Bayes Theorem, Bayesian Stats, Bayes Nets

Kenneth (Kenny) Joseph

University at Buffalo
Department of Computer Science and Engineering
School of Engineering and Applied Sciences

Reminders

- Corrections due today
- PA3 grades out early next week
- Quiz 10 out tonight, due Tuesday night
- ~~PA4 due~~ tonight Tuesday

Review Quiz 9 on Thursday

Bayesian Statistics

$$P(\theta|D) \qquad P(D|\theta) \qquad P(\theta)$$

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior},$$

$$P(\theta|D) = \frac{P(D|\theta)\,P(\theta)}{P(D)}$$

Reminder: Bayes rule != Bayesian stats!

- 1. Set up the full probability model (the **joint**) $P(D,\theta) = P(D|\theta)\,P(\theta)$
- 2. Condition on observed data (estimate the **posterior**)
- 3. Evaluate model fit

Today

- How to set up the model
  - DAGs Directed Acyclic Graphs ["Bayesian Networks"/ D-PGMs]
    - Relationship to conditional probability
    - Conditional Independence w/ the Markov Assumption
    - Relationship to causal modeling / causal inference
  - Generative stories
- How to estimate posterior (i.e. **inference**)
  - MAP estimation
  - Simulation

Setting up the model ... Directed Probabilistic Graphical Models

- Bayesian models can be complex
- How do we easily explain them?
- Two ways
  - **Directed Probabilistic Graphical Models**
    - **These are also called Bayesian Networks. But you can use them for even non-Bayesian models.**
  - **Generative Stories**
- ^ this is an oversimplification, but not by all that much.

Generative Stories for the text message example

Draw $\lambda_{1/2}$ from Exp(α) hyperparameter

Pick a day to switch from λ_1 to λ_2 ~ U(1,70)

If the day i > τ, then $\lambda = \lambda_1$, else $\lambda = \lambda_2$

Pick for day i a count of text messages from Pois(λ)

$$\lambda_1 \sim \text{Exp}(\alpha)$$
$$\lambda_2 \sim \text{Exp}(\alpha)$$

$$\tau \sim \text{DiscreteUniform}(1{,}70)$$

$$\lambda = \begin{cases} \lambda_1 & \text{if } t < \tau \\ \lambda_2 & \text{if } t \geq \tau \end{cases}$$

$$C_i \sim \text{Poisson}(\lambda)$$

Grow your own generative story

Hyperparameters

Decide which foods to buy at grocery store $\sim Mult(\lambda)$

Pick food to eat $\sim Multinomial(\pi)$

Graphical models, Generally

- http://www.cs.cmu.edu/~mgormley/courses/10601/slides/lecture20-bayesnet.pdf

10-301/601 Introduction to Machine Learning

MACHINE LEARNING DEPARTMENT

Machine Learning Department
School of Computer Science
Carnegie Mellon University

~~HMMs~~
~~+~~
Bayesian Networks /

D-PGMs!

Matt Gormley
Lecture 20
Mar. 30, 2022

Bayesian Networks

DIRECTED GRAPHICAL MODELS

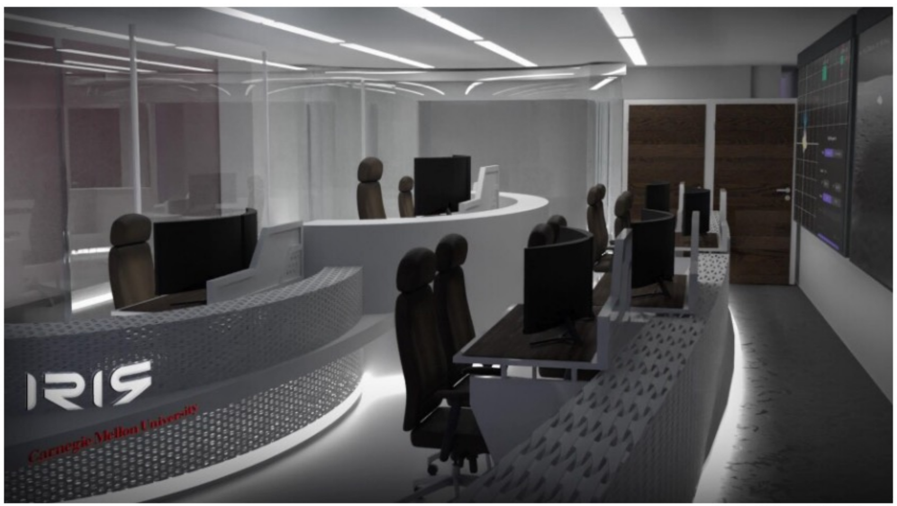
Example: CMU Mission Control

90.5 WESA
Pittsburgh's NPR News Station

WESA
Morning Edition

Pittsburgh's first mission control center to land at CMU ahead of 2022 lunar rover launch

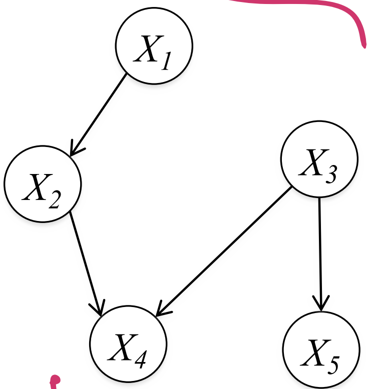
90.5 WESA | By Kiley Koscinski
Published March 29, 2022 at 4:44 PM EDT



Courtesy Of Carnegie Mellon University

Bayesian Network

A way to specify a (joint) distribution graphically!



Pretty!

$$p(X_1, X_2, X_3, X_4, X_5) =$$
$$p(X_5|X_3)p(X_4|X_2, X_3)$$
$$p(X_3)p(X_2|X_1)p(X_1)$$

Yikes

$p(X_1, X_2, X_3, X_4, X_5) = \#$ parameters?

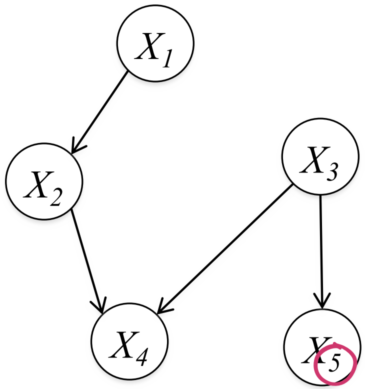
$p(\text{all combinations}) = 2^5 - 1$

$p(X_4 | X_2, X_3) = \#$ parameters? $= p(X_4 = 1 | X_2 = 0, X_3 = 0)$

Bayesian Network

Parent = something w/ an arrow to me.

Definition:



$$P(X_1, \ldots, X_T) = \prod_{t=1}^{T} P(X_t \mid \text{parents}(X_t))$$

- A Bayesian Network is a **directed graphical model**
- It consists of a graph **G** and the conditional probabilities **P**
- These two parts full specify the distribution:
  - "Qualitative" Specification: **G**
  - "Quantitative" Specification: **P**

What do probabilities look like?

Qualitative Specification

- Where does the qualitative specification come from?

  – Prior knowledge of causal relationships
  – Prior knowledge of modular relationships
  – Assessment from experts
  – Learning from data (i.e. structure learning)
  – We simply prefer a certain architecture (e.g. a layered graph)
  – …

Quantitative Specification

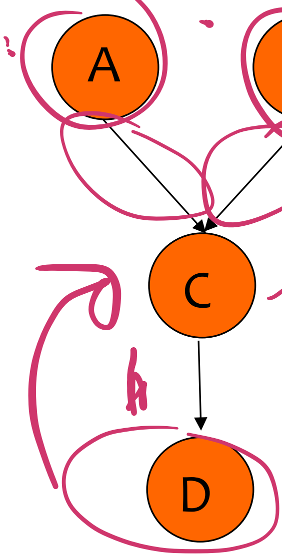
Example: Conditional probability tables (CPTs) for discrete random variables

a^0	0.75
a^1	0.25

b^0	0.33
b^1	0.67

$$P(a,b,c.d) = P(a)P(b)P(c|a,b)P(d|c)$$

$$P(a,b,c,d) = \prod p(x_i \mid \text{Parents}(x_i))$$



	a^0b^0	a^0b^1	a^1b^0	a^1b^1
c^0	0.45	1	0.9	0.7
c^1	0.55	0	0.1	0.3

$p(d|c)$

$p(c|$

	c^0	c^1
d^0	0.3	0.5
d^1	07	0.5

$c = 1$

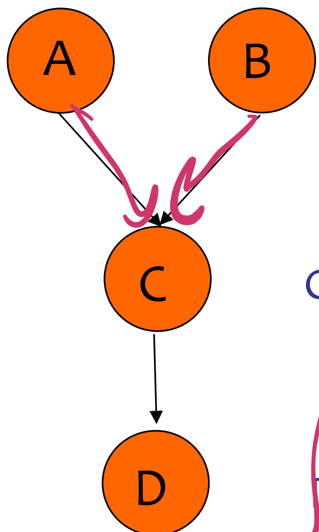
$p(a,b,d \mid c=1)$

Quantitative Specification

Example: Conditional probability density functions (CPDs) for continuous random variables

$A \sim N(\mu_a, \Sigma_a)$ $B \sim N(\mu_b, \Sigma_b)$

$$P(a,b,c.d) = P(a)P(b)P(c|a,b)P(d|c)$$



$C \sim N(A+B, \Sigma_c)$

$D \sim N(\mu_d + C, \Sigma_d)$

$P(D|C)$

$p(c|a,b)$

a b

c c

D

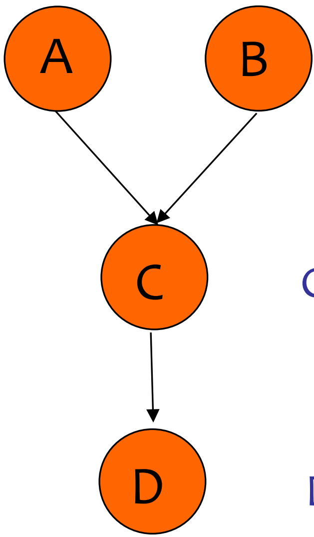
Quantitative Specification

Example: Combination of CPTs and CPDs for a mix of discrete and continuous variables

a^0	0.75
a^1	0.25

b^0	0.33
b^1	0.67

$$P(a,b,c.d) = P(a)P(b)P(c|a,b)P(d|c)$$



$C \sim N(A+B, \Sigma_c)$

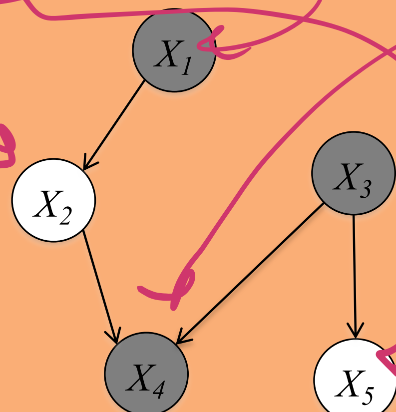
$D \sim N(\mu_d + C, \Sigma_d)$

Observed Variables

- In a graphical model, **shaded nodes** are "**observed**", i.e. their values are given

Example:

$$P(X_2, X_5 \mid X_1 = 0, X_3 = 1, X_4 = 1)$$

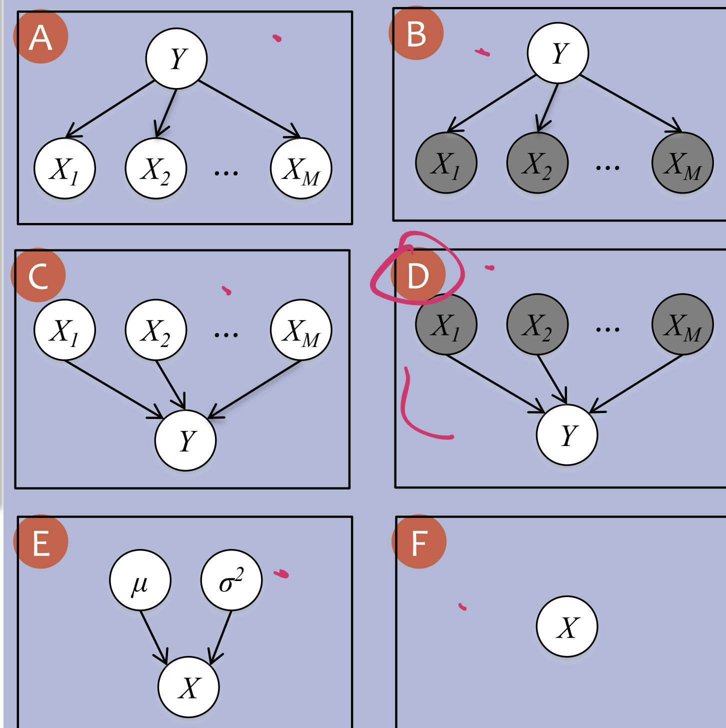


Familiar Models as Bayesian Networks

Question:
Match the model name to the corresponding Bayesian Network

1. Logistic Regression
2. Linear Regression
3. Bernoulli Naïve Bayes
4. Gaussian Naïve Bayes
5. 1D Gaussian

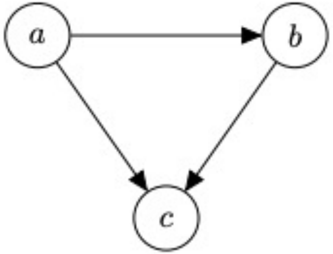
Answer:



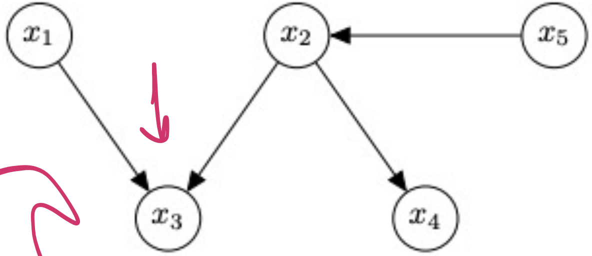
$p(y, x) = p(y|x)\, p(x)$

$y \sim N(w^T x, \sigma)$

Practice: Get Distribution from BayesNet



(a) Fully connected.



(b) Not fully connected.

$$p(a,b,c) =$$

$$p(c|a,b)\,p(b|a)\,p(a)$$

$$p(x_1, x_2, x_3, x_4, x_5) =$$

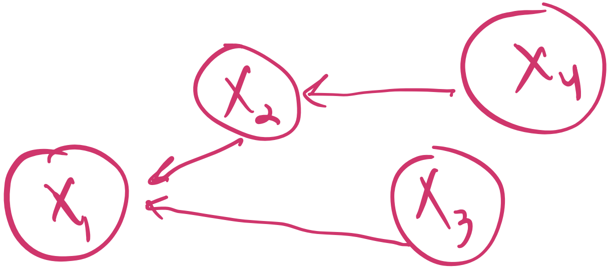
$$p(x_3|x_1,x_2)\,p(x_1)\,p(x_2|x_5)\,p(x_5)\,p(x_4|x_2)$$

Practice: Get Distribution from BayesNet

Practice: Draw Bayes Net from Specified Distribution

$$p(x_1, x_2, x_3, x_4) =$$

$$p(x_1 \mid x_2, x_3)\, p(x_2 \mid x_4)\, p(x_3)\, p(x_4)$$



Practice: Draw Bayes Net from Specified Distribution

Practice: Draw Models we know!

- ~~Logistic Regression~~ near impossible :)
- ~~Linear Regression~~
- Ridge Regression (tricky!)

$$L(h) = \sum_{i=1}^{N} (y_i - w^T x_i)^2 - \lambda \|w\|_2$$

What if we were Bayesian?

$p(\theta) =$

$p(D|\theta) =$

$$p(\theta|D) = p(D|\theta)\, p(\theta) =$$

$$y \sim N(w^T x, \sigma)$$

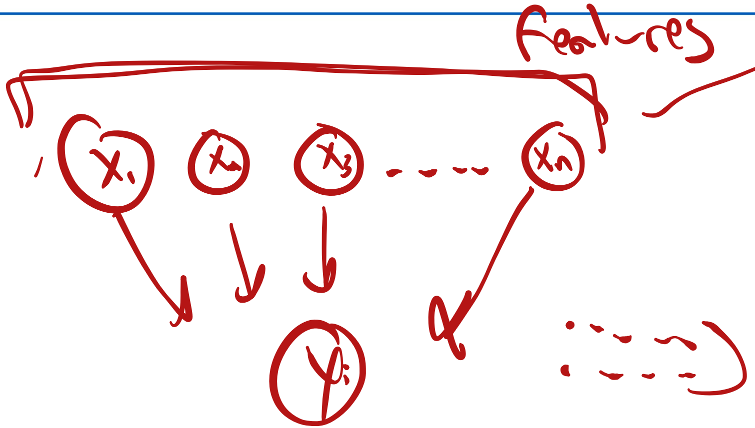
$$w \sim N(0, \lambda I)$$

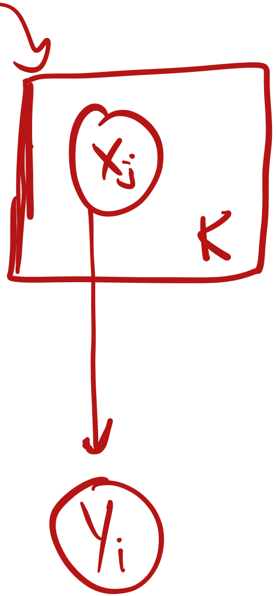
identity matrix

University at Buffalo
Department of Computer Science
and Engineering
School of Engineering and Applied Sciences

@_kenny_joseph

Plate Notation





For one observation

$|X|$ of size K

Graphical Model for text message example

$$\lambda_1 \sim \text{Exp}(\alpha)$$
$$\lambda_2 \sim \text{Exp}(\alpha)$$

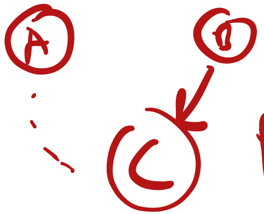
$$\tau \sim \text{DiscreteUniform}(1,70)$$

$$\lambda = \begin{cases} \lambda_1 & \text{if } t < \tau \\ \lambda_2 & \text{if } t \geq \tau \end{cases}$$

$$C_i \sim \text{Poisson}(\lambda)$$

Conditional Independence In Bayes Nets

$$\underbrace{\mathcal{A} \perp\!\!\!\perp \mathcal{B} \,|\, \mathcal{C},}$$



$p(c|b)$

$p(b)$

$p(a)$

- The above is read **A is conditionally independent of B, given C**
- Intuitively, "telling me something about B gives me no new information if I already know C"
- Any examples you can think of?
- Example here: Markov Property

Markov Property

] Example of a conditional independence assumption

$$X_1 \to X_2 \to X_3 \to X_4$$

Monday T W Th

$$X_3 \perp X_1 \mid X_2$$

$$p(x_1, x_2, x_3, x_4)$$

$$p(x_4|x_3)\,p(x_3|x_2)\ldots$$

D-separation

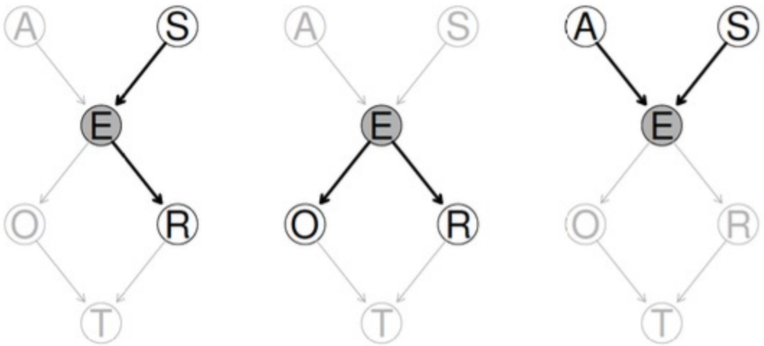


Figure 1.3
Some examples of d-separation covering the three fundamental connections: the *serial connection* (left), the *divergent connection* (centre) and the *convergent connection* (right). Nodes in the conditioning set are highlighted in grey.

- The full treatment of conditional independence in Bayes Nets requires a discussion about **d-separation**

Estimating the Posterior – MAP estimation

$P(D|\theta) = \text{Bin}(n_H, n_T; \theta) = \binom{n_H + n_T}{n_H} \theta^{n_H} (1-\theta)^{n_T}$

$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D;\theta) = \ldots ?$

- n_H = # heads
- n_T = # tails

$\theta = p(\text{heads})$

What if we are Bayesian?!

$m_H = 1000$

$m_T = 1000$

$$\theta = \frac{n_H + m_H}{n_H + n_T + m_H + m_T}$$

$\theta \sim \text{Beta}(m_H, m_T)$

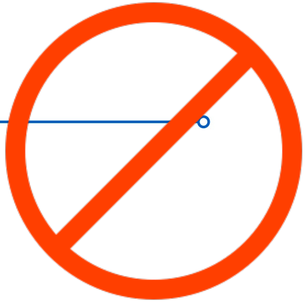
maximum a posteriori ("MAP")

University at Buffalo
Department of Computer Science and Engineering
School of Engineering and Applied Sciences

@_kenny_joseph

Estimating the Posterior – MAP estimation

Estimating the Posterior - Sampling



- Problem w/ MAP
  - Doesn't give us a distribution
  - Doesn't work if we cant do a closed form solution!
  - ^ Intertwined … hard part is the normalizing constant (knowing the whole probability space)
- Solution: Sampling / Simulation-based approaches
  - https://chi-feng.github.io/mcmc-demo/app.html

What to do once we have the posterior?

- Make probabilistic statements about our parameters
- Make predictions averaged over ALL models
- Does this model actually fit? (a wholeee thing)

Where we are at

- We can use these tools to build complex, interesting, but **intuitive** interpretable models
  - But can be hard to fit!
  - And not always super predictive
- Next: deep learning
  - Trade intuition and interpretability for ease of training and predictive power