COMS30035, Machine learning: Probabilistic Graphical Models 1

James Cussens
james.cussens@bristol.ac.uk

Department of Computer Science, SCEEM
University of Bristol

October 7, 2020

Agenda

- Factorising joint probability distributions
- Conditional independence
- Bayesian networks (BNs)
- Plate notation for BNs representing machine learning models

The chain rule

- For any joint distribution $P(x_1, \ldots, x_n)$ we have:

$$P(x_1, \ldots, x_n) = P(x_1)P(x_2|x_1)\ldots P(x_n|x_1, \ldots x_{n-1}) \quad (1)$$

- This just follows from the definition of conditional probability.
- Note that we can re-order the the variables at will e.g. $P(x_1, \ldots, x_n) = P(x_2)P(x_1|x_2)\ldots P(x_n|x_1, \ldots x_{n-1})$

Conditional independence

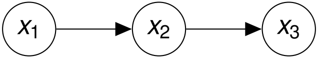
- For any joint distribution over random variables x_1, x_2, x_3 we always have:

$$P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2) \tag{2}$$

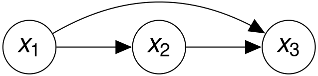
- Now suppose that for some particular probability distribution P we have that: $P(x_3|x_1, x_2) = P(x_3|x_2)$.
- In other words for the distribution P, x_3 is independent of x_1 conditional on x_2.
- Intuition: Once I know the value of x_2 (no matter what that value might be) then knowing x_1 provides no information about x_3.
- Then $P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_2)$
- *Probabilistic graphical models (PGMs)* provide a graphical representation of how a joint distribution factorises when there are conditional independence relations.

Bayesian networks

- The most commonly used PGM is the *Bayesian network*.
- If we have $P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_2)$
- Then this factorisation of the joint distribution is represented by the following directed acyclic graph (DAG):

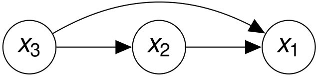


For a distribution with no conditional independence relations a suitable BN representation would be:



$$P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$$

or



$$P(x_1, x_2, x_3) = P(x_3)P(x_2|x_3)P(x_1|x_2, x_3)$$

Bayesian network terminology

- If there is an arrow from A to B in a Bayesian network we say that A is a *parent* of B and B is a *child* of A.
- The set of parents of a node x_k is denoted (by Bishop) like this: pa_k.
- Note that any directed acyclic graph (DAG) determines pa_k for each node x_k in that DAG (and conversely the collection of parent sets determine the DAG).
- A Bayesian network with parent sets pa_k for random variables $x_1, \ldots, x_K$ represents a joint distribution which factorises as follows:

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k | \mathrm{pa}_k) \tag{3}$$

BN structure and parameters

- For a BN to represent a given joint distribution we need to specify:
  1. the DAG (*the structure of the BN*)
  2. the conditional probability distributions $p(x_k|\mathrm{pa}_k)$ (*the parameters of the BN*)
- A given DAG represents a **set** of joint distributions: each distribution in the set corresponds to a choice of values for the conditional distributions $p(x_k|\mathrm{pa}_k)$.
- We will see that it is possible to ‘read off’ conditional independence relations that are true for a distribution represented by a BN, just by using the DAG.

BNs represent machine learning models

- We will use BNs to represent machine learning models.
- Later we will see how to use such a representation to ‘automatically’ do Bayesian machine learning.
- Let’s start with a BN to represent Bayesian polynomial regression [Bis06, §8.1.1].

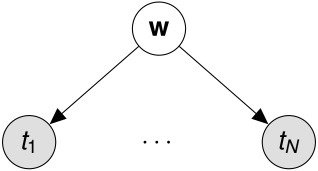
Polynomial regression model

To begin with let's just focus on the joint distribution $p(\mathbf{t}, \mathbf{w})$ where $\mathbf{w}$ is the vector of polynomial coefficients and $\mathbf{t}$ is the observed (output) data.

$p(\mathbf{t}, \mathbf{w})$ can be factorised as follows (since we assume the data is i.i.d.)

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w}) \tag{4}$$

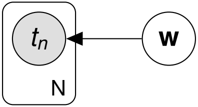
and so has the corresponding BN:



where the dots represent the t_n that have not been explicitly represented in the BN. I have shaded the t_1 and t_n nodes to indicate that the values of these random variables are observed (since they are data).

Plate notation

- Using dots to represent BN nodes we don't wish to explicitly represent is a bit yucky.
- Instead we use *plate notation* to represent BNs with many nodes:



- The plate around t_n represents a set of nodes $t_1, \ldots, t_N$ all of which have **w** as their (single) parent.
- Bishop [Bis06, Fig 8.4] labels the plate with N (the number of nodes 'in' the plate). Other authors label plates with an index (here it would be n). We will stick with Bishop's notation to be consistent with the textbook.

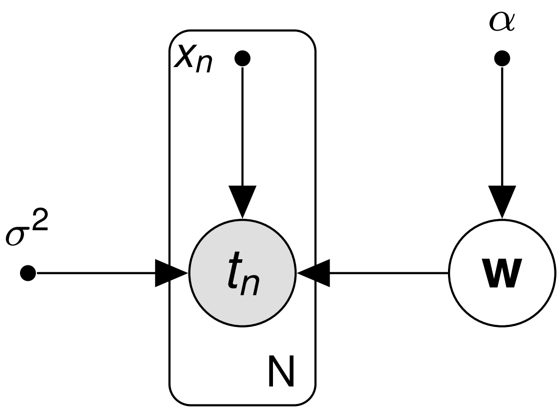
A fuller description

The full Bayesian polynomial regression model contains:

1. The input data $\mathbf{x} = (x_1, \ldots, x_N)^T$
2. The observed ouputs $\mathbf{t} = (t_1, \ldots, t_N)^T$
3. The parameter vector $\mathbf{w}$.
4. A hyperparameter α.
5. The noise variance σ^2.

- We don't care how $\mathbf{x}$ is distributed and we would probably just set α to some value.
- So we would typically consider $\mathbf{x}$, α and also σ^2 as parameters of the model rather than random variables.
- But it is also useful represent these quantities in the BN.
- This leads us to more notation for BNs

A complete BN representation for the polynomial regression model



Christopher M. Bishop.
Pattern Recognition and Machine Learning.
Springer, 2006.