Bayesian Networks - Learning

Dan Weld

Slides adapted from Jack Breese, Dan Klein, Daphne Koller, Stuart Russell, Andrew Moore & Luke Zettlemoyer

Bayes’ Net Semantics

Formally:
- A set of nodes, one per variable X
- A directed, acyclic graph
- A CPT for each node
  - CPT = “Conditional Probability Table”
  - Collection of distributions over X, one for each combination of parents’ values
  
\[ P(X|A_1 \ldots A_n) \]

A Bayes net = Topology (graph) + Local Conditional Probabilities

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!

Example: Alarm Network

Example: Car Diagnosis

\[ P(B | J=true, M=true) = \sum_{e,a} P(b,j,m,e,a) \]
Variable Elimination

\[ P(b|m) = \alpha P(b) \sum \sum P(e) P(a|b,e) P(j|a) P(m,a) \]

Repeated computations \(\rightarrow\) Dynamic Programming

MCMC with Gibbs Sampling

- Fix the values of observed variables
- Set the values of all non-observed variables randomly
- Perform a random walk through the space of complete variable assignments. On each move:
  1. Pick a variable X
  2. Calculate \( P(X=\text{true} \mid \text{Markov blanket}) \)
  3. Set X to true with that probability
- Repeat many times. Frequency with which any variable X is true is its posterior probability.
- Converges to true posterior when frequencies stop changing significantly
  - stable distribution, mixing

The Origin of Bayes Nets

The Origin of Bayes Nets

Parameter Estimation and Bayesian Networks

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

Parameter Estimation and Bayesian Networks

P(B) = ? = 0.4
P(¬B) = 1 - P(B) = 0.6

Parameter Estimation and Bayesian Networks

P(A|E,B) = ?
P(A|E,¬B) = ?
P(A|¬E,B) = ?
P(A|¬E,¬B) = ?

Parameter Estimation and Bayesian Networks

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) = 1/3  P(C_2) = 1/3  P(C_3) = 1/3
Uniform Prior: All hypothesis are equally likely before we make any observations

Parameter Estimation and Bayesian Networks

Experiment 1: Heads

Which coin did I use?
P(C_1|H) = ?  P(C_2|H) = ?  P(C_3|H) = ?
P(C_1|H) = \frac{P(H|C_1)P(C_1)}{P(H)} = \frac{0.1 \times \frac{1}{3}}{P(H)} = \frac{0.033}{P(H)}
P(C_2|H) = \frac{P(H|C_2)P(C_2)}{P(H)} = \frac{0.5 \times \frac{1}{3}}{P(H)} = \frac{0.167}{P(H)}
P(C_3|H) = \frac{P(H|C_3)P(C_3)}{P(H)} = \frac{0.9 \times \frac{1}{3}}{P(H)} = \frac{0.3}{P(H)}

P(H) = \sum_{C_i} P(H|C_i)P(C_i)

P(H|C_1) = 0.1  P(H|C_2) = 0.5  P(H|C_3) = 0.9
P(C_1) = 1/3  P(C_2) = 1/3  P(C_3) = 1/3

Prior: Probability of a hypothesis before we make any observations

Coin Flip

C_1  C_2  C_3
P(H|C_1) = 0.5  P(H|C_2) = 0.9
Which coin will I use?
P(C_1) = 1/3  P(C_2) = 1/3  P(C_3) = 1/3
Uniform Prior: All hypothesis are equally likely before we make any observations
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

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

C1 C2 C3

P(C1) = 1/3 P(C2) = 1/3 P(C3) = 1/3

Experiment 2: Tails

Now, Which coin did I use?

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

\[
P(C_1|HT) = \alpha P(H|C_1) P(C_1) = \alpha P(H|C_2) P(C_2) = \alpha P(H|C_3) P(C_3)
\]

C1 C2 C3

P(H|C1) = 0.1 P(H|C2) = 0.5 P(H|C3) = 0.9

P(C1) = 1/3 P(C2) = 1/3 P(C3) = 1/3

Experiment 2: Tails

Now, Which coin did I use?

\[
P(C_1|HT) = 0.21 \quad P(C_2|HT) = 0.58 \quad P(C_3|HT) = 0.21
\]

\[
P(C_1|HT) = \alpha P(H|C_1) P(C_1) = \alpha P(H|C_2) P(C_2) = \alpha P(H|C_3) P(C_3)
\]

C1 C2 C3

P(H|C1) = 0.1 P(H|C2) = 0.5 P(H|C3) = 0.9

P(C1) = 1/3 P(C2) = 1/3 P(C3) = 1/3

Experiment 2: Tails

Which coin did I use?

\[
P(C_1|HT) = 0.21 \quad P(C_2|HT) = 0.58 \quad P(C_3|HT) = 0.21
\]

C1 C2 C3

P(H|C1) = 0.1 P(H|C2) = 0.5 P(H|C3) = 0.9

P(C1) = 1/3 P(C2) = 1/3 P(C3) = 1/3

Your Estimate?

What is the probability of heads after two experiments?

Most likely coin: C2

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

\[
P(H|C_2) = 0.5
\]
Your Estimate?

Maximum Likelihood Estimate: The best hypothesis that fits observed data assuming uniform prior

Most likely coin: $C_2$

Best estimate for $P(H)$

$P(H|C_2) = 0.5$

$P(H|C_1) = 0.5$

$P(C_1) = 1/3$

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

$P(H|C_2) = 0.5$

$P(H|C_1) = 0.1$

$P(H|C_3) = 0.9$

$P(C_1) = 0.05$

$P(C_2) = 0.25$

$P(C_3) = 0.70$

Using Prior Knowledge

We can encode it in the prior:

$P(C_1) = 0.05$

$P(C_2) = 0.25$

$P(C_3) = 0.70$

$P(H|C_1) = 0.1$

$P(H|C_2) = 0.5$

$P(H|C_3) = 0.9$

Experiment 1: Heads

Which coin did I use?

$P(C_1|H) = ?$

$P(C_2|H) = ?$

$P(C_3|H) = ?$

$P(C_1|H) = 0.006$

$P(C_2|H) = 0.165$

$P(C_3|H) = 0.829$

Compare with ML posterior after Exp 1:

$P(C_1|H) = 0.066$

$P(C_2|H) = 0.333$

$P(C_3|H) = 0.600$

Experiment 2: Tails

Which coin did I use?

$P(C_1|HT) = ?$

$P(C_2|HT) = ?$

$P(C_3|HT) = ?$

$P(C_1|HT) = 0.05$

$P(C_2|HT) = 0.25$

$P(C_3|HT) = 0.70$
Experiment 2: Tails

Which coin did I use?

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

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

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

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

Did We Do The Right Thing?

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

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

C_2 and C_3 are almost equally likely.

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

Did We Do The Right Thing?

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

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

Recall: \[ P(B) = \sum P(C_i) P(C_i) = 0.680 \]

\[ \begin{align*}
P(C_1|HT) &= 0.035 \quad P(C_2|HT) = 0.481 \quad P(C_3|HT) = 0.485 \\
P(H|C_1) &= 0.1 \quad P(H|C_2) = 0.5 \quad P(H|C_3) = 0.9
\end{align*} \]

Bayesian Estimate

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

\[ P(H) = \sum P(U(C_i) P(C_i) = 0.680 \]

\[ \begin{align*}
P(C_1|HT) &= 0.035 \quad P(C_2|HT) = 0.481 \quad P(C_3|HT) = 0.485 \\
P(H|C_1) &= 0.1 \quad P(H|C_2) = 0.5 \quad P(H|C_3) = 0.9
\end{align*} \]

Comparison

After more experiments: HTHHHHHHHHHH

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>

Bayesian Learning

Use Bayes rule:

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

Or equivalently:

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

Parameter Estimation and Bayesian Networks

Now compute either MAP or Bayesian estimate
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

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

Parameter Estimation and Bayesian Networks

Parameter Estimation and Bayesian Networks

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

Prior \( P(B) = 1/(1+4) = 20\% \) with equivalent sample size 5
Parameter Estimation and Bayesian Networks

Bayesian Learning

Use Bayes rule:

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

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