Bayesian Reasoning

6.872/HST950
Peter Szolovits

Simplest Example

- Relationship between a diagnostic conclusion and a diagnostic test

<table>
<thead>
<tr>
<th></th>
<th>Disease Present</th>
<th>Disease Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Positive</td>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td>Test Negative</td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
<tr>
<td></td>
<td>TP+FP</td>
<td>FN+TN</td>
</tr>
</tbody>
</table>

Definitions

- True Positive (TP): Disease present and test positive
- True Negative (TN): Disease absent and test negative
- False Positive (FP): Disease absent but test positive
- False Negative (FN): Disease present but test negative

**Sensitivity (true positive rate):**
\[ TP/(TP+FN) \]

**False negative rate:**
\[ 1 - Sensitivity = FN/(TP+FN) \]

**Specificity (true negative rate):**
\[ TN/(FP+TN) \]

**False positive rate:**
\[ 1 - Specificity = FP/(FP+TN) \]

**Positive Predictive Value:**
\[ TP/(TP+FP) \]

**Negative Predictive Value:**
\[ TN/(FN+TN) \]

Test Thresholds

Test Thresholds Change Trade-off between Sensitivity and Specificity
Receiver Operator Characteristic (ROC) Curve

What makes a better test?

How certain are we after a test?

Bayes' Rule:

Example: Acute Renal Failure

• Choice of a handful (8) of therapies (antibiotics, steroids, surgery, etc.)
• Choice of a handful (3) of invasive tests (biopsies, IVP, etc.)
• Choice of 27 diagnostic “questions” (patient characteristics, history, lab values, etc.)
• Underlying cause is one of 14 diseases
  – We assume one and only one disease

Rationality

• Behavior is a continued sequence of choices, interspersed by the world’s responses
• Best action is to make the choice with the greatest expected value
• … decision analysis

Entropy of a distribution

For example:

\[ H(0.5, 0.5) = 1.0 \]
\[ H(0.1, 0.9) = 0.47 \]
\[ H(0.01, 0.99) = 0.08 \]
\[ H(0.001, 0.999) = 0.01 \]

\[ H(0.33, 0.33, 0.33) = 1.58 (!) \]
\[ H(0.005, 0.455, 0.5) = 1.04 \]
\[ H(0.005, 0.995, 0) = 0.045 \]

(!) -- should use \( \log_2 \)}
Interacting with ARF in 1973

Question 1: What is the patient's age?
1 0-10
2 11-30
3 31-50
4 51-70
5 Over 70
Reply: 5

The current distribution is:

<table>
<thead>
<tr>
<th>Disease</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>FARF</td>
<td>0.58</td>
</tr>
<tr>
<td>IBSTR</td>
<td>0.22</td>
</tr>
<tr>
<td>ATN</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Question 2: What is the patient's sex?
1 Male
2 Pregnant Female
3 Non-pregnant Female
Reply: 1

Local Sensitivity Analysis

Case-specific Likelihood Ratios

Assumptions in ARF

- Exhaustive, mutually exclusive set of diseases
- Conditional independence of all questions, tests, and treatments
- Cumulative (additive) disutilities of tests and treatments
- Questions have no modeled disutility, but we choose to minimize the number asked anyway


- "Idiot Bayes" for appendicitis
  1. Based on expert estimates -- lousy
  2. Statistics -- better than docs
  3. Different hospital -- lousy again
  4. Retrained on local statistics -- good
Probabilistic Models

• What to represent?
  – Disease
  – Finding (signs, symptoms, labs, radiology, …)
  – Syndromes
  – History, predisposing conditions
  – Treatments
    • modify disease, cause new symptoms, …
  – (Outcomes, preferences, …)

State Space

• Set of random variables
• Possible values of each
• Assignment of probability to every possible combination of values of all variables
  • \( p(v_1=a_1, v_2=a_2, v_3=a_3, \ldots) \)

Questions of Interest

• Given a set of values of certain variables, what is the probability that certain other variables have certain other values?
• E.g., \( p(v_1=a_1, v_7=a_7|v_2=a_2, v_4=a_4) \)
  \[ \frac{p(v_1=a_1, v_7=a_7, v_2=a_2, v_4=a_4)}{p(v_2=a_2, v_4=a_4)} \]
• We don’t care about all other variables
  – marginalize; i.e., sum over them all

Computational Cost

• For \( n \) binary variables, we need probability assignments to \( 2^n \) states.
• In programs such as DXPLAIN, \( n \) is on the order of thousands.
• Need to be very careful and clever
  – simple models
  – approximate solution techniques

Independence

• Two random variables are independent iff
  \( p(A&B)=p(A)p(B) \)
• Usually, however, variables may depend on others, but we are still interested whether they have a conditional dependence
• Two random variables are conditionally independent if for a conditioning variable \( D \), \( p(A&B|D)=p(A|D)p(B|D) \)

Independence was crucial to ARF

• Diseases were dependent; mutually exhaustive and exclusive.
• Questions were conditionally independent, given disease.
ARF model convenient

- Odds: \( O(D) = P(D)/P(\neg D) = P(D)/(1-P(D)) \)
- Likelihood ratio: \( L(S|D) = P(S|D)/P(S|\neg D) \)
- Bayes: \( O(D|S) = O(D)L(S|D) \)
- Multiple evidence:
  - \( O(D|S1&S2&\ldots) = O(D)L(S1|D)L(S2|D)\ldots \)
- Log transform:
  - \( W(D|S1&S2&\ldots) = W(D)+W(S1|D)+W(S2|D)+\ldots \)

Side comment on likelihood ratio

- \( L(s|d) = p(s|d)/p(s|\neg d) \) is constant only if \( \neg d \) is a "fixed" entity
- If, as in ARF, we have \( d1, d2, d3, \ldots \), then
  \[
  p(s|\neg dj) = \sum_{j} p(d_j) p(s|d_j)
  \]
  - As probabilities vary over the \( d_j \), \( p(s|\neg dj) \) will vary!

What if we made no assumptions in ARF?

- Any combination of diseases: \( 2^{14} = 16K \)
- Distinct probability for any combination of answers to any questions:
  \[
  p(q_1=a_1\&q_2=a_2\&\ldots) = 3^{27} = 7.7*10^{12}
  \]
  - \( p(q_1=a_1\&q_2=a_2\&\ldots|d_1\&d_2\&\neg d_3\&\ldots) = 2^{14}\cdot3^{27} = 1.25\cdot10^{17} \), just for ARF
- Simplification is essential!

Conditional Independence is not Independence

- \( P(b\&c|a) = P(a\&b\&c)/P(a) \)
- But
  \[
  P(b\&c) = p(a)p(b\&c|a)+p(\neg a)p(b\&c|\neg a)
  \]
- Two variables may be made conditionally dependent when we learn about a common descendant
  \[
  p(A)=.2, p(B)=.1
  p(C|A&B)=.8, p(C|A&\neg B)=.4
  p(C|\neg A&B)=.6, p(C|\neg A&\neg B)=.1
  p(C)=.2*.1*.8+.2*.9*.4+.8*.1*.6+.8*.9*.1=.208
  
  If we observe C, \( p(A&B) = .02*.8/.208 = .077 \), but \( p(A)=.42, p(B)=.31 \). \( p(a\&b\&c) \neq p(a|c)*p(b|c) \)
We don’t want to model high-arity dependence
• $P(C \mid A_1 \& \neg A_2 \& \neg A_3 \& A_4 \& \ldots)$
• too many probabilities
• Can we simplify?
  – Noisy or
  – noisy and
  – noisy max/min
  – ?

Noisy or
• $p(C \mid A \& B) = p(\text{“C happened because of A”})$
  * $p(\text{“C happened because of B”})$
  * $p(\text{“C happened anyway”})$
• $(1-p(C \mid A \& B)) = (1-p c(C \mid A)) \times (1-p c(C \mid B)) \times (1-L)$
• $p c(C \mid A) \leq p(C \mid A)$

Simple Models
(Singly-Connected)

Causality?
• Noisy-or (and, …)
• Bayes arrows are/are not causal
  – reversing an arrow adds new dependencies

How to propagate likelihood information?
Like likelihood ratios.
(Pearl poly-tree algorithm)

General Case Bayes Nets
(Multiply-connected DAGs)

Factoring to Simplify Computation

$p(E = t, B = t) = \sum_{D,E=C,D,E=t} p(A,B,C,D,E)$
$= \sum_{D,E=C,D,E=t} p(A,B,E=t,C,D,E=t) \times p(E=t) \times p(D|E=t,C)$
$= \sum_{D,E=C,D,E=t} p(E=t|C) \times \sum_{D} p(D|B=E=t,C) \times p(B=E=t|C) \times p(E=t|C)$
$= \sum_{E=C} p(E=t) \times p(D|B=E=t,C) \times p(B=E=t|C) \times p(E=t|C)$

In this simple example, 12 instead of 32 multiplications
How to build Bayes Nets?

- David Heckerman
  - Pathfinder/IntelliPath, around 1990