CPS331 Lecture: Bayesian Learning

last revised November 7, 2018

Objectives:

1. To introduce naive Naieve Bayesian learning
2. To introduce Bayesian networks

Materials:

1. Projectable of Bayesian network for Alarm problem
2. Projectable of the above with CPTs (from Russell & Norvig)
3. Projectable of computation of P(Burglary | JohnCalls and MaryCalls)
4. Projectable of data counts for Wisconsin high school senior college plans survey
5. Projectable of two Bayesian networks inferred from data
6. Projectable of Bayesian network with hidden variable

I. Naieve Bayesian Learning

A. The name "Bayesian" is given to an interpretation of probability theory. "Bayesian learning" is the name given to an approach to machine learning that seeks to find Bayesian probabilities based on patterns in data.

B. Consider the following problem:

1. We have available a medical database with 10,000 descriptions of individual patients. Each patient is described in terms of the presence or absence of 20 symptoms (i.e. 20 boolean values) and the patient's diagnosis or an indication that the patient is not sick.

2. Suppose a patient comes along who has a fever, coughing, and chills. What is the likelihood that these symptoms are due to having the flu - i.e. what is

   P(flu | fever ^ coughing ^ chills)

3. The database is of little help to us directly. If there are 20 different boolean values in each row in the database representing the presence or absence of different symptoms, then there are 2^{20} (over 1 million) possible combinations. In a database of 10,000 entries, it is unlikely that this precise combination of symptoms appears - and even if it does, the number of times it appears is likely very small and we can't get a reliable estimate of the probability we want but just counting cases.

4. But we can make use of Bayes' theorem to get information that we can extract from the database:

   $$P(flu \mid fever \wedge cough \wedge chills) = \frac{P(flu) * P(fever \wedge cough \wedge chills \mid flu)}{P(fever \wedge cough \wedge chills)}$$

   a) We can estimate P(flu) by counting rows in the database where the diagnosis is flu, and dividing by the total number of rows.

   b) In similar fashion, we can estimate P(fever ^ cough ^ chills) by counting the rows in the database where these three appear together (perhaps with other symptoms and various diagnoses) and dividing by the total number of rows.

   c) But what about P(fever ^ cough ^ chills | flu)? This would seem to pose the same challenge as our original problem, since we are looking for rows where these three appear with flu - possibly an empty or very small set.

5. P(fever ^ cough ^ chills | flu) is called a joint probability. Ordinarily, we cannot compute a joint probability by simply knowing the individual probabilities if the individual items are not independent - i.e. P(A ^ B) is ordinarily not equal to P(A) * P(B)

Example: In Fall, 2018, 17 students are taking Discrete Math (15 CS majors) and 11 are taking Software Systems (10 CS majors). Five are taking both (all CS majors).

The probability that a given Gordon student is taking Discrete Math is
P(D) = 17/1600 = .01,
The probability that a given Gordon student is taking Software Systems is
P(S) = 11/1600 = .007.

But the probability that a given Gordon student is taking both is **not**
.01 * .007 = .00007
Rather, it is
P(D^S) = 5/1600 = .003

a) In the diagnostic case we have been considering, it is not true that the prior probabilities of symptoms (e.g. cough and fever) are statistically independent. The likelihood of someone who has a fever also having a cough is more than the likelihood of someone just having a cough.

b) However, this dependency virtually goes away in the case where the person has a disease like the flu - i.e. a person who has the flu and has a fever is not more likely to have a cough than is the person who has the flu without a fever.

c) The approach known as Naieve Bayes assumes that statistical dependencies go away for posterior probabilities conditioned on the same variable - i.e. it calculates P(fever ^ cough ^ chills | flu) as P(fever | flu) * P(cough | flu) * P(chills | flu) - quantities that can be found simply by counting rows in the database.

C. Naieve Bayes does not give a totally correct figure, but rather one that generally is close enough to do something like aiding diagnosis. As one statistician put it "All models are wrong, but some are useful".

Continuing our example: The probability that a CS major is taking Discrete is
P(D | CS) = 15 / 51 = 0.29
and that a CS major is taking Software Systems is
P(S | CS) = 10/51 = 0.2
The probability that a CS major is taking both is
P(D ^ S | CS) = 5/51 = 0.1
Naieve Bayes would estimate this as 0.29 * 0.2 = .06 - not exactly correct, but much closer than what we would get using the probabilities without the dependency on being a CS major

D. Naieve Bayes is widely used in machine learning problems - e.g. it is heavily used by Google, is used by many spam filters, etc.

II.Bayesian Networks (also known as Belief Networks)

A. To understand what a Bayesian network is, we will use an example that is **<u>not</u>** machine learning. Then we'll look at using machine learning to learn Bayesian networks - but first we need to know what a Bayesian network is.

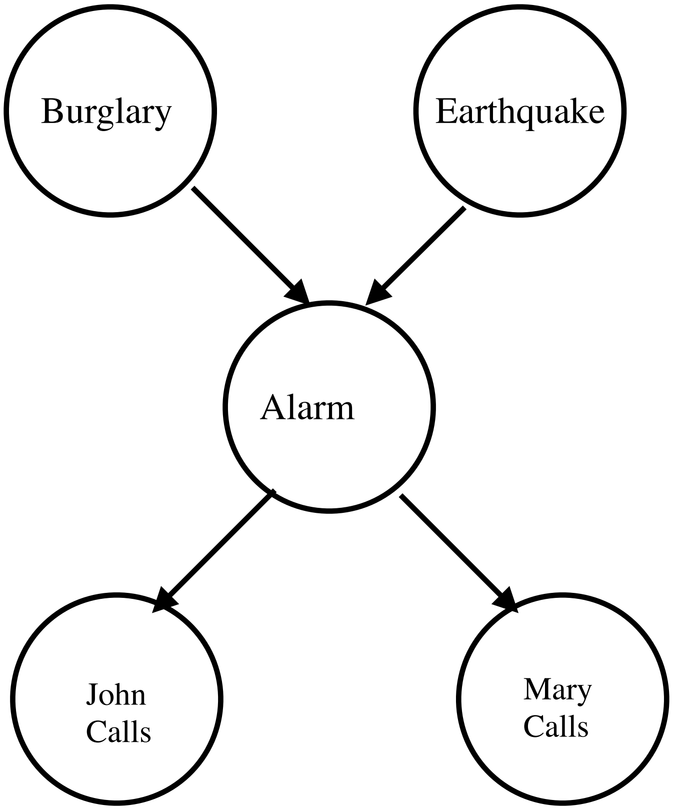
B. Consider the following problem (formulated by Judea Pearl, who coined the term Bayesian network):

1. He is concerned about Burglaries.

2. He installs an Alarm in his house that is supposed to sound if a burglary occurs. It is sometimes also set off by an a minor earthquake. (Pearl lived in the Los Angeles area)

3. Because the individual does not work close to home, he asks two neighbors - John and Mary - to call him if they hear the alarm go off.

   a) This can be described by the following structure of causes, which is called a Bayesian network. (The directed links denote causality)



PROJECT

b) For each of the nodes, we can construct a conditional probability table that shows the conditional probability for a given node having the value true based on various combinations of the values of the parent nodes.

(1)Neither Burglary nor Alarm has a parent, so the table for each contains just one a-prior probability - derivable from police data for one and historical seismic data for the other - say

P(B) = .001
P(E) = .002

(Note how we will abbreviate to B and E to save writing)

(2)Alarm has two parents, so we need conditional probabilities for of the four possible combinations of the values of the parents, based on how well the alarm functions;

B	E	P(A)
T	T	0.95 [estimated that alarm catches 95% of burglaries]
T	F	0.95 [ditto]
F	T	0.29 [false alarm cause by earthquake]
F	F	0.001 [totally false alarm]

(3) John Calls depends only on Alarm. He almost always hears the alarm when it goes off, but sometimes confuses the telephone ringing with the alarm and calls then too. His behavior is described by the following CPT:

A	P(J)
T	.90
F	.05 [false alarm]

(4)Mary Calls also depends only on Alarm. She likes loud music and sometimes misses the alarm. This is the CPT for Mary:

A	P(M)
T	.70
F	.01 [relatively rare false alarm]

(5)This can all be summarized by the following diagram

PROJECT Bayesian network with conditional probability tables

4. Now suppose both John and Mary call. What is the probability that Burglary has occurred?

   a) We want to calculate P(Burglary | JohnCalls ^ MaryCalls). If we used standard probabilistic inference, we would have to make use of 32 joint probabilities, many of which would require considerable effort to find :

      ¬ Burglary ^ ¬ Earthquake ^ ¬ Alarm ^ ¬ John ^ ¬ Mary
      ¬ Burglary ^ ¬ Earthquake ^ ¬ Alarm ^ ¬ John ^ Mary
      ¬ Burglary ^ ¬ Earthquake ^ ¬ Alarm ^ John ^ ¬ Mary
      ....

   b) But the Bayesian network allows us to compute the result more simply, by taking advantage of parent-child (causality) relationships as noted above.

      Thus, instead of needing 32 probabilities, some of which are hard to figure out, the causal relationships reduce the number we need to just the 10 given above, all fairly easy to figure out. For networks with more nodes, the impact of using a Bayesian network will be even greater because the number of joint probabilities grows exponentially with the number of nodes, while the number of causal parent-child relationships tends to grow linearly.

      (1)We wish to compute `P(B | J ^ M )`.

         PROJECT Computation

      (2)This calculation involved only the ten values in the CPT and their complements (1 - the CPT value), all of which were easy to estimate.

(3)(As an aside, note that even if both John and Mary call, the probability that a Burglary is occurring much less than half (but it would still be wise to call the police.) In fact, an implication of the fact that the probability of Burglary is 0.001 is that there would be an expected interval of almost three years between burglaries - but since John giving a false alarm has probability 0.05 we would expect about 50 false alarm calls from him in this same period!

C. In the above example, we were able to manually construct a Bayesian network on the basis of knowledge of the situation. In many cases, though, a structure like this may exist but may not be obvious.

1. It is possible, though to use machine learning to learn the structure of a Bayesian network from the data, by identifying parent-child relationships (hence conditional independence) implied by the data.

   This is done by calculating the extent to which various potential structures exhibit the kind of "shielding" of variables from ancestors of parents that shows up in a Bayesian network.

2. One writer gives the following example: (From Shi, Zongzhi *Advanced Artificial Intelligence* p. 246-247)

   a) The following actual data comes from a study on college plans of Wisconsin high school seniors. The survey included data on

      (1)Sex (SEX) (male, female)
      (2)Socioeconomic status (SES) (low, lower middle, upper middle, high)
      (3)Intelligence quotient (IQ) (low, lower middle, upper middle, high)
      (4)Parental encouragement (PE) (low, high)
      (5)College plans (CP) (yes, no)

      PROJECT

(6)The tables show counts of the number of surveys in which various combinations of values from the variables occur - e.g. the first entry is SEX = male, SES = low, IQ = low, PE = low, CP = yes.

b) Probabilities for various possible structures were calculated, based on the assumption that SEX did not depend on any other variable, and that no other variable depended on CP. Two possible structures were the most probable structures implied by the data.

PROJECT

c) A curious feature is that these two structures differed only in the direction of the arrow connecting IQ and PE. Also unusual was the fact is that they showed SES being the parent of IQ.

The researchers hypothesized the existence of a hidden variable which influenced both SES and IQ. When a new run was done with this additional variable (for which, of course, there were no known values), the following resulted as the most probable structure - also much more probable of either of the other two.

PROJECT

III.Summary

A. Naieve Bayesian learning is used widely in machine learning,

B. Bayesian network structures can be learned from data - and once the network is learned, the necessary conditional probabilities can also be extracted from the data.