Cognitive
Robotics
Lab

Statistical Inference and PGM

A Short Lecture on Statistical Machine Learning

Sajad Azami & Taher Ahmadi

Foundations of Data Mining

CEIT - Amirkabir University of Technology

sajjadaazami@gmail.com
14Taher@gmail.com

Spring 2017

Cognitive Robotics Lab

Outlines

- Context Definition
- General SML Concepts and CDF Estimation
- Models and Statistical Inference
- Conditional Independence
- PGM
- Applications
- References

Why Statistics is Important?

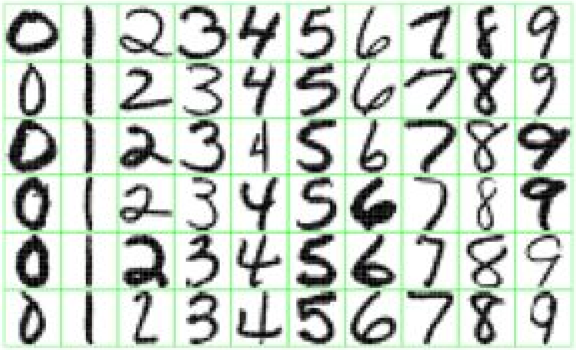


Figure 1.2: *Examples of handwritten digits from U.S. postal envelopes.*

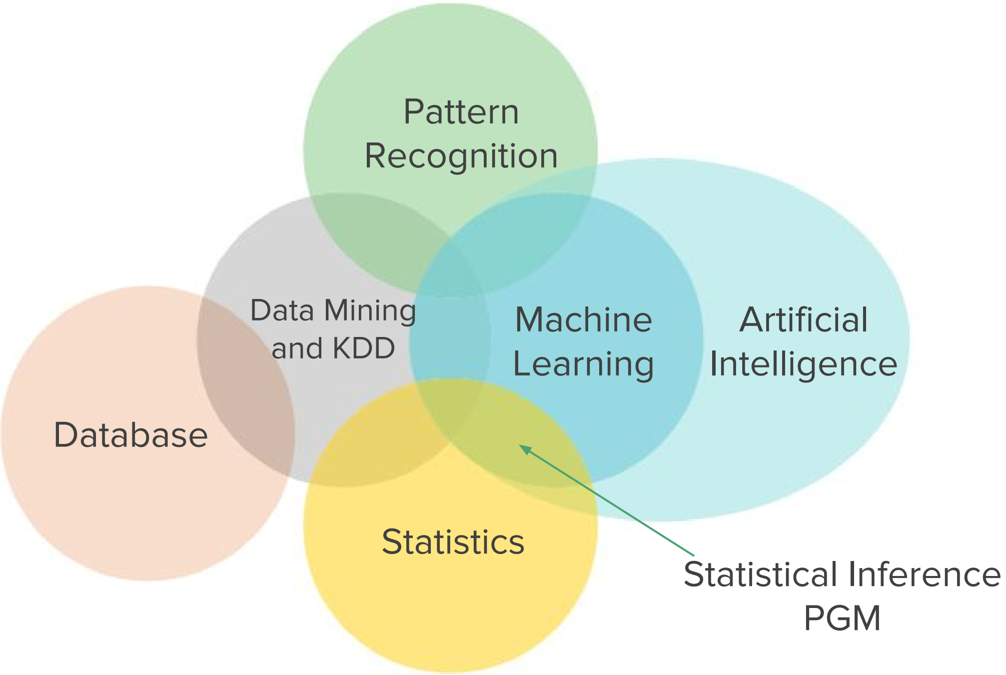
Why Statistics is Important?

TECHNOLOGY

For Today's Graduate, Just One Word: Statistics

By STEVE LOHR AUG. 5, 2009

Where Are We?



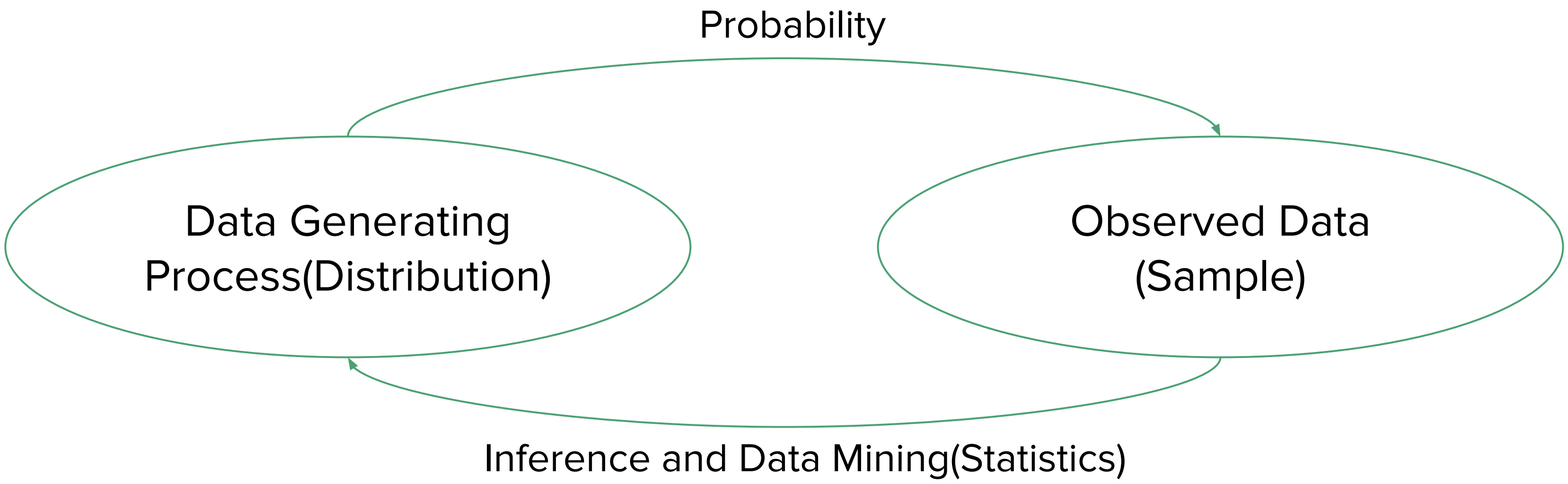
Pattern Recognition
Data Mining and KDD
Machine Learning
Artificial Intelligence
Database
Statistics
Statistical Inference
PGM

Why Statistical Models?

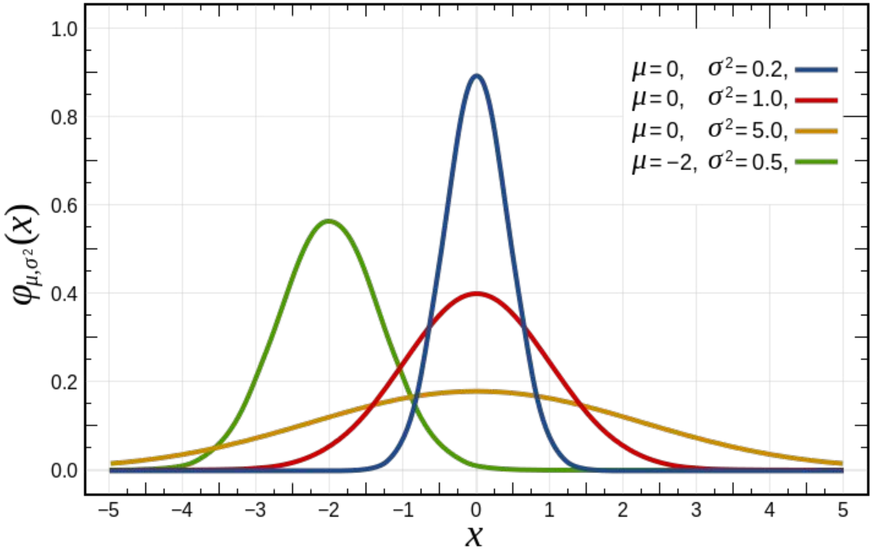
- Partial discovery of state of the world
- Noisy observation(blood test)
- Phenomena not covered by our models(diseases)
- Inherent stochasticity

Probability Theory

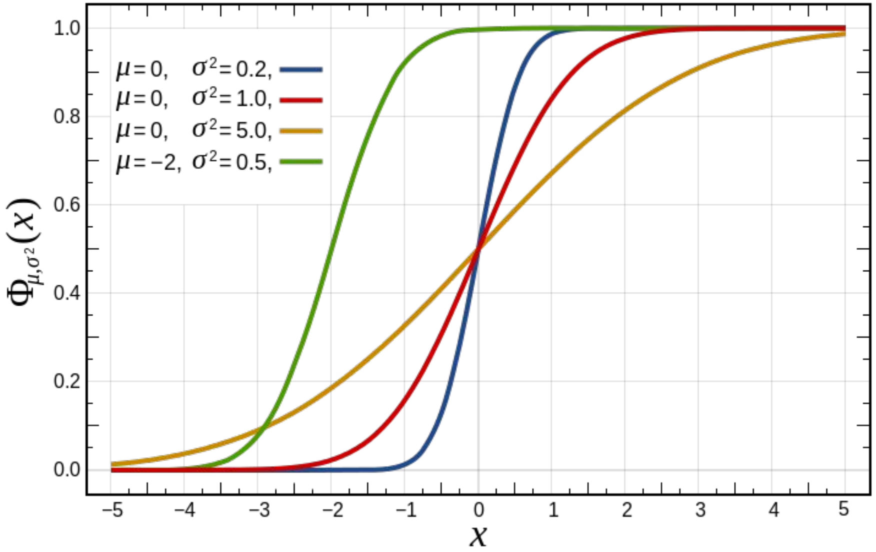
Statistics vs Probability



Distribution(Review)



providing a *relative likelihood* that the value of the random variable would equal that sample



right-continuous, non-decreasing
normalized

Joint Distribution

Intelligence(I): low, high

Difficulty(D): easy, hard

Grade(G): A, B, C

12 Independence Parameters

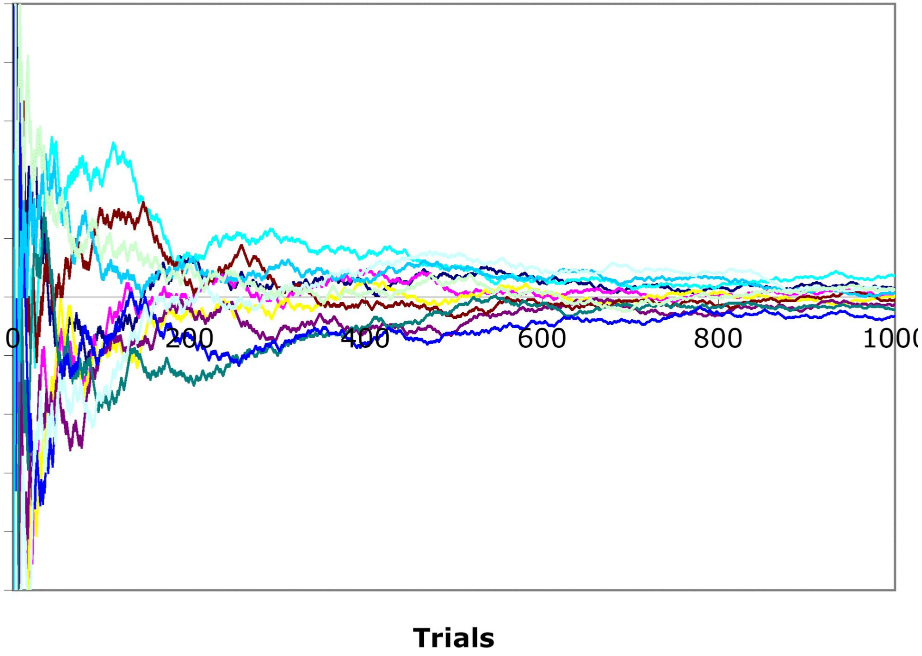
I	D	G	Prob.
i^0	d^0	g^1	0.126
i^0	d^0	g^2	0.168
i^0	d^0	g^3	0.126
i^0	d^1	g^1	0.009
i^0	d^1	g^2	0.045
i^0	d^1	g^3	0.126
i^1	d^0	g^1	0.252
i^1	d^0	g^2	0.0224
i^1	d^0	g^3	0.0056
i^1	d^1	g^1	0.06
i^1	d^1	g^2	0.036
i^1	d^1	g^3	0.024

Some Basic Concepts: WLLN

Sample Mean converges **in Probability** to E(X)

$$\textit{If } X_1, \ldots, X_n \textit{ are } \mathrm{IID}$$

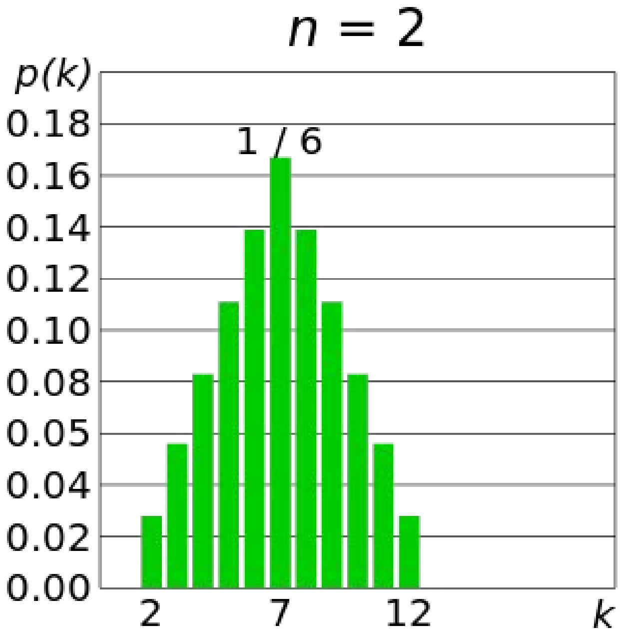
$$\textit{then } \overline{X}_n \xrightarrow{\mathrm{P}} \mu$$



Some Basic Concepts: CLT

Probability statements about Sample Mean can be approximated using a Normal distribution

$$Z_n \equiv \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \rightsquigarrow Z$$



Estimation

We have samples, we want to know about data generating process(CDF)

Estimation

We have samples, we want to know about data generating process(CDF)

Statistical Inference: the process of deducing properties of an underlying distribution by analysis of data

1. Hypothesis Testing and p-values
2. Deriving Estimates(That's why normal dist. is important)

Statistical Inference

Frequentist Inference

- Point Estimation
- Confidence Sets

Bayesian Inference

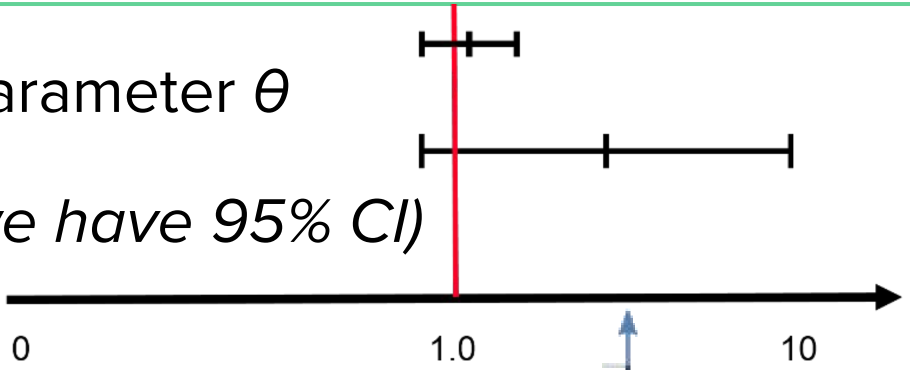
Frequentist Inference: Point Estimation

Providing a single "best guess" of some quantity of interest

Imagine tossing a fair coin and estimate p

Frequentist Inference: Confidence Sets

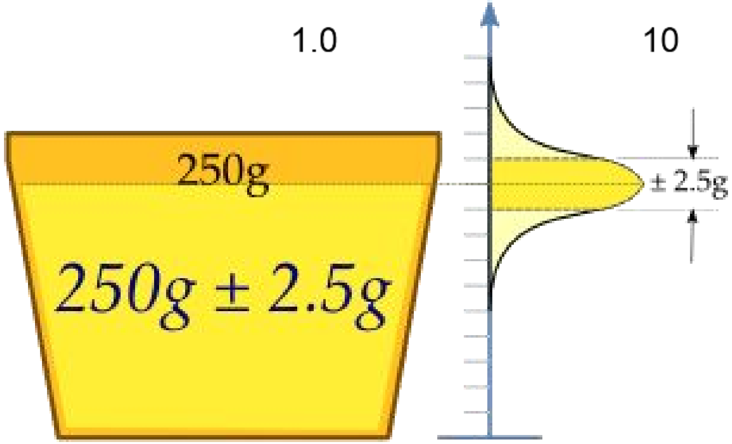
A *1-a* **confidence interval** for a parameter *θ*



(a is usually set to 0.05 so that we have 95% CI)

Interpretation:

- Repeat experiment and CI will contain true values 95% of the time
- Construct CI over time and 95% of CI's will trap the true value



Bootstrap

A nonparametric method for estimating standard errors and computing confidence intervals

1. Draw bootstrap samples n times
2. Compute statistic of interest as T_n
3. Repeat 1 and 2, *B* times to get T_n,1 … T_n,B
4. se <= sqrt(variance(Tboot))

Interval Types

Normal Interval

$$T_n \pm z_{\alpha/2}\, \widehat{\mathsf{se}}_{\,\mathrm{boot}}$$

Percentile Interval

$$C_n = \left(\theta^*_{\alpha/2},\ \theta^*_{1-\alpha/2}\right)$$

Pivotal Interval

$$C_n = \left(2\widehat{\theta}_n - \widehat{\theta}^*_{1-\alpha/2},\ 2\widehat{\theta}_n - \widehat{\theta}^*_{\alpha/2}\right)$$

Interval Types

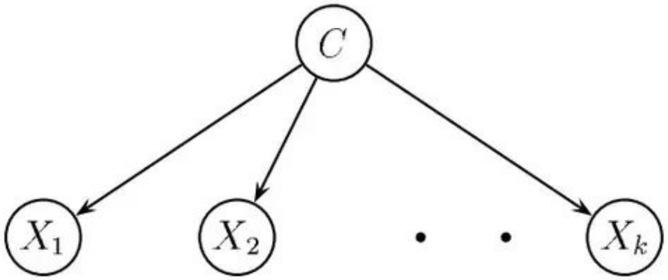
```
Normal     <- (th.hat - 2*se, th.hat + 2*se)
percentile <- (quantile(Tboot,.025), quantile(Tboot,.975))
pivotal    <- ( 2*th.hat-quantile(Tboot,.975), 2*th.hat-quantile(Tboot,.025) )
```

Bayesian Inference

1. Choose a prior distribution(flat, improper)
2. Choose a statistical model that reflects our beliefs about x
3. Update beliefs and form the **posterior** after observing data

Bayesian Inference

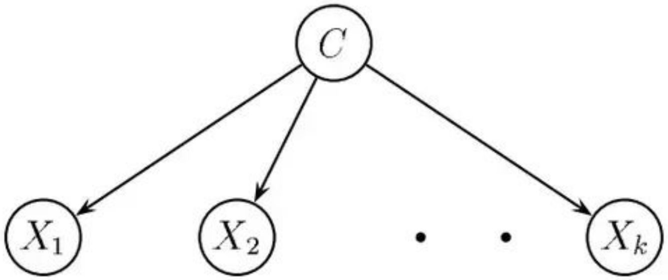
We had **Naive Bayes** as an example of probabilistic models, with strong (naive) independence assumptions between the features



Bayesian Inference

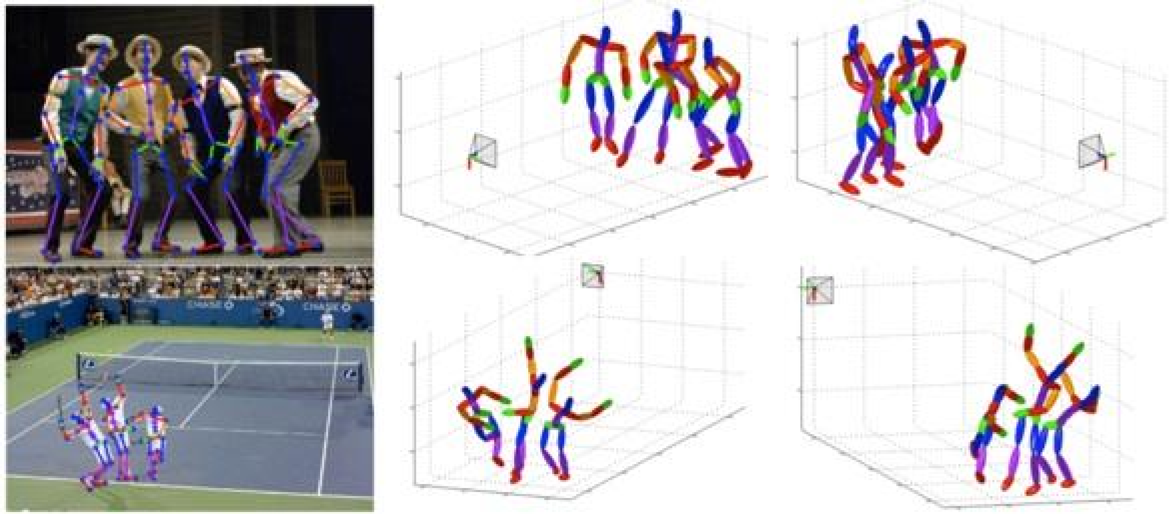
We had **Naive Bayes** as an example of probabilistic models, with strong (naive) independence assumptions between the features

What if we don't want to assume strong independence?

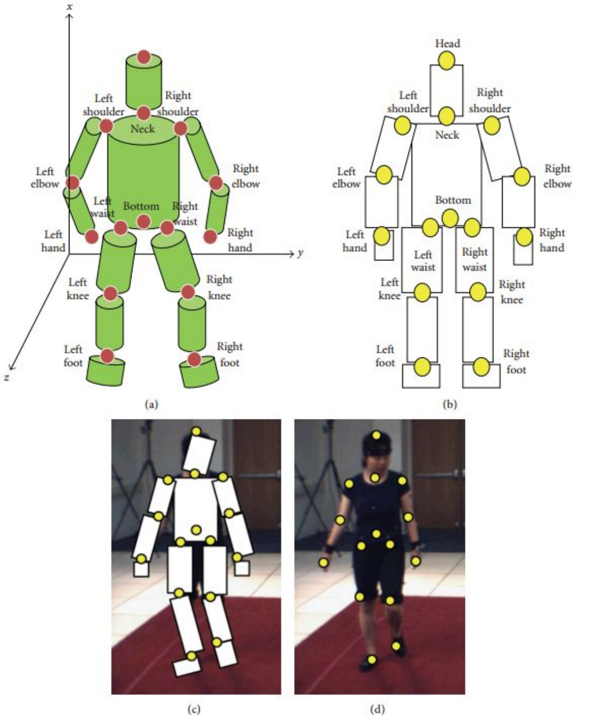


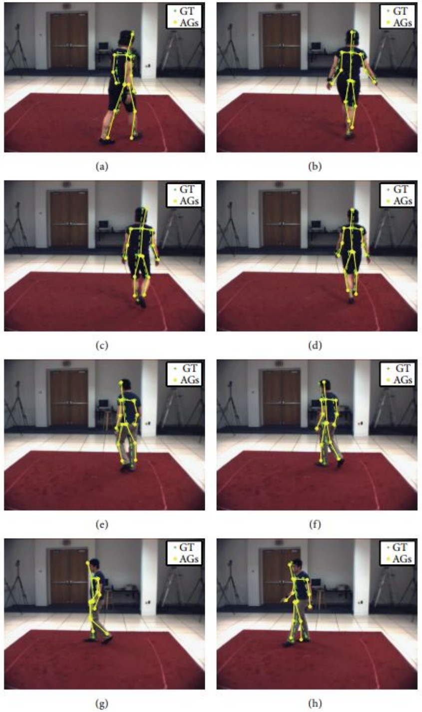
A Case Study ...

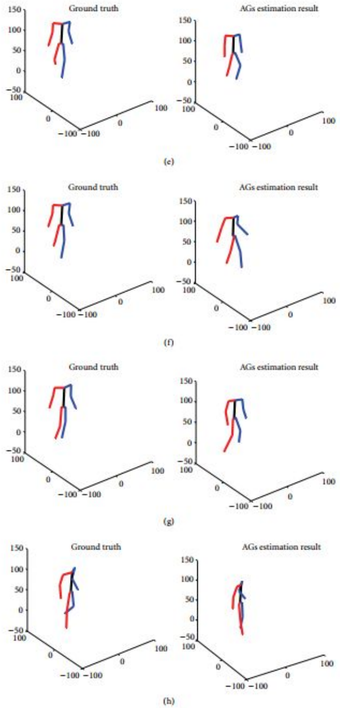
Estimating 3D Human Poses From 2D Images



Reconstructing Articulated 3D Human Poses







Graph Terminology

Node, Edge, Directed/Undirected edge,

Neighbor, Parent-Child, Node degree, Indegree, Outdegree,

Subgraph, Complete subgraph (clique), Maximal clique,

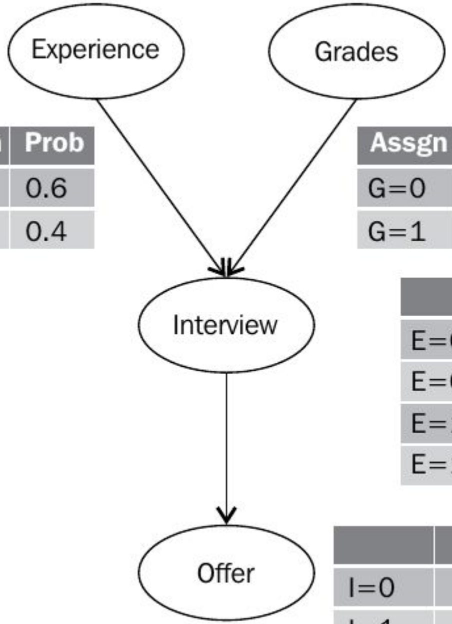
Path, trail, Cycle(Loop), Tree, Forest, DAG, PDAG

Directed Graphical Models or Bayes Networks

- Directed Acyclic Graph
  - A compact and modular representation of the joint distribution using the chain rule for Bayes network
- Conditional Probability Distribution (CPD)
  - The conditional independence assumptions between vertices

Graphical Models

- Representation
  - Directed & Undirected
  - Resoning
- Learning
  - Structure
  - Parameters
- Inference
  - Exact
  - Approximate



Assgn	Prob
E=0	0.6
E=1	0.4

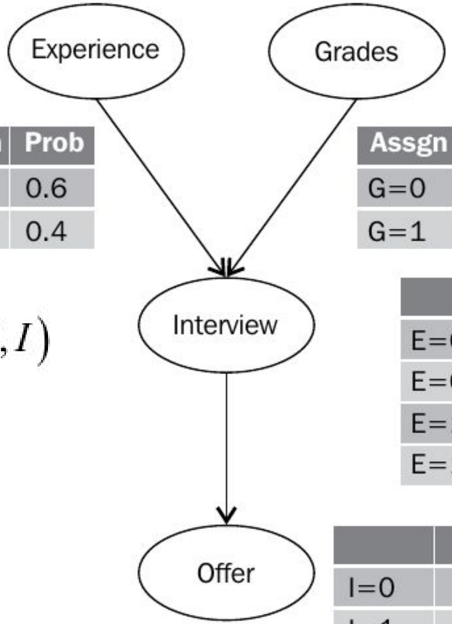
Assgn	Prob
G=0	0.7
G=1	0.3

	I=0	I=1	I=2
E=0, G=0	.8	.18	.02
E=0, G=1	.3	.6	.1
E=1, G=0	.3	.4	.3
E=1, G=1	.1	.2	.7

	O=0	O=1
I=0	.9	.1
I=1	.4	.6
I=2	.01	.99

Graphical Models

- Representation



Assgn	Prob
E=0	0.6
E=1	0.4

Assgn	Prob
G=0	0.7
G=1	0.3

	I=0	I=1	I=2
E=0, G=0	.8	.18	.02
E=0, G=1	.3	.6	.1
E=1, G=0	.3	.4	.3
E=1, G=1	.1	.2	.7

	O=0	O=1
I=0	.9	.1
I=1	.4	.6
I=2	.01	.99

$$P(E,G,I,O)=P(E)\times P(G\mid E)\times P(I\mid E,G)\times P(O\mid E,G,I)$$

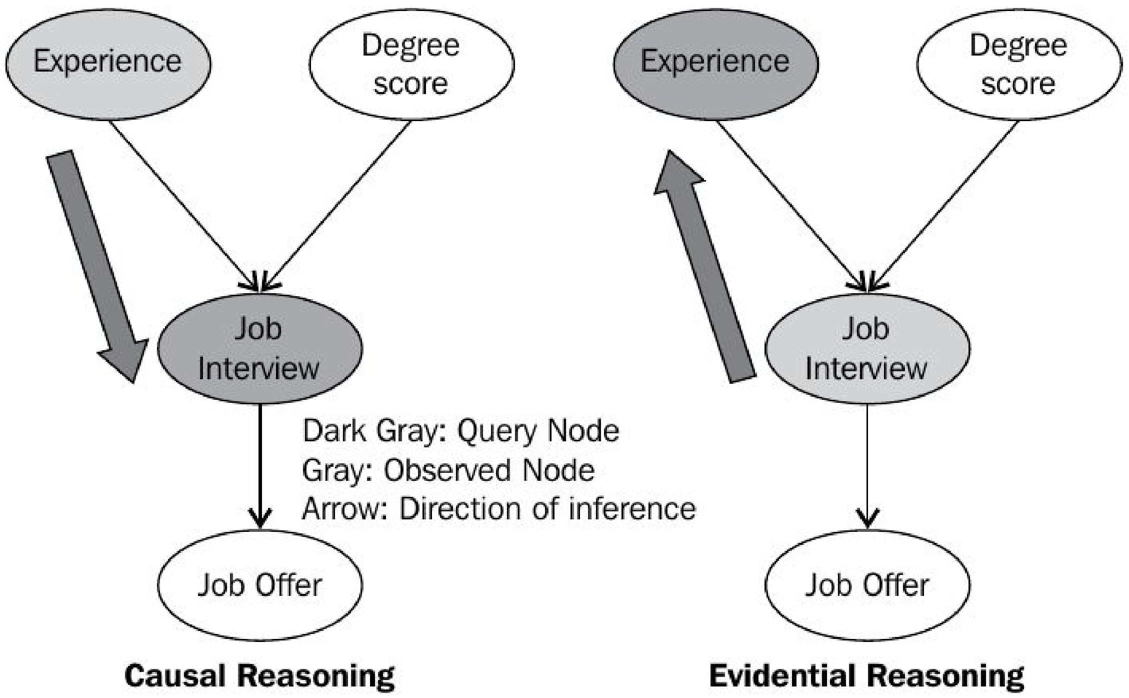
$$P(E,G,I,O)=P(E)\times P(G)\times P(I\mid E,G)\times P(O\mid I)$$

$$P(X_1,X_2,\ldots,X_n)=\prod i\, P(X_i\mid Par_G(X_i))$$

Reasoning Patterns

- Causal
- Evidential
- Inter Causal

Causal vs Evidential Reasoning



Inter-Causal Reasoning

- Explaining Away Phenomenon

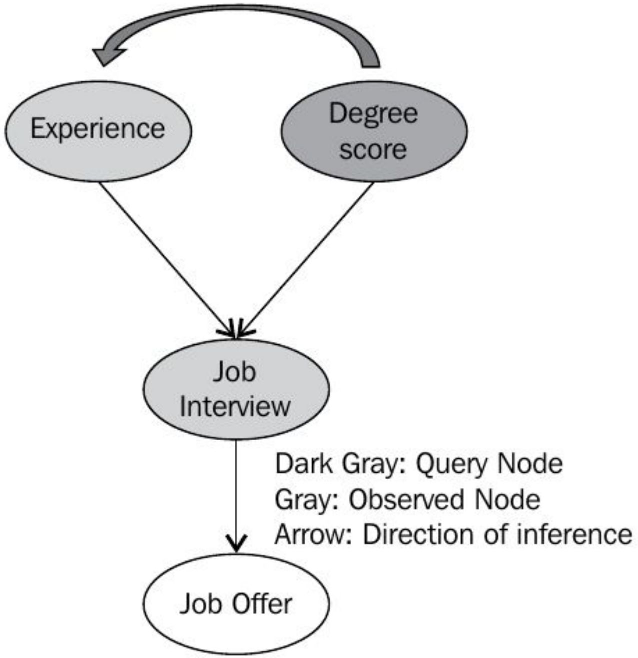
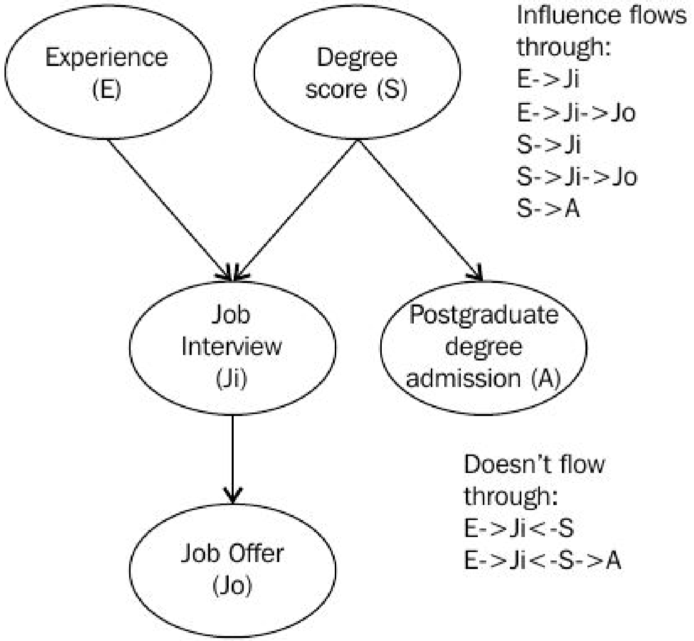


Fig x.x Intercausal Reasoning

Some Concepts

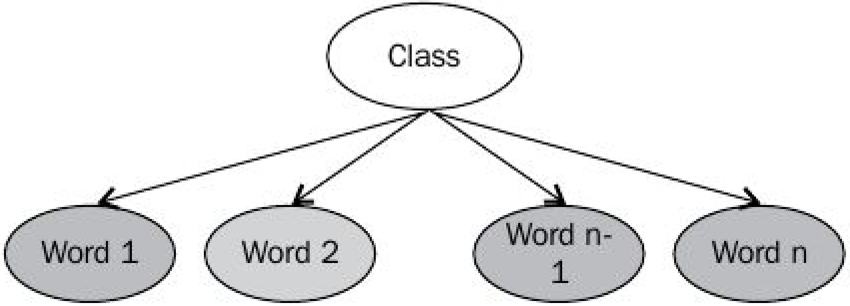
- Factorization
- I-Maps & P-Maps & I-Equivalents
- Active Trail (of influence)
- V-Structure
- D-Separation



Flow of influence without any variable being observed.

Naive-Bayes Example

-



Naive Bayes: N words which have been observed, Class unobserved

$$P(C, X_1, X_2, \ldots X_n) = P(C) \prod_{i=2}^{n} P(X_i \mid C)$$

Structure Learning

- Using:
  - Data set
  - Domain Knowledge
- Constraint-Based
  - null hypothesis testing : Pearson chi-square test
  - Graph skeleton and Finding I-maps
- Score-Based
  - The likelihood score
  - The Bayesian score

$$P(A,B)=P(A)P(B)$$

$$x^2=\sum_{i=1}^{n}\frac{(O_i-E_i)^2}{E_i}$$

$$score_l(G:D)=M\sum_{i=1}^{n} I_{\hat{p}}\left(X_i;Pa_{X_i}^{G}\right)-M\sum_i H_{\hat{p}}(X_i)$$

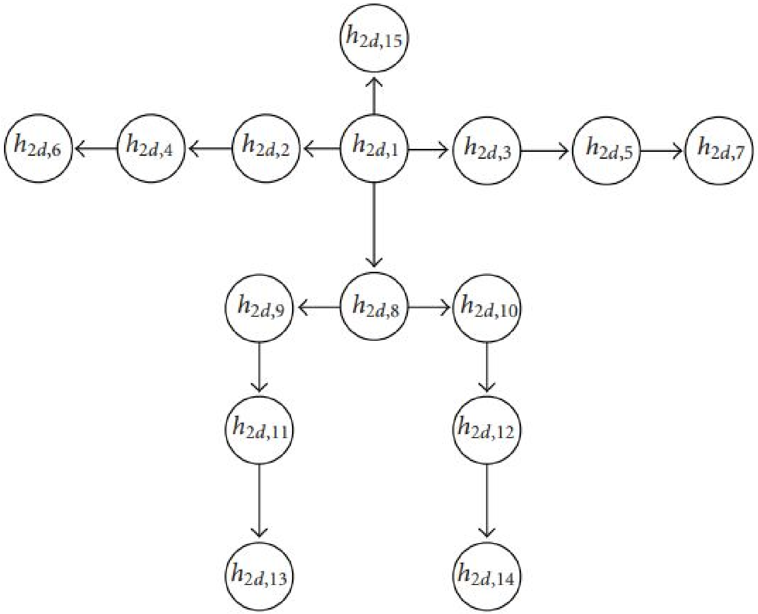
Parameter Learning

Parameters are CPD's of Random Variables in PGM

- Maximum Likelihood Estimation
- Bayesian Statistics

Back to the problem ...

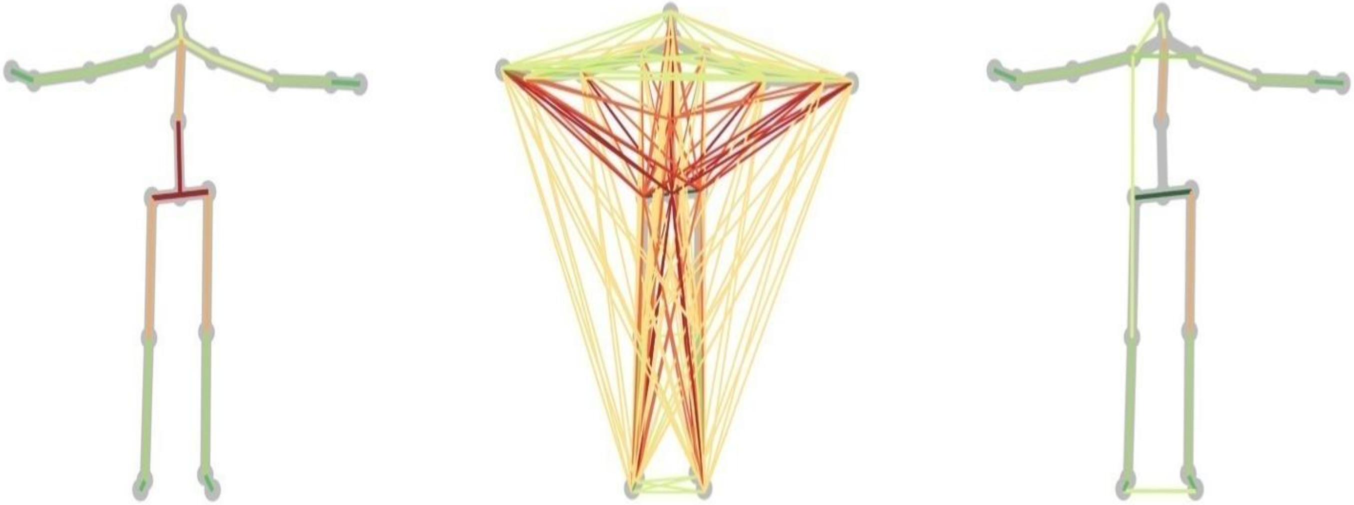
- Structure



(a)

Back to the problem ...

- Structure



Back to the problem ...

- Parameters

$$P(V) = \prod_{i=1}^{n} P(V_i \mid pa(V_i)),$$

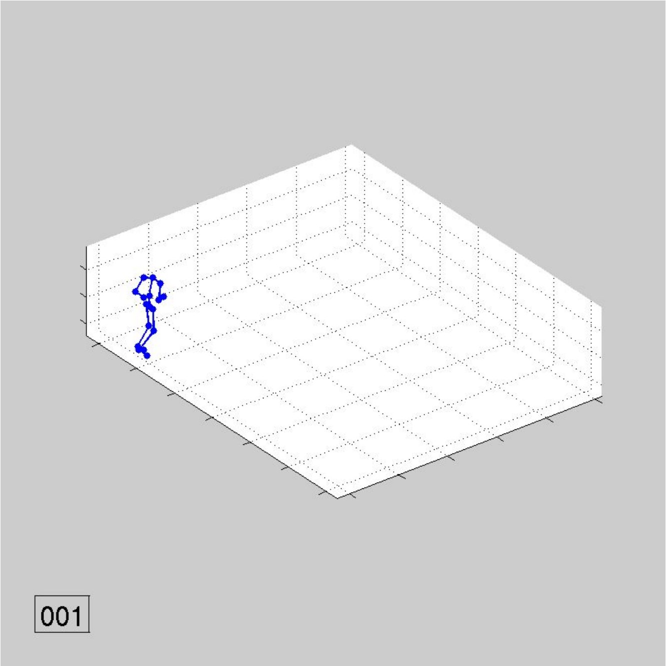
$$L_D(\theta) = \log\left\{\prod_{l=1}^{N} P(V_1[l], \ldots, V_n[l] \mid \theta)\right\}$$

$$= \sum_{i=1}^{n}\sum_{l=1}^{N} \log P(V_i[l] \mid pa_i(V_i(l)), \theta).$$

$$\widehat{\theta} = \arg\max_{\theta} L_D(\theta)0$$

Back to the problem ...

- Inference



Real World Application

References

- Wasserman, L., 2013. All of statistics: a concise course in statistical inference. Springer Science & Business Media.
- Koller, D. and Friedman, N., 2009. Probabilistic graphical models: principles and techniques. MIT press.
- Wang, Y.K. and Cheng, K.Y., 2010. A two-stage Bayesian network method for 3D human pose estimation from monocular image sequences. EURASIP Journal on Advances in Signal Processing, 2010, p.12.
- Ramakrishna, V., Kanade, T. and Sheikh, Y., 2012. Reconstructing 3d human pose from 2d image landmarks. Computer Vision–ECCV 2012, pp.573-586.

Thanks for your attention.
Any Questions?