CS 1675 Introduction to Machine Learning
Lecture 13b

Bayesian belief networks

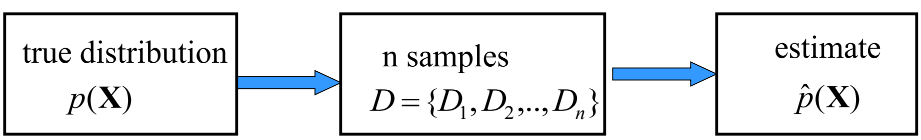
Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square

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



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 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:

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

- **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:
    - the target distribution is approximated as:

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

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 "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?

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 distribution**:

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

One probability per assignment of value to variables:

P(Pneumonia =T, Fever =T, Cought=T, WBC=High, Chest pain=T)

Modeling complex distributions

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

Complication: **Full joint:** $\hat{p}(\mathbf{X})$

- may easily become very complex and hard to learn

Solution: **Bayesian belief networks**

- More compact representation
- Relies on independences among variables/attributes

Preliminaries

Probabilities

$$P(Pneumonia = True) = 0.001$$

$$P(Pneumonia = False) = 0.999$$

$$P(WBCcount = high) = 0.005$$

Probability distribution

- Defines probabilities **for all possible value assignments to a random variable**
- Values are mutually exclusive

$$P(Pneumonia = True) = 0.001$$
$$P(Pneumonia = False) = 0.999$$

$Pneumonia$	$\mathbf{P}(Pneumonia)$
$True$	0.001
$False$	0.999

Probability distribution

Defines probability for **all possible value assignments**

Example 1:

$P(Pneumonia = True) = 0.001$

$P(Pneumonia = False) = 0.999$

$Pneumonia$	$\mathbf{P}(Pneumonia)$
$True$	0.001
$False$	0.999

$$P(Pneumonia = True) + P(Pneumonia = False) = 1$$

Probabilities sum to 1 !!!

Example 2:

$P(WBCcount = high) = 0.005$

$P(WBCcount = normal) = 0.993$

$P(WBCcount = high) = 0.002$

$WBCcount$	$\mathbf{P}(WBCcount)$
$high$	0.005
$normal$	0.993
low	0.002

Joint probability distribution

Joint probability distribution (for a set variables)

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

Example: variables *Pneumonia* and *WBCcount*

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

Is represented by 2×3 array(matrix)

Pneumonia × *WBCcount*

	$high$	$normal$	low
$True$	0.0008	0.0001	0.0001
$False$	0.0042	0.9929	0.0019

Joint probability distribution

Joint probability distribution (for a set variables)

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

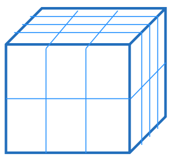
Example 2: Assume variables:

Pneumonia (2 values)

WBCcount (3 values)

Pain (4 values)

$\mathbf{P}(pneumonia, WBCcount, Pain)$ is represented by $2 \times 3 \times 4$ array



Example of an entry in the array

$P(pneumonia = T, WBCcount = high, Pain = severe)$

Joint probabilities: marginalization

Marginalization

- reduces the dimension of the joint distribution
- Sums variables out

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

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 (here summing of columns or rows)

Marginalization

Marginalization

- reduces the dimension of the joint distribution

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

- We can continue doing this

$$P(X_1, \ldots X_{n-2}) = \sum_{\{X_{n-1}, X_n\}} P(X_1, X_2, \ldots X_{n-1}, X_n)$$

What is the maximal joint probability distribution?

- **Full joint probability**

Full joint distribution

- **the joint distribution for all variables in the problem**
  – It defines the complete probability model for the problem

Example: pneumonia diagnosis

- **Variables:** *Pneumonia, Fever, Paleness, WBCcount, Cough*
- Full joint probability: P(*Pneumonia, Fever, Paleness, WBCcount, Cough)*
  – defines the probability for all possible assignments of values to these variables

$$P(Pneumonia = T, WBCcount = High, Fever = T, Cough = T, Paleness = T)$$

$$P(Pneumonia = T, WBCcount = High, Fever = T, Cough = T, Paleness = F)$$

$$P(Pneumonia = T, WBCcount = High, Fever = T, Cough = F, Paleness = T)$$

… etc

- **How many probabilities are there?**

Full joint distribution

- **the joint distribution for all variables in the problem**
  - It defines the complete probability model for the problem

Example: pneumonia diagnosis

Variables: *Pneumonia, Fever, Paleness, WBCcount, Cough*

Full joint probability: P(*Pneumonia, Fever, Paleness, WBCcount, Cough)*

- defines the probability for all possible assignments of values to these variables

$$P(Pneumonia = T, WBCcount = High, Fever = T, Cough = T, Paleness = T)$$

$$P(Pneumonia = T, WBCcount = High, Fever = T, Cough = T, Paleness = F)$$

$$P(Pneumonia = T, WBCcount = High, Fever = T, Cough = F, Paleness = T)$$

… etc

- **How many probabilities are there?** Exponential in the number of variables

Full joint distribution

- **Any joint probability over a subset of variables can be obtained via marginalization**

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

- **Is it possible to recover the full joint from the joint probabilities over a subset of variables?**