What action next?

Environment

Static vs. Dynamic

Fully vs. Partially Observable

Perfect vs. Noisy

Deterministic vs. Stochastic

Instantaneous vs. Durative

Percepts

Actions
Algorithms

- Blind search
- Heuristic search
- Mini-max & Expectimax
- MDPs
- Reinforcement learning
- State estimation
- Variable Elimination

Knowledge Representation

- Problem spaces
- Constraint networks
- HMMs
- Bayesian networks
- First-order logic
- Markov logic networks
- ...

What action next?
Example: Alarm Network

<table>
<thead>
<tr>
<th>B</th>
<th>P(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+b</td>
<td>0.001</td>
</tr>
<tr>
<td>←b</td>
<td>0.999</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E</th>
<th>P(E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+e</td>
<td>0.002</td>
</tr>
<tr>
<td>←e</td>
<td>0.998</td>
</tr>
</tbody>
</table>

| A   | J   | P(J|A) |
|-----|-----|------|
| +a  | +j  | 0.9  |
| +a  | ←j  | 0.1  |
| ←a  | +j  | 0.05 |
| ←a  | ←j  | 0.95 |

| A   | M   | P(M|A) |
|-----|-----|------|
| +a  | +m  | 0.7  |
| +a  | ←m  | 0.3  |
| ←a  | +m  | 0.01 |
| ←a  | ←m  | 0.99 |

Probabilities in BNs

- Bayes’ nets implicitly encode joint distributions
  - As a product of local conditional distributions
  - To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:
    \[ P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\text{parents}(X_i)) \]
  - This lets us reconstruct any entry of the full joint
- Not every BN can represent every joint distribution
  - The topology enforces certain independence assumptions
  - Compare to the exact decomposition according to the chain rule!
P(B | J=true, M=true)

Earthquake

Burglary

Alarm

JohnCalls

MaryCalls

P(blj,m) = α \sum_{e,a} P(b,j,m,e,a)

Variable Elimination

P(blj,m) = αP(b) \sum_{e}P(e) \sum_{a}P(alb,e)P(jla)P(m,a)

Repeated computations \rightarrow Dynamic Programming
Learning

What is Machine Learning?
Machine Learning

Study of algorithms that
- improve their performance
- at some task
- with experience

Exponential Growth in Data
Supremacy of Machine Learning

- Machine learning is preferred approach to
  - Speech recognition, Natural language processing
  - Web search – result ranking
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - Computational biology
  - Sensor networks
  - ...

- This trend is accelerating
  - Improved machine learning algorithms
  - Improved data capture, networking, faster computers
  - Software too complex to write by hand
  - New sensors / IO devices
  - Demand for self-customization to user, environment

Space of ML Problems

<table>
<thead>
<tr>
<th>What is Being Learned?</th>
<th>Labeled Examples</th>
<th>Reward</th>
<th>Nothing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Discrete Function</strong></td>
<td>Classification</td>
<td></td>
<td>Clustering</td>
</tr>
<tr>
<td><strong>Continuous Function</strong></td>
<td>Regression</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Policy</strong></td>
<td>Apprenticeship Learning</td>
<td>Reinforcement Learning</td>
<td></td>
</tr>
</tbody>
</table>
The Origin of Bayes Nets

<table>
<thead>
<tr>
<th>Earthquake</th>
<th>Burglary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radio</td>
<td>Alarm</td>
</tr>
<tr>
<td>Nbr1Calls</td>
<td>Nbr2Calls</td>
</tr>
</tbody>
</table>

Pr(B=t)  Pr(B=f)
0.05    0.95

Pr(A|E,B)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>e,b</td>
<td>0.9 (0.1)</td>
</tr>
<tr>
<td>e,b</td>
<td>0.2 (0.8)</td>
</tr>
<tr>
<td>e,b</td>
<td>0.85 (0.15)</td>
</tr>
<tr>
<td>e,b</td>
<td>0.01 (0.99)</td>
</tr>
</tbody>
</table>

Learning Topics

- Learning Parameters for a Bayesian Network
  - Fully observable
    - Maximum Likelihood (ML)
    - Maximum A Posteriori (MAP)
    - Bayesian
  - Hidden variables (EM algorithm)
- Learning Structure of Bayesian Networks
Parameter Estimation and Bayesian Networks

We have:
- Bayes Net structure and observations
- We need: Bayes Net parameters

<table>
<thead>
<tr>
<th>E</th>
<th>B</th>
<th>R</th>
<th>A</th>
<th>J</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P(B) = ?

P(¬B) = 1 - P(B) = 0.6
Parameter Estimation and Bayesian Networks

\[
P(A|E,B) = ?
\]
\[
P(A|E,\neg B) = ?
\]
\[
P(A|\neg E,B) = ?
\]
\[
P(A|\neg E,\neg B) = 0.5
\]
Parameter Estimation and Bayesian Networks

Coin

Coin Flip

\[ P(H|C_1) = 0.1 \]
\[ P(H|C_2) = 0.5 \]
\[ P(H|C_3) = 0.9 \]

Which coin will I use?

\[ P(C_1) = \frac{1}{3} \]
\[ P(C_2) = \frac{1}{3} \]
\[ P(C_3) = \frac{1}{3} \]

Prior: Probability of a hypothesis before we make any observations
Coin Flip

\[ P(H|C_1) = 0.1 \quad P(H|C_2) = 0.5 \quad P(H|C_3) = 0.9 \]

Which coin will I use?

\[ P(C_1) = 1/3 \quad P(C_2) = 1/3 \quad P(C_3) = 1/3 \]

**Uniform Prior**: All hypotheses are equally likely before we make any observations.

Experiment 1: Heads

Which coin did I use?

\[ P(C_1|H) = ? \quad P(C_2|H) = ? \quad P(C_3|H) = ? \]

\[ P(C_1|H) = \frac{P(H|C_1)P(C_1)}{P(H)} \]

\[ P(H) = \sum_{i=1}^{3} P(H|C_i)P(C_i) \]
Experiment 1: Heads

Which coin did I use?

\[ P(C_1|H) = 0.066 \quad P(C_2|H) = 0.333 \quad P(C_3|H) = 0.6 \]

**Posterior:** Probability of a hypothesis given data

<table>
<thead>
<tr>
<th>Coin</th>
<th>Prior Probability</th>
<th>Posterior Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C_1)</td>
<td>1/3</td>
<td>0.1</td>
</tr>
<tr>
<td>(C_2)</td>
<td>1/3</td>
<td>0.5</td>
</tr>
<tr>
<td>(C_3)</td>
<td>1/3</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Using Prior Knowledge

- **Should we always use a *Uniform Prior*?**
- **Background knowledge:**
  Heads => we have to buy Dan chocolate
  Dan *likes* chocolate…
  => Dan is more likely to use a coin biased in his favor

<table>
<thead>
<tr>
<th>Coin</th>
<th>Prior Probability</th>
<th>Posterior Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C_1)</td>
<td>1/3</td>
<td>0.1</td>
</tr>
<tr>
<td>(C_2)</td>
<td>1/3</td>
<td>0.5</td>
</tr>
<tr>
<td>(C_3)</td>
<td>1/3</td>
<td>0.9</td>
</tr>
</tbody>
</table>
Using Prior Knowledge

We can encode it in the prior:

\[
P(C_1) = 0.05 \quad P(C_2) = 0.25 \quad P(C_3) = 0.70
\]

\begin{align*}
P(H|C_1) & = 0.1 \quad P(H|C_2) = 0.5 \quad P(H|C_3) = 0.9
\end{align*}

Experiment 1: Heads

Which coin **did** I use?

\[
P(C_1|H) = ? \quad P(C_2|H) = ? \quad P(C_3|H) = ?
\]

\[
P(C_1|H) = \alpha P(H|C_1)P(C_1)
\]

\begin{align*}
P(H|C_1) & = 0.1 \quad P(H|C_2) = 0.5 \quad P(H|C_3) = 0.9
\end{align*}

\[
P(C_1) = 0.05 \quad P(C_2) = 0.25 \quad P(C_3) = 0.70
\]
Experiment 1: Heads
Which coin did I use?

P(C₁|H) = 0.006  P(C₂|H) = 0.165  P(C₃|H) = 0.829

Compare with ML posterior after Exp 1:
P(C₁|H) = 0.066  P(C₂|H) = 0.333  P(C₃|H) = 0.600

Experiment 2: Tails
Which coin did I use?

P(C₁|HT) = ?  P(C₂|HT) = ?  P(C₃|HT) = ?

P(C₁|HT) = αP(HT|C₁)P(C₁) = αP(H|C₁)P(T|C₁)P(C₁)

P(H|C₁) = 0.1  P(H|C₂) = 0.5  P(H|C₃) = 0.9
P(C₁) = 0.05  P(C₂) = 0.25  P(C₃) = 0.70
Experiment 2: Tails

Which coin did I use?

\[
P(C_1|HT) = 0.035 \quad P(C_2|HT) = 0.481 \quad P(C_3|HT) = 0.485
\]

\[
P(C_1|HT) = \alpha P(HT|C_1)P(C_1) = \alpha P(H|C_1)P(T|C_1)P(C_1)
\]

\[
\begin{align*}
P(H|C_1) &= 0.1 & P(H|C_2) &= 0.5 & P(H|C_3) &= 0.9 \\
P(C_1) &= 0.05 & P(C_2) &= 0.25 & P(C_3) &= 0.70
\end{align*}
\]
Your Estimate?

What is the probability of heads after two experiments?

Most likely coin: \( C_3 \)

Best estimate for \( P(H) \):

\[ P(H|C_3) = 0.9 \]

\[ \begin{align*}
C_1 & \quad P(H|C_1) = 0.1 \\
& \quad P(C_1) = 0.05 \\
C_2 & \quad P(H|C_2) = 0.5 \\
& \quad P(C_2) = 0.25 \\
C_3 & \quad P(H|C_3) = 0.9 \\
& \quad P(C_3) = 0.70
\end{align*} \]

Your Estimate?

**Maximum A Posteriori (MAP) Estimate:**
The best hypothesis that fits observed data assuming a non-uniform prior

Most likely coin: \( C_3 \)

Best estimate for \( P(H) \):

\[ P(H|C_3) = 0.9 \]

\[ \begin{align*}
C_3 & \quad P(H|C_3) = 0.9 \\
& \quad P(C_3) = 0.70
\end{align*} \]
### Did We Do The Right Thing?

| Event (C) | P(H|C)  |
|-----------|---------|
| C₁        | 0.1     |
| C₂        | 0.5     |
| C₃        | 0.9     |

- P(C₁|HT) = 0.035
- P(C₂|HT) = 0.481
- P(C₃|HT) = 0.485

C₂ and C₃ are almost equally likely.

12/1/14
A Better Estimate

Recall: \[ P(H) = \sum_{i=1}^{3} P(H|C_i)P(C_i) = 0.680 \]

\[
P(C_1|HT)=0.035 \quad P(C_2|HT)=0.481 \quad P(C_3|HT)=0.485
\]

\[
P(H|C_1) = 0.1 \quad P(H|C_2) = 0.5 \quad P(H|C_3) = 0.9
\]

Bayesian Estimate

Bayesian Estimate: Minimizes prediction error, given data assuming an arbitrary prior

\[ P(H) = \sum_{i=1}^{3} P(H|C_i)P(C_i) = 0.680 \]

\[
P(C_1|HT)=0.035 \quad P(C_2|HT)=0.481 \quad P(C_3|HT)=0.485
\]

\[
P(H|C_1) = 0.1 \quad P(H|C_2) = 0.5 \quad P(H|C_3) = 0.9
\]
Comparison
After more experiments: \textbf{HTHHHHHHHH}

ML (Maximum Likelihood):
\[ P(H) = 0.5 \]
after 10 experiments: \( P(H) = 0.9 \)

MAP (Maximum A Posteriori):
\[ P(H) = 0.9 \]
after 10 experiments: \( P(H) = 0.9 \)

Bayesian:
\[ P(H) = 0.68 \]
after 10 experiments: \( P(H) = 0.9 \)

Summary

<table>
<thead>
<tr>
<th>Prior</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>The most likely</td>
</tr>
<tr>
<td>Any</td>
<td>The most likely</td>
</tr>
<tr>
<td>Any</td>
<td>Weighted combination</td>
</tr>
</tbody>
</table>

Easy to compute

Minimum error
Great when data is scarce
Potentially much harder to compute

Still easy to compute
Incorporates prior knowledge
Bayesian Learning

Use Bayes rule:

\[
P(Y \mid X) = \frac{P(X \mid Y) P(Y)}{P(X)}
\]

Data Likelihood

Prior

Normalization

Or equivalently:

\[
P(Y \mid X) \propto P(X \mid Y) P(Y)
\]

Posterior

Parameter Estimation and Bayesian Networks

Earthquake

Burglary

Radio

Alarm

Nbr1Calls

Nbr2Calls

Now compute either MAP or Bayesian estimate

\[
P(B) = \text{Prior} + \text{data}
\]
What Prior to Use?

- Prev, you *knew*: it was one of only three coins
  - Now more complicated…
- The following are two common priors
  - **Binary variable Beta**
    - Posterior distribution is binomial
    - Easy to compute posterior
  - **Discrete variable Dirichlet**
    - Posterior distribution is multinomial
    - Easy to compute posterior

© Daniel S. Weld
Beta Distribution

- Example: Flip coin with Beta distribution as prior over $p$ [prob(heads)]
  1. Parameterized by two positive numbers: $a$, $b$
  2. Mode of distribution ($E[p]$) is $a/(a+b)$
  3. Specify our prior belief for $p = a/(a+b)$
  4. Specify confidence in this belief with high initial values for $a$ and $b$
- Updating our prior belief based on data
  - incrementing $a$ for every heads outcome
  - incrementing $b$ for every tails outcome

One Prior: Beta Distribution

$$\beta(x) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1},$$

$$0 \leq x \leq 1 \text{ and } a, b > 0$$

Here $\Gamma(y) = \int_0^\infty x^{y-1} e^{-x} dx$

For any positive integer $y$, $\Gamma(y) = (y-1)!$
Parameter Estimation and Bayesian Networks

Prior
\[ P(B|\text{data}) = ? \] Beta(1,4) “+ data” = (3,7) 

\[ \begin{array}{c|cc}
B & \neg B \\
\hline
F & .3 \\
T & .7 \\
\end{array} \]

Prior \( P(B) = 1/(1+4) = 20\% \) with equivalent sample size 5

Parameter Estimation and Bayesian Networks

\[ \begin{array}{c|cc|c|c}
E & B & A \\
\hline
T & F & T \\
F & F & F \\
F & T & T \\
F & F & T \\
F & T & F \\
\ldots & \ldots & \ldots \\
\end{array} \]

\[
\begin{align*}
P(A|E,B) &= ? \\
P(A|E,\neg B) &= ? \\
P(A|\neg E,B) &= ? \\
P(A|\neg E,\neg B) &= ?
\end{align*}
\]
Parameter Estimation and Bayesian Networks

P(A|E,B) = ?  Prior
P(A|E,¬B) = ?
P(A|¬E,B) = ?  Beta(2,3)
P(A|¬E,¬B) = ?

Parameter Estimation and Bayesian Networks

P(A|E,B) = ?  Prior
P(A|E,¬B) = ?
P(A|¬E,B) = ?  Beta(2,3)  + data= (3,4)
P(A|¬E,¬B) = ?
Bayesian Learning

Use Bayes rule:

\[ P(Y | X) = \frac{P(X | Y) P(Y)}{P(X)} \]

Or equivalently: \( P(Y | X) \propto P(X | Y) P(Y) \)

Naïve Bayes

\[ P(Y, F_1 \ldots F_n) = P(Y) \prod_i P(F_i | Y) \]

Assume that features are conditionally independent given class variable

Works well in practice

But forces probabilities towards 0 and 1
Naïve Bayes

- Naïve Bayes assumption:
  - Features are independent given class:
    \[
    P(X_1, X_2 | Y) = P(X_1 | X_2, Y) P(X_2 | Y) = P(X_1 | Y) P(X_2 | Y)
    \]
  - More generally:
    \[
    P(X_1 \ldots X_n | Y) = \prod_{i} P(X_i | Y)
    \]

- How many parameters now?
  - Suppose \( X \) is composed of \( n \) binary features

NB with Bag of Words for text classification

- Learning phase:
  - Prior \( P(Y) \)
    - Count how many documents from each topic (prior)
  - \( P(X_i | Y) \)
    - For each of \( m \) topics, count how many times you saw word \( X_i \) in documents of this topic (+ k for prior)
    - Divide by number of times you saw the word (+ k×m)

- Test phase:
  - For each document
    - Use naïve Bayes decision rule
    \[
    h_{NB}(x) = \arg \max_y P(y)^{LengthDoc} \prod_{i=1}^{LengthDoc} P(x_i | y)
    \]
Probabilities: Important Detail!

- \( P(\text{spam} \mid X_1 \ldots X_n) = \prod P(\text{spam} \mid X_i) \)
  
  Any more potential problems here?

- We are multiplying lots of small numbers
  Danger of underflow!
  - 0.5^{57} = 7 \times 10^{-18}

- Solution? Use logs and add!
  - \( p_1 \times p_2 = e^{\log(p_1) + \log(p_2)} \)
  - Always keep in log form