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

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

true distribution \( p(X) \) \( \rightarrow \) n samples \( D = \{D_1, D_2, \ldots, D_n\} \) \( \rightarrow \) estimate \( \hat{p}(X) \)

Standard (iid) assumptions: Samples
• are independent of each other
• come from the same (identical) distribution (fixed \( p(X) \))
Learning via parameter estimation

In this lecture we consider **parametric density estimation**

**Basic settings:**

- A set of random variables \( X = \{X_1, X_2, \ldots, X_d \} \)
- **A model of the distribution** over variables in \( X \) with parameters \( \Theta \):
  \[
  \hat{p}(X | \Theta)
  \]
- **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 \( p(D \mid \Theta, \xi) \)

  – yields: one set of parameters \( \Theta_{ML} \)

  – the target distribution is approximated as:

  \[ \hat{p}(X) = p(X \mid \Theta_{ML}) \]

• **Bayesian parameter estimation**

  – uses the posterior distribution over possible parameters

  \[ p(\Theta \mid D, \xi) = \frac{p(D \mid \Theta, \xi) p(\Theta \mid \xi)}{p(D \mid \xi)} \]

  – Yields: all possible settings of \( \Theta \) (and their “weights”)

  – The target distribution is approximated as:

  \[ \hat{p}(X) = p(X \mid D) = \int p(X \mid \Theta) p(\Theta \mid D, \xi)d\Theta \]
Parameter estimation

Other possible criteria:

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

- **Expected value of the parameter**
  
  \[
  \hat{\Theta} = E(\Theta) \quad \text{(mean of the posterior)}
  \]
  
  - 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 using 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
Modeling uncertainty with probabilities

- **Full joint distribution:**
  - Assume \( X = \{X_1, X_2, \ldots, X_d\} \) are all random variables that define the domain
  - Full joint: \( P(X) \) or \( P(X_1, X_2, \ldots, X_d) \)

**Full joint it is sufficient** to do any type of probabilistic inference:

- Computation of joint probabilities for sets of variables
  \[
P(X_1, X_2, X_3) \quad P(X_1, X_{10})
\]
- Computation of conditional probabilities
  \[
P(X_1 \mid X_2 = True, X_3 = False)
\]
## Marginalization

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

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

<table>
<thead>
<tr>
<th>Pneumonia</th>
<th>high</th>
<th>normal</th>
<th>low</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>0.0008</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>False</td>
<td>0.0042</td>
<td>0.9929</td>
<td>0.0019</td>
</tr>
</tbody>
</table>

\[
P(WBCcount) \quad \text{Marginalization (summing of rows, or columns)}
\]
- Summing out variables

\[
P(Pneumonia) \quad 0.001 \quad 0.999
\]
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

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

<table>
<thead>
<tr>
<th>Pneumonia</th>
<th>high</th>
<th>normal</th>
<th>low</th>
<th>P(Pneumonia)</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>0.0008</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.001</td>
</tr>
<tr>
<td>False</td>
<td>0.0042</td>
<td>0.9929</td>
<td>0.0019</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td>0.005</td>
<td>0.993</td>
<td>0.002</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>WBCcount</th>
<th>P(WBCcount)</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>0.005</td>
</tr>
<tr>
<td>normal</td>
<td>0.993</td>
</tr>
<tr>
<td>low</td>
<td>0.002</td>
</tr>
</tbody>
</table>
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

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

\[ P(X_1, X_2, \ldots X_n) = \prod_{i=1}^{n} P(X_i \mid X_1, \ldots X_{i-1}) \quad \text{(chain rule)} \]

• Conditional probability – is useful for various probabilistic inferences

\[ P(\text{Pneumonia} = \text{True} \mid \text{Fever} = \text{True}, \text{WBC count} = \text{high}, \text{Cough} = \text{True}) \]
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

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

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

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

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

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

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

• 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), WBC count (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

\[ P(\text{Pneumonia} = T) = \sum_{i \in T,F} \sum_{j \in T,F} \sum_{k = h,n,l} \sum_{u \in T,F} P(\text{Fever} = i, \text{Cough} = j, \text{WBC count} = k, \text{Pale} = u) \]

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