CMPUT463/563
Probabilistic Graphical Models

Introduction

Lili Mou

Dept. Computing Science, University of Alberta

lmou@ualberta.ca

Consider the environment before printing.
Please print double-sided.

Introduction

Probability

Graph

Model

$P(X)$: The probability of event X

- Kolmogorov Axioms
  - Normalizing: $P(\Omega) = 1$ (Ω: sample space)
  - Nonnegative: $P(X) \geq 0$ for every event $X \subseteq \Omega$
  - σ-additive: For disjoint events $E_1, E_2, \cdots$

    $$P\left(\bigcup_i E_i\right) = \sum_i P(E_i)$$

    σ-additive means countably additive

    (Natural numbers are countable; real numbers are not)
- Interpretation
  - Frequentist: the frequency of X if #trials goes to infinity
  - Bayesian: Subjective belief (Is it science? Yes, science is inevitably subjective)

Introduction

Probability

Graph

Model

Cheatsheet

- Joint probability $p(X, Y)$
- Conditional probability $p(X \mid Y) = p(X, Y)/p(Y)$ when $p(Y)$ is non-zero
- Marginal probability $p(X) = \sum_y p(X, y)$
- Bayes' rule $p(X \mid Y) = \dfrac{p(Y \mid X)p(X)}{\sum_x p(Y \mid x)p(x)}$
- Expectation $\mathbb{E}_{x \sim p(X)}[f(x)] = \sum_x p(x)f(x)$

These are supposed to be known prerequisite knowledge

Introduction

Probability

Graph

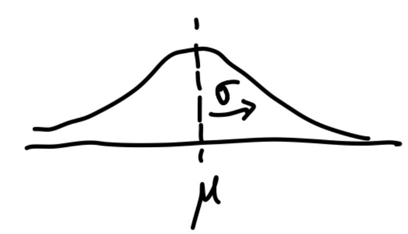
Model

Specifying a probabilistic distribution

Continuous variable

- Oftentimes, a parametric form is assumed
- E.g., 1-D Gaussian

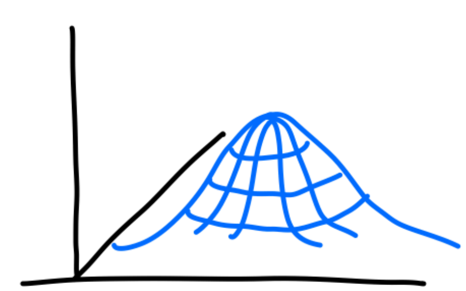
$$p(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$



Concern 1: A parametric form may not reflect true data

- E.g., Multi-dimensional Gaussian

$$p(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{1/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$



Concern 2 (Curse of dimensionality):

- #Para increases quadratically
- 1-D Gaussian distribution may be a good approximation to real data, but high-dimensional Gaussian may be very poor approximation.

Introduction

Specifying a probabilistic distribution

Discrete variable (with finite values)

- Any finite-value discrete variable can be modeled by the multinomial distribution
- A variable with K values requires $K-1$ free parameters

Value

X	$p(X)$
1	π_1
2	π_2
$\vdots$	
K	$\pi_K = 1 - \pi_1 - \pi_2 - \cdots - \pi_{K-1}$

- **Multiple** finite-value discrete variables can be modeled by a joint probability table
- Consider two variables, each taking value 0 or 1

Row #	X_1	X_2	$P(X_1, X_2)$
1	0	0	π_1
2	0	1	π_2
3	1	0	π_3
4	1	1	$\pi_4 = 1 - \pi_1 - \pi_2 - \pi_3$

How about N variables, each taking K values?

- $K^N - 1$ free variables (again, curse of dimensionality)

What if we know they are independent?

- $N(K-1)$ free variables

Introduction

Probability

Graph

Model

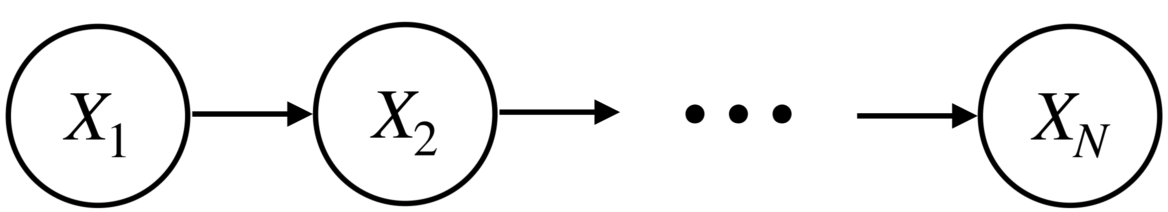
Still N variables, each taking K values

- No independencies are known
  - $K^N - 1$ free parameters
- All variables are independent
  - $N(K-1)$ free parameters
- What if we know X_i depends on X_{i-1} only for $i = 2, \cdots, N$?

$$p(X_1, \cdots X_n) = p(X_1)p(X_2 | X_1) \cdots p(X_n | X_n - 1)$$

  - For X_1, we have $K - 1$ parameters
  - For $X_i, i = 2, \cdots, N$, we have $K(K-1)$ parameters

  In total, how many parameters do we have?



Introduction

Probability

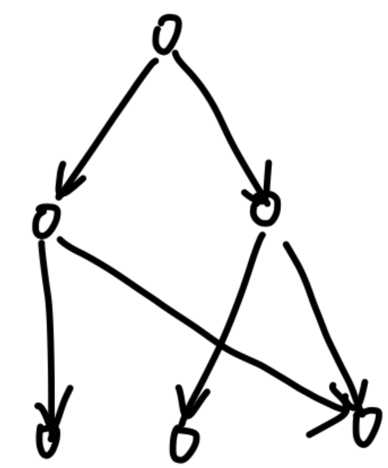
Graph

Model

Graph $G = \langle V, E \rangle$, where $E \subseteq V \times V$

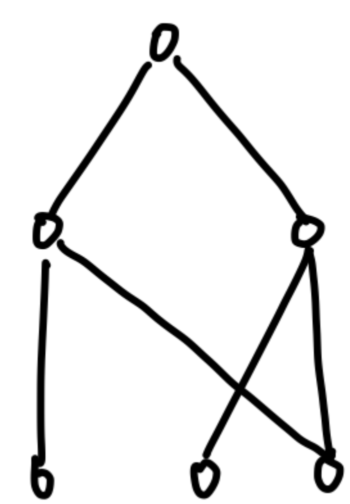
Directed graph

Relationship in a sense of "cause-and-effect"



Undirected graph

General correlation



Introduction

Probability

Graph

Model

Machine learning model

- Supervised learning
  - Training: Learn h from data $\{(x^{(i)}, y^{(i)})\}_{i=1}^{M}$
  - Inference: Given x_*, predict $\widehat{y}_* = h(x_*)$
- Unsupervised learning
  - Data are unlabeled
  - E.g., clustering, representation learning

Examples

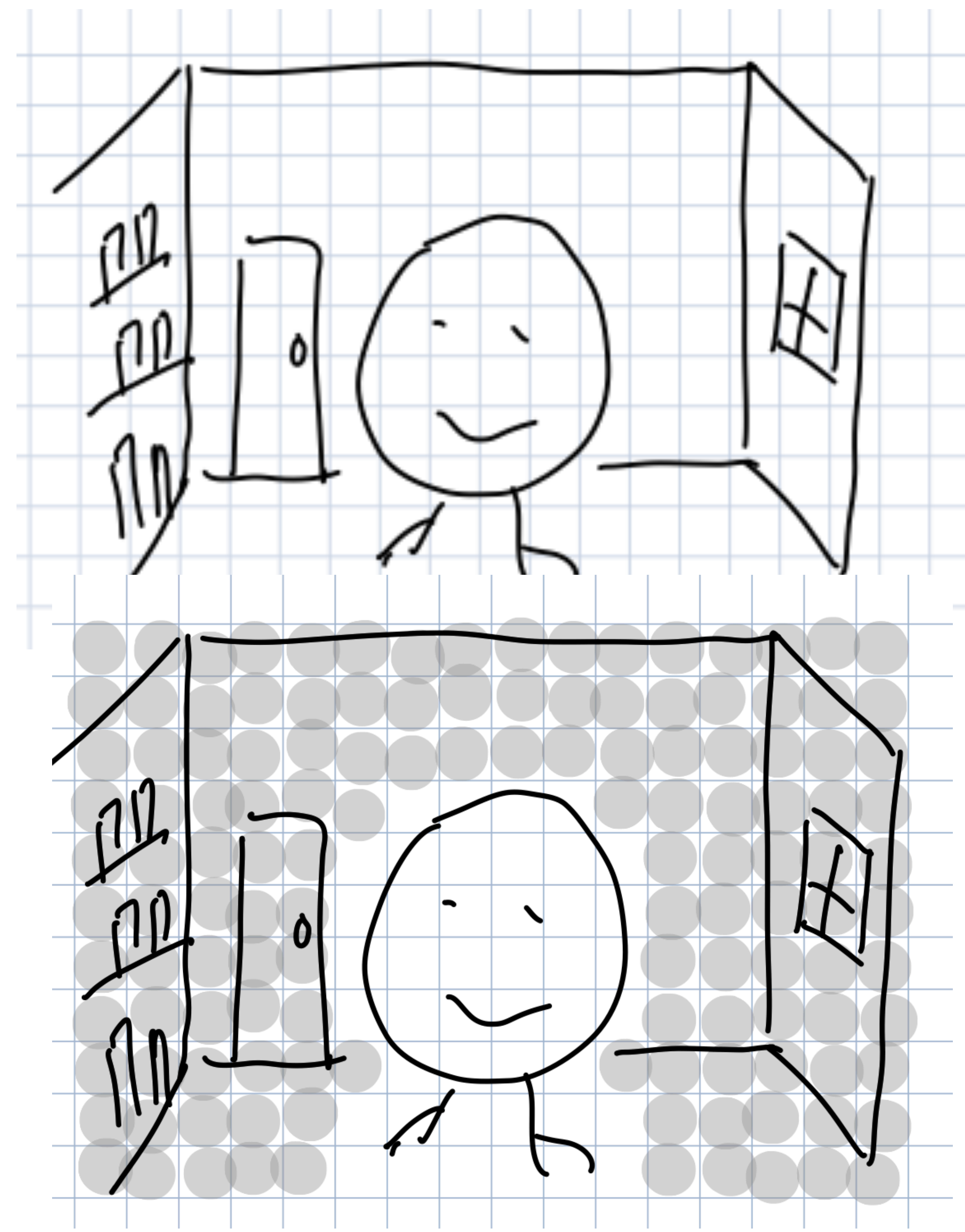
Part-of-speech (POS) tagging in Natural Language Processing

- POS tags: Pronoun, verb, determiner, noun

Pron — Verb — DT — Noun / Verb ?

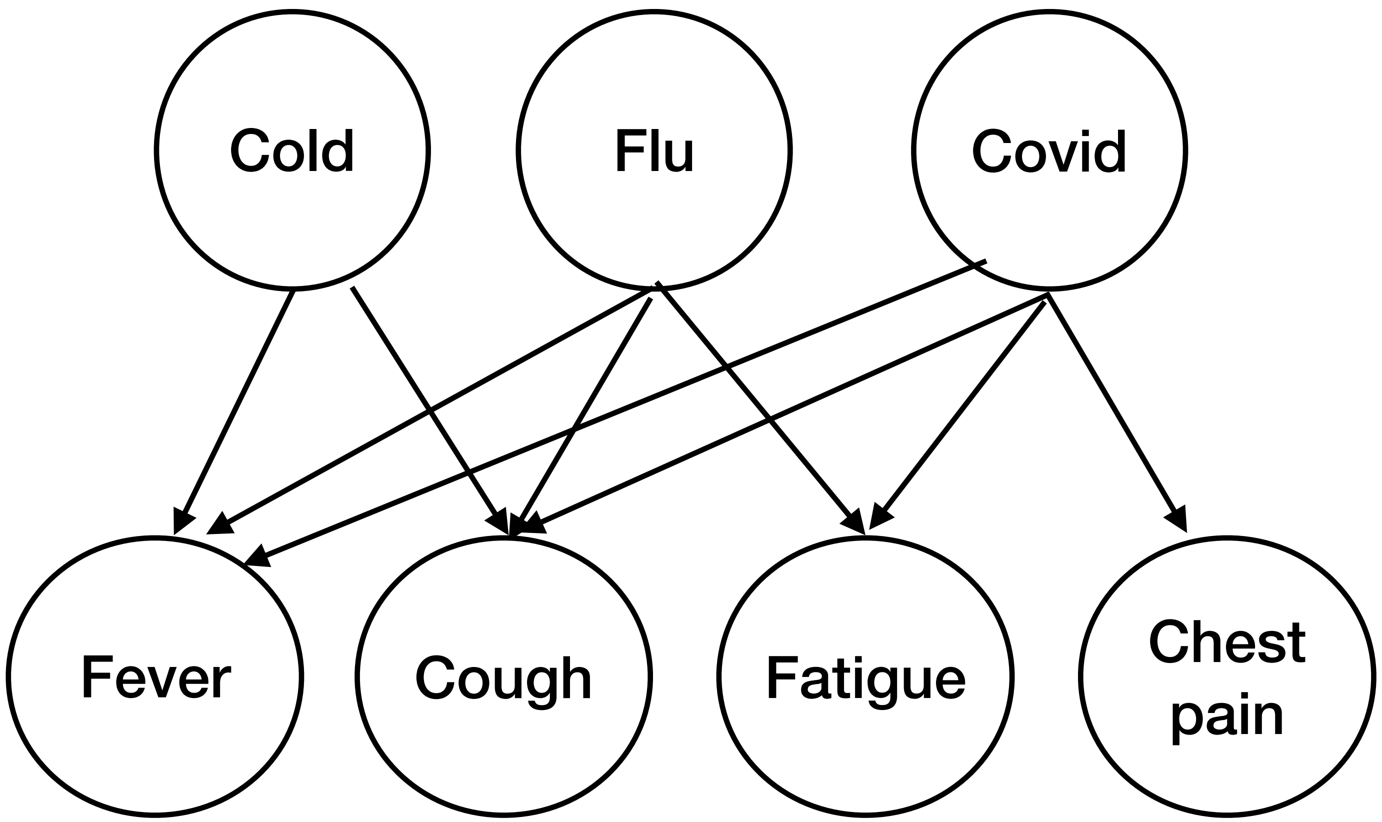
This. is a book

Salient object detection in Computer Vision



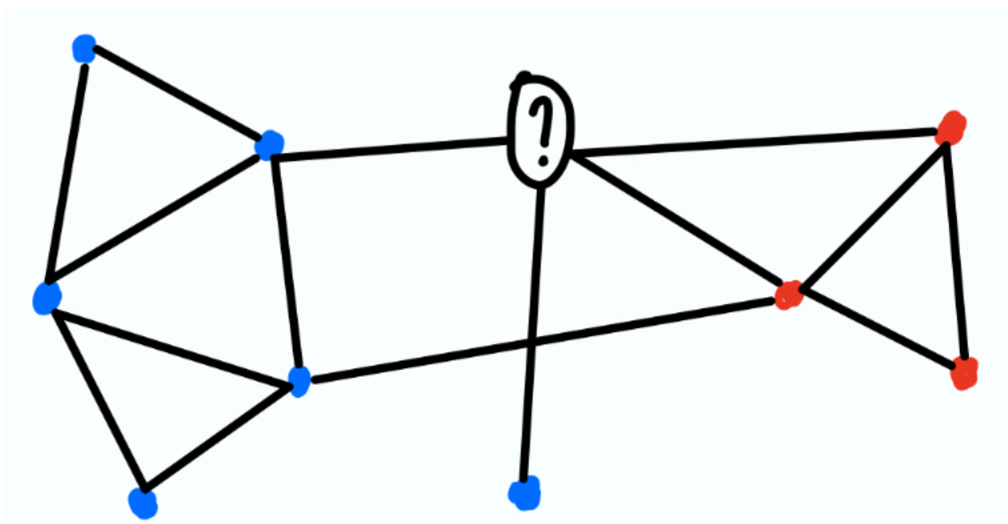
Examples

Medical Diagnosis



Network analysis

- Blue nodes: people of one party
- Red nodes: people of another party
- What's the political polarity of the person in question?



PGM in a nutshell

In PGM, the machine learning system **models** the **probability** of data variables, which are oftentimes related in **graphical** structures.

- Capture (important) dependencies
- Establish independencies
  - Variables are independent by physical laws
  - Ignore unimportant dependencies

- Which is more important?
  - In deep learning models, most variables are connected (dependencies captured). Thus, DL achieves remarkable performance compared with old-day shallow models that emphasize on independencies
  - Nevertheless, certain dependencies in a standard DL model may not be adequately captured, so PGM is still important in the DL era.

Key Problems in PGM

Representation

- What does it mean by a (directed or undirected) graph?
- What is the probability defined by a graph?

Inference

- What is $p(x_1, \cdots, x_n)$ for given values?
- What is the most likelihood

  argmax p(variables in question | evidence)

Learning

- Model parameters
  - Fully observed VS partially observed
- Graph structures