Bayesian belief networks

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

Density estimation

Data: \( D = \{D_1, D_2, \ldots, D_n\} \)
\[ D_i = x_i \quad \text{a vector of attribute values} \]

Attributes:
- modeled by random variables \( X = \{X_1, X_2, \ldots, X_d\} \) with:
  - Continuous values
  - Discrete values
E.g. blood pressure with numerical values
  or chest pain with discrete values
  [no-pain, mild, moderate, strong]

Underlying true probability distribution:
\[ p(X) \]
Density estimation

Data: \( D = \{D_1, D_2, \ldots, D_n\} \)
\( D_i = x_i \) 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 \( p(\mathbf{X}) \))

Learning via parameter estimation

In this lecture we consider parametric density estimation

Basic settings:
- A set of random variables \( \mathbf{X} = \{X_1, X_2, \ldots, X_d\} \)
- A model of the distribution over variables in \( \mathbf{X} \) with parameters \( \Theta \):
  \[ \hat{p}(\mathbf{X} | \Theta) \]
- Data \( D = \{D_1, D_2, \ldots, D_n\} \)

Objective: find the parameters \( \Theta \) that explain best the observed data
Parameter estimation

• **Maximum likelihood (ML)**
  - maximize $p(D | \Theta, \xi)$
  - yields: one set of parameters $\Theta_{ML}$
  - the target distribution is approximated as:
    $$\hat{p}(X) = p(X | \Theta_{ML})$$

• **Bayesian parameter estimation**
  - uses the posterior distribution over possible parameters
    $$p(\Theta | D, \xi) = \frac{p(D | \Theta, \xi) p(\Theta | \xi)}{p(D | \xi)}$$
  - Yields: all possible settings of $\Theta$ (and their “weights”)
  - The target distribution is approximated as:
    $$\hat{p}(X) = p(X | D) = \int p(X | \Theta) p(\Theta | D, \xi) d\Theta$$

Parameter estimation.

Other possible criteria:

• **Maximum a posteriori probability (MAP)**
  - maximize $p(\Theta | D, \xi)$ (mode of the posterior)
  - Yields: one set of parameters $\Theta_{MAP}$
  - Approximation:
    $$\hat{p}(X) = p(X | \Theta_{MAP})$$

• **Expected value of the parameter**
  - $\hat{\Theta} = E(\Theta)$ (mean of the posterior)
  - Expectation taken with regard to posterior $p(\Theta | D, \xi)$
  - Yields: one set of parameters
  - Approximation:
    $$\hat{p}(X) = p(X | \hat{\Theta})$$
Density estimation

- So far we have covered density estimation for “simple” distribution models:
  - Bernoulli
  - Binomial
  - Multinomial
  - Gaussian
  - Poisson

But what if:
- The dimension of \( X = \{X_1, X_2, \ldots, X_d\} \) is large
  - Example: patient data
- Compact parametric distributions do not seem to fit the data
  - E.g.: multivariate Gaussian may not fit
- We have only a “small” number of examples to do accurate parameter estimates

How to learn complex distributions

How to learn complex multivariate distributions \( \hat{p}(X) \) with large number of variables?

One solution:
- Decompose the distribution along conditional independence relations
- Decompose the parameter estimation problem to a set of smaller parameter estimation tasks

Decomposition of distributions under conditional independence assumption is the main idea behind Bayesian belief networks
Example

**Problem description:**
- **Disease:** pneumonia
- **Patient symptoms (findings, lab tests):**
  - Fever, Cough, Paleness, WBC (white blood cells) count, Chest pain, etc.

**Representation of a patient case:**
- Symptoms and disease are represented as random variables

**Our objectives:**
- Describe a multivariate distribution representing the relations between symptoms and disease
- Design of inference and learning procedures for the multivariate model

---

### Joint probability distribution

**Joint probability distribution (for a set variables)**
- Defines probabilities for all possible assignments to values of variables in the set

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

\[ P(\text{Pneumonia}) \]

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

- Marginalization (summing of rows, or columns) - summing out variables

\[ \begin{array}{c|ccc}
    \text{Pneumonia} & \text{high} & \text{normal} & \text{low} \\
    \hline
    \text{True} & 0.0008 & 0.0001 & 0.0001 \\
    \text{False} & 0.0042 & 0.9929 & 0.0019 \\
    \hline
    & 0.005 & 0.993 & 0.002 \\
\end{array} \]

\[ P(\text{Pneumonia}) = 0.001 \quad 0.999 \]

\[ P(\text{WBC count}) = 0.001 \quad 0.999 \]
Variable independence

- The joint distribution over a subset of variables can be always computed from the joint distribution through marginalization
- Not the other way around !!!
  - Only exception: when variables are independent

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

Pneumonia & WBCcount & P(WBCcount) & P(Pneumonia)
\hline
True & high & 0.0008 & 0.0001 & 0.0001 & 0.001 & 0.999
False & normal & 0.0042 & 0.9929 & 0.0019
 & low & 0.005 & 0.993 & 0.002
\hline

Conditional probability

Conditional probability:
- Probability of A given B

\[ P(A | B) = \frac{P(A, B)}{P(B)} \]
- Conditional probability is defined in terms of joint probabilities
- Joint probabilities can be expressed in terms of conditional probabilities

\[ P(A, B) = P(A | B)P(B) \quad \text{(product rule)} \]

\[ P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1}) \quad \text{(chain rule)} \]
- Conditional probability – is useful for various probabilistic inferences

\[ P(\text{Pneumonia} = \text{True} | \text{Fever} = \text{True}, \text{WBCcount} = \text{high}, \text{Cough} = \text{True}) \]
Modeling uncertainty with probabilities

- **Full joint distribution:**
  - joint distribution over all random variables that define the domain
  - it is sufficient to do any type of probabilistic inferences

### Inference

Any query can be computed from the full joint distribution !!!

- **Joint over a subset of variables** is obtained through marginalization

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

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

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

Any query can be computed from the full joint distribution !!!

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

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

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

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

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

Modeling uncertainty with probabilities

- **Full joint distribution**: joint distribution over all random variables defining the domain
  - it is sufficient to represent the complete domain and to do any type of probabilistic inferences

Problems:

- **Space complexity**. To store full joint distribution requires to remember \( O(d^n) \) numbers.
  - \( n \) – number of random variables, \( d \) – number of values
- **Inference complexity**. To compute some queries requires \( O(d^n) \) steps.
- **Acquisition problem**. Who is going to define all of the probability entries?
Pneumonia example. Complexities.

• **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 probability of Pneumonia=T from the full joint

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

Bayesian belief networks (BBNs)

**Bayesian belief networks.**

• 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

• **A and B are independent**

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

• **A and B are conditionally independent given C**

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

\[
P(A \mid C, B) = P(A \mid C)
\]
**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**

  ![Causal relations diagram]

**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

  ![Bayesian belief network diagram]
Bayesian belief network.

2. Local conditional distributions
   • relate variables and their parents
Bayesian belief networks (general)

Two components: \( B = (S, \Theta_S) \)

- **Directed acyclic graph**
  - Nodes correspond to random variables
  - (Missing) links encode independences

- **Parameters**
  - Local conditional probability distributions for every variable-parent configuration

\[
P(A | B, E)
\]

\[
P(A | B, E)
\]

\[
P(A | B, E)
\]

<table>
<thead>
<tr>
<th>B</th>
<th>E</th>
<th>T</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>0.95</td>
<td>0.05</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>0.94</td>
<td>0.06</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>0.29</td>
<td>0.71</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>0.001</td>
<td>0.999</td>
</tr>
</tbody>
</table>

Then its probability is:

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

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
   - JohnCalls
   - MaryCalls
1. JohnCalls is independent of Burglary given Alarm

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

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

3. MaryCalls is independent of JohnCalls given Alarm

\[
P(J \mid A, M) = P(J \mid A)\\
P(J, M \mid A) = P(J \mid A)P(M \mid A)
\]
Undirected path blocking

- 1. With linear substructure

- 2. With wedge substructure

- 3. With vee substructure

Independences in BBNs

- 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  \( F \)
- Burglary and MaryCalls are independent (not knowing Alarm)  \( F \)
- Burglary and RadioReport are independent given Earthquake  \( T \)
- Burglary and RadioReport are independent given MaryCalls  \(?\)

CS 2750 Machine Learning

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  \( T \)
- Burglary and RadioReport are independent given MaryCalls  \( F \)

CS 2750 Machine Learning
Full joint distribution in BBNs

Rewrite the full joint probability using the product rule:

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

\[ = P(J=T | B=T, E=T, A=T, M=F)P(B=T, E=T, A=T, M=F) \]
\[ = P(J=T | A=T)P(B=T, E=T, A=T, M=F) \]
Rewrite the full joint probability using the product rule:

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

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

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

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

\[
P(M=\text{F} \mid A=\text{T})P(B=\text{T}, E=\text{T}, A=\text{T})
\]
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 expressed as a product of conditionals (of smaller) complexity

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

Parameters:
full joint: 2⁵ = 32
BBN: 2³ + 2(2²) + 2(2) = 20

Parameters to be defined:
full joint: 2⁵ − 1 = 31
BBN: 2² + 2(2) + 2(1) = 10
Model acquisition problem

The structure of the BBN typically reflects causal relations
- BBNs are also sometime referred to as causal networks
- Causal structure is very intuitive in many applications domain
  and it is relatively easy to obtain from the domain expert

Probability parameters of BBN correspond to conditional
  distributions relating a random variable and its parents only
- Their complexity much smaller than the full joint
- Easier to come up (estimate) the probabilities from expert or
  automatically by learning from data

BBNs built in practice

- In various areas:
  - Intelligent user interfaces (Microsoft)
  - Troubleshooting, diagnosis of a technical device
  - Medical diagnosis:
    - Pathfinder (Intellipath)
    - CPSC
    - Munin
    - QMR-DT
  - Collaborative filtering
  - Military applications
  - Insurance, credit applications
Diagnosis of car engine

- Diagnose the engine start problem

Car insurance example

- Predict claim costs (medical, liability) based on application data
(ICU) Alarm network

CPCS

- Computer-based Patient Case Simulation system (CPCS-PM) developed by Parker and Miller (at University of Pittsburgh)
- 422 nodes and 867 arcs
QMR-DT

- **Medical diagnosis in internal medicine**

  Bipartite network of disease/findings relations

  QMR-DT derived from Internist-1/QMR KB

  534 diseases

  40740 arcs

  4040 findings