Bayesian multimodeling: graphical models

MIPT

2023

Graphical models

Conditional independence

Events X, Y are conditionally independint w.r.t. Z : $\quad X \perp Y|Z$, if

$$P(X|Y,Z) = P(X|Z).$$

Conditional dependence

Events X, Y are conditionally dependent w.r.t. $\mathfrak{S}$: $X, Y \in \mathfrak{S}$, if

$$X \not\perp Y|\mathfrak{S} \setminus \{X, Y\}.$$

Graphical models

A probabiliy model is graphical, if it can be represented as a graph, where the edges correspond to conditionally dependent events.

Non-graphical models

- MLP, decision trees, etc.
- Models with complex behaviour.

Types of graphical models

- Directed models (aka Bayesian networks)
  - Easy to desing
- Undirected (Markov models)
- Factor-graphs
  - Easy to infer and optimize

Plate notation

Plate notation is an alternative visuzliation for graphical models.

Elements:

- White circles (random variables);
- Grey circels(observed variables);
- Small circles (deterministic values);
- Plates (batching).

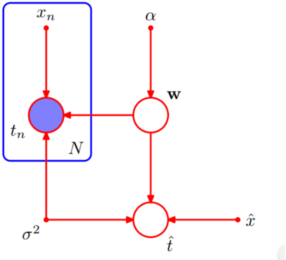


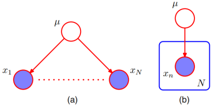
Plate notation for linear regression (Bishop)

Bayesian networks

- Models are set using directed acyclic graphs
- Joint distribution for the graph with K vertices:

$$p(v_1, \ldots, v_k) = \prod_{i=1}^{K} p(v_i | \text{parent}(v_i))$$

- Example: linear regresssion



DAG and Plate notation (Bishop)

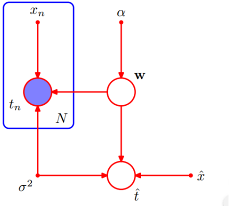


Plate notation for regression model (Bishop)

Causality graph elements

$$X \to Y \to Z \text{ — } \textbf{chain}$$

Example:

- X — school budget
- Y — average student score
- Z — unviersity acceptance ratio

Properties:

1. X and Y, Y and Z are dependent:
   $\exists x, y : \mathbf{P}(Y = y \,|X = x\,) \neq p\,(Y = y)$
   $\exists y, z : \mathbf{P}(Z = z \,|\, Y = y\,) \neq p\,(Z = z)$
2. Z and X: are (probably) dependent
3. $Z \perp X|Y$: are conditionally independent: $\forall x, y, z$

$$\mathbf{P}(Z = z \,|X = x, Y = y\,) = \mathbf{P}(Z = z \,|\, Y = y\,)$$

(if Y is fixed, then X and Z are independent)

Causality graph elements

$$X \leftarrow Y \rightarrow Z \text{ — } \textbf{fork}$$

Example:

- X — ice cream sells
- Y — average temperature
- Z — crime ratio

Properties:

1. X and Y, Y and Z are dependent
2. X and Z are (probably) dependent
3. $X \perp Z | Y$ are conditionally independent

Causality graph elements

$$Y \to X \leftarrow Z - \textbf{collider}$$

Example (illnes):

- X — bad symptoms
- Y — age
- Z — chronical diseases

Properties:

1. Y and X, Z and X are dependent
2. Y and Z are independent
3. $Y \not\perp Z | X$ are conditionally dependent

d-separation

The path P **is blocked** by Z, if:

1. P contains $A \rightarrow B \rightarrow C$, $A \leftarrow B \rightarrow C$, $B \in Z$
2. P contains $A \rightarrow B \leftarrow C$, $B \notin Z$ and all children of $B \notin Z$

If Z blocks all the paths from X to Y, then X and Y are **d-separated**:

$$X \perp Y | Z.$$

d-separation

The path P is blocked by Z, if:

1. P contains $A \rightarrow B \rightarrow C$, $A \leftarrow B \rightarrow C$, $B \in Z$
2. P contains $A \rightarrow B \leftarrow C$, $B \notin Z$ and all children of $B \notin Z$

If Z blocks all the paths from X to Y, then X and Y are **d-separated**.

Example:

$$
\begin{tikzcd}
Z_1 \arrow[dd] \arrow[dr] & & Z_2 \arrow[dl] \arrow[dd] \\
& Z_3 \arrow[dl] \arrow[dr] & \\
X \arrow[r] & W \arrow[r] & Y
\end{tikzcd}
$$

Pair	d-separation set
(Z_1, W)	X

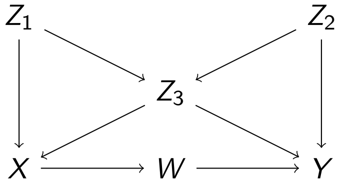
d-separation

The path P is blocked by Z, if:

1. P contains $A \to B \to C$, $A \leftarrow B \to C$, $B \in Z$
2. P contains $A \to B \leftarrow C$, $B \notin Z$ and all children of $B \notin Z$

If Z blocks all the paths from X to Y, then X and Y are **d-separated**.

Example:



Pair	d-separation set
(Z_1, W)	X
(Z_1, Y)	$\{Z_3, X, Z_2\}, \{Z_3, W, Z_2\}$

Model selection for Bayesian networks

- Generally, NP-hard problem
- Reduces to optimization problem with predefined search space or sampling problem
- Independence determination:
  - ML and MAP
  - Evidence
  - Information criteria

Simple algorithm for predefined vertices

1. $\forall A, B \in V$ search a set $S_{AB}: A \perp B | S_{AB},\ A, B \notin S_{AB}$. If S_{AB} does not exist, make an edge AB.
2. $\forall A, B$, not connected by edge and having a common neighbor C, check: $C \in S_{AB}$? If not, replace $A - C, C - B$ by $A \to C, C \leftarrow B$
3. Recursively:
   - if there is an oriented path from A to B $A \to \cdots \to B$, then replace $A - B$ by $A \to B$;
   - if A and B are not connected, $A \to C$, $C - B$, then replace $C - B$ by $C \to B$.

Markov random fields

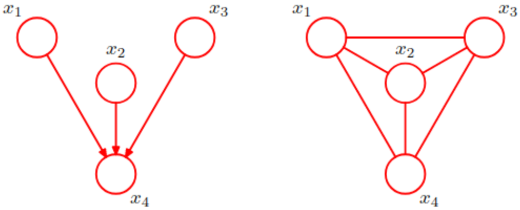
Models are represented as undirected graphs.

Difference from Bayesian networks:

- No direction $\rightarrow$ cannot infer causality.
- The likelihood is factorized as follows:

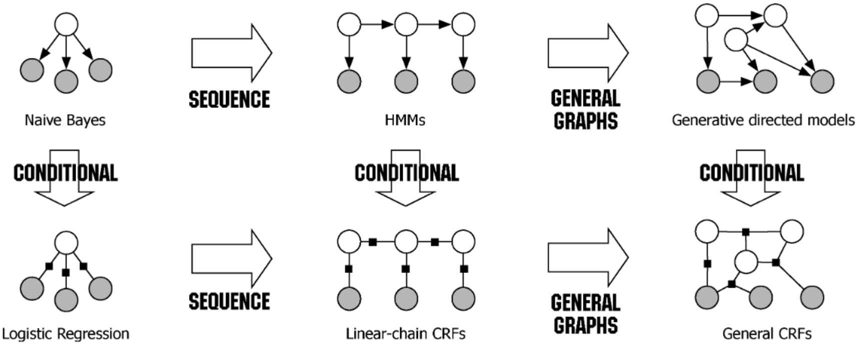
$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \psi(\mathbf{X}_C),$$

  where $\mathbf{X}_C$ is a maximal clicque, $\psi \geq 0$ is a potential function.
- Conditional indepdence: if all the paths from A to B go throught C, then $A \perp B | C$.



(Bishop)

Example: CRF and HMM



(Sutton, McCallum)

Inference in chains

(Bishop)

Naive likelihood calculation for x_n:

$$p(x_n) = \sum_{x_1} \sum_{x_2} \ldots, \sum_{x_{n-1}} \sum_{x_{n+1}} \cdots \sum_{x_N} p(\mathbf{x}),$$

For N discrete variables with K values the complexity is $O(K^N)$

Inference in chains: regroupping

$$p(x_n) = \sum_{x_1} \sum_{x_2} \dots, \sum_{x_{n-1}} \sum_{x_{n+1}} \cdots \sum_{x_N} p(\mathbf{x}),$$

$$p(\mathbf{x}) = \psi(x_1, x_2)\psi(x_2, x_3) \dots \psi(x_{N-1}, x_N).$$

Regroup the sum:

$$p(x_n) = \sum_{x_{n-1}} \psi(x_{n-1}, x_n) \dots \big(\sum_{x_1} \psi(x_1, x_2)\big) \times$$

$$\times \big(\sum_{x_{n+1}} \psi(x_n, x_{n+1}) \dots \big(\sum_{x_N} \psi(x_{N-1}, x_N)\big)\big).$$

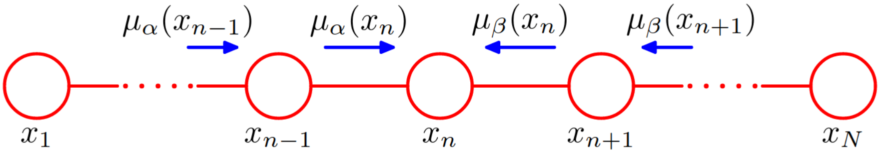
Now complexity is $O(NK^2)$.

Message passing

$$p(x_n) = \underbrace{\sum_{x_{n-1}} \psi(x_{n-1}, x_n) \ldots \Big(\sum_{x_1} \psi(x_1, x_2)\Big)}_{\mu_a(x_n)} \times$$

$$\times \underbrace{\Big(\sum_{x_{n+1}} \psi(x_n, x_{n+1}) \ldots \Big(\sum_{x_N} \psi(x_{N-1}, x_N)\Big)\Big)}_{\mu_b(x_n)} .$$

Interpretation: $\mu_a(x_n)$ is a message transferred from x_{n-1} to x_n, $\mu_b(x_n)$ is a backward message from x_{n+1}.



Inference in chains: details

The inference is iterative:

- calculate $\sum_{x_1} \psi(x_1, x_2) = \boldsymbol{\mu}_a(\mathbf{x}_2)$, that stores $\mu_a(x_2)$ for each value of x_2;
- calculate $\sum_{x_2} \psi(x_2, x_3)\big(\sum_{x_1} \psi(x_1, x_2)\big) = \sum_{x_2} \psi(x_2, x_3)\mu_a(x_2) = \boldsymbol{\mu}_a(\mathbf{x}_3)$;
- ...
- calculate$\sum_{x_{n+1}} \psi(x_n, x_{n+1})\mu_b(x_{n+1}) = \boldsymbol{\mu}_b(\mathbf{x}_n)$.
- for directed variables, where

$$\psi(x_1, x_2) = p(x_1)p(x_2|x_1), \quad \psi(x_i, x_{i+1}) = p(x_{i+1}|x_i),$$

  μ_b should not be calculated:

$$\mu_b(x_n) = \sum_{x_{n+1}} \psi(x_n, x_{n+1}) \dots \big(\sum_{x_N} \psi(x_{N-1}, x_N)\big) =$$

$$= \sum_{x_{n+1}} p(x_{n+1}|x_n) \dots \big(\sum_{x_N} p(x_N|x_{N-1})\big) = 1.$$

Factor graph

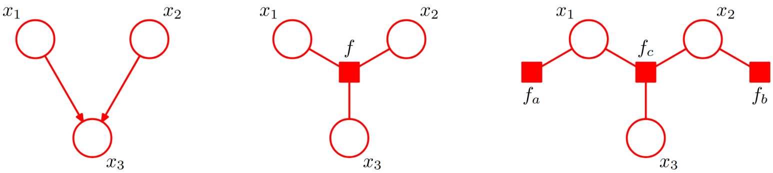
Definition

Factor-graph is a bipartite graph with two types of vertivces: variables and factors.
The likelihood is a production of factors:

$$p(\mathbf{x}) = \prod_i f_i.$$

Example: model $p(x_1)p(x_2)p(x_3|x_2,x_1)$ has two variants of factorization:

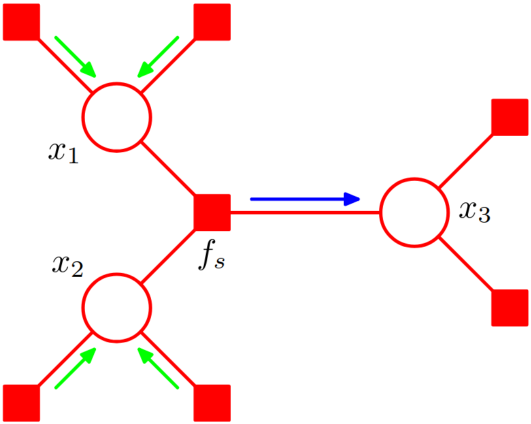
$$f = p(x_1)p(x_2)p(x_3|x_2,x_1), \quad f_a = p(x_1), f_b = p(x_2), f_3 = p(x_1)p(x_2)p(x_3|x_2,x_1).$$



(Bishop)

Inference in factor-graphs: example

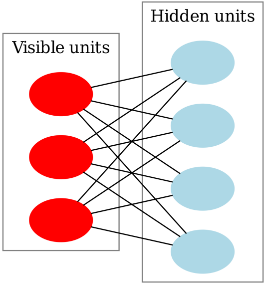
Sum-product: likelihood is a composition of messages from factors to variables.



Model examples: RBM

$$p(\mathbf{x}, \mathbf{h}) = \frac{1}{Z}\exp(-E(\mathbf{x}, \mathbf{h})),$$

$$E = -\mathbf{w}_1^\top \mathbf{x} - \mathbf{w}_2^\top \mathbf{h} - \mathbf{x}^\top \mathbf{W}_3 \mathbf{h}.$$



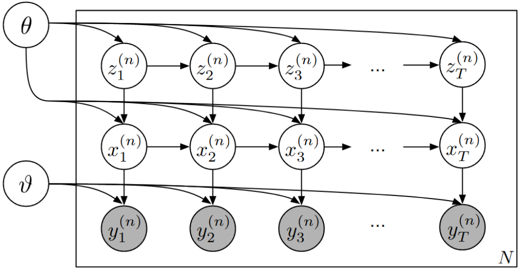
Model examples: Structured VAEs

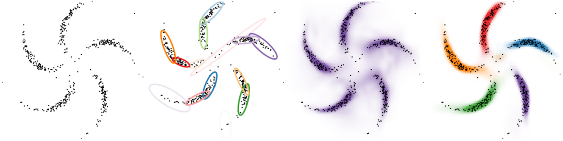
Based on SLDS:

$$z_{t+1} | z_t \sim \pi^{t+1},$$

$$\mathbf{y}_t \sim \mathcal{N}(\mathbf{MLP}^{z_t}(\mathbf{x}_t)).$$

Optimization: optimize ELBO.
Inference: message-passing.





(a) Data (b) GMM (c) Density net (VAE) (d) GMM SVAE

References

- Bishop C. M. Pattern recognition //Machine learning. – 2006. – T. 128. – №. 9.
- Edwards D. Introduction to graphical modelling. – Springer Science & Business Media, 2012.
- Pearl J., Glymour M., Jewell N. P. Causal inference in statistics: A primer. – John Wiley & Sons, 2016.
- Hinton G. E., Salakhutdinov R. R. Reducing the dimensionality of data with neural networks //science. – 2006. – T. 313. – №. 5786. – C. 504-507.
- https://en.wikipedia.org/wiki/Restricted_Boltzmann_machine
- Johnson M. J. et al. Structured VAEs: Composing probabilistic graphical models and variational autoencoders //arXiv preprint arXiv:1603.06277. – 2016. – T. 2. – C. 2016.
- Johnson M. J. et al. Composing graphical models with neural networks for structured representations and fast inference //Advances in neural information processing systems. – 2016. – T. 29. – C. 2946-2954.
- Højsgaard, S., Edwards, D., Lauritzen, S. (2012). Graphical models with R. Springer Science & Business Media.
- Sutton, C, McCallum, A 2012, 'An Introduction to Conditional Random Fields', Foundations and Trends in Machine Learning, vol. 4, no. 4, pp. 267-373

$$
\begin{aligned}
\nabla_\theta \mathbb{E}_{p_\theta(z)}[f_\theta(z)] &= \nabla_\theta \Big[\int_z p_\theta(z) f_\theta(z) dz \Big] \\
&= \int_z \nabla_\theta \Big[p_\theta(z) f_\theta(z) \Big] dz \\
&= \int_z f_\theta(z) \nabla_\theta p_\theta(z) dz + \int_z p_\theta(z) \nabla_\theta f_\theta(z) dz \\
&= \underbrace{\int_z f_\theta(z) \nabla_\theta p_\theta(z) dz}_{\text{What about this?}} + \mathbb{E}_{p_\theta(z)} \Big[\nabla_\theta f_\theta(z) \Big]
\end{aligned}
$$

https://gregorygundersen.com/blog/2018/04/29/reparameterization/