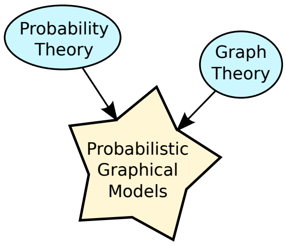
HAVSS Summer School

Probabilistic Graphical Models Introduction

Rémi Emonet – 2012-10-01

Graphical Models: What? why?

- Graphical representations of probability distributions
  - Probability theory + graphs theory
  - Visualization of the structure of probability distributions
  - New insights into existing models (e.g. conditional independence)
  - Computation (learning and inference)
    using graph-based algorithms

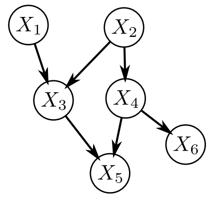


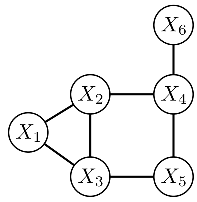
Why Probabilistic Graphical Models

- A way to consolidate advances
  - Good communication tool
  - Clear representations
  - Reuse learning approaches and algorithms
- Probabilities are a sound way of modeling uncertainty
- Many applications
  - Lot of models in the wild
  - Lot of models in this summer school
  - Various application domains

Joint Distributions as Graphs

- Joint distribution: $p(x_1, \ldots, x_k) = ?$
- Nodes
  - random variables (RV)
  - continuous or discrete
- Edges
  - relations between RVs
  - directed or undirected
- Resulting graphs
  - Directed Acyclic Graphs (Bayesian Network)
  - Undirected with cycles (Markov Random Fields)



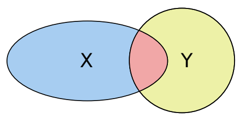


Lecture Content (unordered)

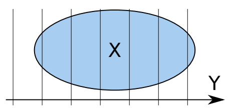
- Graphical models formalisms
  - Directed Graphs: Bayesian networks
  - Undirected Graphs: Markov Random Fields
  - Factor Graphs
- Tasks around graphical models
- Example: Gaussian Mixture Models (more in other lectures)
- Expectation/Maximization algorithm overview

Probabilities, Measure Theory

- Product rule
  - $p(X, Y) = p(X|Y)\ p(Y) = p(Y|X)\ p(X)$



- Marginalization, Sum rule
  - $p(X) = \sum_{y} p(X, Y)$
- Bayes rule
  - $p(Y|X) = \dfrac{p(X|Y)\ p(Y)}{p(X)}$
  - $p(X) = \sum_{Y} p(X|Y)\ p(Y)$

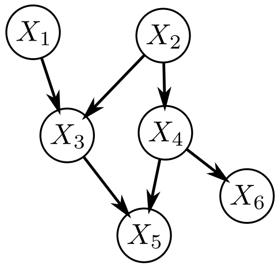


HAVSS Summer School

Bayesian Networks

Bayesian Networks: DAGs

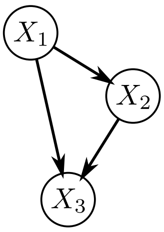
- Bayesian Network:
  Directed Acyclic Graphs
  - oriented edges
  - no loops (directed cycles)
  - concepts: "parents" and "children"
    - $x3$ is a child of $x1$ and $x2$
    - $x3$ is a parent of $x5$
    - $x1$ and $x2$ have no parent



- Spoiler alert: represents a decomposition of $p(x_{1..6})$

Probability Distribution Decomposition

- Consider a joint distribution $p(x_{1..3})$
- Using product rule ($p(a, b) = p(a|b)\ p(b)$)
  - $p(x_{1..3}) = p(x_2, x_3|x_1)\ p(x_1)$
  - $p(x_{1..3}) = p(x_3|x_1, x_2)\ p(x_2|x_1)\ p(x_1)$



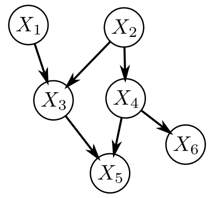
- Modeling decision
  - different decompositions
    ⇒ different representations
    ⇒ different graphs
  - documented decisions
    - as the mathematical decomposition
    - as the Bayesian Network graph

Equations from a Bayesian Network

- Joint distribution for $x_1, \dots, x_K$

$$p(x_1, \dots, x_K) = \prod_{k=1}^{K} p(x_k|\mathrm{par}_k)$$

where par_k is the set of parents of x_k



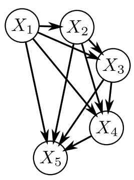
- In the example
  $p(x_{1..6}) = p(x_1)\ p(x_2)\ p(x_3|x_1, x_2)\ p(x_4|x_2)\ p(x_5|x_3, x_4)\ p(x_6|x_4)$
- Factorized representation
  product of "local" conditional distributions

Bayesian Networks: Absence Matters

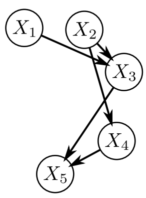
- Any joint distribution can be factorized

$$p(x_1, \dots, x_K) = p(x_K|x_1, \dots, x_{K-1})\ \dots\ p(x_2|x_1)\ p(x_1)$$

⇒ fully connected graph (ignoring edge direction)

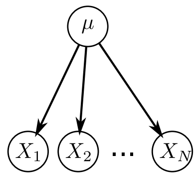


- Absence of links is key
  - Encoding of conditional independence
  - Many problems modeled with sparse links
  - Simplified dependencies
  - Fewer links ⇒ easier computations

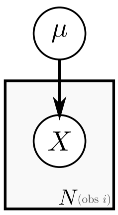


Bayesian Networks: Plate Notation

- Repetition of N nodes with exact same links
- Example: $p(x_{1..N}, \mu) = p(\mu) \prod_{i=1}^{N} p(x_i|\mu)$

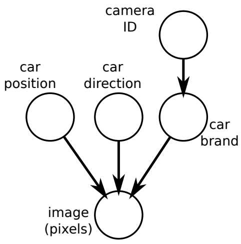


- Plate notation
  - Number of repetitions (N)
  - Optional explicit plate index (i)
  - Plates can be nested



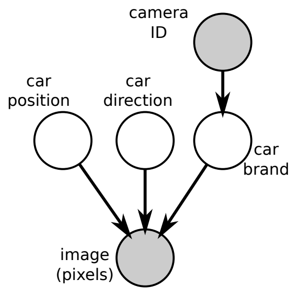
Bayesian Networks as Generative Models

- Generative process
  - A Bayesian network describes the process by which the observations are (supposed to be) generated
- Example
  - Camera ID: an identifier for the camera
  - Car brand: a brand in the list of existing car brands. The probability of a brand depends on the actual camera
  - Image: the colors of all pixels in the image, supposing there is a car in the image. The image depends on the brand, position and parking direction of the car



Bayesian Networks: Types of Variables

- Variables can be
  - Visible, observed (grayed-out)
  - Hidden, latent (empty)
- Visible variables: evidence, knowledge
  - Observed measurements
  - Known context
- Hidden variables
  - Increase richness of models
  - Often with clear (physical) interpretation
- "Visibility" depends on context

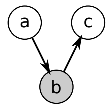


Conditional Independence: Definition Reminder

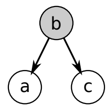
- Considering 3 random variables a, b, c
- a and c are conditionally independent given b
  - $a \perp c \mid b$
  - iif $p(a, c|b) = p(a|b)p(c|b)$
  - iif $p(c|a, b) = p(c|b)$
  - iif $p(a|c, b) = p(a|b)$
- Pervasively used to simplify probabilistic expressions
- Easily derived from a Bayesian network representation

Conditional Independence in Bayesian Networks

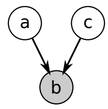
- b Head-to-Tail: $a \perp c \mid b$



- b Tail-to-Tail: $a \perp c \mid b$



- b Head-to-Head: $a \perp c \mid b$? No!



- D-separation: rule to assess conditional independence in more complex cases

HAVSS Summer School

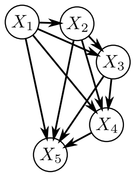
Graphical Models: Modeling and Use

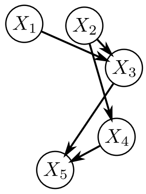
Modeling: Specifying the Model

- Encoding design/modeling decision
- Structure
  - Involved variables
  - Dependencies (conditional independence)
- Parameters
  - Form of the dependencies (e.g., "Categorical", "Normal")
  - Parametrization (e.g., "Gaussian mean, fixed variance")

Modeling: Number of Parameters

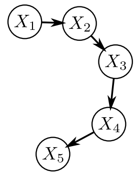
- Supposing discrete variables with K values
- Fully connected case
  - $p(x_1)$: $K - 1$ parameters
  - $p(x_2|x_1)$: $K\ (K-1)$ parameters
  - $p(x_3|x_1, x_2)$: $K^2\ (K-1)$ parameters
  - …
- Less links, less parameters
  - $p(x_1)$: $K - 1$ parameters
  - $p(x_2)$: $K - 1$ parameters
  - $p(x_3|x_1, x_2)$: $K^2\ (K-1)$ parameters
  - …

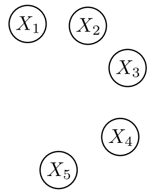




Modeling: Number of Parameters

- Simple chain
  - $p(x_1)$: $K - 1$ parameters
  - $p(x_2|x_1)$: $K\ (K-1)$ parameters
  - $p(x_3|x_2)$: $K\ (K-1)$ parameters
- Indep
  - $p(x_1)$: $K - 1$ parameters
  - $p(x_2)$: $K - 1$ parameters
- Removing links
  - Less parameters
  - More restricted/limited models



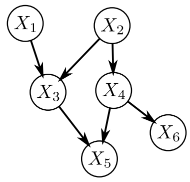


Graphical Models Use

- Generating samples form joint distribution
- Learning
  - finding the “best” parameters
  - given some observation
- Inference
  - finding most probable hidden variable values
  - given some parameters and observations
- Model selection: “best” among multiple models (diff. structures)?
- Recognition
  - What learned model explains best some observations? (competing candidate models)

Generative Models: Ancestral Sampling

- Goal: draw a sample $\widehat{x_1}, ..., \widehat{x_K}$ from $p(x_{1..K})$
- Step 1: define an ancestral ordering such that each node comes after its parents
  - e.g.: $x_1, x_2, x_3, x_4, x_5, x_6$
  - e.g.: $x_2, x_4, x_1, x_3, x_6, x_5$
- Step 2: draw successively following the order
  - Parent values are always available
  - e.g.
    Sample first $\widehat{x_2}$ from $p(x_2)$
    Then $\widehat{x_4}$ from $p(x_4|x_2 = \widehat{x_2})$
- To sample from a marginal (e.g., $p(x_1, x_4)$) just keep $\widehat{x_1}$ and $\widehat{x_4}$



HAVSS Summer School

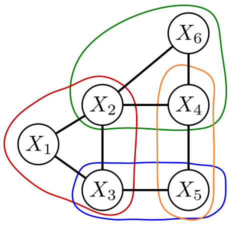
Graphical Models Zoo

Examples of Bayesian Networks

- Nature of Variables: discrete, continuous, mixed, static, dynamic
- Examples
  - Gaussian Mixture Models (GMM)
  - Hidden Markov Models (HMM)
  - Kalman Filters (KM)
  - Particle Filters (PF)
  - Probabilistic Principal Component Analysis (PPCA)
  - Factor Analysis (FA)
  - Transformed Component Analysis (TCA)
  - Probabilistic Topic Models (PTM)

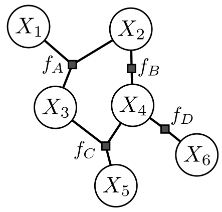
Markov Random Fields: Undirected Graphs

- Joint distribution in MRF
  - Product over non-negative functions over the (maximal) cliques of the graph
  - $p(X) = \frac{1}{Z} \prod_C \psi_C(X_C)$
  - $\psi_C(X_C)$: clique potentials
  - Z is a normalization constant
- Example
  - $p(x_{1..6}) = \frac{1}{Z} \psi_A(x_1, x_2, x_3) \ \psi_B(x_2, x_4, x_6) \ \psi_C(x_3, x_5) \ \psi_D(x_4, x_5)$

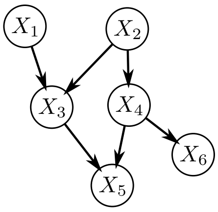


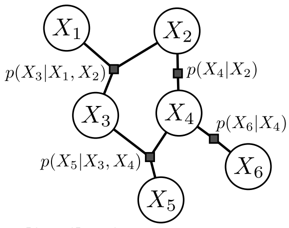
Factor Graphs

- Undirected bipartite graph
  - Random variables
  - Factors
- Function value (joint distribution) in a factor graph
  - Product of factors
  - $f(x_{1..N}) = \prod_{i=1}^{N} f_i(S_i)$
  - S_i: neighborhood of node f_i in the graph



Bayesian Networks as Factor Graphs





- Factor graph for the joint probability distribution
- Any Bayesian network can be expressed as a factor graph
  - More generic
  - More explicit (shows distributions)
  - Loss of direction information (visually)

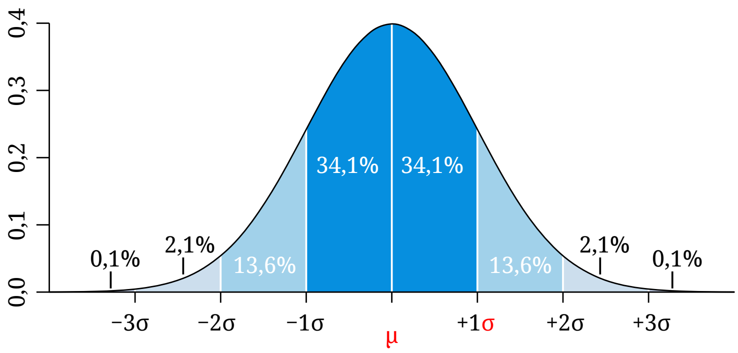
Many Graphical Representations

- Bayesian networks
  - mixture models
  - hierarchical structures
- MRF
  - image denoising
- Factor Graphs
  - generic
  - explicit
  - verbose
  - message passing algorithm

HAVSS Summer School

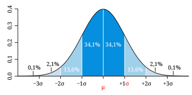
Use Case With GMM

Gaussian/Normal Distribution: basics



Gaussian/Normal Distribution: basics

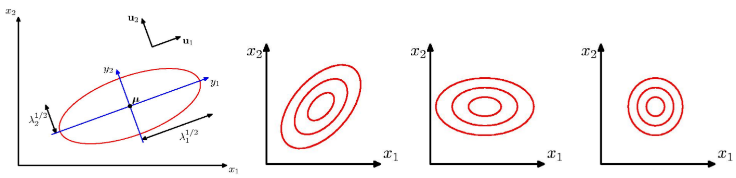
- Normal Distrbution or Gaussian Distribution
  - $N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
  - Is-a probability density
    - $N(x|\mu,\sigma^2) > 0$
    - $\int_{-\infty}^{+\infty} N(x|\mu,\sigma^2)dx = 1$



- Parameters
  - μ: mean, $E[x] = \mu$
  - σ^2: variance, $E\left[x^2 - E[X]\right] = \sigma^2$

Multivariate Normal Distribution

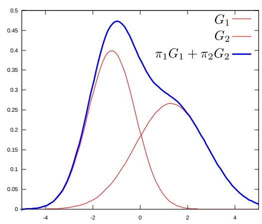
- D-dimensional space: $x = \{x_1, \ldots, x_D\}$
- Probability distribution
  - $N(x|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^D|\Sigma|}} \; exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right)$
  - Σ: covariance matrix



Elliptical equiprobability surfaces. Major axes = eigenvectors of Σ | full cov. matrix ($\frac{D(D+1)}{2}$ parameters) | diagonal cov. matrix (D parameters) | identity cov. matrix (1 parameter)

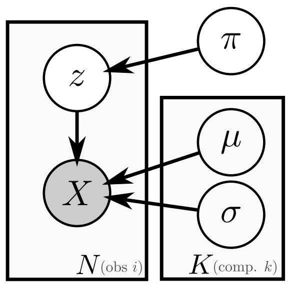
Gaussian Mixture Models (GMM)

- Weighted sum of Gaussians
- Parameters with K Gaussians
  - $\pi_{1..K}$: weights such that $\sum_{k=1}^{K} \pi_k = 1$
  - $\mu_{1..K}$: means of the Gaussians
  - $\sigma^2_{1..K}$: variances of the Gaussians

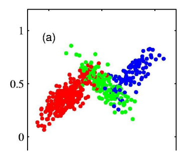


GMM Bayesian Network

- Generative process
  - $\forall i = 1..N$
    - draw z_i from $Categorical(\pi)$
    - draw X_i for $Normal\left(\mu_{z_i}, \sigma^2_{z_i}\right)$



- Example samples
  - $K = 3$ components
  - N draws
  - show complete data (color encodes the z_i)

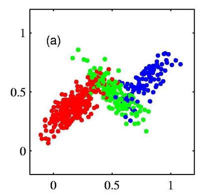


Learning with Expectation Maximization (EM)

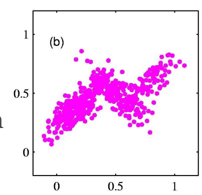
- Goal: given some observations, find the "best" parameters
  - best = maximum likelihood estimator (MLE)
  - Parameters $\theta = \{\pi_k, \mu_k, \sigma_k\}_{k=1..K}$
  - Find $\theta_{ML} = \text{argmax}_\theta \; L(\theta|D) = \text{argmax}_\theta \; p(x|\theta)$
- Log-likelihood function:
  $\ln L(\theta|x) = \ln p(x|\theta) = \ln \sum_z p(x,z|\theta)$
- Problem
  - Incomplete data
  - z unknown $\Rightarrow \sum_z$ in the likelihood $\Rightarrow$ difficult to optimize

Complete/Incomplete Data: Illustration

- Complete data
  - Supposing we know z
  - z: known labels (each point: red, green or blue)
  - Estimating θ is easy



- Incomplete data (actually observed)
  - Case of learning the model
  - Difficult to estimate θ
  - Use of Expectation Maximization algorithm



EM Intuition

- Complete data log-likelihood
  - $\ln\ L_C(\theta|x,z) = \ln\ p(x,z|\theta)$
  - easier to maximize to *get* θ_{ML}
  - but need to know z
- If we knew θ_{ML} (but not z)
  - we could estimate the posterior of z, i.e., $p(z|x,\theta)$
  - i.e., how probable are the different values of z
  - i.e., for each point i and component k, $p(z_i = k\,|\,\ldots)$: the "responsibility" of component k for x_i

EM: The E and M steps

- Iterative algorithm
  - Random initialization: $\theta^0 = rand()$
  - Local optimum ⇒ needs multiple initializations
- E step
  - use the current estimate θ^{old}
  - to find the posterior/responsibilities $p\big(z|x,\theta^{old}\big)$
- M step
  - use the computed $p\big(z|x,\theta^{old}\big)$
  - to find a new best estimate $\theta^{new} = \operatorname{argmax}_\theta \sum_z p\big(z|x,\theta^{old}\big) \ln p(x,z|\theta)$

EM: Remarks and Limitations

- EM divides a difficult problem (learning) into two steps that *might* be simpler to implement
- E-step or M-step might be intractable
  - *intractable M-step (generalized EM)*: instead of maximizing wrt θ, just modify θ to increase the value (non-linear optimization method)
  - *intractable E-step*: perform a partial (rather than full) optimization of $L(q,\theta)$ (wrt $q(Z)$)
- EM requires some initialization values
- EM can get trapped into non-global maxima

HAVSS Summer School

Wrap up

Summary

- Graphical models
  - Different representations
  - Communication tool
  - Inference support
- Tasks around graphical models
- Gaussian Mixture Models
  - introduction
  - EM algorithm overview
- Inference methods

HAVSS Summer School

Thank you for your attention

Probabilistic Graphical Models Introduction