

**CS 2750 Machine Learning
Lecture 14**

Bayesian belief networks

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

Term projects

Goal: focus on an analysis of a nontrivial ML problem

- **Can be:**
 - **Application of various ML methods to nontrivial data**
 - **A theoretical ML problem**
 - **Proposal and testing of a new ML method**
 - **Term project:**
 - **Groups (2 students)**
 - **If somebody wants to work alone please let me know**
-

Term projects

ML datasets:

- your own (collected, or extracted)
- www.kaggle.com
 - A good source of ML datasets used in various ML and data mining competitions
 - **Example:** <https://www.kaggle.com/c/instacart-market-basket-analysis/data>
- other websites and resources

Note: If your project focuses on the application of ML to existing data please avoid UCI repository datasets (they are too simple)

Term project proposals

Proposals due date: Thursday, March 15, 2018

- 1 page long

Written proposal:

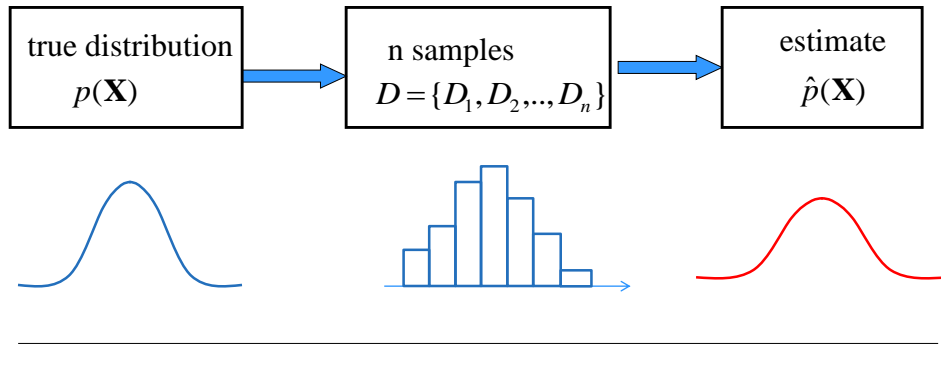
1. Outline of a learning problem, type of data you have available. Why is the problem important?
2. Learning methods you plan to start with for the problem. References to previous work.
3. How do you plan to test, compare solutions
4. Brief schedule of work (approximate timeline of work)

Project due date: April 26, 2018

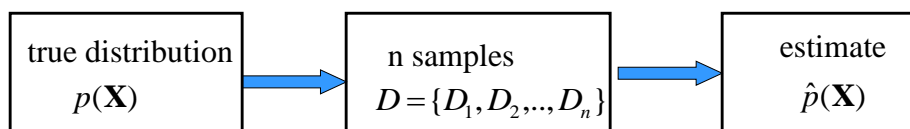
Density estimation

Data: $D = \{D_1, D_2, \dots, D_n\}$
 $D_i = \mathbf{x}_i$ a vector of attribute values

Objective: estimate the model of the underlying probability distribution over variables \mathbf{X} , $p(\mathbf{X})$, using examples in D

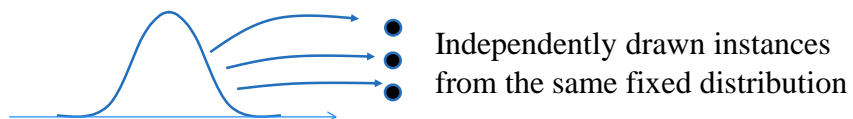


Density estimation



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 $\mathbf{X} = \{X_1, X_2, \dots, X_d\}$
- **A model of the distribution** over variables in \mathbf{X}
with parameters Θ :

$$\hat{p}(\mathbf{X} | \Theta)$$

- **Data** $D = \{D_1, D_2, \dots, D_n\}$

Objective: Find the parameters Θ that explain the observed data the best

Parameter estimation

- **Maximum likelihood (ML)**

maximize $p(D | \Theta, \xi)$

- yields: one set of parameters Θ_{ML}
- the target distribution is approximated as:

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} | \Theta_{ML})$$

- **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 Θ (and their “weights”)
- The target distribution is approximated as:

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} | D) = \int_{\Theta} p(\mathbf{X} | \Theta) p(\Theta | D, \xi) d\Theta$$

Parameter estimation

Other possible criteria:

- **Maximum a posteriori probability (MAP)**

maximize $p(\Theta | D, \xi)$ (mode of the posterior)

– Yields: one set of parameters Θ_{MAP}

– Approximation:

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} | \Theta_{MAP})$$

- **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:

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} | \hat{\Theta})$$

Distribution models

- **So far we have covered density estimation for “simple” distribution models:**

- Bernoulli
- Binomial
- Multinomial
- Gaussian
- Poisson

But what if:

- The dimension of $\mathbf{X} = \{X_1, X_2, \dots, X_d\}$ 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

Modeling complex distributions

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

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
-

Representation complexity

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:** $\hat{p}(\mathbf{X})$
 $P(\text{Pneumonia}, \text{Fever}, \text{Cough}, \text{Paleness}, \text{WBC}, \text{Chest pain})$

One probability per assignment of values to variables:

$P(\text{Pneumonia} = T, \text{Fever} = T, \text{Cough} = T, \text{WBC} = \text{High}, \text{Chest pain} = T)$

$P(\text{Pneumonia} = T, \text{Fever} = T, \text{Cough} = T, \text{WBC} = \text{High}, \text{Chest pain} = F)$

$P(\text{Pneumonia} = T, \text{Fever} = T, \text{Cough} = T, \text{WBC} = \text{Norm}, \text{Chest pain} = T)$

- **How many probabilities are there?**

Marginalization

Joint probability distribution (for a set variables)

- Defines probabilities for all possible assignments to values of variables in the set

$P(\text{pneumonia}, \text{WBCcount})$ 2×3 table

		<i>WBCcount</i>			
		<i>high</i>	<i>normal</i>	<i>low</i>	
<i>Pneumonia</i>	<i>True</i>	0.0008	0.0001	0.0001	0.001 0.999
	<i>False</i>	0.0042	0.9929	0.0019	
		0.005	0.993	0.002	

$P(\text{WBCcount})$

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

Joint distribution over a subset variables

- Full joint distribution is defined over all variables we use in the model

E.g. $P(\text{Pneumonia}, \text{Fever}, \text{Cough}, \text{Paleness}, \text{WBC}, \text{Chest pain})$

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

E.g.

$P(\text{Pneumonia}, \text{WBCcount}, \text{Fever}) =$

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

- 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?

$P(\text{pneumonia}, \text{WBCcount})$ 2×3 matrix

		WBCcount			$P(\text{Pneumonia})$
		high	normal	low	
Pneumonia	True	?	?	?	0.001 0.999
	False	?	?	?	
		0.005	0.993	0.002	

$P(\text{WBCcount})$ →

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 !!!

$\mathbf{P}(pneumonia, WBCcount)$ 2×3 matrix

		WBCcount			$\mathbf{P}(Pneumonia)$
		high	normal	low	
Pneumonia	True	?	?	?	0.001
	False	?	?	?	0.999
		0.005	0.993	0.002	

$\mathbf{P}(WBCcount)$ →

Variable independence

- The two events **A, B** are said to be independent if:

$$\mathbf{P}(A, B) = \mathbf{P}(A)\mathbf{P}(B)$$
- The variables **X, Y** are said to be independent if their joint probabilities can be expressed as a product of marginal probabilities:

$$\mathbf{P}(X, Y) = \mathbf{P}(X)\mathbf{P}(Y)$$

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

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

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i | X_1, \dots, X_{i-1}) \quad (\text{chain rule})$$

- Conditional probability – is useful for **various probabilistic inferences**

$$P(\text{Pneumonia} = \text{True} | \text{Fever} = \text{True}, \text{WBCcount} = \text{high}, \text{Cough} = \text{True})$$

Conditional probabilities

Conditional probability

- Is defined in terms of the joint probability:

$$P(A|B) = \frac{P(A, B)}{P(B)} \quad \text{s.t. } P(B) \neq 0$$

- **Example:**

$$P(\text{pneumonia} = \text{true} | \text{WBCcount} = \text{high}) =$$

$$\frac{P(\text{pneumonia} = \text{true}, \text{WBCcount} = \text{high})}{P(\text{WBCcount} = \text{high})}$$

$$P(\text{pneumonia} = \text{false} | \text{WBCcount} = \text{high}) =$$

$$\frac{P(\text{pneumonia} = \text{false}, \text{WBCcount} = \text{high})}{P(\text{WBCcount} = \text{high})}$$

Conditional probabilities

Conditional probability distribution

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

$$P(\text{Pneumonia} = \text{true} \mid \text{WBCcount} = \text{high})$$

$\mathbf{P}(\text{Pneumonia} \mid \text{WBCcount})$ 3 element vector of 2 elements

		<i>Pneumonia</i>		
		<i>True</i>	<i>False</i>	
WBCcount	<i>high</i>	0.08	0.92	1.0
	<i>normal</i>	0.0001	0.9999	1.0
	<i>low</i>	0.0001	0.9999	1.0

Variable we
condition on

$$P(\text{Pneumonia} = \text{true} \mid \text{WBCcount} = \text{high}) + P(\text{Pneumonia} = \text{false} \mid \text{WBCcount} = \text{high})$$

Inference

Any probability (joint or conditional) 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

$$\begin{aligned}
 P(D = d \mid A = a, C = c) &= \frac{P(A = a, C = c, D = d)}{P(A = a, C = c)} \\
 &= \frac{\sum_i P(A = a, B = b_i, C = c, D = d)}{\sum_i \sum_j P(A = a, B = b_i, C = c, D = d_j)}
 \end{aligned}$$

Inference

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

$$\begin{aligned}P(X_1, X_2, \dots, X_n) &= P(X_n | X_1, \dots, X_{n-1})P(X_1, \dots, X_{n-1}) \\ &= P(X_n | X_1, \dots, X_{n-1})P(X_{n-1} | X_1, \dots, X_{n-2})P(X_1, \dots, X_{n-2}) \\ &= \prod_{i=1}^n P(X_i | X_1, \dots, X_{i-1})\end{aligned}$$

Why this may be important?

- It is often easier to define the distribution in terms of conditional probabilities:

– E.g. $\mathbf{P}(\text{Fever} | \text{Pneumonia} = T)$
 $\mathbf{P}(\text{Fever} | \text{Pneumonia} = F)$

Probabilistic inference

Various probabilistic inference tasks:

- **Diagnostic task. (from effect to cause)**

$$\mathbf{P}(\text{Pneumonia} | \text{Fever} = T)$$

- **Prediction task. (from cause to effect)**

$$\mathbf{P}(\text{Fever} | \text{Pneumonia} = T)$$

- **Other probabilistic queries** (queries on joint distributions).

$$\mathbf{P}(\text{Fever})$$

$$\mathbf{P}(\text{Fever}, \text{ChestPain})$$

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 – number of random variables, d – 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

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

- 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**

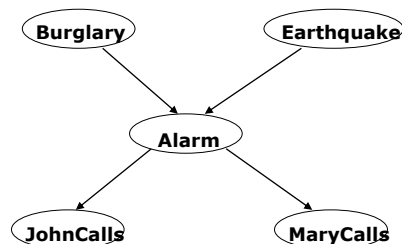
$$P(X, Y | Z) = P(X | Z)P(Y | Z)$$

$$P(X | Y, Z) = P(X | Z)$$

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

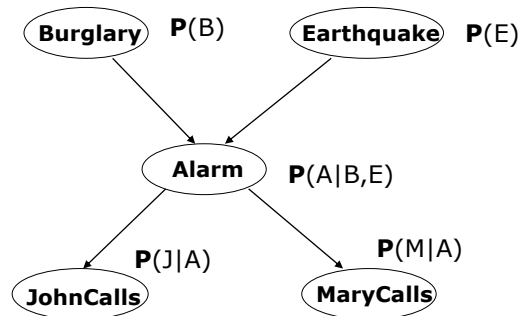
Causal relations



Bayesian belief network

1. Directed acyclic graph

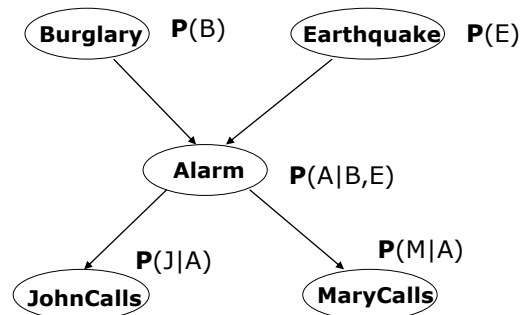
- **Nodes** = random variables
Burglary, Earthquake, Alarm, Mary calls and John calls
- **Links** = direct (causal) dependencies between variables.
The chance of Alarm being is influenced by Earthquake,
The chance of John calling is affected by the Alarm



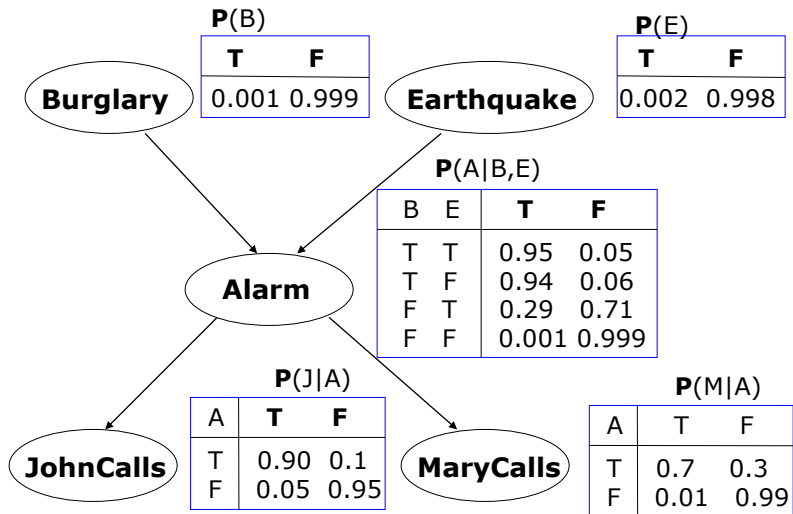
Bayesian belief network

2. Local conditional distributions

- relating variables and their parents



Bayesian belief network



Full joint distribution in BBNs

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

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1, \dots, n} P(X_i \mid pa(X_i))$$

Example:

Assume the following assignment of values to random variables

$$B = T, E = T, A = T, J = T, M = F$$

Then its probability is:

$$P(B = T, E = T, A = T, J = T, M = F) =$$

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

