CS 2750 Machine Learning
Lecture 12

Bayesian belief networks

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

Midterm exam

When: Wednesday, March 2, 2011

Midterm is:

- **In-class (75 minutes)**
- **closed book**
- **material covered during the semester including lecture today**

Project proposals

Due: **Wednesday, March 16, 2011**

- **1 page long**

Proposal

- **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 try and implement for the problem. References to previous work.
  3. How do you plan to test, compare learning approaches
  4. Schedule of work (approximate timeline of work)

Project proposals

Where to find the data:

- From your research
- UC Irvine data repository
- Various text document repositories
- I have some bioinformatics data I can share but other data can be found on the NIH or various university web sites

  (e.g. microarray data, proteomic data)
- Synthetic data that are generated to demonstrate your algorithm works

Project proposals

Problems to address:

- Get the ideas for the project by browsing the web
- It is tempting to go with simple classification but definitely try to add some complexity to your investigations
- Multiple, not just one method, try some more advanced methods, say those that combine multiple classifiers to learn a model (ensemble methods) or try to modify the existing methods

Project proposals

Interesting problems to consider:

- Advanced methods for learning multi-class problems
- Learning the parameters and structure of Bayesian Belief networks
- Clustering of data – how to group examples
- Dimensionality reduction/feature selection – how to deal with a large number of inputs
- Learning how to act – Reinforcement learning
- Anomaly detection – how to identify outliers in data

Density estimation

Data: $D = \{D_1, D_2, .., D_n\}$

$D_i = \mathbf{x}_i$ a vector of attribute values

Attributes:

- modeled by random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$ 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:

$p(\mathbf{X})$

Density estimation

Data: $D = \{D_1, D_2, .., D_n\}$

$D_i = \mathbf{x}_i$ 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

true distribution $p(\mathbf{X})$	→	n samples $D = \{D_1, D_2, .., D_n\}$	→	estimate $\hat{p}(\mathbf{X})$

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 $\boldsymbol{X}$ with parameters Θ :

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

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

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

Parameter estimation

- **Maximum likelihood (ML)**

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

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

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \boldsymbol{\Theta}_{ML})$$

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

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

Parameter estimation

Other possible criteria:

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

  maximize $p(\mathbf{\Theta} \mid D, \xi)$ (mode of the posterior)

  – Yields: one set of parameters $\mathbf{\Theta}_{MAP}$

  – Approximation:

  $$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \mathbf{\Theta}_{MAP})$$

- **Expected value of the parameter**

  $\hat{\mathbf{\Theta}} = E(\mathbf{\Theta})$ (mean of the posterior)

  – Expectation taken with regard to posterior $p(\mathbf{\Theta} \mid D, \xi)$

  – Yields: one set of parameters

  – Approximation:

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

Density estimation

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

But what if:

- The dimension of $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$ 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 $\hat{p}(\mathbf{X})$ 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 $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$ are all random variables that define the domain
  - Full joint: $P(\mathbf{X})$ 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

  $P(X_1, X_2, X_3)$ $\qquad P(X_1, X_{10})$
- 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

$\mathbf{P}(pneumonia, WBCcount)$ $\quad 2 \times 3$ table

WBCcount

		high	*normal*	*low*	$\mathbf{P}(Pneumonia)$
Pneumonia	*True*	0.0008	0.0001	0.0001	0.001
	False	0.0042	0.9929	0.0019	0.999
$\mathbf{P}(WBCcount)$		0.005	0.993	0.002	

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

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**

$$P(A,B) = P(A)P(B)$$

$\mathbf{P}(pneumonia, WBCcount)$

		$WBCcount$			$\mathbf{P}(Pneumonia)$
		$high$	$normal$	low	
$Pneumonia$	$True$	0.0008	0.0001	0.0001	0.001
	$False$	0.0042	0.9929	0.0019	0.999
$\mathbf{P}(WBCcount)$		0.005	0.993	0.002	

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

$$P(A,B) = P(A \mid B)P(B) \quad \textbf{(product rule)}$$

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

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

$$P(Pneumonia = True \mid Fever = True, WBCcount = high, Cough = True)$$

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

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

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

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. $\mathbf{P}(Fever \mid Pneumonia=T)$
    $\mathbf{P}(Fever \mid Pneumonia=F)$

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

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

  - Sum over 2*2*3*2=24 combinations

Bayesian belief networks (BBNs)

Bayesian belief networks.

- 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**

$$P(A,B) = P(A)P(B)$$

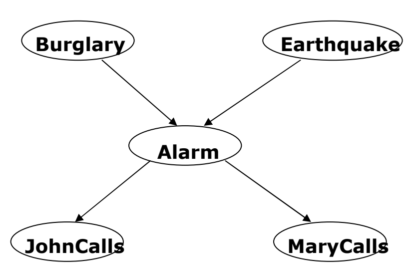
- **A and B are conditionally independent given C**

$$P(A,B \mid C) = P(A \mid C)P(B \mid C)$$
$$P(A \mid C,B) = P(A \mid C)$$

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

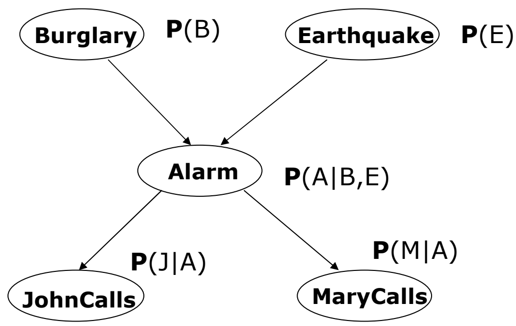
Causal relations



Bayesian belief network

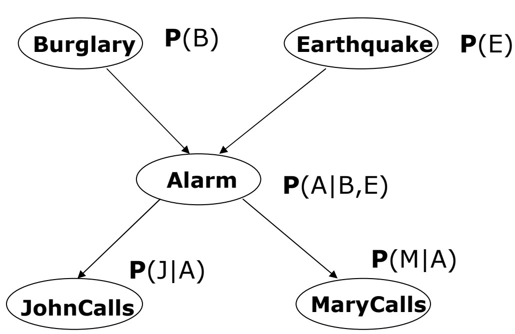
1. **Directed acyclic graph**
   - **Nodes** = random variables
     Burglary, Earthquake, Alarm, Mary calls and John calls
   - **Links** = direct (causal) dependencies between variables.

     The chance of Alarm being is influenced by Earthquake,
     The chance of John calling is affected by the Alarm

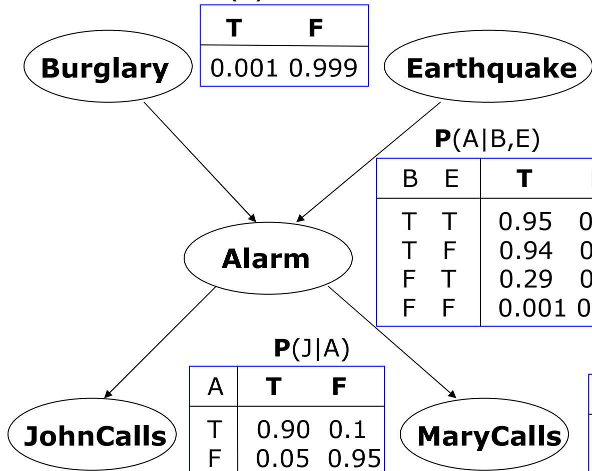


Bayesian belief network

2. **Local conditional distributions**
   - relate variables and their parents



Bayesian belief network



P(B)

T	**F**
0.001	0.999

P(E)

T	**F**
0.002	0.998

P(A|B,E)

B	E	**T**	**F**
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

P(J|A)

A	**T**	**F**
T	0.90	0.1
F	0.05	0.95

P(M|A)

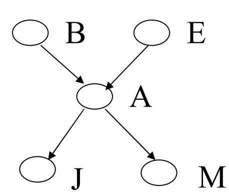
A	T	F
T	0.7	0.3
F	0.01	0.99

Full joint distribution in BBNs

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

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

Example:



Assume the following assignment of values to random variables

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

Then its probability is:

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

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

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**

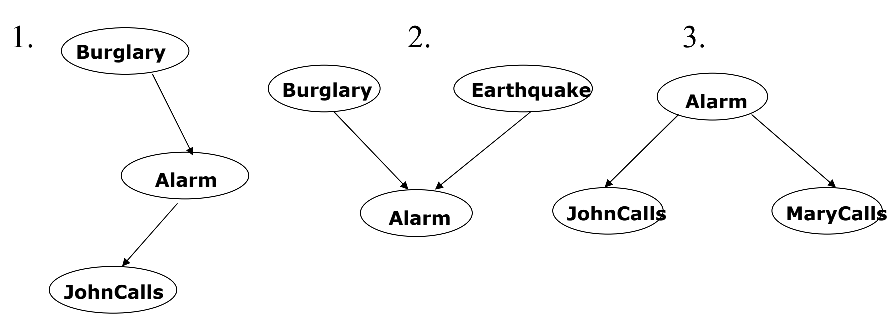
$$P(A \mid C,B) = P(A \mid C)$$

$$P(A,B \mid C) = P(A \mid C)P(B \mid C)$$

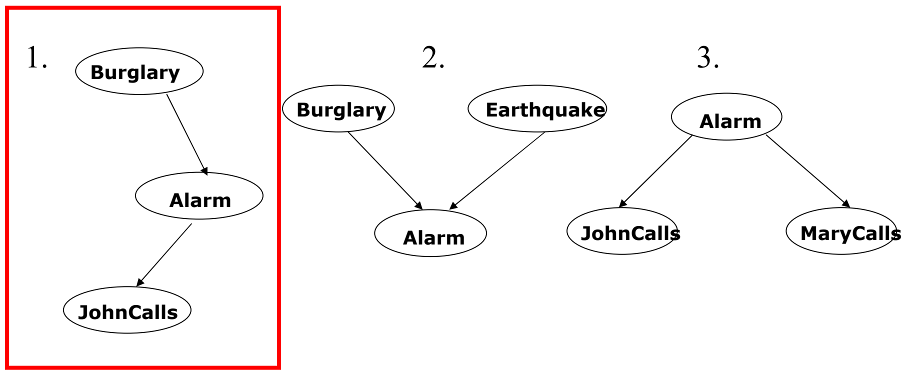
- **<u>The graph structure implies the decomposition !!!</u>**

Independences in BBNs

3 basic independence structures:



Independences in BBNs

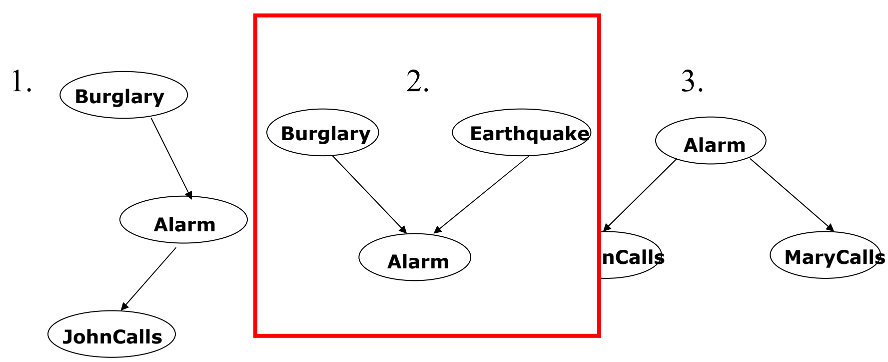


1. JohnCalls **is independent** of Burglary given Alarm

$$P(J \mid A, B) = P(J \mid A)$$

$$P(J, B \mid A) = P(J \mid A) P(B \mid A)$$

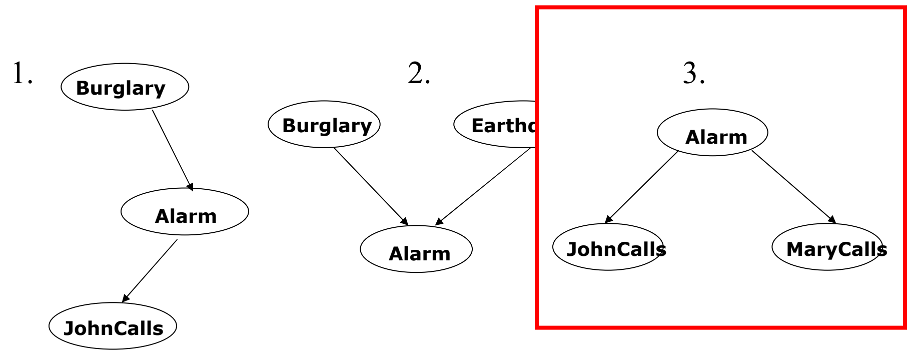
Independences in BBNs



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



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**
- **Path blocking**
  – 3 cases that expand on three basic independence structures