Machine Learning

Lecture 8 – Graphical models

“A graphical model is a way to represent a joint distribution by making use of conditional independence assumptions.“

UPPSALA UNIVERSITET

Thomas Schön
Division of Systems and Control
Department of Information Technology
Uppsala University.

Email: thomas.schon@it.uu.se,
www: `user.it.uu.se/~thosc112`

Contents – lecture 8

(Chapter 10.7 and Chapter 8)

Summary – lecture 7

Variational inference is a type of approximate Bayesian inference where factorial approximations like

$$p(Z|X) \approx q(Z) = \prod_i q_i(Z_i)$$

are used for the posterior.

The Kullback-Leibler (KL) distance is used to measure the distance and hence to find an optimization problem.

Variational Bayes (VB) is a form of variational inference where $\text{KL}(q||p)$ is used for the optimization. We fix all but one of the factors and optimize as follows,

$$\hat{q}_j(Z_j) = \arg\min_{q_j} \left(q_j(Z_j) \prod_{i \neq j} \hat{q}_i(Z_i) \middle\| p(X, Z) \right)$$

Probabilistic graphical models – motivation

“Graphical models bring together graph theory and probability theory in a powerful formalism for multivariate statistical modeling.” [1]

We can of course always handle probabilistic models using pure algebraic manipulation. Some reasons for using probabilistic graphical models,

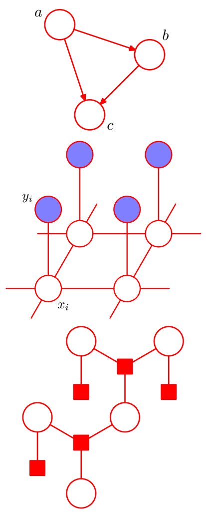
1. A simple way to **visualize** the structure of a probabilistic model.
2. Knowledge about **model properties** directly from the graph.
3. A different way of **performing and structuring** calculations.

[1] Wainwright, M. J. and Jordan, M. I. **Graphical models, exponential families, and variational inference**, *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

3 types of graphical models

We are going to consider three types of graphs,

1. **Directed graphs** (a.k.a. Bayesian networks) Represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG).
2. **Undirected graphs** (a.k.a. Markov random fields) represents a set of random variables having a Markov property by an undirected graph.
3. **Factor graphs** are a more convenient form that can be obtained from the above two for the purposes of inference and learning.



Observing conditional independence (I/III)

Example: Polynomial regression.

- Let $t_{1:N}$ be the values of a function at the points $x_{1:N}$.
- We would like to find the Kth degree polynomial approximating this function whose coefficients are shown as $\mathbf{w} \in \mathbb{R}^{K+1}$.
- $\mathbf{w} \sim \mathcal{N}(0, \Sigma)$
- Then the model can be written as

$$t_n = \phi(x_n)\mathbf{w} + v_n$$

  where $\phi(x) = \left[1, x, x^2, \dots, x^K\right]$.
- $\{v_n\}_{n=1}^N$ is i.i.d. and $v_n \sim \mathcal{N}(0, R)$.

Observing conditional independence (II/III)

Example: Polynomial regression

$$t_n = \phi(x_n)\mathbf{w} + v_n$$

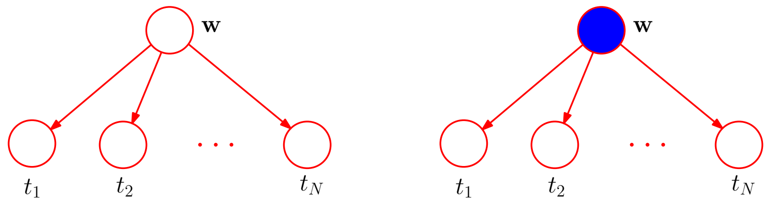
- The joint density for the problem can be written as

$$p(t_{1:N}, \mathbf{w}) = p(t_{1:N} | \mathbf{w}) p(\mathbf{w}) = p(\mathbf{w}) \prod_{i=1}^{N} p(t_i | \mathbf{w})$$

- What is the reason for the equality $p(t_{1:N} | \mathbf{w}) = \prod_{i=1}^N p(t_i | \mathbf{w})$?

Observing conditional independence (III/III)

Example: Polynomial regression



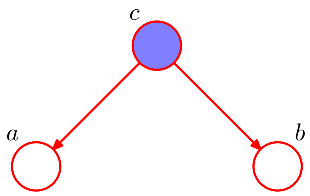
When $\mathbf{w}$ is assumed known it is said to "**block** the path", rendering all the variables $\{t_n\}_{n=1}^N$ conditionally independent.

Important question: Can this be formalized, i.e., can we discern CI properties directly from the graph?

CI from DAGs – Ex 1

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a|c)p(b|c)p(c)}{p(c)} = p(a|c)p(b|c)$$

$$\Longrightarrow \quad a \perp b|c$$

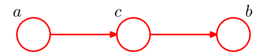


CI rule for tail-to-tail nodes

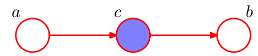
For conditional independence of two nodes, the tail-to-tail nodes between them must be observed, which blocks the path.

CI from DAGs – Ex 2

Head-to-tail nodes:



- Are a and b independent $a \overset{?}{\perp} b$?
- How about when c is given; $a \overset{?}{\perp} b|c$?



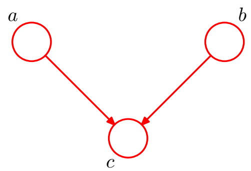
CI rule for head-to-tail nodes

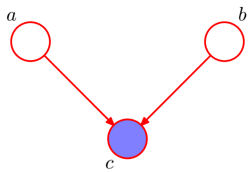
For conditional independence of two nodes, the head-to-tail nodes between them must be observed, which blocks the path.

CI from DAGs – Ex 3

Head-to-head nodes:

- Are a and b independent $a \overset{?}{\perp} b$? Yes, since $\int p(a,b,c)\mathrm{d}c = \int p(a)p(b)p(c \mid a,b)\mathrm{d}c = p(a)p(b)$.
- How about when c is given; $a \overset{?}{\perp} b|c$? No, since $p(a,b \mid c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(b)p(c|a,b)}{p(c)} \neq p(a \mid c)p(b \mid c)$.





CI rule for head-to-head nodes

For conditional independence of two nodes, the head-to-head nodes between them must be unobserved, which blocks the path.

D-separation

D-separation for Directed Acyclic Graphs

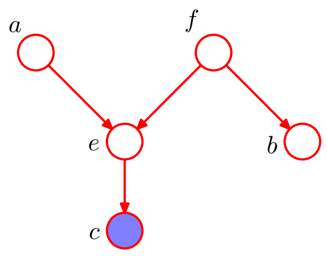
Consider a directed acyclic graph in which A, B and C are arbitrary non-intersecting sets of nodes. We have the property

$$A \perp B|C$$

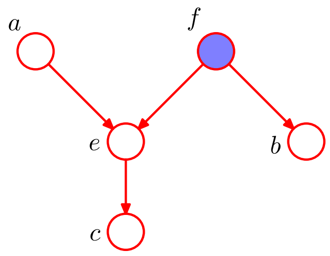
if, on all possible paths from any node in A to any node in B,

- all tail-to-tail and head-to-tail nodes are in C;
- neither head-to-head nodes nor any of their descendants are in C.

D-separation examples



- The path from a to b is not blocked by f, since it is a tail-to-tail node and f not observed.
- Nor is it blocked by e, which is a head-to-head node, with an observed node c as descendant.
- Hence, CI $(a \perp b \mid c)$ does **not follow** from this graph.



- The path from a to b is blocked by f, since it is a tail-to-tail node and f is observed.
- It is also blocked by e, head-to-head node and neither it not its descendants are observed.
- Hence, CI $(a \perp b \mid c)$ **follows** from this graph.

A few concepts to summarize lecture 8

Probabilistic graphical model: Offers a compact way of encoding the conditional dependency structure of a set of random variables.

Bayesian network: A probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG).

Markov random field: A probabilistic graphical model that represents a set of random variables having a Markov property by an undirected graph.

D-separation: Checking for conditional independence is somewhat troublesome for directed graphs requiring a condition called D-separation to be satisfied.