the Bayes net in Figure 14.23?

e. Suppose we want to add the variable P = PresidentialPardon to the network; draw the new network and briefly explain any links you add.

Answers:

Part a:

(i) It is not true, because event I is dependent to B & M.

(ii) P(J | G) = P(J | G, I) is true because J is independent of I, which means P(J | G, I) can be replaced by P(J | G).

(iii) P(M | G, B, I) = P(M | G, B, I, J) is true because M is independent of J.

Part b:

$P(b, i, \neg m, g, j) = P(j|g)P(b, i, \neg m, g)$

$= P(j|g)P(g|b, i, \neg m)P(b, i, \neg m)$

$= P(j|g)P(g|b, i, \neg m)P(i|b, \neg m)P(b, \neg m)$

$= P(j|g)P(g|b, i, \neg m)P(i|b, \neg m)P(b)P(\neg m)$

$= 0.9 \times 0.8 \times 0.5 \times (1 - 0.1) \times 0.9$

$= 0.2916$

Part c:

First, we rewrite the expression "someone goes to jail given that they broke the law, have been indicted, and face a politically motivated prosecutor" as probability:

$P(j|b, i, m) = P(j|g)P(g|b, i, m) + P(j|\neg g)P(\neg g|b, i, m)$

$= 0.9 \times 0.9 + 0 \times 0.1 = 0.81$

Part d:

The context-specific independence variable is G. Given I is false, the variable G is independent of B and M. Since despite what B and M are, the probability that the value for G is true is always going to be 0.

Part e:

I would put P in between G and J, assuming Presidential Pardon is not given when someone has been committed guilty. There should be a relationship as G is the parent of P. Since if pardoned or not directly related to if someone should be send to jail, J should be children of P.

COSI 101A PS2
Zexi Josh Jin
Feb 14, 2017