

Readings: 18.1, 18.2, 18.3

EM for BNs

Graphical Models – 10708
Carlos Guestrin
Carnegie Mellon University

November 24th, 2008

10-708 – ©Carlos Guestrin 2006-2008

1

Thus far, fully supervised learning

- We have assumed fully supervised learning:
- Many real problems have missing data:

10-708 – ©Carlos Guestrin 2006-2008

2

The general learning problem with missing data

- Marginal likelihood – \mathbf{x} is observed, \mathbf{z} is missing:

$$\begin{aligned}\ell(\mathcal{D} : \theta) &= \log \prod_{j=1}^m P(x^{(j)} | \theta) \\ &= \sum_{j=1}^m \log P(x^{(j)} | \theta) \\ &= \sum_{j=1}^m \log \sum_z P(z, x^{(j)} | \theta)\end{aligned}$$

10-708 – ©Carlos Guestrin 2006-2008

3

E-step

- \mathbf{x} is observed, \mathbf{z} is missing
- Compute probability of missing data given current choice of θ
 - $Q(\mathbf{z} | \mathbf{x}^{(i)})$ for each $\mathbf{x}^{(i)}$
 - e.g., probability computed during classification step
 - corresponds to “classification step” in K-means

$$Q^{(t+1)}(z | x^{(j)}) = P(z | x^{(j)}, \theta^{(t)})$$

10-708 – ©Carlos Guestrin 2006-2008

4

Jensen's inequality

$$\ell(\mathcal{D} : \theta) = \sum_{j=1}^m \log \sum_z P(z, x^{(j)} | \theta)$$

■ **Theorem:** $\log \sum_z P(z) f(z) \geq \sum_z P(z) \log f(z)$

Applying Jensen's inequality

■ Use: $\log \sum_z P(z) f(z) \geq \sum_z P(z) \log f(z)$

$$\ell(\mathcal{D} : \theta^{(t)}) = \sum_{j=1}^m \log \sum_z Q^{(t+1)}(z | x^{(j)}) \frac{P(z, x^{(j)} | \theta^{(t)})}{Q^{(t+1)}(z | x^{(j)})}$$

The M-step maximizes lower bound on weighted data

- Lower bound from Jensen's:

$$\ell(\mathcal{D} : \theta^{(t)}) \geq \sum_{j=1}^m \sum_z Q^{(t+1)}(z | x^{(j)}) \log P(z, x^{(j)} | \theta^{(t)}) + H(Q^{(t+1)})$$

- Corresponds to weighted dataset:

- ☐ $\langle \mathbf{x}^{(1)}, \mathbf{z}=1 \rangle$ with weight $Q^{(t+1)}(\mathbf{z}=1 | \mathbf{x}^{(1)})$
- ☐ $\langle \mathbf{x}^{(1)}, \mathbf{z}=2 \rangle$ with weight $Q^{(t+1)}(\mathbf{z}=2 | \mathbf{x}^{(1)})$
- ☐ $\langle \mathbf{x}^{(1)}, \mathbf{z}=3 \rangle$ with weight $Q^{(t+1)}(\mathbf{z}=3 | \mathbf{x}^{(1)})$
- ☐ $\langle \mathbf{x}^{(2)}, \mathbf{z}=1 \rangle$ with weight $Q^{(t+1)}(\mathbf{z}=1 | \mathbf{x}^{(2)})$
- ☐ $\langle \mathbf{x}^{(2)}, \mathbf{z}=2 \rangle$ with weight $Q^{(t+1)}(\mathbf{z}=2 | \mathbf{x}^{(2)})$
- ☐ $\langle \mathbf{x}^{(2)}, \mathbf{z}=3 \rangle$ with weight $Q^{(t+1)}(\mathbf{z}=3 | \mathbf{x}^{(2)})$
- ☐ ...

10-708 – ©Carlos Guestrin 2006-2008

7

The M-step

$$\ell(\mathcal{D} : \theta^{(t)}) \geq \sum_{j=1}^m \sum_z Q^{(t+1)}(z | x^{(j)}) \log P(z, x^{(j)} | \theta^{(t)}) + H(Q^{(t+1)})$$

- Maximization step:

$$\theta^{(t+1)} \leftarrow \arg \max_{\theta} \sum_{j=1}^m \sum_z Q^{(t+1)}(z | x^{(j)}) \log P(z, x^{(j)} | \theta)$$

- Use expected counts instead of counts:

- ☐ If learning requires $\text{Count}(\mathbf{x}, \mathbf{z})$
- ☐ Use $E_{Q^{(t+1)}}[\text{Count}(\mathbf{x}, \mathbf{z})]$

10-708 – ©Carlos Guestrin 2006-2008

8

Convergence of EM

- Define potential function $F(\theta, Q)$:

$$\ell(\mathcal{D} : \theta^{(t)}) \geq F(\theta, Q) = \sum_{j=1}^m \sum_z Q(z | x^{(j)}) \log \frac{P(z, x^{(j)} | \theta)}{Q(z | x^{(j)})}$$

- EM corresponds to coordinate ascent on F
 - Thus, maximizes lower bound on marginal log likelihood
 - As seen in Machine Learning class last semester

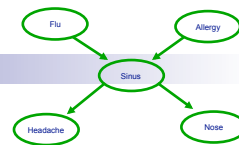
10-708 – ©Carlos Guestrin 2006-2008

9

Data likelihood for BNs

- Given structure, log likelihood of fully observed data:

$$\log P(\mathcal{D} | \theta_{\mathcal{G}}, \mathcal{G})$$



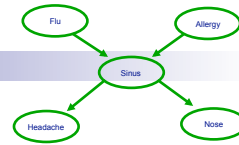
10-708 – ©Carlos Guestrin 2006

10

Marginal likelihood

- What if S is hidden?

$$\log P(\mathcal{D} \mid \theta_{\mathcal{G}}, \mathcal{G})$$



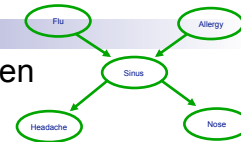
10-708 – ©Carlos Guestrin 2006

11

Log likelihood for BNs with hidden data

- Marginal likelihood – **O** is observed, **H** is hidden

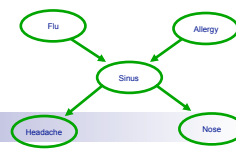
$$\begin{aligned} \ell(\theta : \mathcal{D}) &= \sum_{j=1}^m \log P(\mathbf{o}^{(j)} \mid \theta) \\ &= \sum_{j=1}^m \log \sum_{\mathbf{h}} P(\mathbf{h}, \mathbf{o}^{(j)} \mid \theta) \end{aligned}$$



10-708 – ©Carlos Guestrin 2006

12

E-step for BNs



- E-step computes probability of hidden vars \mathbf{h} given \mathbf{o}

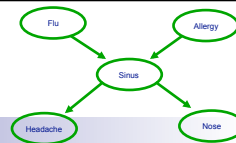
$$Q^{(t+1)}(\mathbf{h} \mid \mathbf{o}) \leftarrow P(\mathbf{h} \mid \mathbf{o}, \theta^{(t)})$$

- Corresponds to inference in BN

10-708 – ©Carlos Guestrin 2006

13

The M-step for BNs



- Maximization step:

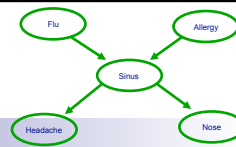
$$\theta^{(t+1)} \leftarrow \arg \max_{\theta} \sum_{j=1}^m \sum_{\mathbf{h}} Q^{(t+1)}(\mathbf{h} \mid \mathbf{o}^{(j)}) \log P(\mathbf{h}, \mathbf{o}^{(j)} \mid \theta)$$

- Use expected counts instead of counts:
 - If learning requires $\text{Count}(\mathbf{h}, \mathbf{o})$
 - Use $E_{Q^{(t+1)}}[\text{Count}(\mathbf{h}, \mathbf{o})]$

10-708 – ©Carlos Guestrin 2006

14

M-step for each CPT



■ M-step decomposes per CPT

□ Standard MLE:

$$P(X_i = x_i \mid \text{Pa}_{X_i} = \mathbf{z}) = \frac{\text{Count}(X_i = x_i, \text{Pa}_{X_i} = \mathbf{z})}{\text{Count}(\text{Pa}_{X_i} = \mathbf{z})}$$

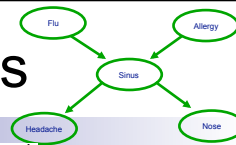
□ M-step uses expected counts:

$$P(X_i = x_i \mid \text{Pa}_{X_i} = \mathbf{z}) = \frac{\text{ExCount}(X_i = x_i, \text{Pa}_{X_i} = \mathbf{z})}{\text{ExCount}(\text{Pa}_{X_i} = \mathbf{z})}$$

10-708 – ©Carlos Guestrin 2006

15

Computing expected counts



$$P(X_i = x_i \mid \text{Pa}_{X_i} = \mathbf{z}) = \frac{\text{ExCount}(X_i = x_i, \text{Pa}_{X_i} = \mathbf{z})}{\text{ExCount}(\text{Pa}_{X_i} = \mathbf{z})}$$

■ M-step requires expected counts:

- Observe $\mathbf{O}=\mathbf{o}$
- For a set of vars \mathbf{A} , must compute $\text{ExCount}(\mathbf{A}=\mathbf{a})$
- Some of \mathbf{A} in example j will be observed
 - denote by $\mathbf{A}_O = \mathbf{a}_O^{(j)}$
- Some of \mathbf{A} will be hidden
 - denote by \mathbf{A}_H

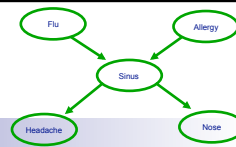
■ Use inference (E-step computes expected counts):

- $\text{ExCount}^{(t+1)}(\mathbf{A}_O = \mathbf{a}_O, \mathbf{A}_H = \mathbf{a}_H)$

10-708 – ©Carlos Guestrin 2006

16

Data need not be hidden in the same way



- When data is fully observed
 - A data point is
- When data is partially observed
 - A data point is
- But unobserved variables can be different for different data points
 - e.g.,
- Same framework, just change definition of expected counts
 - Observed vars in point j ,
 - Consider set of vars \mathbf{A}
 - $\text{ExCount}^{(t+1)}(\mathbf{A} = \mathbf{a})$

10-708 – ©Carlos Guestrin 2006

17

Poster printing

- Poster session:
 - Friday Dec 1st, 3-6pm in the NSH Atrium.
 - There will be a popular vote for best poster. Invite your friends!
 - please be ready to set up your poster at 2:45pm sharp.
- We will provide posterboards, easels and pins.
 - The posterboards are 30x40 inches
 - We don't have a specific poster format for you to use.
 - You can either bring a big poster or a print a set of regular sized pages and pin them together.
- Unfortunately, we don't have a budget to pay for printing. If you are an SCS student, SCS has a poster printer you can use which prints on a 36" wide roll of paper.
 - If you are a student outside SCS, you will need to check with your department to see if there are printing facilities for big posters (I don't know what is offered outside SCS), or print a set of regular sized pages.
- We are looking forward to a great poster session!

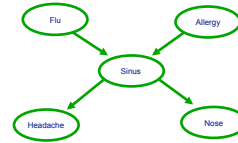
10-708 – ©Carlos Guestrin 2006-2008

18

Learning structure with missing data

[K&F 18.4]

- Known BN structure: Use expected counts, learning algorithm doesn't change
- Unknown BN structure:
 - Can use expected counts and score model as when we talked about structure learning
 - But, very slow...
 - e.g., greedy algorithm would need to redo inference for every edge we test...
- (Much Faster) **Structure-EM**: Iterate:
 - compute expected counts
 - do a some structure search (e.g., many greedy steps)
 - repeat
- **Theorem**: Converges to local optima of marginal log-likelihood
 - details in the book



10-708 – ©Carlos Guestrin 2006

19

What you need to know about learning BNs with missing data

- EM for Bayes Nets
- E-step: inference computes expected counts
 - Only need expected counts over X_i and \mathbf{Pa}_{x_i}
- M-step: expected counts used to estimate parameters
- Which variables are hidden can change per datapoint
 - Also, use labeled and unlabeled data → some data points are complete, some include hidden variables
- Structure-EM:
 - iterate between computing expected counts & many structure search steps

10-708 – ©Carlos Guestrin 2006

20

MNs & CRFs with missing data

- MNs with missing data
 - Models $P(\mathbf{X})$, part of \mathbf{X} hidden
 - Use EM to optimize
 - Same ideas as BN

- CRFs with missing data
 - Models $P(\mathbf{Y}|\mathbf{X})$
 - What's hidden?
 - Part of \mathbf{Y} :

 - All of \mathbf{Y} :

 - Part of \mathbf{X} :

10-708 – ©Carlos Guestrin 2006-2008

21

Readings:

K&F: 6.1, 6.2, 6.3, 14.1, 14.2

Kalman Filters Gaussian BNs

Graphical Models – 10708
Carlos Guestrin
Carnegie Mellon University
November 24th, 2008

22

Adventures of our BN hero

- Compact representation for probability distributions 1. Naïve Bayes
- Fast inference
- Fast learning
- Approximate inference 2 and 3.
Hidden Markov models (HMMs)
Kalman Filters
- But... Who are the most popular kids?

23

The Kalman Filter

- An HMM with Gaussian distributions
- Has been around for at least 60 years
- Possibly the most used graphical model ever
- It's what
 - does your cruise control
 - tracks missiles
 - controls robots
 - ...
- And it's so simple...
 - Possibly explaining why it's so used
- Many interesting models build on it...
 - An example of a Gaussian BN (more on this later)

24

Example of KF – SLAT

Simultaneous Localization and Tracking

[Funiak, Guestrin, Paskin, Sukthankar '06]

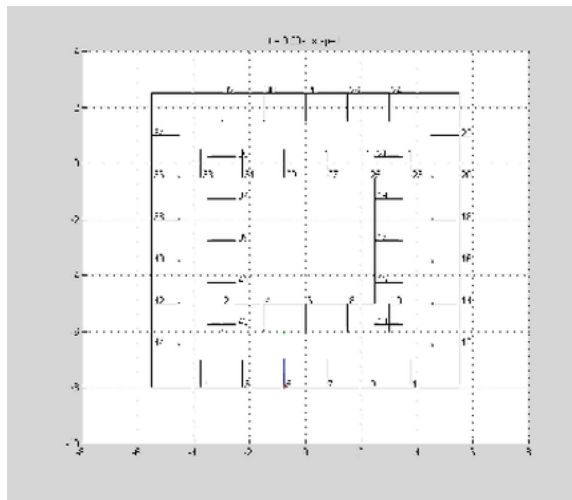
- Place some cameras around an environment, don't know where they are
- Could measure all locations, but requires lots of grad. student (Stano) time
- Intuition:
 - A person walks around
 - If camera 1 sees person, then camera 2 sees person, learn about relative positions of cameras

25

Example of KF – SLAT

Simultaneous Localization and Tracking

[Funiak, Guestrin, Paskin, Sukthankar '06]



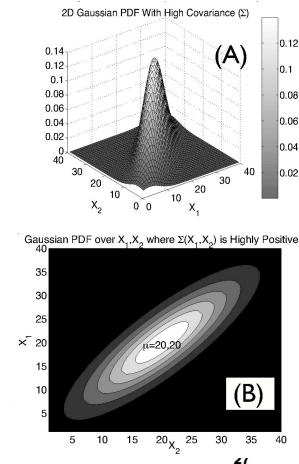
26

Multivariate Gaussian

$$p(X_1, \dots, X_n) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right\}$$

Mean vector:

Covariance matrix:



Conditioning a Gaussian

■ Joint Gaussian:

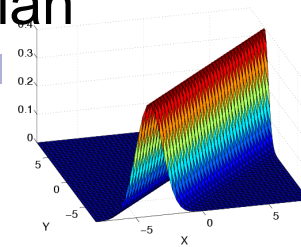
□ $p(X, Y) \sim N(\mu; \Sigma)$

■ Conditional linear Gaussian:

□ $p(Y|X) \sim N(\mu_{Y|X}, \sigma_{Y|X}^2)$

$$\mu_{Y|X} = \mu_Y + \frac{\sigma_{YX}}{\sigma_X^2} (x - \mu_x)$$

$$\sigma_{Y|X}^2 = \sigma_Y^2 - \frac{\sigma_{YX}^2}{\sigma_X^2}$$



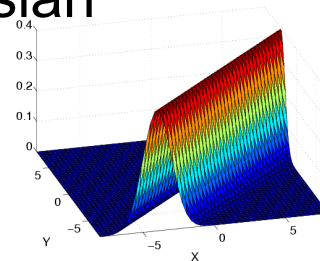
Gaussian is a “Linear Model”

- $\mu_{Y|X} = \mu_Y + \frac{\sigma_{YX}}{\sigma_X^2}(x - \mu_x)$
- Conditional linear Gaussian:
 - $p(Y|X) \sim N(\beta_0 + \beta X; \sigma^2)$
 - $\sigma_{Y|X}^2 = \sigma_Y^2 - \frac{\sigma_{YX}^2}{\sigma_X^2}$

29

Conditioning a Gaussian

- Joint Gaussian:
 - $p(X, Y) \sim N(\mu; \Sigma)$
- Conditional linear Gaussian:
 - $p(Y|X) \sim N(\mu_{Y|X}; \Sigma_{YY|X})$



$$\mu_{Y|X} = \mu_Y + \Sigma_{YX} \Sigma_{XX}^{-1} (x - \mu_x)$$

$$\Sigma_{YY|X} = \Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY}$$

30

Conditional Linear Gaussian (CLG) – general case

■ Conditional linear Gaussian:

□ $p(Y|X) \sim N(\beta_0 + BX; \Sigma_{YY|X})$

$$\mu_{Y|X} = \mu_Y + \Sigma_{YX} \Sigma_{XX}^{-1} (x - \mu_x)$$

$$\Sigma_{YY|X} = \Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY}$$

31

Understanding a linear Gaussian – the 2d case

- Variance increases over time (motion noise adds up)
- Object doesn't necessarily move in a straight line

32

Tracking with a Gaussian 1

- $p(X_0) \sim N(\mu_0, \Sigma_0)$
- $p(X_{i+1}|X_i) \sim N(B X_i + \beta; \Sigma_{X_{i+1}|X_i})$

33

Tracking with Gaussians 2 – Making observations

- We have $p(X_i)$
- Detector observes $O_i=o_i$
- Want to compute $p(X_i|O_i=o_i)$
- Use Bayes rule:
- Require a CLG observation model
 - $p(O_i|X_i) \sim N(W X_i + v; \Sigma_{O_i|X_i})$

34