Bayesian belief networks

Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square

Midterm exam

Midterm Thursday, March 2, 2017
• in-class (75 minutes)
• closed book

• No programing questions
• Know/understand well what is in lecture notes and assignments
• Know key ideas/principle behind different models, learning algorithms
• Know principles for ML and Error function optimization
• Know how to differentiate simple functions
• True/false questions will require justification
Density estimation

Data: \( D = \{D_1, D_2, ..., D_n\} \)
\( D_i = x_i \) a vector of attribute values

Objective: try to estimate the underlying true probability distribution over variables \( \mathbf{X} \), \( p(\mathbf{X}) \), using examples in \( D \)

Standard (iid) assumptions: Samples
- are independent of each other
- come from the same (identical) distribution (fixed \( p(\mathbf{X}) \))

Modeling complex distributions

Question: How to model and learn complex multivariate distributions \( \hat{p}(\mathbf{X}) \) with a large number of variables?

Example: modeling of disease – symptoms relations
- Disease: pneumonia
- Patient symptoms (findings, lab tests):
  - Fever, Cough, Paleness, WBC (white blood cells) count, Chest pain, etc.
- Model of the full joint distribution:
  \( P(\text{Pneumonia, Fever, Cough, Paleness, WBC, Chest pain}) \)

One probability per assignment of values to variables:
\( P(\text{Pneumonia=T, Fever=T, Cough=T, WBC=High, Chest pain=T}) \)

- How many probabilities are there?
Full joint distribution

- Any joint probability over a subset of variables can be obtained via marginalization from the full joint

\[ P(\text{Pneumonia}, \text{WBCcount}, \text{Fever}) = \sum_{c, p = \{T, F\}} P(\text{Pneumonia}, \text{WBCcount}, \text{Fever}, \text{Cough} = c, \text{Paleness} = p) \]

- **Question:** Is it possible to recover the full joint from the joint probabilities over a subset of variables?

Joint probabilities

- Is it possible to recover the full joint from the joint probabilities over a subset of variables?

\[ \text{P}(\text{pneumonia, WBCcount}) \quad 2 \times 3 \text{ matrix} \]

<table>
<thead>
<tr>
<th>Pneumonia</th>
<th>high</th>
<th>normal</th>
<th>low</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>False</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

\[ \text{P}(\text{WBCcount}) = \begin{bmatrix} 0.005 & 0.993 & 0.002 \end{bmatrix} \]

\[ \text{P}(\text{Pneumonia}) = \begin{bmatrix} 0.001 & 0.999 \end{bmatrix} \]
Joint probabilities and independence

- Is it possible to recover the full joint from the joint probabilities over a subset of variables?
- Only if the variables are independent !!!

\[
P(pneumonia, WBC\text{count}) \quad 2 \times 3 \text{ matrix}
\]

\[
Pneumonia \quad \begin{array}{c|ccc} 
\text{high} & \text{normal} & \text{low} \\
\hline 
\text{True} & ? & ? & ? \\
\text{False} & ? & ? & ? \\
\end{array}
\]

\[
P(WBC\text{count})
\]

\[
P(Pneumonia)
\]

\[
0.005 \quad 0.993 \quad 0.002
\]

Variable independence

- The two events A, B are said to be independent if:
  \[P(A, B) = P(A)P(B)\]

- The variables X, Y are said to be independent if their joint can be expressed as a product of marginals:
  \[P(X, Y) = P(X)P(Y)\]
Conditional probability

Conditional probability:
- Probability of A given B
  \[ P(A \mid B) = \frac{P(A, B)}{P(B)} \]

- Conditional probability is defined in terms of joint probabilities
- Joint probabilities can be expressed in terms of conditional probabilities
  \[ P(A, B) = P(A \mid B)P(B) \text{ (product rule)} \]
  \[ P(X_1, X_2, \ldots X_n) = \prod_{i=1}^{n} P(X_i \mid X_1, \ldots X_{i-1}) \text{ (chain rule)} \]

- Conditional probability – is useful for various probabilistic inferences
  \[ P(\text{Pneumonia} = \text{True} \mid \text{Fever} = \text{True}, \text{WBC count} = \text{high}, \text{Cough} = \text{True}) \]

Conditional probabilities

Conditional probability
- Is defined in terms of the joint probability:
  \[ P(A \mid B) = \frac{P(A, B)}{P(B)} \text{ s.t. } P(B) \neq 0 \]

- Example:
  \[ P(\text{pneumonia} = \text{true} \mid \text{WBC count} = \text{high}) = \]
  \[ \frac{P(\text{pneumonia} = \text{true}, \text{WBC count} = \text{high})}{P(\text{WBC count} = \text{high})} \]
  \[ P(\text{pneumonia} = \text{false} \mid \text{WBC count} = \text{high}) = \]
  \[ \frac{P(\text{pneumonia} = \text{false}, \text{WBC count} = \text{high})}{P(\text{WBC count} = \text{high})} \]
Conditional probabilities

Conditional probability distribution
- Defines probabilities for all possible assignments of values to target variables, given a fixed assignment of other variable values

\[ P(Pneumonia = true | WBCcount = high) \]

\[ P(Pneumonia | WBCcount) \] 3 element vector of 2 elements

<table>
<thead>
<tr>
<th>Pneumonia</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>0.08</td>
<td>0.92</td>
</tr>
<tr>
<td>normal</td>
<td>0.0001</td>
<td>0.9999</td>
</tr>
<tr>
<td>low</td>
<td>0.0001</td>
<td>0.9999</td>
</tr>
</tbody>
</table>

Variable we condition on \[ P(Pneumonia = true | WBCcount = high) \] + \[ P(Pneumonia = false | WBCcount = high) \]

Inference

Any query can be computed from the full joint distribution !!!
- Joint over a subset of variables is obtained through marginalization

\[ P(A = a, C = c) = \sum_i \sum_j P(A = a, B = b_i, C = c, D = d_j) \]

- Conditional probability over a set of variables, given other variables’ values is obtained through marginalization and definition of conditionals

\[ P(D = d | A = a, C = c) = \frac{P(A = a, C = c, D = d)}{P(A = a, C = c)} \]

\[ = \frac{\sum_i P(A = a, B = b_i, C = c, D = d_j)}{\sum_i \sum_j P(A = a, B = b_i, C = c, D = d_j)} \]
**Inference**

Any joint probability can be expressed as a product of conditionals via the **chain rule**.

\[
P(X_1, X_2, \ldots, X_n) = P(X_n \mid X_1, \ldots, X_{n-1}) P(X_{n-1} \mid X_1, \ldots, X_{n-2}) \cdots P(X_1 \mid X_2)
\]

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

**Why this may be important?**

- It is often easier to define the distribution in terms of conditional probabilities:
  - E.g. \( P(\text{Fever} \mid \text{Pneumonia} = T) \)
  - \( P(\text{Fever} \mid \text{Pneumonia} = F) \)

---

**Probabilistic inference**

Various probabilistic inference tasks:

- **Diagnostic task.** (from effect to cause)
  \( P(\text{Pneumonia} \mid \text{Fever} = T) \)

- **Prediction task.** (from cause to effect)
  \( P(\text{Fever} \mid \text{Pneumonia} = T) \)

- **Other probabilistic queries** (queries on joint distributions).
  \( P(\text{Fever}) \)
  \( P(\text{Fever}, \text{Chest Pain}) \)
Modeling complex distributions

• Defining the **full joint distribution** makes it possible to represent and reason with the probabilities
• We are able to handle an arbitrary inference problem

Problems:
– **Space complexity.** To store a full joint distribution we need to remember $O(d^n)$ numbers.
  \[ n \text{ – number of random variables, } d \text{ – number of values} \]
– **Inference (time) complexity.** To compute some queries requires $O(d^n)$ steps.
– **Acquisition problem.** How to acquire/learn all these probabilities?

Pneumonia example

• **Space complexity.**
  – Pneumonia (2 values: T,F), Fever (2: T,F), Cough (2: T,F), WBCcount (3: high, normal, low), paleness (2: T,F)
  – Number of assignments: 2*2*2*3*2=48
  – We need to define at least 47 probabilities.

• **Time complexity.**
  – Assume we need to compute the marginal of Pneumonia=T from the full joint

\[
P(Pneumonia = T) = \sum_{i \in T,F} \sum_{j \in T,F} \sum_{k = h,n,l} \sum_{u \in T,F} P(Fever = i, Cough = j, WBCcount = k, Pale = u)
\]
  – Sum over: 2*2*3*2=24 combinations
Bayesian belief networks (BBNs)

Bayesian belief networks (late 80s, beginning of 90s)
- Give solutions to the space, acquisition bottlenecks
- Partial solutions for time complexities

Key features:
- Represent the full joint distribution over the variables more compactly with a smaller number of parameters.
- Take advantage of conditional and marginal independences among random variables
- X and Y are independent
  \[ P(X, Y) = P(X)P(Y) \]
- X and Y are conditionally independent given Z
  \[ P(X, Y | Z) = P(X | Z)P(Y | Z) \]
  \[ P(X | Y, Z) = P(X | Z) \]

Alarm system example

- Assume your house has an alarm system against burglary. You live in the seismically active area and the alarm system can get occasionally set off by an earthquake. You have two neighbors, Mary and John, who do not know each other. If they hear the alarm they call you, but this is not guaranteed.
- We want to represent the probability distribution of events:
  - Burglary, Earthquake, Alarm, Mary calls and John calls

Causal relations
Bayesian belief network

1. Directed acyclic graph
   • Nodes = random variables
     Burglary, Earthquake, Alarm, Mary calls and John calls
   • Links = direct (causal) dependencies between variables.
     The chance of Alarm being is influenced by Earthquake,
     The chance of John calling is affected by the Alarm

2. Local conditional distributions
   • relating variables and their parents
Bayesian belief network

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>E</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T</td>
<td>T</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>F</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>T</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>F</td>
<td>0.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>P(B)</th>
<th>P(E)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.001</td>
<td>0.002</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>0.05</td>
</tr>
</tbody>
</table>

|       | P(A|B,E) |       |
|-------|----------|-------|
|       | T | T | 0.95  |
|       | T | F | 0.94  |
|       | F | T | 0.29  |
|       | F | F | 0.001 |

|       | P(J|A) | P(M|A) |
|-------|------|-------|
|       | T | 0.7   |
|       | F | 0.01  |

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Full joint distribution in BBNs

**Full joint distribution** is defined in terms of local conditional distributions (obtained via the chain rule):

\[
P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | pa(X_i))
\]

Example:

Assume the following assignment of values to random variables

\(B = T, E = T, A = T, J = T, M = F\)

Then its probability is:

\[
\]
Bayesian belief networks (BBNs)

Bayesian belief networks
- Represent the full joint distribution over the variables more compactly using the product of local conditionals.
- But how did we get to local parameterizations?

Answer:
- Graphical structure encodes conditional and marginal independences among random variables
- A and B are independent \[ P(A, B) = P(A)P(B) \]
- A and B are conditionally independent given C
  \[ P(A \mid C, B) = P(A \mid C) \]
  \[ P(A, B \mid C) = P(A \mid C)P(B \mid C) \]
- The graph structure implies the decomposition !!!

Independences in BBNs

3 basic independence structures:

1. Burglary
   - Alarm
     - JohnCalls
2. Burglary
   - Earthquake
   - Alarm
     - JohnCalls
3. Alarm
   - MaryCalls
Independences in BBNs

1. JohnCalls is independent of Burglary given Alarm

\[
P(J \mid A, B) = P(J \mid A) \\
P(J, B \mid A) = P(J \mid A)P(B \mid A)
\]

Independences in BBNs

2. Burglary is independent of Earthquake (not knowing Alarm) 
Burglary and Earthquake become dependent given Alarm !!

\[
P(B, E) = P(B)P(E)
\]
Independences in BBNs

1. Burglary
   \[ \text{Alarm} \]
   \[ \text{JohnCalls} \]

2. Burglary
   \[ \text{Earthquake} \]
   \[ \text{Alarm} \]
   \[ \text{JohnCalls} \]
   \[ \text{MaryCalls} \]

3. MaryCalls is independent of JohnCalls given Alarm
   \[
P(J \mid A, M) = P(J \mid A)
   \]
   \[
P(J, M \mid A) = P(J \mid A)P(M \mid A)
   \]

Independence in BBN

- BBN distribution models many conditional independence relations relating distant variables and sets
- These are defined in terms of the graphical criterion called d-separation
- D-separation in the graph
  - Let X, Y and Z be three sets of nodes
  - If X and Y are d-separated by Z then X and Y are conditionally independent given Z
- D-separation:
  - A is d-separated from B given C if every undirected path between them is blocked with C
- Path blocking
  - 3 cases that expand on three basic independence structures
Undirected path blocking

A is d-separated from B given C if every undirected path between them is **blocked**.
Undirected path blocking

A is d-separated from B given C if every undirected path between them is blocked

• 1. Path blocking with a linear substructure

• 2. Path blocking with the wedge substructure
Undirected path blocking

A is d-separated from B given C if every undirected path between them is blocked

• 3. Path blocking with the vee substructure

\[
\begin{align*}
X & \quad \quad \quad \quad \quad \quad Y \\
X \quad Z & \quad \quad \quad \quad \quad \quad Y \\
& \quad \quad \quad \quad \quad \quad Z \text{ or any of its descendants not in } C
\end{align*}
\]

Independences in BBNs

\[
\begin{align*}
\text{Burglary} & \quad \quad \text{Earthquake} \\
\text{Alarm} & \quad \quad \text{RadioReport} \\
\text{JohnCalls} & \quad \quad \text{MaryCalls}
\end{align*}
\]

• Earthquake and Burglary are independent given MaryCalls

?
Independences in BBNs

• Earthquake and Burglary are independent given MaryCalls  F
• Burglary and MaryCalls are independent (not knowing Alarm)  ?

Independences in BBNs

• Earthquake and Burglary are independent given MaryCalls  F
• Burglary and MaryCalls are independent (not knowing Alarm)  F
• Burglary and RadioReport are independent given Earthquake  ?
Independences in BBNs

- Earthquake and Burglary are independent given MaryCalls
- Burglary and MaryCalls are independent (not knowing Alarm)
- Burglary and RadioReport are independent given Earthquake
- Burglary and RadioReport are independent given MaryCalls

Independences in BBNs

- Earthquake and Burglary are independent given MaryCalls
- Burglary and MaryCalls are independent (not knowing Alarm)
- Burglary and RadioReport are independent given Earthquake
- Burglary and RadioReport are independent given MaryCalls
Rewrite the full joint probability using the product rule:

\[ P(B = T, E = T, A = T, J = T, M = F) = \]

**Product rule**

\[ = P(J = T \mid B = T, E = T, A = T, M = F)P(B = T, E = T, A = T, M = F) \]
Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:

\[ P(B = T, E = T, A = T, J = T, M = F) = \]

\[ = \frac{P(J = T | B = T, E = T, A = T, M = F)}{P(B = T, E = T, A = T, M = F)} P(B = T, E = T, A = T, M = F) \]

\[ = P(J = T | A = T) P(B = T, E = T, A = T, M = F) \]

---

CS 2750 Machine Learning
Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:

\[ P(B = T, E = T, A = T, J = T, M = F) = \]

\[ = P(J = T \mid B = T, E = T, A = T, M = F)P(B = T, E = T, A = T, M = F) \]

\[ = P(J = T \mid A = T)P(B = T, E = T, A = T, M = F) \]

\[ \underline{P(M = F \mid B = T, E = T, A = T)}P(B = T, E = T, A = T) \]

\[ \underline{P(M = F \mid A = T)}P(B = T, E = T, A = T) \]

Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:

\[ P(B = T, E = T, A = T, J = T, M = F) = \]

\[ = P(J = T \mid B = T, E = T, A = T, M = F)P(B = T, E = T, A = T, M = F) \]

\[ = P(J = T \mid A = T)P(B = T, E = T, A = T, M = F) \]

\[ \underline{P(M = F \mid B = T, E = T, A = T)}P(B = T, E = T, A = T) \]

\[ \underline{P(M = F \mid A = T)}P(B = T, E = T, A = T) \]

\[ \underline{P(A = T \mid B = T, E = T)}P(B = T, E = T) \]
Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:

\[ P(B = T, E = T, A = T, J = T, M = F) = \]

\[ = P(J = T \mid B = T, E = T, A = T, M = F)P(B = T, E = T, A = T, M = F) \]

\[ = P(J = T \mid A = T)P(B = T, E = T, A = T, M = F) \]

\[ P(M = F \mid B = T, E = T, A = T)P(B = T, E = T, A = T) \]

\[ P(M = F \mid A = T)P(B = T, E = T, A = T) \]

\[ P(A = T \mid B = T, E = T)P(B = T, E = T) \]

\[ P(B = T)P(E = T) \]
Parameter complexity problem

• In the BBN the full joint distribution is defined as:
  \[ P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | pa(X_i)) \]

• What did we save?

Alarm example: binary (True, False) variables

# of parameters of the full joint: ?

Parameter complexity problem

• In the BBN the full joint distribution is defined as:
  \[ P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | pa(X_i)) \]

• What did we save?

Alarm example: binary (True, False) variables

# of parameters of the full joint: ?

One parameter is for free:

\[ 2^5 - 1 = 31 \]

# of parameters of the BBN: ?
Bayesian belief network: parameters count

Parameter complexity problem

- In the BBN the **full joint distribution** is defined as:
  \[
P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | \text{pa}(X_i))
\]
- What did we save?
  Alarm example: 5 binary (True, False) variables
  
  **# of parameters of the full joint:**
  \[2^5 = 32\]
  One parameter is for free:
  \[2^5 - 1 = 31\]
  **# of parameters of the BBN:**
  \[2^3 + 2(2^2) + 2(2) = 20\]
  One parameter in every conditional is for free: ?
Bayesian belief network: free parameters

<table>
<thead>
<tr>
<th>Variable</th>
<th>P(B)</th>
<th>P(E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td>F</td>
<td>0.999</td>
<td>0.998</td>
</tr>
</tbody>
</table>

Parameter complexity problem

- In the BBN the **full joint distribution** is defined as:
  \[ P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid pa(X_i)) \]

- **What did we save?**

Alarm example: 5 binary (True, False) variables

# of parameters of the full joint:

\[ 2^5 = 32 \]

One parameter is for free:

\[ 2^5 - 1 = 31 \]

# of parameters of the BBN:

\[ 2^3 + 2(2^2) + 2(2) = 20 \]

One parameter in every conditional is for free:

\[ 2^2 + 2(2) + 2(1) = 10 \]