National Science Foundation
WHERE DISCOVERIES BEGIN

Biosphere 2
THE UNIVERSITY OF ARIZONA.

Workshop on
Computationally Intensive Modeling of Social Interaction

A (very brief) Introduction to Probabilistic Generative Models

Clay Morrison
University of Arizona
clayton@sista.arizona.edu

Interdisciplinary Visual Intelligence Lab
http://ivilab.org

8 November 2014

Models

- "Essentially, all models are wrong, but some are useful" – George E. P. Box
- What is *useful*?
  - Prediction
    - Basis for evaluation
    - Estimating generalization *error* (e.g., cross validation)
  - Explanation
    - Extracting an underlying pattern (taxonomy)
    - Reducing the number of independent phenomena
- Why Probabilistic Models?
  - Principled way of modeling uncertainty
  - Characterize lack of knowledge

Probabilistic Graphical Models

- PGM = Multivariate Statistics + Structure
- Provide simple way to visualize the structure of a probabilistic model and can be used to design and motivate new models
- Insights into the properties of the model, including conditional independence properties, can be obtained by inspection of the graph
- Complex computations (inference, learning) can be expressed in terms of graphical manipulations, in which the underlying math is carried along implicitly

(Christopher Bishop, 2007)

Probabilities and Distributions

	Green socks	Red socks	other socks	
CS student	8	4	3	15
Other student	2	15	8	25
	10	19	11	40

This is a joint frequency distribution, used for intuitive demonstration; we can estimate probabilities from the joint (green), marginal (yellow) and total (purple) frequencies.

- The "Big Three"
  - Marginal probability (or "plain old probability")
  - Joint probability (of two or more things together)
  - Conditional probability (of one thing ***given*** another)

Conditional probabilities

Probability in *context*

	Green socks	Red socks	other socks	
CS student	8	4	3	15
Other student	2	15	8	25
	10	19	11	40

"Prob. of *X given Y*"

$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$

joint / marginal

P(CS | GreenSocks) = 8/10 = .8
P(GreenSocks | CS) = 8/15 = .53

The Building Blocks of Probability Calculus

- Sum Rule $p(X) = \sum_Y p(X,Y)$
- Product Rule $p(X,Y) = p(Y|X)p(X)$

- More generally, can always write any joint distribution as an incremental product of conditional distributions

(This is called the Chain Rule, but it's just the generalization of the Product Rule)

$$P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$$

$$P(x_1, x_2, \ldots x_n) = \prod_i P(x_i|x_1 \ldots x_{i-1})$$

Bayes' Rule

- Two ways to factor a joint distribution by product rule

$$p(X,Y) = p(X|Y)\ p(Y) = p(Y|X)\ p(X)$$

- Dividing, we get:

$$p(X|Y) = \frac{p(Y|X)\ p(X)}{p(Y)}$$

- Why is this a **big deal ?**
  - Let's us build one conditional from its reverse.
  - Often one conditional is hard to get , but the other components are available.

Bayes' Rule

Relating Hypotheses to Evidence

hypothesis

evidence

Likelihood

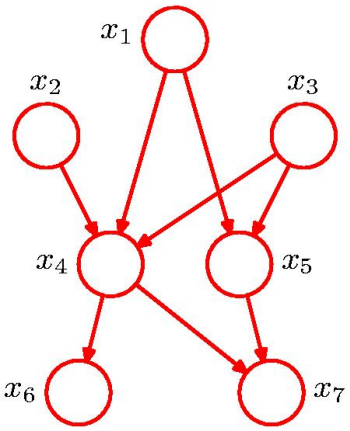
Prior

$$\underbrace{P(H \mid E)}_{\text{Posterior}} = \frac{P(E \mid H)P(H)}{\underbrace{P(E)}_{\text{Marginal probability of Evidence (features)}} = \underbrace{\sum_H p(E|H)\ p(H)}_{\text{= Marginal likelihood}}}$$

Bayesian Networks

Factorization of Bayesian Network into product of conditional & marginal probabilities (chain rule)

$$p(x_1,\ldots,x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)$$
$$p(x_5|x_1,x_3)p(x_6|x_4)p(x_7|x_4,x_5)$$



General Factorization

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$$

Sum $\quad p(X) = \sum_Y p(X,Y)$

Product $\quad p(X,Y) = p(Y|X)p(X)$

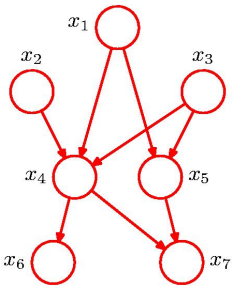
(Christopher Bishop, 2007)

Probabilistic Queries

Bold face because these are vectors of variables

Organize variables into

- Evidence (observed), **E**
- Query (what you want to know), **Y**
- Hidden (leftover), **X** (for completeness)



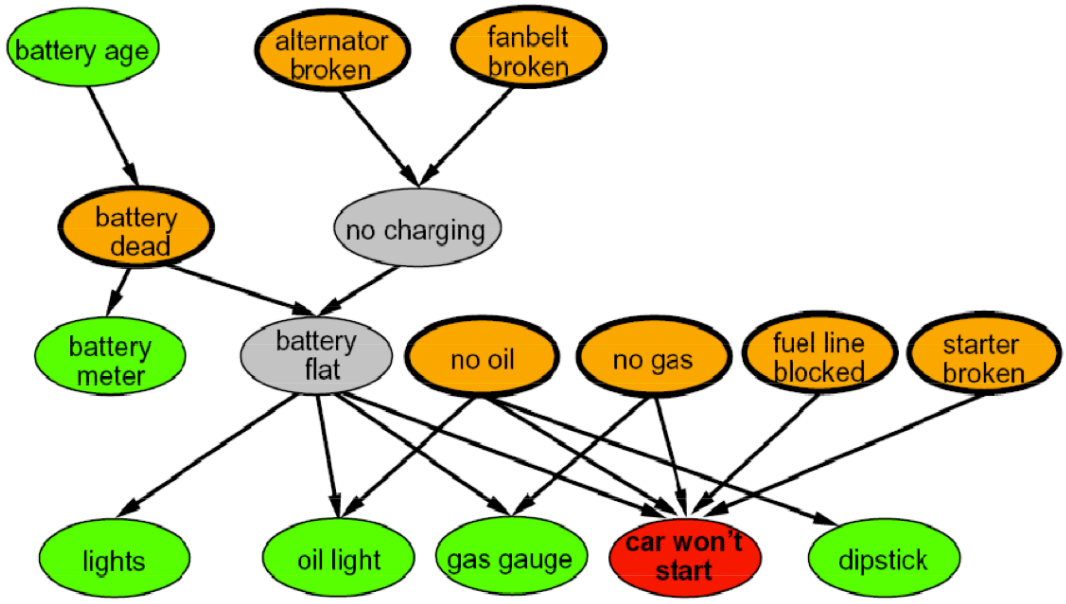
Generic Query: P(**Y**|**E**)

- This leads to a distribution over Y given the evidence
- Note that X is marginalized out
- We can use this to make a decision
- Simplest is most probable, i.e., $\underset{\mathbf{Y}}{\mathrm{Argmax}}\ P(\mathbf{Y},\mathbf{E})$

MAP Query (most probably configuration of variables):

$$MAP(\mathbf{W}\,|\,\mathbf{E}) = \underset{\mathbf{W}}{\mathrm{Argmax}}\ P(\mathbf{W},\mathbf{E}) \qquad (\mathbf{W}=\mathbf{Y}\cup\mathbf{X})$$

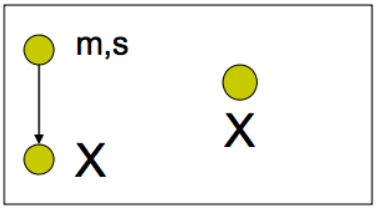
Example Bayesian Network



PGMs are your old friends

Density estimation

Parametric and nonparametric methods

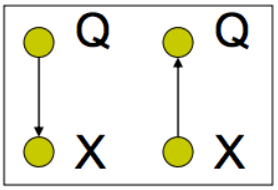


Regression

Linear, conditional mixture, nonparametric

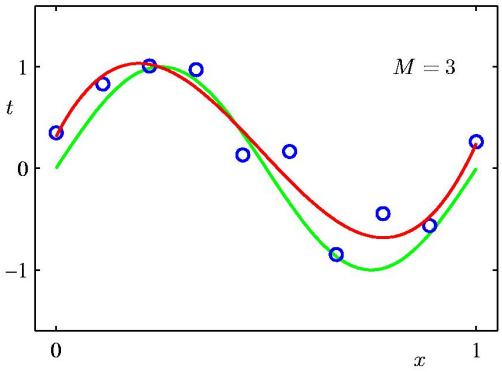
Classification

Generative and discriminative approach



Clustering

Bayesian Linear Regression



Polynomial

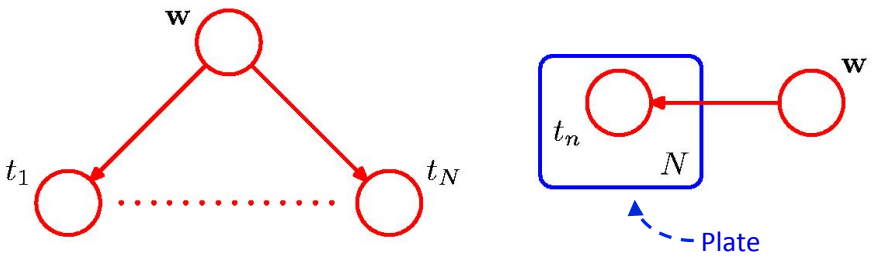
$$y(x, \mathbf{w}) = \sum_{j=0}^{M} w_j x^j$$

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n | y(\mathbf{w}, x_n))$$

(Christopher Bishop, 2007)

Bayesian Linear Regression

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n | y(\mathbf{w}, x_n))$$

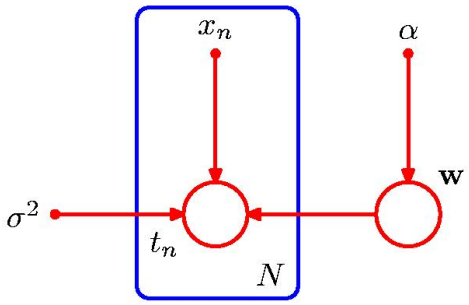


(Christopher Bishop, 2007)

Bayesian Linear Regression

- Input variables and explicit hyperparameters

$$p(\mathbf{t}, \mathbf{w}|\mathbf{x}, \alpha, \sigma^2) = p(\mathbf{w}|\alpha) \prod_{n=1}^{N} p(t_n|\mathbf{w}, x_n, \sigma^2).$$

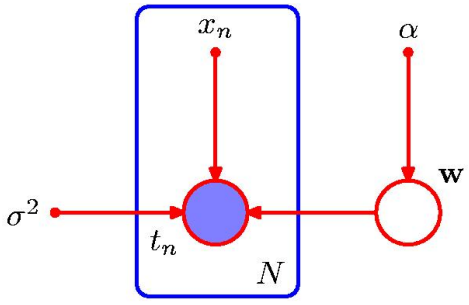


(Christopher Bishop, 2007)

Bayesian Linear Regression Learning

- Condition on data

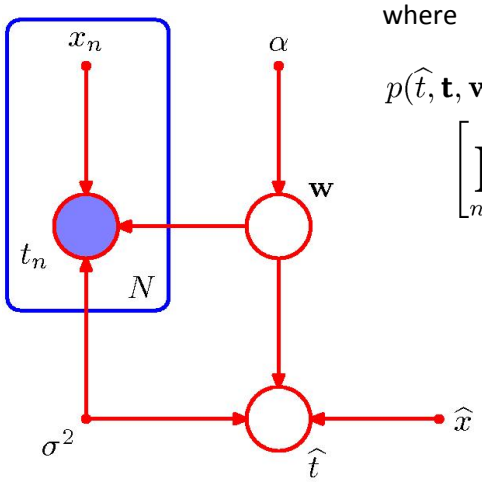
$$p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w})$$



(Christopher Bishop, 2007)

Bayesian Linear Regression Prediction

Predictive distribution: $p(\widehat{t}|\widehat{x}, \mathbf{x}, \mathbf{t}, \alpha, \sigma^2) \propto \int p(\widehat{t}, \mathbf{t}, \mathbf{w}|\widehat{x}, \mathbf{x}, \alpha, \sigma^2)\, \mathrm{d}\mathbf{w}$

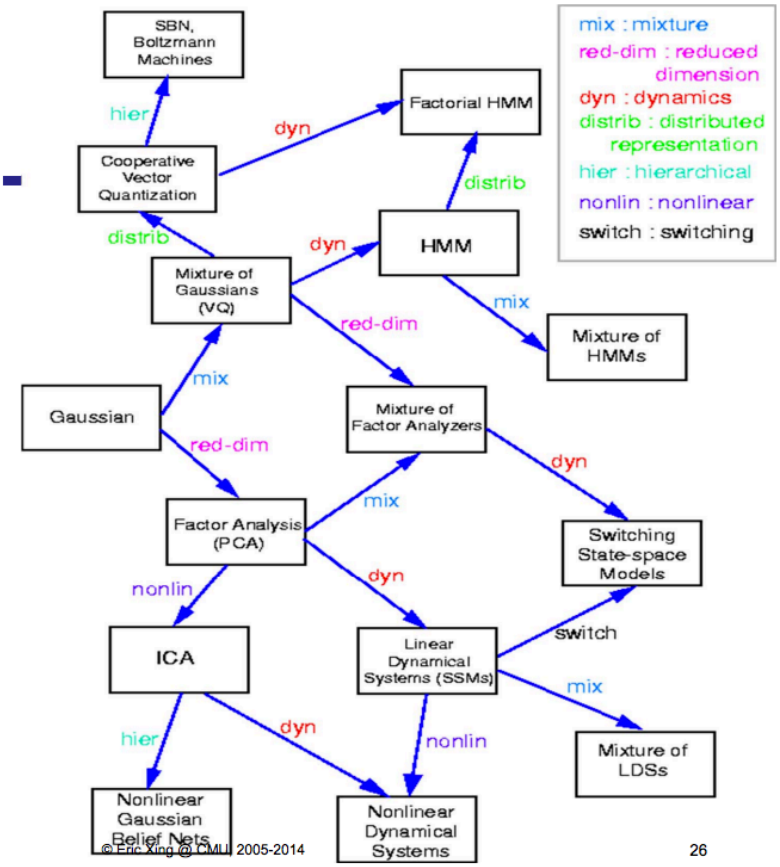


where

$$p(\widehat{t}, \mathbf{t}, \mathbf{w}|\widehat{x}, \mathbf{x}, \alpha, \sigma^2) = \left[\prod_{n=1}^{N} p(t_n|x_n, \mathbf{w}, \sigma^2)\right] p(\mathbf{w}|\alpha) p(\widehat{t}|\widehat{x}, \mathbf{w}, \sigma^2)$$

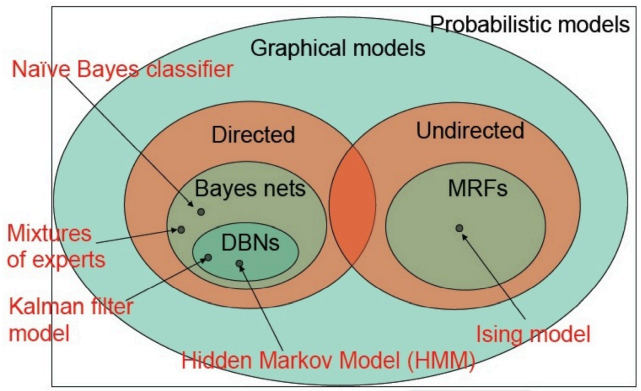
(Christopher Bishop, 2007)

An (incomplete) genealogy of graphical models



(Picture by Zoubin Ghahramani and Sam Roweis)

The PGM Zoo



Factorization (local functions) — **Conditional Independence**

$p(\mathbf{x})$
$p(\mathbf{y}|\mathbf{x})$

Graphical Structure
(relational structure of factors)

Directed Graphical Models

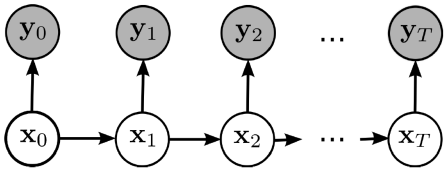
$$p(\mathbf{y}) = \prod_{s=1}^{S} p(y_s|\mathbf{y}_{\pi(s)})$$

Undirected Graphical Model

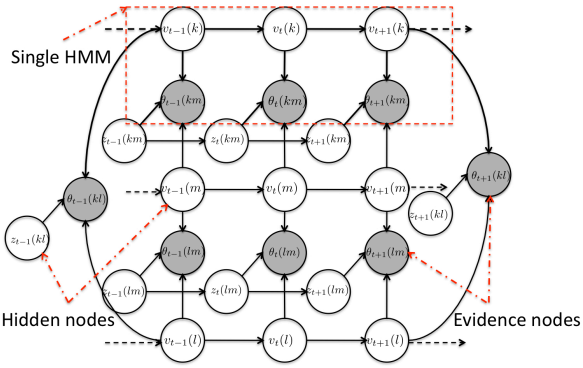
$$p(\mathbf{y}) = \frac{1}{Z}\prod_{a=1}^{A}\Psi_a(\mathbf{y}_a)$$

$$Z = \sum_{\mathbf{y}}\prod_{a=1}^{A}\Psi_a(\mathbf{y}_a)$$

Modeling Sequences

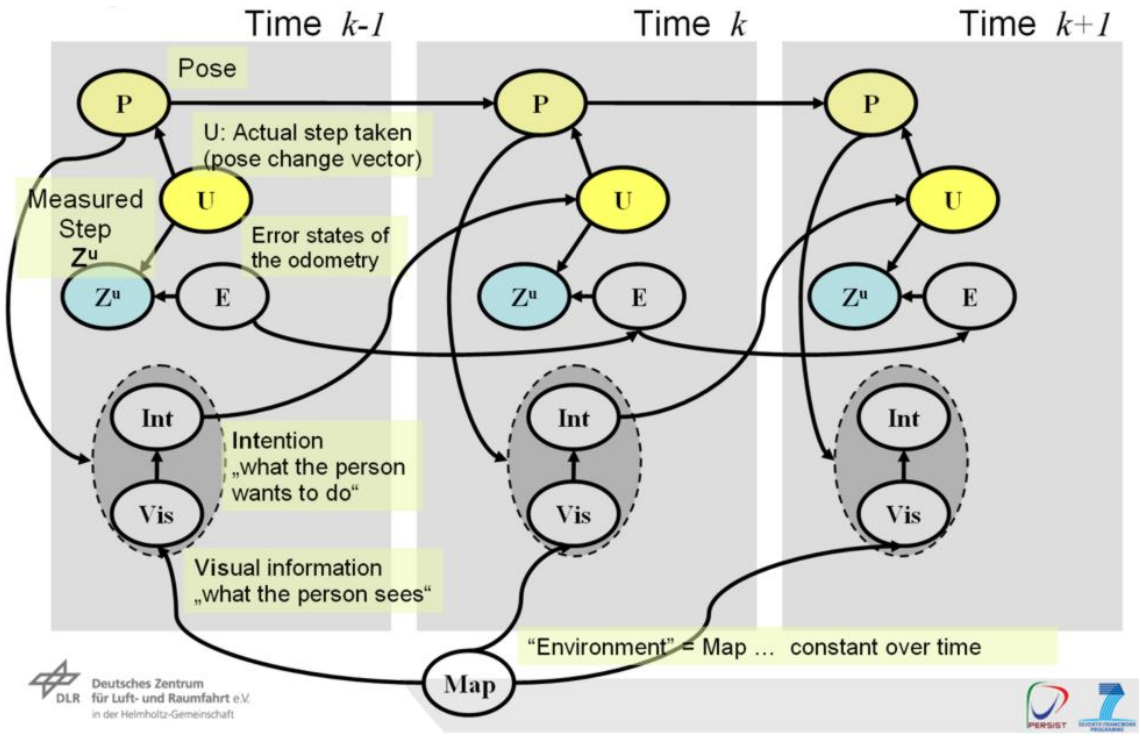


Hidden Markov Model (HMM)



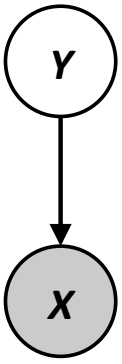
Coupled HMM

Dynamic Bayesian Network (DBN)



Generative vs. Discriminative

The **Generative** Picture



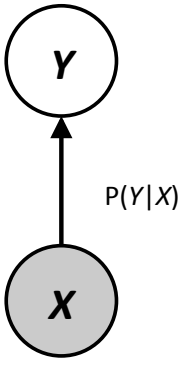
Can infer [label, latent state, cause] from evidence using **Bayes Thrm**

$P(Y|X) = P(X|Y)\,P(Y) / P(X)$

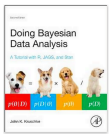
Model the Joint of X and Y

$P(X,Y) = P(X|Y)\,P(Y)$

The **Discriminative** Picture



Digging Deeper

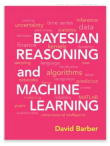


John Kruschke. (2014). Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan. Academic Press. (Release: Nov 17, 2014)

- Not about general PGMs, but an excellent introduction to Bayesian modeling, esp. coming from a basic statistics background (although even that is not necessary).

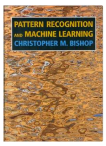
Daphne Koller and Nir Friedman. (2009). Probabilistic Graphical Models: Principles and Techniques. MIT Press.

- Encyclopedic presentation of the PGM framework.



David Barber. (2012). Bayesian Reasoning and Machine Learning. Cambridge University Press.

- Upper undergrad / early grad presentation of machine learning, but emphasizing PGMs (introducing from the beginning). Available for free online:
- http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/090310.pdf



Christopher Bishop. (2007). Pattern Recognition and Machine Learning. Springer.

- Another excellent machine learning text that also introduces graphical models, although a little more advanced. Ch. 8, which introduces PGMs, is available online:
- http://research.microsoft.com/en-us/um/people/cmbishop/PRML/pdf/Bishop-PRML-sample.pdf

- Kevin Murphy's online introduction
  - http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html