

Statistical Learning (part II)

July 5, 2005
CS 486/686
University of Waterloo

Outline

- Learning from complete Data
 - EM algorithm
- Reading: R&N Ch 20.3

CS486/686 Lecture Slides (c) 2005 P. Pasupat

2

Incomplete data

- So far...
 - Values of all attributes are known
 - Learning is relatively easy
- But many real-world problems have **hidden variables** (a.k.a **latent variables**)
 - Incomplete data
 - Values of some attributes missing

CS486/686 Lecture Slides (c) 2005 P. Pasupat

3

Unsupervised Learning

- Incomplete data → unsupervised learning
- Examples:
 - Categorisation of stars by astronomers
 - Categorisation of species by anthropologists
 - Market segmentation for marketing
 - Pattern identification for fraud detection
 - Research in general!

CS486/686 Lecture Slides (c) 2005 P. Pasupat

4

Maximum Likelihood Learning

- ML learning of Bayes net parameters:
 - For $\theta_{V=\text{true},pa(V)=v} = \Pr(V=\text{true} | \text{par}(V) = v)$
 - $\theta_{V=\text{true},pa(V)=v} = \frac{\#[V=\text{true},pa(V)=v]}{\#[V=\text{true},pa(V)=v] + \#[V=\text{false},pa(V)=v]}$
 - Assumes all attributes have values...
- What if values of some attributes are missing?

CS486/686 Lecture Slides (c) 2005 P. Pasupat

5

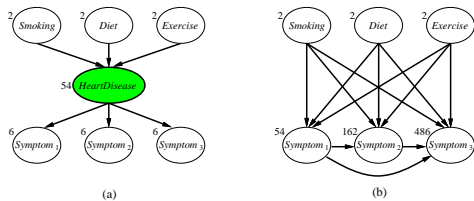
"Naive" solutions for incomplete data

- Solution #1: Ignore records with missing values
 - But what if all records are missing values (i.e., when a variable is hidden, none of the records have any value for that variable)
- Solution #2: Ignore hidden variables
 - Model may become significantly more complex!

CS486/686 Lecture Slides (c) 2005 P. Pasupat

6

Heart disease example



- a) simpler (i.e., fewer CPT parameters)
- b) complex (i.e., lots of CPT parameters)

CS486/686 Lecture Slides (c) 2005 P. Poupart

7

"Direct" maximum likelihood

- Solution 3: maximize likelihood directly
 - Let \mathbf{Z} be hidden and \mathbf{E} observable
 - $h_{ML} = \operatorname{argmax}_h P(\mathbf{e}|h)$
 - $= \operatorname{argmax}_h \sum_{\mathbf{Z}} P(\mathbf{e}, \mathbf{Z}|h)$
 - $= \operatorname{argmax}_h \sum_{\mathbf{Z}} \prod_i \text{CPT}(V_i)$
 - $= \operatorname{argmax}_h \log \sum_{\mathbf{Z}} \prod_i \text{CPT}(V_i)$
- Problem: can't push log past sum to linearize product

CS486/686 Lecture Slides (c) 2005 P. Poupart

8

Expectation-Maximization (EM)

- Solution #4: EM algorithm
 - Intuition: if we knew the missing values, computing h_{ML} would be trivial
- Guess h_{ML}
- Iterate
 - **Expectation**: based on h_{ML} , compute expectation of the missing values
 - **Maximization**: based on expected missing values, compute new estimate of h_{ML}

CS486/686 Lecture Slides (c) 2005 P. Poupart

9

Expectation-Maximization (EM)

- More formally:
 - Approximate maximum likelihood
 - Iteratively compute:

$$h_{i+1} = \operatorname{argmax}_h \underbrace{\sum_{\mathbf{Z}} P(\mathbf{Z}|h_i, \mathbf{e}) \log P(\mathbf{e}, \mathbf{Z}|h)}_{\text{Expectation}}$$

$$\underbrace{\hspace{10em}}_{\text{Maximization}}$$

CS486/686 Lecture Slides (c) 2005 P. Poupart

10

Expectation-Maximization (EM)

- Derivation
 - $\log P(\mathbf{e}|h) = \log [P(\mathbf{e}, \mathbf{Z}|h) / P(\mathbf{Z}|\mathbf{e}, h)]$
 - $= \log P(\mathbf{e}, \mathbf{Z}|h) - \log P(\mathbf{Z}|\mathbf{e}, h)$
 - $= \sum_{\mathbf{Z}} P(\mathbf{Z}|\mathbf{e}, h) \log P(\mathbf{e}, \mathbf{Z}|h)$
 - $= \sum_{\mathbf{Z}} P(\mathbf{Z}|\mathbf{e}, h) \log P(\mathbf{Z}|\mathbf{e}, h)$
 - $\geq \sum_{\mathbf{Z}} \tilde{P}(\mathbf{Z}|\mathbf{e}, h) \log P(\mathbf{e}, \mathbf{Z}|h)$
- EM finds a local maximum of $\sum_{\mathbf{Z}} \tilde{P}(\mathbf{Z}|\mathbf{e}, h) \log P(\mathbf{e}, \mathbf{Z}|h)$ which is a lower bound of $\log P(\mathbf{e}|h)$

CS486/686 Lecture Slides (c) 2005 P. Poupart

11

Expectation-Maximization (EM)

- Log inside sum can linearize product
 - $h_{i+1} = \operatorname{argmax}_h \sum_{\mathbf{Z}} P(\mathbf{Z}|h_i, \mathbf{e}) \log P(\mathbf{e}, \mathbf{Z}|h)$
 - $= \operatorname{argmax}_h \sum_{\mathbf{Z}} P(\mathbf{Z}|h_i, \mathbf{e}) \log \prod_j \text{CPT}_j$
 - $= \operatorname{argmax}_h \sum_{\mathbf{Z}} P(\mathbf{Z}|h_i, \mathbf{e}) \sum_j \log \text{CPT}_j$
- Monotonic improvement of likelihood
 - $P(\mathbf{e}|h_{i+1}) \geq P(\mathbf{e}|h_i)$

CS486/686 Lecture Slides (c) 2005 P. Poupart

12

Candy Example

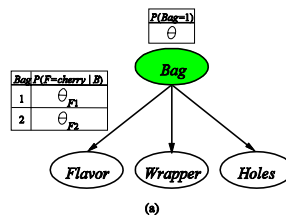
- Suppose you buy two bags of candies of unknown type (e.g. flavour ratios)
- You plan to eat sufficiently many candies of each bag to learn their type
- Ignoring your plan, your roommate mixes both bags...
- How can you learn the type of each bag despite being mixed?

CS486/686 Lecture Slides (c) 2005 P. Pasupat

13

Candy Example

- "Bag" variable is hidden

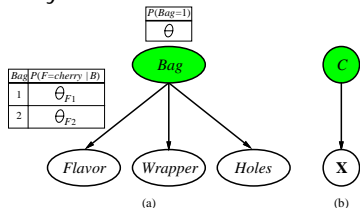


CS486/686 Lecture Slides (c) 2005 P. Pasupat

14

Unsupervised Clustering

- "Class" variable is hidden
- Naïve Bayes model



CS486/686 Lecture Slides (c) 2005 P. Pasupat

15

Candy Example

- Unknown Parameters:
 - $\theta_i = P(\text{Bag}=i)$
 - $\theta_{F_i} = P(\text{Flavour}=\text{cherry}|\text{Bag}=i)$
 - $\theta_{W_i} = P(\text{Wrapper}=\text{red}|\text{Bag}=i)$
 - $\theta_{H_i} = P(\text{Hole}=\text{yes}|\text{Bag}=i)$
- When eating a candy:
 - F, W and H are observable
 - B is hidden

CS486/686 Lecture Slides (c) 2005 P. Pasupat

16

Candy Example

- Let true parameters be:
 - $\theta=0.5, \theta_{F_1}=\theta_{W_1}=\theta_{H_1}=0.8, \theta_{F_2}=\theta_{W_2}=\theta_{H_2}=0.3$
- After eating 1000 candies:

	W=red		W=green	
	H=1	H=0	H=1	H=0
F=cherry	273	93	104	90
F=lime	79	100	94	167

CS486/686 Lecture Slides (c) 2005 P. Pasupat

17

Candy Example

- EM algorithm
- Guess h_0 :
 - $\theta=0.6, \theta_{F_1}=\theta_{W_1}=\theta_{H_1}=0.6, \theta_{F_2}=\theta_{W_2}=\theta_{H_2}=0.4$
- Alternate:
 - Expectation: expected # of candies in each bag
 - Maximization: new parameter estimates

CS486/686 Lecture Slides (c) 2005 P. Pasupat

18

Candy Example

- Expectation: expected # of candies in each bag
 - $\#[\text{Bag}=i] = \sum_j P(B=i|f_j, w_j, h_j)$
 - Compute $P(B=i|f_j, w_j, h_j)$ by variable elimination (or any other inference alg.)
- Example:
 - $\#[\text{Bag}=1] = 612$
 - $\#[\text{Bag}=2] = 388$

CS486/686 Lecture Slides (c) 2005 P. Poupart

19

Candy Example

- Maximization: relative frequency of each bag
 - $\theta_1 = 612/1000 = 0.612$
 - $\theta_2 = 388/1000 = 0.388$

CS486/686 Lecture Slides (c) 2005 P. Poupart

20

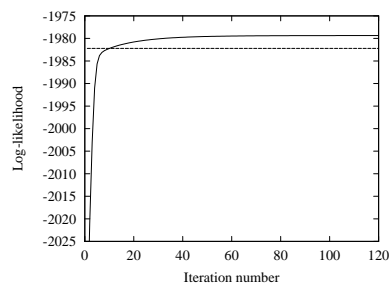
Candy Example

- Expectation: expected # of cherry candies in each bag
 - $\#[B=i, F=\text{cherry}] = \sum_j P(B=i|f_j=\text{cherry}, w_j, h_j)$
 - Compute $P(B=i|f_j=\text{cherry}, w_j, h_j)$ by variable elimination (or any other inference alg.)
- Maximization:
 - $\theta_{F1} = \#[B=1, F=\text{cherry}] / \#[B=1] = 0.668$
 - $\theta_{F2} = \#[B=2, F=\text{cherry}] / \#[B=2] = 0.389$

CS486/686 Lecture Slides (c) 2005 P. Poupart

21

Candy Example



CS486/686 Lecture Slides (c) 2005 P. Poupart

22

Bayesian networks

- EM algorithm for general Bayes nets
- Expectation:
 - $\#[V_i=v_{ij}, \text{Pa}(V_i)=\text{pa}_{ik}] = \text{expected frequency}$
- Maximization:
 - $\theta_{v_{ij}, \text{pa}_{ik}} = \#[V_i=v_{ij}, \text{Pa}(V_i)=\text{pa}_{ik}] / \#[\text{Pa}(V_i)=\text{pa}_{ik}]$

CS486/686 Lecture Slides (c) 2005 P. Poupart

23

Next Class

- Next Class:
 - Neural networks
 - Russell and Norvig Sect. 20.5

CS486/686 Lecture Slides (c) 2005 P. Poupart

24