Probabilistic Graphical Models in Machine Learning

Sargur N. Srihari
University at Buffalo, The State University of New York
USA

ICDAR Plenary, Beijing, China
September 2011

Plan of Discussion

- Machine Learning (ML)
  - History and Problem types solved
- Probabilistic Graphical Models (PGMs)
  - Tutorial
    - Specialized models
- Computational Forensics Application
  - Handwriting

What is Machine Learning?

- Automatic construction of programs from examples of input-output behavior
- Marriage of Computer Science and Probability /Statistics

1. Computer Science:
   - Artificial Intelligence
     - Tasks performed by humans not well described algorithmically
   - Data Explosion
     - User and thing generated
2. Statistics:
   - Methods that learn from data (MLE or Bayesian)

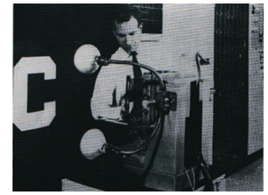
When is Machine Learning Needed

- Problems involving uncertainty
  - Perceptual data (images, text, speech, video)
- Information overload
  - Large Volumes
  - Limitations of time, cognitive ability
- Constantly Changing Data Streams
  - Search engine adaptation
- Principled design
  - High performance systems

Problem Types and Methods

1. Classification
   - OCR, Spam Filter (Logistic Regression)
   - Text Categorization (SVM)
2. Regression:
   - LeToR (GP)
3. Collective Classification
   - Speech, Handwriting (HMM)
   - PoS, NE (MEMM, CRF)
4. Inferring a Probability Distribution
   - Computational Forensics (BN, Sampling)
5. Clustering Data Mining (EM, BIC)

History of ML



20 x 20 cell

Adaptive Wts

- First Generation (1960-1980)
  - Perceptrons, Nearest-neighbor, Naïve Bayes
  - Special Hardware, Limited performance

USPS-MLOCR

- Second Generation (1980-2000)
  - ANNs, Kalman, HMMs, SVMs
    - HW addresses, speech reco, postal words
  - Difficult to include domain knowledge
    - Black box models fitted to large data sets

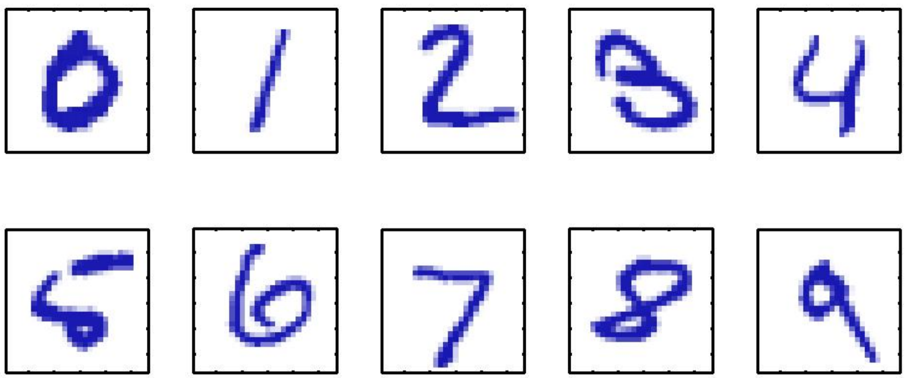
USPS-RCR

- Third Generation (2000-Present)
  - PGMs, Fully Bayesian (including GP)
    - Image segmentation, Text analytics (NE Tagging)
  - Expert prior knowledge with statistical models

Classification: OCR

Input x={$x_1 \ldots x_{12}$}: Image Features

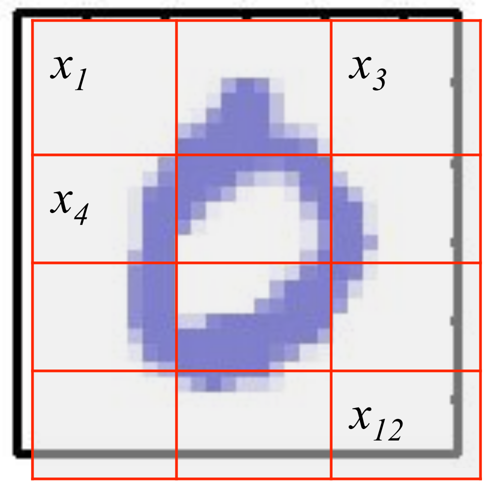
Output (y): Class Labels {$y^0, y^1, ..y^9$}



Wide variability of same numeral

- Handcrafted rules will result in large no of rules and exceptions
- Better to have a machine that learns from a large training set

1,000 chars/page,
1,000s of pages



Handwritten Digits

Features (x_i):

Values: Proportion of black pixels in each of 12 cells x_i $i=1,..,12$

x_i^0=0-10%
x_i^1=10-20%
….

$|Val(x_i)|=10$

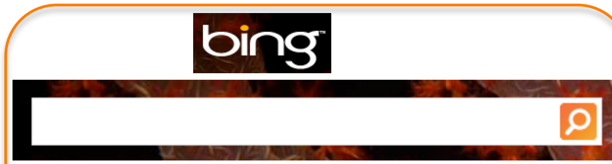
No of parameters$=10^{12}- 1$
Or 1 trillion

Per class

No of samples needed=??

Regression: Learning To Rank

$(d > 200)$



In LETOR 4.0 dataset
46 query-document features
Maximum of 124 URLs/query

Yahoo! data set has $d=700$

Input (x_i):

(d Features of Query-URL pair)

- Log frequency of query in anchor text
- Query word in color on page
- # of images on page
- # of (out) links on page
- PageRank of page
- URL length
- URL contains "~"
- Page length

Traditional IR uses TF/IDF

Output (y):

Relevance Value

Target Variable

- Point-wise (0,1,2,3)
- Regression returns continuous value

-Allows fine-grained ranking of URLs

Role of PGMs in ML

- Dozens of ML models, Large Data sets
  - PGMs Provide understanding of:
    - model relationships (theory)
    - problem structure (practice)
  - Allow including human knowledge
- Nature of PGMs
  1. Represent joint distributions
     1. Many variables without full independence
     2. Expressive
     3. Declarati
  2. Inference: Separate model/algorithm errors
  3. Learning

Representation

Probabilistic Graphical Models

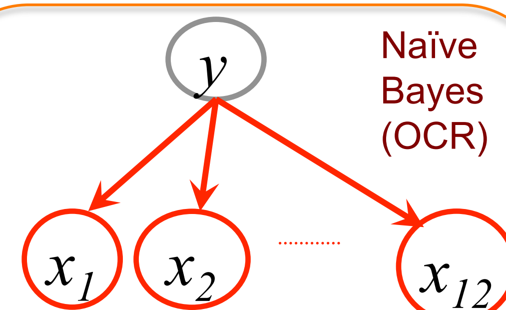
1. Bayesian Network (BN)

Directed Acyclic Graph

Nodes: variables Edges: direct causality

(correlation irrespective of others)

Naïve Bayes (OCR)



Joint distribution: product of CPDs

$$P(\mathrm{x},y) = P(y)\prod_{i=1}^{12} P(x_i \mid y)$$

CPDs

$P(y)$

y^0	y^1	y^2	y^3	y^4	y
0.1	0.1	0.1	0.1	0.1	0

$P(x_i|y)$

y	x_i^0	x_i^1	x_i^2	x_i^3	x_i^4	x_i^5	x_i^6	x_i^7	x_i^8	x_i^9
y^0	0	0	0.1	0.2	0.3	0.2	0.1	0.1	0	0
y^1										
y^9										

12 CPDs

No of parameters=*100 x 12=1,200*

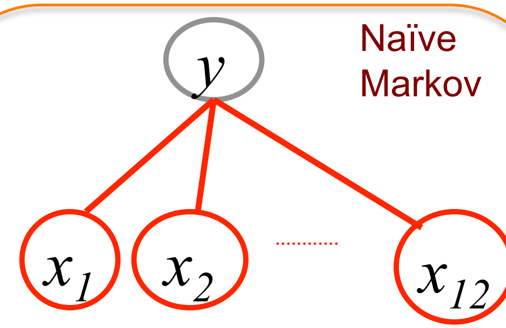
(instead of 1 trillion)

2. Markov Network (or MRF)

Undirected

– Edge: influence (non-directional)

Naïve Markov



Joint distribution: product of factors

$$P(\mathrm{x},y) = \frac{1}{Z}\prod_{i=1}^{12} \phi_i(x_i, y)$$

where Z is normalizing constant: Partition function

$$Z = \sum_{\mathrm{x}}\prod_{i=1}^{12} \phi_i(x_i, y)$$

Factors $\phi(\mathrm{x}_1, y) \rightarrow R$

(potential)

x_1^0	y^0	100
x_1^0	y^1	1
x_1^9	y^9	100

CRF: MN for conditional $P(y|\mathrm{x})$

y target x: observed

$$\tilde{P}(y \mid \mathrm{x}) = \prod_{i=1}^{m} \phi_i(x_i, y)$$

$$P(y \mid \mathrm{x}) = \frac{1}{Z(\mathrm{x})}\tilde{P}(y \mid \mathrm{x})$$

$$Z(\mathrm{x}) = \sum_{Y}\tilde{P}(y \mid \mathrm{x})$$

$\tilde{P}(y \mid \mathrm{x})$ Is unnormalized

$Z(\mathrm{x})$ Partition function of x

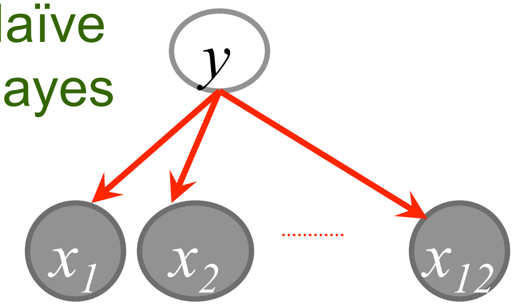
m=no of factors

Discriminative vs Generative Training

Independent variables $\mathrm{x} = \{x_1,..x_{12}\}$ and binary target y

1. Generative: estimate CPD parameters

Naïve Bayes



$$P(y,\mathrm{x}) = P(y)\prod_{i=1}^{12} P(x_i \mid y)$$

From joint get required conditional

Low-dimensional estimation

independently estimate 12x 10 parameters

But pixel independence is false

For sparse data generative is better

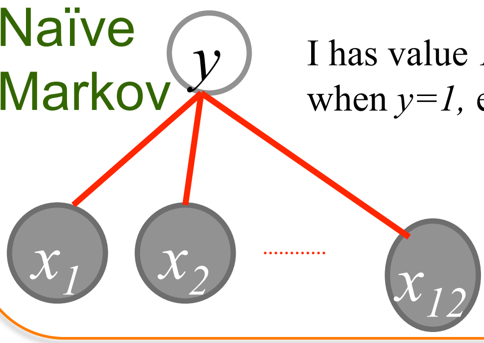
2. Discriminative: estimate CRF parameters w_i

Potential Functions (log-linear)

$\phi_i(x_i,y)=\exp\{\mathrm{w}_i x_i \,\mathrm{I}\{y=1\}\}$,

$\phi_0(y)=\exp\{\mathrm{w}_0 \,\mathrm{I}\{y=1\}\}$

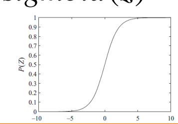
Naïve Markov



I has value 1 when $y=1$, else 0

Unnormalized $\tilde{P}(y=1 \mid \mathrm{x}) = \exp\left\{\mathrm{w}_0 + \sum_{i=1}^{12} \mathrm{w}_i x_i\right\}$ $\tilde{P}(y=0 \mid \mathrm{x}) = \exp\{0\} = 1$

Normalized $P(y=1 \mid \mathrm{x}) = sigmoid\left\{\mathrm{w}_0 + \sum_{i=1}^{12} \mathrm{w}_i x_i\right\}$ where $sigmoid(z) = \dfrac{e^z}{1+e^z}$



Logistic Regression

Jointly optimize *12* parameters

High dimensional estimation

but correlations accounted for

Can use much richer features:

Edges, image patches sharing same pixels

multiclass

$$p(y_i \mid \phi) = y_i(\phi) = \frac{\exp(a_i)}{\sum_j \exp(a_j)}$$

where $a_j = \mathrm{w}_j^{\mathrm{T}}\phi$

Collective Labeling: Three Models

Sequence of observations $X=\{X_1,..X_k\}$
Need a joint label $Y=\{Y_1,..Y_k\}$
Both CRF and MEMM are Discriminative Models
  Directly obtain $P(Y|X)$
HMM is generative
  Needs $P(X,Y)$

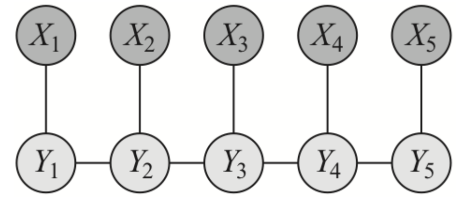
Model Trade-offs in expressive power and learnability

1.MEMM and HMM are more easily learned
- Directed models: ML parameter estimates have closed-form
- CRF requires expensive iterative gradient-based approach

2.Ability to use rich feature sets
- HMM needs explicit modeling over features
- CRF and MEMM are discriminative models and avoid this

3.Independence Assumptions made
- MEMM assumes Y_1 independent of X_2 not given Y_2
- Later observation has no effect on current state
  - In activity recognition in video sequence,
    Frames labelled as running/walking.
    Earlier frames may be blurry but later ones clearer
    Model incapable of going back

CRF

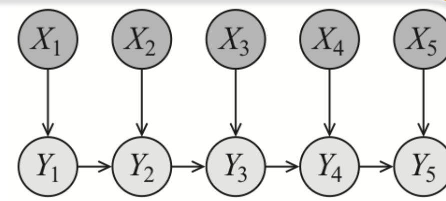


$$P(Y \mid X) = \frac{1}{Z(X)} \tilde{P}(Y,X)$$

$$\tilde{P}(Y,X) = \prod_{i=1}^{k-1} \phi_i(Y_i, Y_{i+1}) \prod_{i=1}^{k} \phi_i(Y_i, X_i)$$

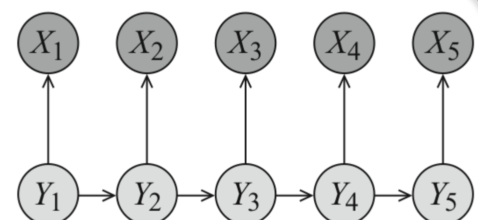
$$Z(X) = \sum_{Y} \tilde{P}(Y,X)$$

MEMM



$$P(Y \mid X) = \prod_{i=1}^{k} P(Y_i \mid X_i) P(Y_i \mid Y_{i-1})$$

HMM

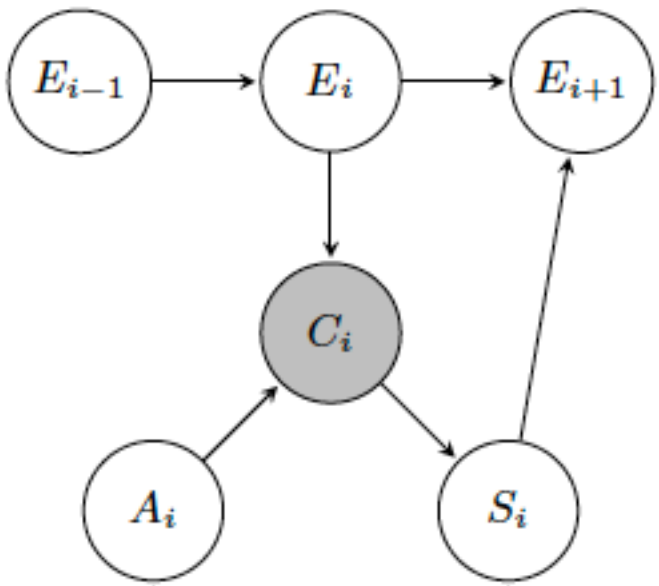


Joint distribution → $P(X,Y) = \prod_{i=1}^{k} P(X_i / Y_i) P(Y_i \mid Y_{i-1})$

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

Dynamic BN: Training Data for LeToR

- Dynamic BN can model Time Trajectory
- LeToR relevance values are assigned by human editors
  - Expensive
  - Can change over time
- Click Logs:
  - provide implicit feedback
  - cheap proxy for editorial labels



C_i: Click on i^{th} URL in *retrieved* list

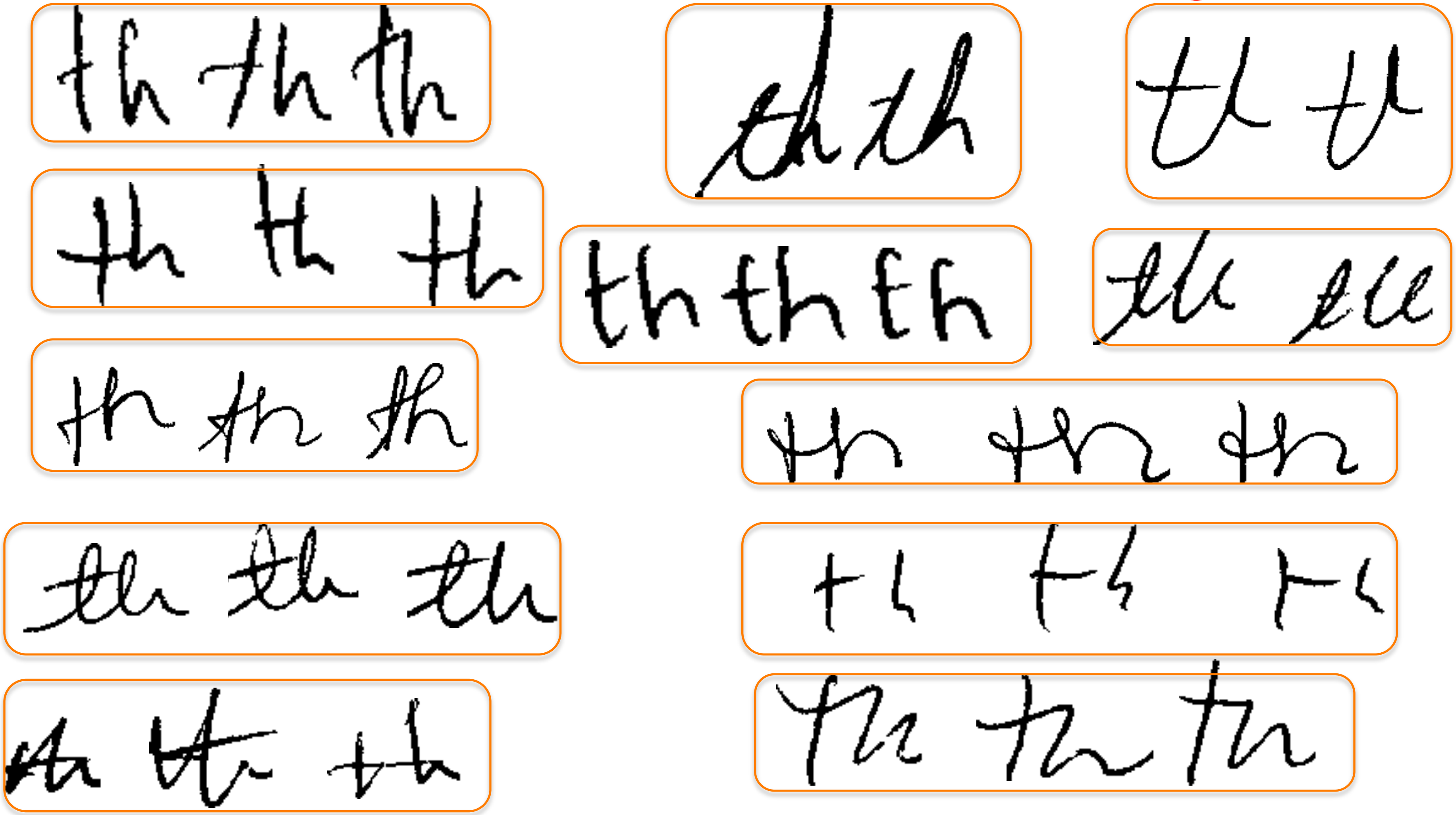
Hidden Variables:

E_i: did the user *examine* the url?
A_i: was the user *attracted* by the url?
S_i: was the user *satisfied* by the landing page?

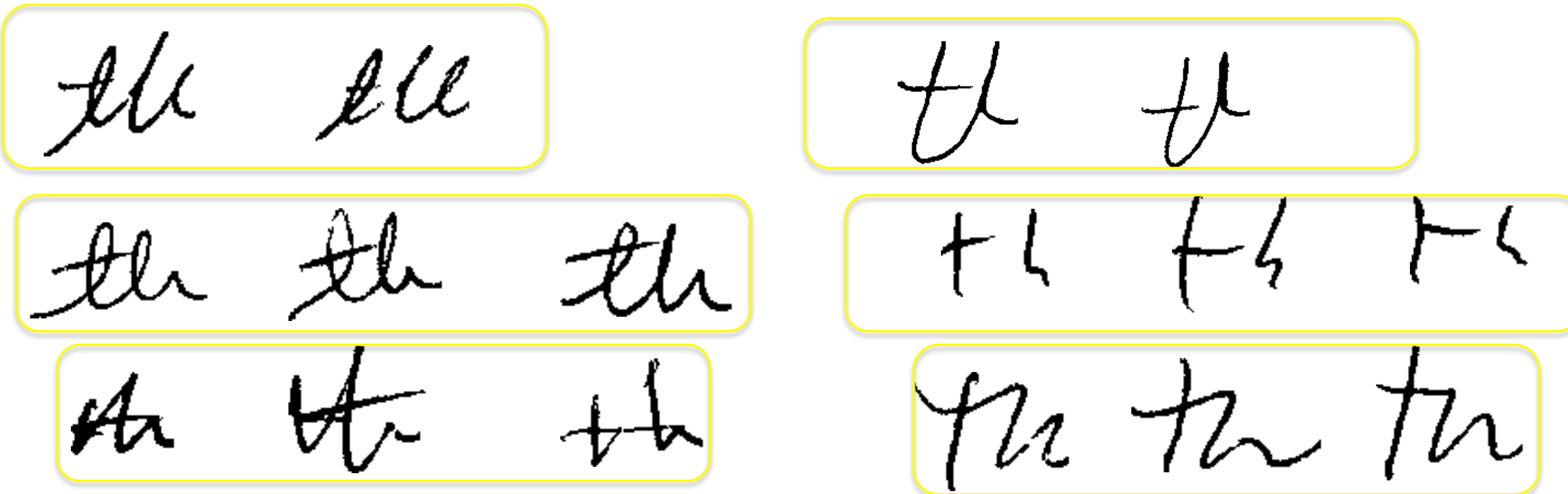
Inference: posterior probabilities of E_i, A_i and S_i

$$r \equiv P(S_i = 1 \mid E_i = 1)$$
$$= P(S_i = 1, E_i = 1) / P(E_i = 1)$$
$$= P(S_i = 1, E_i = 1, C_i = 0) / P(E_i = 1) + P(S_i = 1, E_i = 1, C_i = 1) / P(E_i = 1)$$
$$= 0 + P(S_i = 1, C_i = 1 \mid E_i = 1) \quad \text{[satisfaction only upon click]}$$
$$= P(S_i = 1 \mid C_i = 1)\, P(C_i = 1 \mid E_i = 1)$$

Unusualness of Handwriting



Probabilistic Model for Handwriting Style



QDE Features

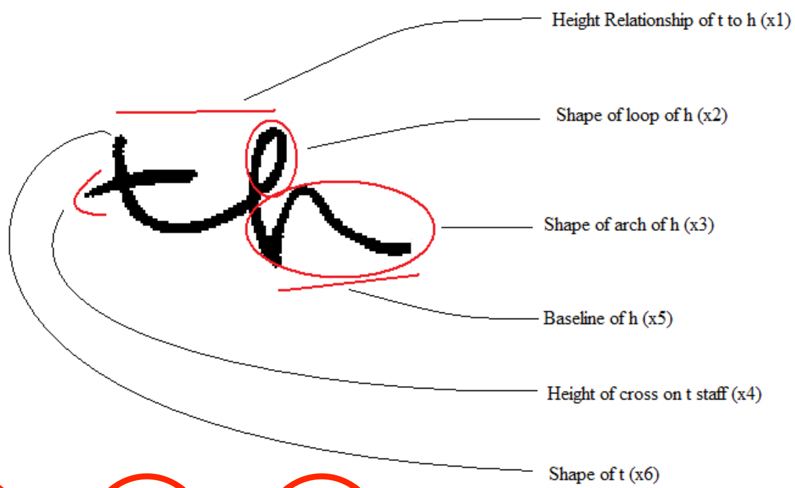
$P(R,L,A,C,B,T)$

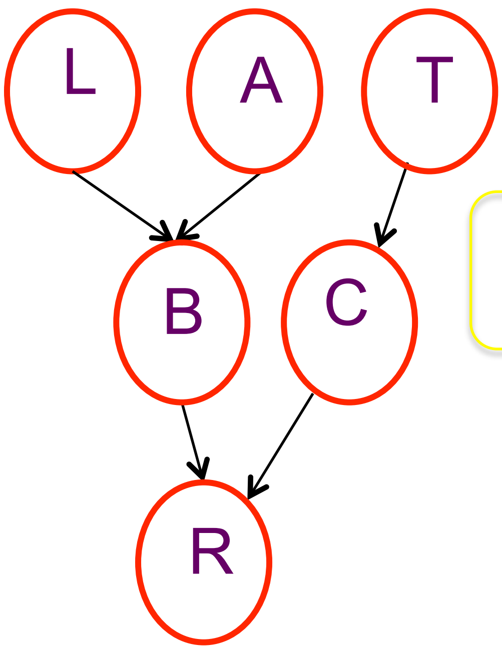
R = **Height Relationship of *t* to *h***	L = **Shape of Loop of *h***	A = **Shape of Arch of *h***	C = **Height of Cross on *t* staff**	B = **Baseline of *h***	S = **Shape of *t***
r^0= *t* shorter than *h*	l^0 = retraced	a^0 = rounded arch	c^0 = upper half of staff	b^0 = slanting upward	s^0 = tented
r^1 = *t* even with *h*	l^1 = curved right side and straight left side	a^1 = pointed	c^1 = lower half of staff	b^1 = slanting downward	s^1 = single stroke
r^2 = *t* taller than *h*	l^2 = curved left side and straight right side	a^2 =no set pattern	c^2 = above staff	b^2 = baseline even	s^2 = looped
r^3 = no set pattern	l^3 = both sides curved		c^3 = no fixed pattern	b^3 = no set pattern	s^3 = closed
	l^4 = no fixed pattern				s^4 = mixture of shapes

$$|Val(X)| = 4 \times 5 \times 3 \times 4 \times 4 \times 5 = 4{,}800$$

No of parameters = 4,799

BN for "th"





$$P(L,A,T,B,C,R) = P(L)P(A)P(T)$$
$$P(B \mid A,L)P(C \mid T)P(R \mid B,C)$$

Joint Probability

No of parameters=

$$4+2+59+4+19+63=151$$

Instead of 4,800

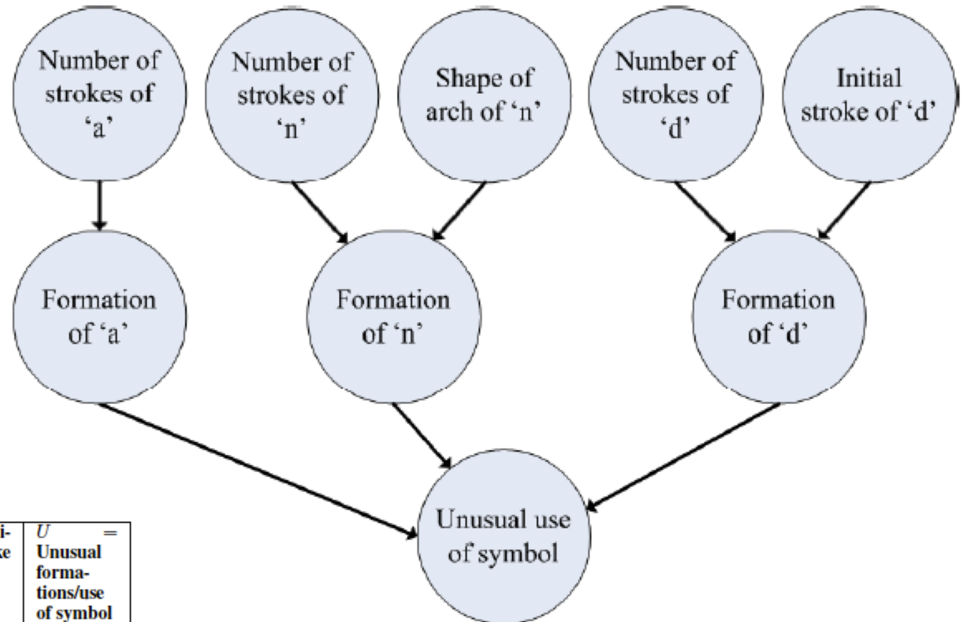
P(C|T)

x_6 \ x_4	*a*	*b*	*c*	*d*
a	0.80	0.08	0.04	0.08
b	0.47	0.37	0.01	0.15
c	0.59	0.29	0.06	0.06
d	0.76	0.12	0.03	0.09
e	0.45	0.18	0.02	0.36

P(B|L, A)

x_5	*a*	*b*	*c*	*d*
$x_2=a, x_3=a$	0.097	0.091	0.59	0.22
$x_2=a, x_3=b$	0.11	0.13	0.47	0.29
$x_2=a, x_3=c$	0.09	0.16	0.31	0.43
$x_2=b, x_3=a$	0.22	0.11	0.44	0.22
$x_2=b, x_3=b$	0.29	0.16	0.33	0.21
$x_2=b, x_3=c$	0.14	0.14	0.43	0.29
$x_2=c, x_3=a$	0.33	0.17	0.17	0.33
$x_2=c, x_3=b$	0.25	0.25	0.25	0.25
$x_2=c, x_3=c$	0.20	0.20	0.40	0.20
$x_2=d, x_3=a$	0.12	0.12	0.41	0.35
$x_2=d, x_3=b$	0.18	0.15	0.51	0.15
$x_2=d, x_3=c$	0.17	0.17	0.33	0.33
$x_2=e, x_3=a$	0.04	0.11	0.61	0.25
$x_2=e, x_3=b$	0.24	0.12	0.24	0.39
$x_2=e, x_3=c$	0.09	0.03	0.42	0.45

BN for “and”



AN = No. of strokes for formation of a	AS = Formation of Staff of a	NN = No. of strokes for formation of n	NS = Formation of staff of n	NA = Shape of Arch of n	DN = No. of strokes for formation of d	DS = Formation of staff of d	DI = Initial Stroke of d	U = Unusual formations/use of symbol
an^0 = one continuous	as^0 = tented	nn^0 = one continuous	ns^0 = tented	na^0 = pointed	dn^0 = one continuous	ds^0 = tented	di^0 = top of staff	u^0 = formation
an^1 = two strokes	as^1 = retraced	nn^1 = two strokes	ns^1 = retraced	na^1 = rounded	dn^1 = two strokes	ds^1 = retraced	di^1 = bulb	u^1 = symbol
an^2 = three strokes	as^2 = looped	nn^2 = three strokes	ns^2 = looped	na^2 = no fixed pattern	dn^2 = three strokes	ds^2 = looped	di^2 = undetermined	u^2 = none
an^3 = upper case	as^3 = no staff	nn^3 = uppercase	ns^3 = no staff		dn^3 = upper case	ds^3 = single down	di^3 = no fixed pattern	
an^4 = no fixed pattern	as^4 = single line down	nn^4 = no fixed pattern	ns^4 = no fixed pattern		dn^4 = no fixed pattern	ds^4 = single up		
	as^5 = no fixed pattern					ds^5 = no fixed pattern		

No of parameters needed= 688
(Less than 1%)

Nine variables
No of parameters needed= 809,999

Inference

Inference and Queries with PGMs

- Inference: Probabilistic Models used to answer queries
- Query Types
  1. Probability Queries
     - Query has two parts
       - *Evidence*: a subset E of variables and their instantiation e
       - *Query Variables*: a subset Y of random variables in network
  2. MAP Queries
     - Maximum a posteriori probability
     - Also called MPE (Most Probable Explanation)

Inferring the Probability of Evidence

Probability Distribution of Evidence

$$P(L,C) = \sum_{A,T,B,R} P(L,A,T,B,C,R)$$ Sum Rule of Probability

$$= \sum_{A,T,B,R} P(L)P(A)P(T)P(B \mid A,L)P(C \mid T)P(R \mid B,C)$$ From the Graphical Model

Probability of Evidence

$$P(L=l^0, C=c^1) = \sum_{A,T,B,R} P(L=l^0)P(A)P(T)P(B \mid A,L)P(C=c^1 \mid T)P(R \mid B,C=c^1)$$

More Generally

$$P(E=e) = \sum_{X \setminus E} \prod_{i=1}^{n} P(X_i \mid pa(X_i)) |_{E=e}$$

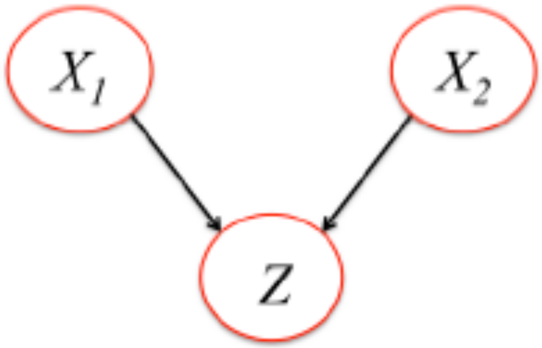
- An intractable problem
  - #P complete

P: solution in polynomial time
NP: verified in polynomial time
#P complete: how many solutions

- Tractable when tree-width is less than 25
- Approximations are usually sufficient (hence sampling)
  - When $P(Y=y|E=e)=0.29292$, approximation yields 0.3

Inference: Rarity

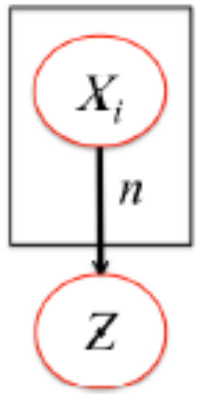
PRC



$$\rho = P(z^0) = \sum\sum P(z^0|X_1, X_2)P(X_1)P(X_2)$$

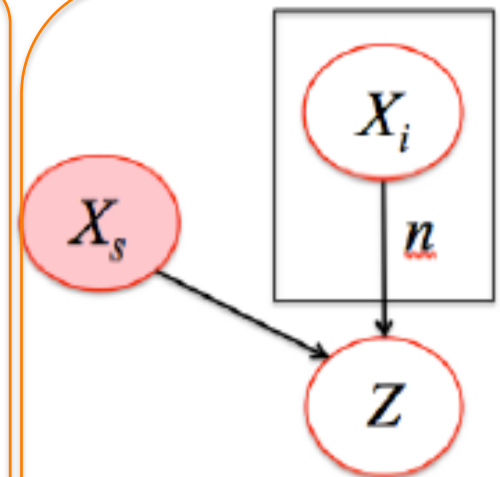
$$P(z^0|X_1, X_2) = \begin{cases} 1 \text{ if } d(X_1, X_2) \le \epsilon \\ 0, otherwise, \end{cases}$$

nPRC



$$\rho[n] = 1 - (1-\rho)^{\frac{n(n-1)}{2}}.$$

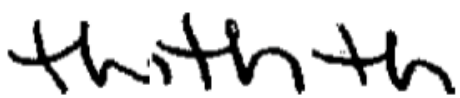
Conditional nPRC



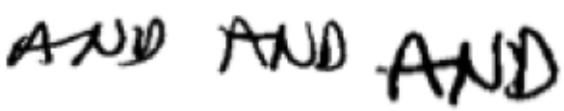
$$p(Z = 1|X_s) = \sum_{\mathbf{X}} p(Z = 1|X_s, \mathbf{X})p(\mathbf{X})$$

For identical match $1 - (1 - P(X_s))^n$

Rare



nPRC=1.17 x10^{-5}



nPRC=2.14 x 10^{-8}

Common

nPRC=0.156

nPRC=0.166

Learning

Learning Problems with PGMs

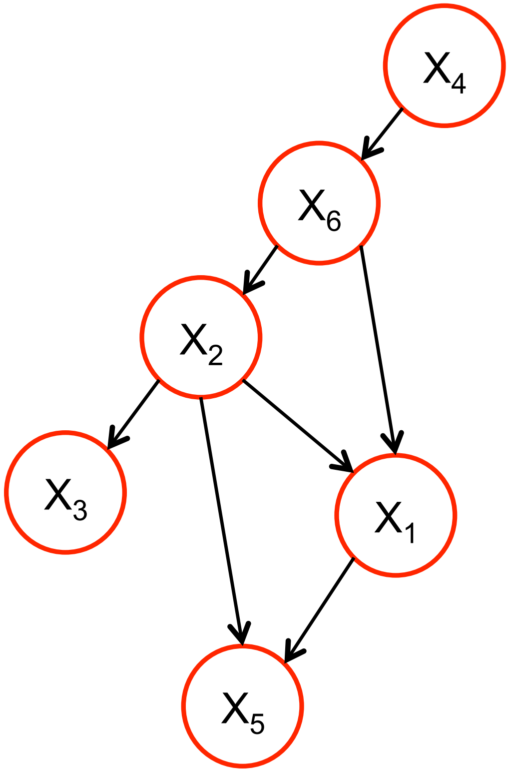
- Parameter Learning (given structure)

Bayesian Networks	Markov Networks
Local normalization within each CPD	Global normalization constant (the partition function)
Estimate local groups of parameters separately	Global parameter coupling across the network (even MLE has no closed form)

Data Collection

- Structure Learning
  - Search through network space
- Partial Data
  - EM

Parameter Learning For BN



Max Likelihood Est $P(x_5|x_1,x_2)$

	$X_5 = 0$	$X_5 = 1$	$X_5 = 2$	$X_5 = 3$
$X_1 = 0, X_2 = 0$	0.50	0	0	0.50
$X_1 = 0, X_2 = 1$	0	1.00	0	0
$X_1 = 0, X_2 = 2$	0.18	0.36	0.27	0.18
$X_1 = 0, X_2 = 3$	0.27	0.40	0.30	0.03
$X_1 = 0, X_2 = 4$	0.22	0.45	0.28	0.05
$X_1 = 1, X_2 = 0$	0.43	0	0.28	0.29
$X_1 = 1, X_2 = 1$	NaN	NaN	NaN	NaN
$X_1 = 1, X_2 = 2$	0.39	0.06	0.33	0.22
$X_1 = 1, X_2 = 3$	0.33	0.17	0.33	0.17
$X_1 = 1, X_2 = 4$	0.42	0.11	0.29	0.18
……	……	……	……	……

Bayesian Estimate

	$X_5 = 0$	$X_5 = 1$	$X_5 = 2$	$X_5 = 3$
$X_1 = 0, X_2 = 0$	0.29	0.14	0.29	0.29
$X_1 = 0, X_2 = 1$	0.25	0.25	0.25	0.25
$X_1 = 0, X_2 = 2$	0.25	0.38	0.25	0.12
$X_1 = 0, X_2 = 3$	0.22	0.41	0.31	0.06
$X_1 = 0, X_2 = 4$	0.16	0.52	0.25	0.07
$X_1 = 1, X_2 = 0$	0.29	0.14	0.29	0.29
$X_1 = 1, X_2 = 1$	0.25	0.25	0.25	0.25
$X_1 = 1, X_2 = 2$	0.37	0.05	0.47	0.11
$X_1 = 1, X_2 = 3$	0.33	0.22	0.33	0.11
$X_1 = 1, X_2 = 4$	0.38	0.13	0.29	0.20
……	……	……	……	……

Bayesian Estimation

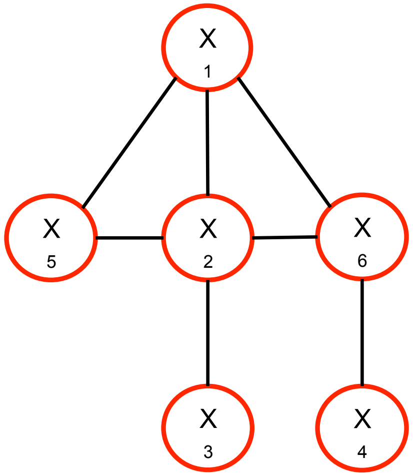
Prior $\boldsymbol{\theta} \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k) \quad \alpha_1 = \ldots = \alpha_k = 1$

Likelihood $O = \{o_1, \ldots, o_k\} \sim \text{Multinomial}(\theta_1, .., \theta_k)$

Posterior $\boldsymbol{\theta}|O \sim \text{Dirichlet}(\alpha'_1, \ldots, \alpha'_k)$

$$\alpha'_i = \alpha_i + o_i, \text{for } i = 1, \ldots, k$$

Parameter Learning for MN



Joint distribution for pairwise MN

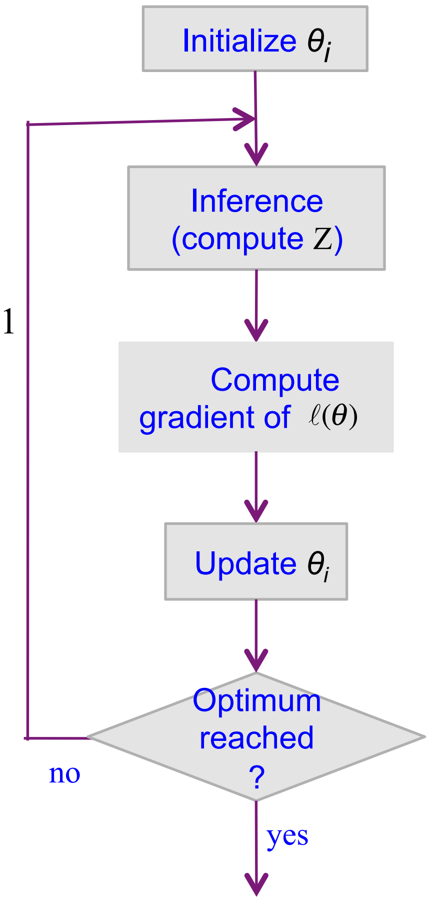
$$p(\mathrm{X}) = \frac{1}{Z}\phi_1(X_1,X_2)\cdot\phi_2(X_1,X_5)\cdot\phi_3(X_1,X_6)$$
$$\cdot\phi_4(X_2,X_5)\cdot\phi_5(X_2,X_6)\cdot\phi_6(X_2,X_3)\cdot\phi_7(X_4,X_6)$$
$$\cdot\varphi_1(X_1)\cdot\varphi_2(X_2)\cdot\varphi_3(X_3)\cdot\varphi_4(X_4)\cdot\varphi_5(X_5)\cdot\varphi_6(X_6)$$

No of Parameters θ_i:

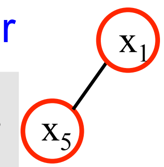
$$20+16+20+20+25+15+20+4+5+3+4+4+5=161$$

e.g.,

$$\theta_{21} = \log\phi_1(X_1=0, X_5=0)$$
$$\theta_{22} = \log\phi_1(X_1=0, X_5=1)$$
$$\vdots$$
$$\theta_{36} = \log\phi_1(X_1=3, X_5=3)$$



Estimated edge potential for



Edge Potential	$X_5 = 0$	$X_5 = 1$	$X_5 = 2$	$X_5 = 3$
$X_1 = 0$	3.35	26.10	5.99	1.42
$X_1 = 1$	1.54	2.14	1.76	0.94
$X_1 = 2$	20.01	69.75	33.49	14.90
$X_1 = 3$	2.99	9.12	4.71	2.25

Log-linear model

$$P(x_1,\ldots,x_n:\theta) = \frac{1}{Z(\theta)}\exp\left\{\sum_{i=1}^{k}\theta_i f_i(D_i)\right\}$$

n: # variables, k: # cliques θ_i: parameters

Log-likelihood of M i.i.d. samples

$$\ell(\theta) = \sum_{i=1}^{k}\theta_i\left(\sum_m f_i(\xi[m])\right) - M\ln\sum_{\xi}\exp\left(\sum_{i=1}^{k}\theta_i f_i(\xi)\right)$$

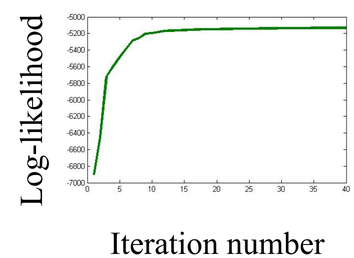
Gradient of log-likelihood

$$\frac{\partial}{\partial\theta_i}\frac{1}{M}\ell(\theta) = \frac{\sum_m f_i(\xi[m])}{M} - \frac{\sum_\xi f_i(\xi)\exp\left(\sum_{i=1}^{k}\theta_i f_i(\xi)\right)}{\sum_\xi \exp\left(\sum_{i=1}^{k}\theta_i f_i(\xi)\right)}$$

Concave, ***BUT*** no analytical maximum
=> Use iterative gradient ascent

Inference step for Z: Computes unnormalized prob for every setting of X => expensive

- Approximate inference
  - particle-based methods (MCMC sampling)
  - global algorithm (belief prop, mean-field)
- Approximate objective
  - Not as much inference
  - Pseudo-likelihood, maxent

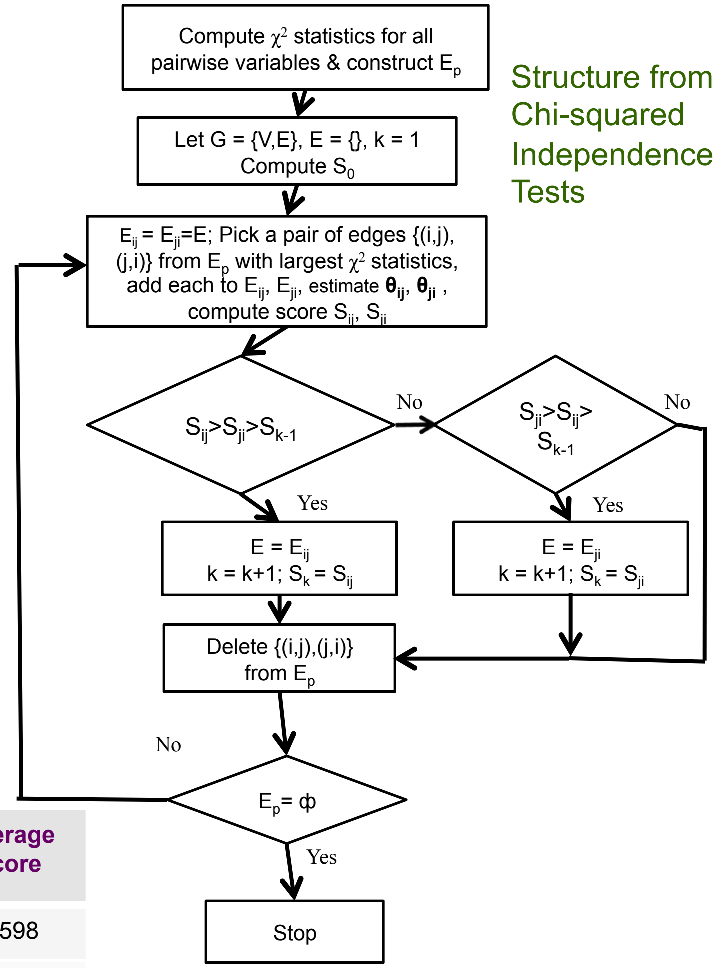


Structure Learning of BNs

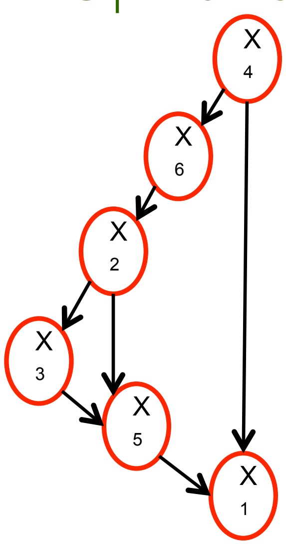
- Problem: Many perfect maps for distribution P*
- Goal: Asymptotically recover G*'s equivalence class
- Search through space of BNs
  - Score function for each BN
  - $Score_L (G : D) = \text{log-likelihood} (\theta_G : D)$
    - θ_G are parameters of G

X_1	X_2	X_3	X_4	X_5	X_6
Height Relation	Shape of loop of 'h'	Shape of arch of 'h'	Height of 't' cross	Baseline of 'h'	Shape of 't'

Structure from Chi-squared Independence Tests



G_1=Human



G_2=Based on Chi-sq tests

Dependency

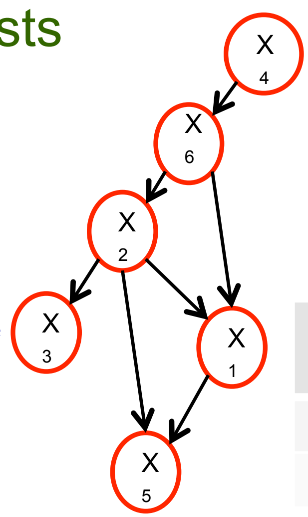
$\chi 2\ (X_4, X_6) = 224$

$\chi 2\ (X_6, X_2) = 167$

Conditional independence

$\chi 2\ (H_0\!: X_4 \perp X_2 \mid X_6) = 42$

$\chi 2\ (H_0\!: X_4 \perp X_1 \mid X_6) = 43$



3-fold Cross Validation	Average Score
G_1	-2598
G_2	**-2591**

$$\chi^2(D) = \sum_{i,j} \frac{(\text{Observed count of } [x_i, y_j] - \text{Expected count of } [x_i, y_j])^2}{\text{Expected count of } [x_i, y_j]}$$

Structure Learning of MNs

- Information-theoretic Chow-Liu algorithm

Algorithm for structure learning:

1. Estimate empirical probability:

$$P(X_1 = N) = \frac{\sum_D 1[X_1 = N]}{\sum_D 1}$$

P	X_1	X_2	X_3	X_4	X_5	X_6
0	0.20	0.05	0.10	0.00	0.29	0.17
1	0.16	0.00	0.62	0.36	0.24	0.04
2	0.37	0.12	0.28	0.30	0.30	0.07
3	0.27	0.08		0.34	0.17	0.71
4		0.75				0.00

2. Calculate all marginal entropies:

$$H(X_1) = -\sum_{X_1} P(X_1)\log(P(X_1))$$

and all pair-joint entropies:

$$H(X_1,X_2) = -\sum_{X_1,X_2} P(X_1,X_2)\log(P(X_1,X_2))$$

H	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1.34	2.14	2.20	2.41	2.65	2.17
X_2	2.14	0.84	1.69	1.92	2.16	1.66
X_3	2.20	1.69	0.89	1.99	2.23	1.74
X_4	2.42	1.92	1.99	1.10	2.44	1.89
X_5	2.66	2.16	2.23	2.44	1.36	2.21
X_6	2.17	1.66	1.75	1.89	2.20	0.87

3. Calculate mutual information:

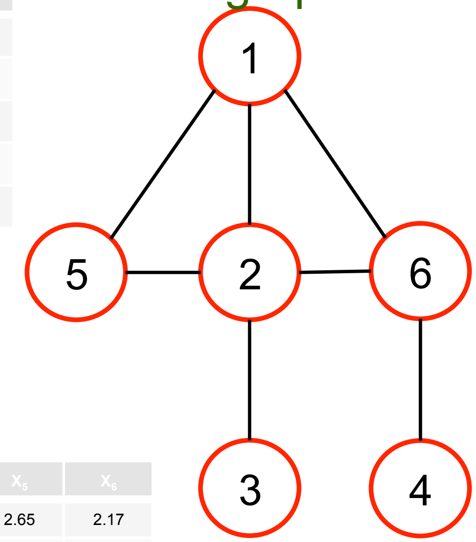
$$I(X_1,X_2) = H(X_1) + H(X_2) - H(X_1,X_2)$$

4. Include edges (X_i, X_j) in to the structure if $I(X_1,X_2) \geq threshold$

I	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1.33	0.03	0.02	0.02	0.03	0.04
X_2	0.03	0.83	0.03	0.01	0.03	0.05
X_3	0.02	0.03	0.89	0.01	0.02	0.02
X_4	0.02	0.01	0.01	1.10	0.02	0.09
X_5	0.03	0.03	0.02	0.02	1.36	0.03
X_6	0.04	0.05	0.02	0.08	0.03	0.87

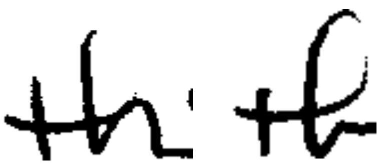
I(X1,X2)≥ threshold

Gives graph



Rare and Common Style Inferences from PGMs

Rare Styles : Looped or tented ‘t’, loop of ‘h’ with both sides curved



Doc: 199a
Score : -12

Doc: 409c
Score : -12

Doc: 124c
Score : -11

Doc: 1434b
Score : -11

Common Styles: Single stroke ‘t’, retraced ‘h’, pointed arch of ‘h’, baseline of ‘h’ slanting down, ‘t’ taller, cross of ‘t’ below

Doc: 40b
Score : -4

Doc: 130b
Score : -4

Doc: 1007c
Score : -4

Doc: 685a
Score : -4

All scores in log-likelihood

Summary and Conclusion

- Machine Learning
  - Several generations, with beginnings in DAR field
  - Necessary for changing high volume data
    - To classify, regress, infer, collectively label
- PGMs able to handle complexity
  - BN and MN are expressive
  - Allow incorporating domain knowledge
  - Provide relationships between models
- Computational Forensics Application
  - Handwriting rarity is inferred from PGMs