Review

* Inference in BNs
  - evidence nodes E
  - query node Q

  How to compute $P(Q|E)$?

* Polytree
  - singly connected networks
  - polynomial time inference

* Loopy BNs
  - Exact inference: node clustering

Learning

- as a form of uncertain reasoning from observations
- Agents can handle uncertainty using probability, but must learn
  probabilistic theories of the world from experience

* BN = DAG + CPTs, not always available from experts

  How to learn from data (examples / observations)?
- Structure (DAG): known or unknown?
- Evidence: complete data vs. "incomplete" data
  \( \rightarrow \) partial instantiation of nodes in BN
- Optimization:
  - Combinatorial vs. continuous
    (e.g. learning DAGs) (e.g. learning CPTs)
- Algorithms: non-iterative vs. iterative
  (loop many times over data set)
- Solution: local vs. global optima in model estimation

**Maximum likelihood estimation (ML, MLE)**
- Simplest form of learning
- Choose ("estimate") model (DAG + CPTs) to maximize
  \[ P(\text{observed data | model}) \]
  "likelihood"
- We'll focus on parameter learning: finding numerical parameters for a probabilistic model whose structure is fixed
Ex: biased coin

\[ X \in \{ \text{heads, tails} \} \]

\[ P(X = \text{heads}) = p \]
\[ P(X = \text{tails}) = 1 - p \]

• How to estimate \( p \) from \( T \) examples (i.e., results of \( T \) coin tosses)?

\[ \text{Intuition: } p = \frac{\# \text{ heads}}{\# \text{ flips} \to T} \]

• i.i.d. assumption:
  samples are independently, identically distributed according to \( P(X) \)

\[ \{ X^{(1)}, X^{(2)}, \ldots, X^{(T)} \} \text{ \( T \) samples} \]

• Probability of i.i.d. data:

\[ P(\text{data}) = P(X = x^{(1)}, X = x^{(2)}, \ldots, X = x^{(T)}) \]

\[ = P(X = x^{(1)}) \cdot P(X = x^{(2)}) \cdot \cdots \cdot P(X = x^{(T)}) \text{ \( \text{coin tosses independent} \)} \]

\[ = \prod_{t=1}^{T} P(X = x^{(t)}) \]

• Log-probability \( \mathcal{L} \)

\[ \mathcal{L} = \log P(\text{data}) \]

\[ \Rightarrow \log \text{-likelihood} = \sum_{t=1}^{T} \log P(X = x^{(t)}) \]
Notation:

Let \( N_+ = \text{count}(X = \text{heads}) \)
\( N_- = \text{count}(X = \text{tails}) \)

Clearly: \( N_+ + N_- = T \) (total # samples)

In terms of counts:

\[
L = N_+ \log p + N_- \log (1-p)
\]

Maximum likelihood estimation

\[
\frac{\partial L}{\partial p} = \frac{N_+}{p} + \frac{N_-}{1-p} = 0 \quad \text{at maximum}
\]

\[
N_+(1-p) - N_- p = 0
\]

\[
N_+ - p(N_+ + N_-) = 0
\]

\[
p = \frac{N_+}{N_+ + N_-} = \frac{N_+}{T}
\]

Intuitively, ML estimate is relative empirical frequency of heads
Discrete BNs with "complete data"

- Given: fixed DAG over discrete nodes \( \{X_1, X_2, \ldots, X_n\}\)
- CPTs enumerate \( P(X_i = x \mid \text{pa}(X_i) = \pi) \) as lookup tables

\[ \log(\text{configuration of parents of } X_i) \]

- Data is \( T \) complete instantiations of all nodes in BN

\[ \frac{1}{T} \sum_{t=1}^{T} (x_1(t), x_2(t), \ldots, x_n(t)) \]

Ex:

![Diagram of a simple Bayesian network with variables X1, X2, and X3, and a table of data with columns X1, X2, and X3.]

<table>
<thead>
<tr>
<th>t</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

- Each \( n \)-tuple of values is called an "example"

Goal: learn from examples

- estimate CPTs \( P(X_i = x \mid \text{pa}(X_i) = \pi) \)
- that maximize probability of data likelihood

* i.i.d. assumption:

Examples are independently, identically distributed according to \( P(X_1, X_2, \ldots, X_n) \)
Probability of data

\[ P(\text{data}) = \prod_{t=1}^{T} P(X_1 = x_1(t), X_2 = x_2(t), \ldots, X_n = x_n(t)) \]

Work out \( t \)-th term:

\[ P(X_1 = x_1(t), \ldots, X_n = x_n(t)) = P(X_1 = x_1(t)) P(X_2 = x_2(t) \mid X_1 = x_1(t)) \ldots \]

\[ P(X_n = x_n(t) \mid X_1 = x_1(t), \ldots, X_{n-1} = x_{n-1}) \]

\[ = \prod_{i=1}^{n} P(X_i = x_i(t) \mid X_1 = x_1(t), \ldots, X_{i-1} = x_{i-1}) \]

\[ = \prod_{i=1}^{n} P(X_i = x_i(t) \mid \text{pa}(X_i) = \text{pa}_i(t)) \]

Log-likelihood

\[ L = \log P(\text{data}) \]

\[ = \log \prod_{t=1}^{T} P(X_1(t), X_2(t), \ldots, X_n(t)) \]

\[ = \log \prod_{t=1}^{T} \prod_{i=1}^{n} P(X_i(t) \mid \text{pa}_i(t)) \]

\[ = \sum_{t=1}^{T} \sum_{i=1}^{n} \log P(X_i(t) \mid \text{pa}_i(t)) \]

\[ L = \sum_{i=1}^{n} \sum_{t=1}^{T} \log P(X_i = x_i(t) \mid \text{pa}(X_i) = \text{pa}_i(t)) \]


case order

\[ \frac{\text{order of sums}}{\sum_{i=1}^{n} \sum_{t=1}^{T} \log P(X_i = x_i(t) \mid \text{pa}(X_i) = \text{pa}_i(t))} \]
* Let \( \text{count}(X_i = x, pa_i = \pi) \) denote the examples for which \( X_i = x \) and \( pa_i = \pi \).

![Diagram](image.png)

<table>
<thead>
<tr>
<th>( t )</th>
<th>( X_1 )</th>
<th>( X_2 )</th>
<th>( X_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ \text{count}(X_1 = 1) = 3 \]
\[ \text{count}(X_3 = 1, X_1 = 0) = 2 \]
\[ \text{count}(X_2 = 1, X_1 = 1) = 1 \]

* Log-likelihood:

\[
L = \sum_{i=1}^{n} \sum_{x \in \pi} \sum_{\pi \in \pi'} \text{count}(X_i = x, pa_i = \pi) \log P(X_i = x | pa_i = \pi)
\]

Values of \( X_i \) values of \( pa(X_i) \)

* \( \pi \) completely determined by data

* ML estimation:

How to choose \( P(X_i = x | pa_i = \pi) \) to maximize \( L(data) \)?

* ML solution (w/out proof):

\[
P_{ML}(X_i = x | pa_i = \pi) = \frac{\text{count}(X_i = x, pa_i = \pi)}{\sum_{x'} \text{count}(X_i = x', pa_i = \pi)}
\]

\[
= \frac{\text{count}(X_i = x, pa_i = \pi)}{\text{count}(pa_i = \pi)}
\]
Ex: 30-day experiment, set alarm to wake up for class

CPIs

\[
\begin{align*}
P(\text{Alarm}) & \quad \text{Alarm} \\
P(\text{Late} | \text{Alarm}) & \quad \text{Late}
\end{align*}
\]

\[
\begin{align*}
\text{count}(\text{Alarm} = 0) &= 10 \\
\text{count}(\text{Alarm} = 1) &= 20 \\
\text{count}(\text{Late} = 0, \text{Alarm} = 0) &= 2 \\
\text{count}(\text{Late} = 1, \text{Alarm} = 0) &= 8 \\
\text{count}(\text{Late} = 0, \text{Alarm} = 1) &= 17 \\
\text{count}(\text{Late} = 1, \text{Alarm} = 1) &= 3
\end{align*}
\]

\[
\begin{align*}
P_{\text{ML}} (\text{Alarm} = 0) &= \frac{10}{30} \\
P_{\text{ML}} (\text{Alarm} = 1) &= \frac{20}{30}
\end{align*}
\]

\[
\begin{align*}
P_{\text{ML}} (\text{Late} = 0 | \text{Alarm} = 0) &= \frac{2}{10} \\
P_{\text{ML}} (\text{Late} = 1 | \text{Alarm} = 0) &= \frac{8}{10} \\
P_{\text{ML}} (\text{Late} = 0 | \text{Alarm} = 1) &= \frac{17}{20} \\
P_{\text{ML}} (\text{Late} = 1 | \text{Alarm} = 1) &= \frac{3}{20}
\end{align*}
\]

* Properties of MLE

- Asymptotically correct: \( P_{\text{ML}} (X_1, X_2, \ldots, X_n) \rightarrow P(X_1, X_2, \ldots, X_n) \) as \( T \rightarrow \infty \)

with enough data, you recover the true model

- Problematic for sparse data:

\[
\begin{align*}
P_{\text{ML}} (X_i = x | \text{pa}_i = \pi) &= 0 \quad \text{if} \quad \text{count}(X_i = x, \text{pa}_i = \pi) = 0 \\
P_{\text{ML}} (X_i = x | \text{pa}_i = \pi) &= \text{undefined} \quad \text{if} \quad \text{count}(\text{pa}_i = \pi) = 0
\end{align*}
\]
Other useful notation:

Indicator function:

\[ I(x, x') = \begin{cases} 0 & \text{if } x \neq x' \\ 1 & \text{if } x = x' \end{cases} \]

\[
\text{count}(p_{ai} = \pi) = \sum_{t=1}^{T} I(p_{ai}^{(t)}, \pi)
\]

\[
\text{count}(X_i = x, p_{ai} = \pi) = \sum_{t=1}^{T} I(p_{ai}^{(t)}, \pi) I(X_i^{(t)}, x)
\]