Reasoning Under Uncertainty

“It's tough to make predictions, especially about the future” — Yogi Berra

(Some of these notes are from Andrew Moore, CMU)

Dealing with Uncertainty

- Statistics is completely based on it
- Machine learning is a kind of neo-statistics
- Early symbolic AI completely ignored it
- In logic, any contradiction leads to nonsense; hence, noisy observations can cause trouble
- In search, we assume perfect predictability of moves
  - E.g., what would have to change if, in solving a traveling salesman problem, airplanes sometimes landed in cities other than their declared destinations
  - E.g., what if in the simple shopping planning problem, actions didn’t always work? The store could be out of bread, I could get a flat tire, the milk could be sour, etc.
- We will study methods of treating uncertainty in terms of probability

Ways to Approach Uncertainty

- Uncertainty
  - Probability Theory
    - Bayes’ Rule
    - Bayesian (Belief) Network
  - Fuzzy Set Theory
  - Dempster-Shafer Theory of Evidence
  - Modal logic of likelihood; e.g., $\phi \models L\phi$, $L\phi \models LL\phi$, etc.
  - Other Qualitative approaches
- Decision Making under Uncertainty
  - Utility Theory
  - Preferences

Discrete Random Variables

- $E$ is a (boolean) random variable if it denotes an uncertain event
- You will receive a grade of “A” in this class
- Global warming will cause Florida to be under water by 2100
- Your fever is caused by malaria
- We can extend this to discrete variables with more than two possible values
- Random variables can also be continuous
- Your first child will be 6’3” in adulthood
- Sources of uncertainty
  - Lack of knowledge (“Is population of Bhutan > 1M?”)
  - Imperfection of models (Florida)
  - Physics (radioactive decay)
Meaning(s) of Probability

- $P(E)$: fraction of "possible worlds" in which $E$ is true
  - The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability, which is thus simply a fraction whose numerator is the number of favorable cases and whose denominator is the number of all the cases possible. —Laplace

- Interpretations of probability (semester course in philosophy...)
  - Physical/Objective/ — related to random events
    - Frequentist
      - measure long runs of experimental trials; e.g., coin flipping
    - Propensity/Inductive
      - Quantum Mechanics; predicted by models, supported by evidence
  - Subjective/Belief

Multi-Valued Random Variables

A is a random variable with arity $k$ if it can take on exactly one value out of {$v_1, v_2, ..., v_k$}

$P(A = v_i \text{ and } A = v_j) = 0 \text{ if } i \neq j$

$P(A = v_i \text{ or } A = v_2 \text{ or } ... \text{ or } A = v_k) = 1$

$P(A = v_1 \text{ or } A = v_2 \text{ or } ... \text{ or } A = v_i) = \sum_{j=1..i} P(A = v_j)$

thus, $\sum_{j=1..k} P(A = v_j) = 1$

$P(B \text{ and } (A = v_1 \text{ or } A = v_2 \text{ or } ... \text{ or } A = v_i)) = \sum_{j=1..i} P(B \text{ and } A = v_j)$

thus, $P(B) = \sum_{j=1..k} P(B \text{ and } A = v_j)$

Probability Axioms

- $0 \leq P(A) \leq 1$
- $P(\text{true}) = 1$
- $P(\text{false}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Useful theorems:

- $P(\neg A) = 1 - P(A)$
- $P(A) = P(A \text{ and } B) + P(A \text{ and } \neg B)$

Notation

- Most writers abuse notation
- $P(A=\text{true})$ is a scalar, the fraction of times when $A$ is true
- $P(A)$ is used sometimes as $P(A=\text{true})$, sometimes as a vector of two components (in the Boolean case), $[P(A=\text{true}), P(A=\text{false})]$

- Of course, for Boolean $A$, $P(A=\text{false}) = 1 - P(A=\text{true})$, so this is not a big stretch
- More confusing for random variables with multiple possible values
- Sometimes, we use \textbf{bold face} to indicate the vector

- $P(A) = [P(A=v_1), P(A=v_2), ... P(A=v_k)]$
- ... sorry 'bout that!
Joint Distribution

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.30</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.10</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Marginalizing: Sum over all rows in which a condition of interest is true:

\[ P(A) = 0.05 + 0.10 + 0.25 + 0.10 = 0.50 \]

\[ P(A \text{ and } \neg C) = 0.05 + 0.25 = 0.30 \]

Joint distribution has all information! Alas, it contains 2^3 rows!

Conditional Probability

- \( P(A|B) = \frac{\text{fraction of possible worlds in which B is true in which A is also true}}{\text{possible worlds in which B is true}} \)
- \( P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \)
- \( \therefore P(A \text{ and } B) = P(B)P(A|B) \)

Definitions

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

Sensitivity (true positive rate): \( TP/(TP+FN) \)
False negative rate: \( 1 - \text{Sensitivity} = FN/(TP+FN) \)

Specificity (true negative rate): \( TN/(FP+TN) \)
False positive rate: \( 1 - \text{Specificity} = FP/(FP+TN) \)

Positive Predictive Value: \( TP/(TP+FP) \)
Negative Predictive Value: \( TN/(FN+TN) \)
How certain are we after a test?

Imagine \( P(D^+) = 0.001 \) (it’s a rare disease)
Accuracy of test \( P(T^+|D^+) = P(T^+|D^-) = 0.95 \)

Bayes’ Rule:

\[
P(D_i) \cdot P(S|D_i) \sum_{k=1}^{n} P(D_k) \cdot P(S|D_k)
\]

Monty Hall Problem

Three doors, Red, Blue, Green
One has a car, two have donkeys
\( P(R) = P(B) = P(G) = 1/3 \)
Suppose you choose a door, say Red. Monty opens the remaining door that does not hold the car, say Blue.
You are given the chance to change your choice, in our case from Red to Green. Should you?

Independence

- We say that two variables, A and B, are independent iff
  \( P(A \text{ and } B) = P(A) \cdot P(B) \)
- or, alternatively, iff \( P(A|B) = P(A) \)
  - remember, \( P(A|B) = P(A \text{ and } B) / P(B) = P(A) \cdot P(B) / P(B) = P(A) \)
- E.g., coin flips, but not rain in Boston and in New York

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.30</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.10</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.10</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.25</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.10</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Are A and B independent?
\( P(A) = 0.50, P(B) = 0.50, P(A \text{ and } B) = 0.35 \)

What about B and C?
\( P(B) = 0.50, P(C) = 0.30, P(B \text{ and } C) = 0.15 \)

Conditional Independence

- Two variables, \( x \) and \( y \) may be dependent, but made independent if we know the value of some other variable.
- \( x, y \) are conditionally independent given \( z \) iff
  \( P(x, y|z) = P(x|z) \cdot P(y|z) \)
- E.g., are coughing and fever independent?
  - What if we know that you have a cold?
    \( P(C \text{ and } F|D) = P(C|D) \cdot P(F|D) \)
- E.g., Being burgled and seeing the police rush toward your house are independent, but what if you know the burglar alarm failed to go off?
  \( P(B \text{ and } P|\neg A) = P(B|\neg A) \cdot P(P|\neg A) \)

\[
\begin{array}{ccc}
D & C & F \\
\hline
B & A & P
\end{array}
\]
Two variables, \( x \) and \( y \), can be independent, but made dependent if we know the value of some other variable.

- E.g., the probability that your house is burgled is independent of the probability that it’s hit by a meteorite.
- But, if we hear an emergency vehicle siren near your house, a dependency is induced: the more likely the burglary, the less we “need” the meteorite to explain the siren.

### Diagnostic Reasoning with Naive Bayes

- We saw Naive Bayes when we talked about Bayesian classification.
- Here, we use the same ideas for sequential diagnostic reasoning.
- State of knowledge: a probability distribution over a set of possible diseases
  
  \[
  \sum_i P(D_i) = 1, \quad P(D_i) = 0 \text{ for } i \neq j
  \]
- In binary case, it’s just \( P(D) \) and \( P(\neg D) \).
- We can observe \( n \) different symptoms, \( S_1, \ldots, S_n \), any one of which can have \( n_j \) possible values, e.g., \( S_j = v_{j,k} \), where \( j \in \{1, \ldots, n\} \) and \( k \in \{1, \ldots, n_j\} \).

### Entropy Redux

- How to choose which observation to make next?
- Compute the expected entropy of \( P(D) \) after requesting each possible observation
  
  \[
  q = \arg\min_j E(H(P(D|S_j)))
  \]
- For each observation, \( S_j \), we can get \( n_j \) possible answers
  
  - For each answer, we can compute the revised (by Bayes rule) posterior probability distribution
  
  - For that distribution, we compute its entropy
  
  - The expected entropy weights these entropies by the probability that we would get that answer if we asked that question, namely
    
    \[
    P(S_j = v_k) = \sum_{i=1}^n P(D_i) P(S_j = v_k | D_i)
    \]
Odds-Likelihood

- In gambling, “3-to-1” odds means 75% chance of success. 
  \[ O = \frac{P}{1 - P} = \frac{P}{\sim P} \]
- \( P = 0.5 \) means \( O = 1 \)
- Likelihood ratio: 
  \[ L(S|D) = \frac{P(S \cap D)}{P(S \cap \sim D)} \]
- Odds-likelihood form of Bayes rule:
  \[ O(D|S_1, \ldots, S_n) = O(D)L(S_1|D) \ldots L(S_n|D) \]
- Log transform:
  \[
  \log O(D|S_1, \ldots, S_n) = \log [O(D)L(S_1|D) \ldots L(S_n|D)] \\
  = \log [O(D)] + \log [O(S_1|D)] + \ldots + \log [O(S_n|D)] \\
  = W(D) + W(S_1|D) + \ldots + W(S_n|D)
  \]

Utility Theory and Decision Analysis

- Goal:
  - In search, binary
  - In real world, better or worse
- Utility measures the value of an outcome:
  - $$$ in investments
  - Years of (quality-adjusted) life in healthcare
- Principle of rationality:
  - Choose the action that maximizes expected utility.
  - No guarantee of instant "win", but in long run, maximizes rewards.

Case of a Man with Gangrene

- From Pauker’s “Decision Analysis Service” at New England Medical Center Hospital, late 1970’s.
- Man with gangrene of foot
- Choose to amputate foot or treat medically
- If medical treatment fails, patient may die or may have to amputate whole leg.
- What to do? How to reason about it?

Decision Tree for Gangrene Case

(Different sense of “Decision Tree” from ML/Classification)

- amputate foot
  - live (.99) 880
  - die (.01) 0
- medicine
  - full recovery (.7) 1000
  - worse (.25) 686
  - die (.05) 995

- amputate leg
  - live (.98) 700
  - die (.02) 0

- Choice
- Chance
“Folding back” a Decision Tree

- The value of an outcome node is its utility
- The value of a chance node is the expected value of its alternative branches; i.e., their values weighted by their probabilities
- The value of a choice node is the maximum value of any of its branches

Where Do Utilities Come From?

- Standard gamble
  - Would you prefer (choose one of the following two):
    1. I chop off your foot
    2. We play a game in which a fair process produces a random number \( r \) between 0 and 1
      - If \( r > 0.8 \), I kill you; otherwise, you live on, healthy
    - If you're indifferent, that's the value of living without your foot!
    - I vary the 0.8 threshold until you are indifferent.
  - Alas, difficult ascertainment problems!