

Bayesian Inferencing

Dr Pradipta Biswas, *PhD (Cantab)*
 Assistant Professor
 Indian Institute of Science
<https://cambum.net/index.htm>

Bayes' rule

- Relates prior knowledge (prior to present experiment) for taking present inference
- **Intuition:** Before we observe the data, the parameters are described by a *prior* which is typically very broad. Once we observed the data, we can make use of Bayes' formula to find *posterior*. Since now we can utilize some more facts in the form of evidences, the *posterior* is narrower than *prior*.

8/7/2018

Bayes' Rule

Posterior Probability

↓

Prior Probability

$$P(Y|X) = P(X|Y) * P(Y) / P(X)$$

Y = Cause
 X = Evidence or Symptoms

- Bayes' rule with more than one evidence

$$P(Y|X, e) = P(X|Y, e)P(Y|e) / P(X|e)$$

8/7/2018

Example

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

1% of people have a certain *genetic defect*.
 90% of tests for the gene defect the defect (true positives).
 9.5% of the tests are *false positives*.
 If a person gets a positive test result, what are the odds they actually have the genetic defect?

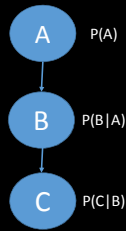
The first step into solving Bayes' theorem problems is to assign letters to events:
 A = chance of having the faulty gene. That was given in the question as 1%. That also means the probability of not having the gene (~A) is 99%.
 B = A positive test result.

P(A|B) = Probability of having the gene given a positive test result.
 P(B|A) = Chance of a positive test result given that the person actually has the gene. That was given in the question as 90%.
 P(B|~A) = Chance of a positive test if the person doesn't have the gene. That was given in the question as 9.6%.

Now we have all of the information we need to put into the equation:
 $P(A|B) = (0.9 * 0.01) / (0.9 * 0.01 + 0.99 * 0.096) = 0.0865$ (8.65%).
 The probability of having the faulty gene on the test is 8.65%.

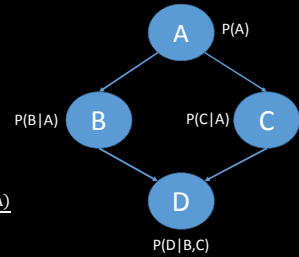
Example

$$\begin{aligned}
 P(A|C) &= \frac{P(C|A)P(A)}{P(C)} \\
 &= \frac{P(C|A,B)P(B|A)P(A)}{P(C)} \\
 &= \frac{P(C|B)P(B|A)P(A)}{P(C)}
 \end{aligned}$$



Bayesian Network

$$\begin{aligned}
 P(A|D) &= \frac{P(D|A)P(A)}{P(D)} \\
 &= \frac{P(D|A,B,C)P(B,C|A)P(A)}{P(D)} \\
 &= \frac{P(D|B,C)P(B,C|A)P(A)}{P(D)} \\
 &= \frac{P(D|B,C)P(B|A)P(C|A)P(A)}{P(D)}
 \end{aligned}$$

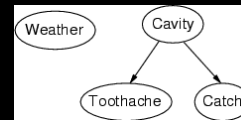


Bayesian networks

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions
- Syntax:
 - a set of nodes, one per variable
 - a directed, acyclic graph (link = "directly influences")
 - a conditional distribution for each node given its parents: $P(X_i | \text{Parents}(X_i))$
- In the simplest case, conditional distribution represented as a **conditional probability table (CPT)** giving the distribution over X_i for each combination of parent values

Example

- Topology of network encodes conditional independence assertions:

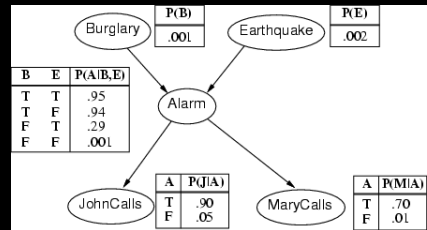


- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity*

Example

- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?
- Variables: *Burglary, Earthquake, Alarm, JohnCalls, MaryCalls*
- Network topology reflects "causal" knowledge:
 - A burglar can set the alarm off
 - An earthquake can set the alarm off
 - The alarm can cause Mary to call
 - The alarm can cause John to call

Example contd.



Example

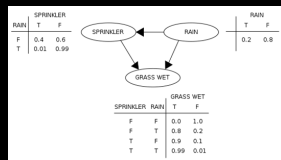
What is the probability of it is raining given the grass is wet?

$$P(R|G) = \frac{P(G|R)P(R)}{P(G)}$$

$$= \frac{P(G|R,S)P(S|R)P(R)}{P(G)}$$

$$\Pr(G, S, R) = \Pr(G|S, R) \Pr(S|R) \Pr(R)$$

$$\sum_{S,R \in \{T,F\}} \Pr(G = T, S, R)$$



Compactness

- A CPT for Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values
- Each row requires one number p for $X_i = \text{true}$ (the number for $X_i = \text{false}$ is just $1-p$)
- If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ numbers
- I.e., grows linearly with n , vs. $O(2^n)$ for the full joint distribution
- For burglary net, $1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$)



Constructing Bayesian networks

- 1. Choose an ordering of variables X_1, \dots, X_n
- 2. For $i = 1$ to n
 - add X_i to the network
 - select parents from X_1, \dots, X_{i-1} such that $P(X_i | \text{Parents}(X_i)) = P(X_i | X_1, \dots, X_{i-1})$

This choice of parents guarantees:

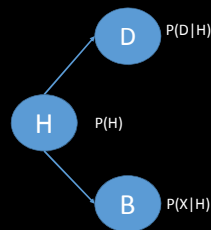
$$\begin{aligned}
 P(X_1, \dots, X_n) &= \prod_{i=1}^n P(X_i | X_1, \dots, X_{i-1}) \\
 \text{(chain rule)} & \\
 &= \prod_{i=1}^n P(X_i | \text{Parents}(X_i)) \\
 \text{(by construction)} &
 \end{aligned}$$

Summary of Bayesian Network

- Bayesian networks provide a natural representation for (causally induced) conditional independence
- Topology + CPTs = compact representation of joint distribution
- Generally easy for domain experts to construct

Bayesian Inference

$$\begin{aligned}
 P(X|D) &= P(X|D,H)P(H|D) \\
 &= P(X|H)P(H|D) \\
 &= P(X|H) \frac{P(D|H)P(H)}{P(D)}
 \end{aligned}$$



$P(h_i | D)$ for $j=1$

$P(h_i)$	$P(d_1 h_i)$	$P(d_1 h_i) \cdot P(h_i)$	$P(h_i d_1)$	$P(X h_i)$
0.1	0	0	0	0
0.2	0.25	0.05	0.1	0.25
0.4	0.5	0.20	0.4	0.5
0.2	0.75	0.15	0.3	0.75
0.1	1	0.10	0.2	1
		$\Sigma = 0.5$		

Suppose there are five kinds of bags of candies.

- 10% are h_1 : 100% cherry candies
- 20% are h_2 : 75% cherry candies + 25% lime candies
- 40% are h_3 : 50% cherry candies + 50% lime candies
- 20% are h_4 : 25% cherry candies + 75% lime candies
- 10% are h_5 : 100% lime candies

Then we observe candies drawn from some bag: ●●●●●●●●●●

What kind of bag is it? What flavour will the next candy be?

$$\begin{aligned}
 \sum_{i=1}^n P(X | h_i) \cdot P(h_i | d_1) &= 0 + 0.025 + 0.2 + 0.225 + 0.2 \\
 &= 0.65
 \end{aligned}$$

P(h_i|D) for j=2

P(h _i)	P(d _{2,1} h _i)	P(d _{2,2} h _i), P(h _i d ₂)	P(h _i d ₂)	P(X h _i)
0.1	0	0	0	0
0.2	0.25*0.25=0.0625	0.0125	0.038	0.25
0.4	0.5*0.5=0.25	0.1	0.308	0.5
0.2	0.75*0.75=0.5625	0.1125	0.346	0.75
0.1	1	0.1	0.308	1
		Σ = 0.325		

$$\sum_{i=1}^n P(X|h_i) \cdot P(h_i|d_{2,j}) = 0 + 0.01 + 0.154 + 0.26 + 0.308 = 0.732$$

P(h_i|D) for j=3

P(h _i)	P(d _{3,1} h _i)	P(d _{3,2} h _i), P(h _i d ₃)	P(h _i d _{3,1})	P(X h _i)
0.1	0	0	0	0
0.2	0.0156	0.0031	0.0131	0.25
0.4	0.125	0.05	0.210	0.5
0.2	0.422	0.0844	0.355	0.75
0.1	1	0.1	0.421	1
		Σ = 0.2375		

$$\sum_{i=1}^n P(X|h_i) \cdot P(h_i|d_{3,2j}) = 0 + 0.003275 + 0.105 + 0.2662 + 0.421 = 0.795525$$

MAP Learning

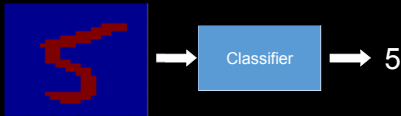
- Summing over the hypothesis space is often intractable
- (e.g., 18,446,744,073,709,551,616 Boolean functions of 6 attributes)
- Maximum a posteriori (MAP) learning: choose h MAP maximizing P(h_i|d)
- I.e., maximize P(d|h_i)P(h_i) or logP(d|h_i) + logP(h_i)
- Log terms can be viewed as (negative of)
 - bits to encode data given hypothesis + bits to encode hypothesis
 - This is the basic idea of minimum description length (MDL) learning

Bayesian Classification

- Problem statement:
 - Given features X₁, X₂, ..., X_n
 - Predict a label Y

Another Application

Digit Recognition



- $X_1, \dots, X_n \in \{0,1\}$ (Black vs. White pixels)
- $Y \in \{5,6\}$ (predict whether a digit is a 5 or a 6)

The Bayes Classifier

- A good strategy is to predict:

$$\arg \max_Y P(Y|X_1, \dots, X_n)$$

- (for example: what is the probability that the image represents a 5 given its pixels?)
- So ... How do we compute that?

The Bayes Classifier

- Use Bayes Rule!

$$P(Y|X_1, \dots, X_n) = \frac{\overset{\text{Likelihood}}{P(X_1, \dots, X_n|Y)} \overset{\text{Prior}}{P(Y)}}{\underset{\text{Normalization Constant}}{P(X_1, \dots, X_n)}}$$

- Why did this help? Well, we think that we might be able to specify how features are "generated" by the class label

The Bayes Classifier

- Let's expand this for our digit recognition task:

$$P(Y=5|X_1, \dots, X_n) = \frac{P(X_1, \dots, X_n|Y=5)P(Y=5)}{P(X_1, \dots, X_n|Y=5)P(Y=5) + P(X_1, \dots, X_n|Y=6)P(Y=6)}$$

$$P(Y=6|X_1, \dots, X_n) = \frac{P(X_1, \dots, X_n|Y=6)P(Y=6)}{P(X_1, \dots, X_n|Y=5)P(Y=5) + P(X_1, \dots, X_n|Y=6)P(Y=6)}$$

- To classify, we have to compute these two probabilities and predict based on which one is greater

Model Parameters

- The problem with explicitly modeling $P(X_1, \dots, X_n | Y)$ is that there are usually way too many parameters:
 - We'll run out of space
 - We'll run out of time
 - And we'll need tons of training data (which is usually not available)

The Naïve Bayes Model

- The *Naïve Bayes Assumption*: Assume that all features are independent **given the class label Y**
- Equationally speaking:

$$P(X_1, \dots, X_n | Y) = \prod_{i=1}^n P(X_i | Y)$$

Why is this useful?

- # of parameters for modeling $P(X_1, \dots, X_n | Y)$:
 - $2^{(2^n - 1)}$
- # of parameters for modeling $P(X_1 | Y), \dots, P(X_n | Y)$
 - $2n$

Take Away Points

- Bayesian Inferencing
- Bayesian Network
- Inferencing using Bayesian Network
- Introduction to Statistical Learning
- Naïve Bayes' Classifier