CS 1675 Intro to Machine Learning
Lecture 17

Bayesian belief networks

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

Midterm exam: reminder

Midterm exam

- **Thursday, March 7, 2018**
- **In-class**
- **Closed book**
- **Material covered by the end of last week**

Density estimation

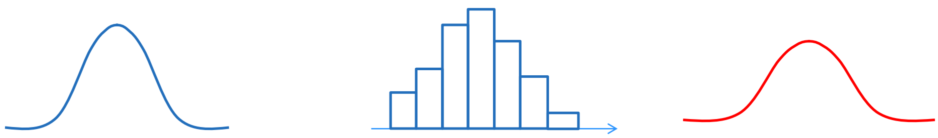
Data: $D = \{D_1, D_2, .., D_n\}$

$D_i = \mathbf{x}_i$ 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

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

$\rightarrow$ ... $\rightarrow$

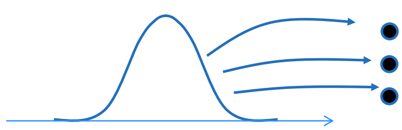


Density estimation

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 $p(\mathbf{X})$)**



Independently drawn instances
from the same fixed distribution

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 ***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 the observed data the best

Parameter estimation

- **Maximum likelihood (ML)**
  maximize $p(D \mid \Theta, \xi)$
  - yields: one set of parameters Θ_{ML}
  - the target distribution is approximated as:
$$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \mathbf{\Theta}_{ML})$$
- **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:
$$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid D) = \int_{\mathbf{\Theta}} p(X \mid \mathbf{\Theta}) p(\mathbf{\Theta} \mid D, \xi) d\mathbf{\Theta}$$

Parameter estimation

Other possible criteria:

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

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

  - Yields: one set of parameters $\boldsymbol{\Theta}_{MAP}$
  - Approximation:

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

- **Expected value of the parameter**

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

  - Expectation taken with regard to posterior $p(\boldsymbol{\Theta} \mid D, \xi)$
  - Yields: one set of parameters
  - Approximation:

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

Distribution models

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

But what if:

- The dimension of $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$ 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 relatively "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 inference and learning procedures for the multivariate model

Representation complexity

Example: modeling of disease – symptoms relations

- **Disease:** pneumonia (T/F)
- **Patient symptoms (findings, lab tests)**:
  - Fever (T/F) Cough (T/F), Paleness (T/F), WBC (white blood cells) count (High/Normal/Low), Chest pain (T/G), etc.

Model of the full joint distribution: $\hat{p}(\mathbf{X})$

P(Pneumonia, Fever, Cough, Paleness, WBC, Chest pain)

Representation complexity

P(Pneumonia, Fever, Cough, Paleness, WBC, Chest pain

Pneumonia	Fever	Cough	Paleness	WBC	Chest pain	Probability
T	T	T	T	High	T	0.02
T	T	T	T	High	F	0.005
T	T	T	T	Normal	T	0.004
…						

- **How many probabilities are there?**

Representation complexity

P(Pneumonia, Fever, Cough, Paleness, WBC, Chest pain

Pneumonia	Fever	Cough	Paleness	WBC	Chest pain	Probability
T	T	T	T	High	T	0.02
T	T	T	T	High	F	0.005
T	T	T	T	Normal	T	0.004
…						

- **How many probabilities are there?** 2^5*3 =32*3=96
  $O(a^k)$ where k is the number of variables

Marginalization

Joint probability distribution (for a set variables)

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

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

Joint distribution over a subset variables

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

  E.g. **P**(Pneumonia, Fever, Cough, Paleness, WBC, Chest pain)

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

E.g.

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

- **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?**

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

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

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

$\mathbf{P}(Pneumonia)$

		WBCcount			
		high	*normal*	*low*	$\mathbf{P}(Pneumonia)$
Pneumonia	*True*	?	?	?	0.001
	False	?	?	?	0.999
$\mathbf{P}(WBCcount)$		0.005	0.993	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 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 $$P(A \mid B) = \frac{P(A,B)}{P(B)}$$
- 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**

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

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:**

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

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

Conditional probabilities

Conditional probability distribution

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

$$P(Pneumonia = true \mid WBCcount = high)$$

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

Pneumonia

	True	*False*	
high	0.08	0.92	1.0
normal	0.0001	0.9999	1.0
low	0.0001	0.9999	$\boxed{1.0}$

$\boxed{WBCcount}$

Variable we condition on

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

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

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

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

Inference

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

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

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

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

Why this may be important?

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

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

Probabilistic inference

Various probabilistic inference tasks:

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

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

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

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

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

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

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

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 $O(d^k)$ numbers.

  k – number of random variables, *d* – number of values
- **Inference (time) complexity.** To compute some queries requires $O(d^k)$ 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), Chest-pain (2:T,F)
  - Number of assignments: 2*2*2*3*2*2=96
  - We need to define at least 95 probabilities.
- **Time complexity**.
  - Assume we need to compute the marginal of Pneumonia=T from the full joint

$$P(Pneumonia = T) =$$

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

  - Sum over: 2*2*3*2*2=48 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**

$$P(X,Y \mid Z) = P(X \mid Z)P(Y \mid Z)$$

$$P(X \mid Y,Z) = P(X \mid Z)$$

Alarm system example

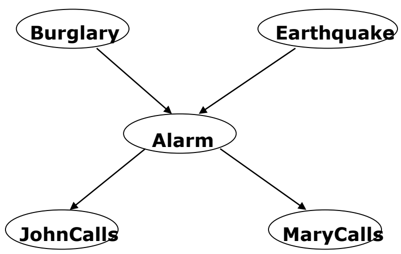
Story: Assume your house has an **alarm system** against **burglary**. You live in the seismically active area and the alarm system can get occasionally set off by an **earthquake**. You have two neighbors, **Mary** and **John**, who do not know each other. If they hear the alarm they call you, but this is not guaranteed.

We want to represent the relations among the events:

- Burglary, Earthquake, Alarm, Mary calls and John calls

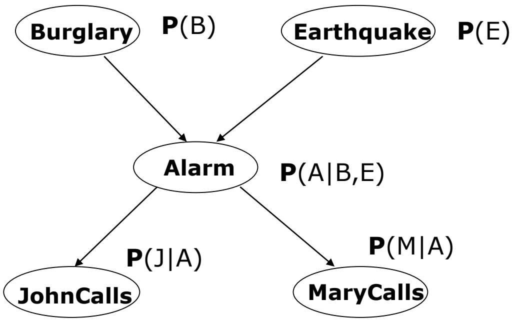
From the story we can extract (typically causal) relations among the events

Causal relations



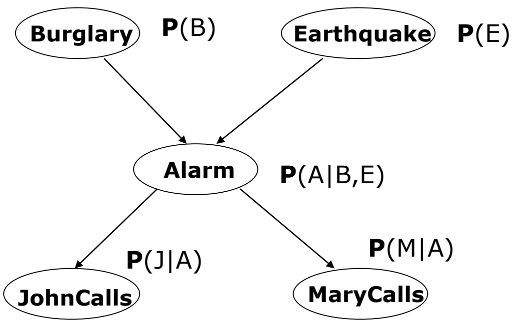
Bayesian belief network

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

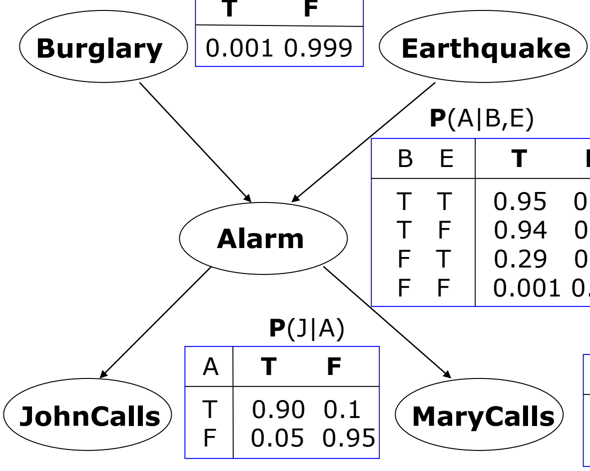


Bayesian belief network

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



Bayesian belief network



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)

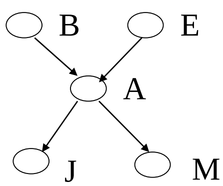
A	T	F
T	0.7	0.3
F	0.01	0.99

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:

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

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