# COMS30035, Machine learning: Probabilistic Graphical Models 

James Cussens<br>james.cussens@bristol.ac.uk<br>School of Computer Science<br>University of Bristol

5th October 2023

## The chain rule

- For any joint distribution $P\left(x_{1}, \ldots, x_{n}\right)$ we have:

$$
\begin{equation*}
P\left(x_{1}, \ldots, x_{n}\right)=P\left(x_{1}\right) P\left(x_{2} \mid x_{1}\right) \ldots P\left(x_{n} \mid x_{1}, \ldots x_{n-1}\right) \tag{1}
\end{equation*}
$$

- This just follows from the definition of conditional probability.
- Note that we can re-order the the variables at will e.g. $P\left(x_{1}, \ldots, x_{n}\right)=P\left(x_{2}\right) P\left(x_{1} \mid x_{2}\right) \ldots P\left(x_{n} \mid x_{1}, \ldots x_{n-1}\right)$


## Conditional independence

- For any joint distribution over random variables $x_{1}, x_{2}, x_{3}$ we always have:

$$
\begin{equation*}
P\left(x_{1}, x_{2}, x_{3}\right)=P\left(x_{1}\right) P\left(x_{2} \mid x_{1}\right) P\left(x_{3} \mid x_{1}, x_{2}\right) \tag{2}
\end{equation*}
$$

- Now suppose that for some particular probability distribution $P$ we have that: $P\left(x_{3} \mid x_{1}, x_{2}\right)=P\left(x_{3} \mid x_{2}\right)$.
- 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 $P\left(x_{1}, x_{2}, x_{3}\right)=P\left(x_{1}\right) P\left(x_{2} \mid x_{1}\right) P\left(x_{3} \mid x_{2}\right)$
- 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 $P\left(x_{1}, x_{2}, x_{3}\right)=P\left(x_{1}\right) P\left(x_{2} \mid x_{1}\right) P\left(x_{3} \mid x_{2}\right)$
- 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:

or

$$
P\left(x_{1}, x_{2}, x_{3}\right)=P\left(x_{1}\right) P\left(x_{2} \mid x_{1}\right) P\left(x_{3} \mid x_{1}, x_{2}\right)
$$



$$
P\left(x_{1}, x_{2}, x_{3}\right)=P\left(x_{3}\right) P\left(x_{2} \mid x_{3}\right) P\left(x_{1} \mid x_{2}, x_{3}\right)
$$

## 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: $\mathrm{pa}_{k}$.
- 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 $\mathrm{pa}_{k}$ for random variables $x_{1}, \ldots, x_{K}$ represents a joint distribution which factorises as follows:

$$
\begin{equation*}
p(\mathbf{x})=\prod_{k=1}^{K} p\left(x_{k} \mid \mathrm{pa}_{k}\right) \tag{3}
\end{equation*}
$$

## BN structure and parameters

- For a BN to represent a given joint distribution we need to specify:

1. the DAG (the structure of the $B M$ )
2. the conditional probability distributions $p\left(x_{k} \mid \mathrm{pa}_{k}\right)$ (the parameters of the $B M)$

- A given DAG represents a set of joint distributions: each distribution in the set corresponds to a choice of values for the conditional distributions $p\left(x_{k} \mid \mathrm{pa}_{k}\right)$.
- 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 [Bisobi §8.1.1].
- In a Bayesian approach we have to define a prior probability distribution over parameters which (is supposed to) represent our beliefs about their values prior to observing the data.


## 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.)

$$
\begin{equation*}
p(\mathbf{t}, \mathbf{w})=p(\mathbf{w}) \prod_{n=1}^{N} p\left(t_{n} \mid \mathbf{w}\right) \tag{4}
\end{equation*}
$$

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 $\mathbf{w}$ as their (single) parent.
- Bishop [Bisoub] 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 $\mathbf{x}=\left(x_{1}, \ldots, x_{N}\right)^{T}$
2. The observed ouputs $\mathbf{t}=\left(t_{1}, \ldots, t_{N}\right)^{T}$
3. The parameter vector $\mathbf{w}$.
4. A hyperparameter $\alpha$.
5. The noise variance $\sigma^{2}$.

- We don't care how $\mathbf{x}$ is distributed and we would probably just set $\alpha$ to some value.
- So we would typically consider $\mathbf{x}, \alpha$ and also $\sigma^{2}$ 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



## Using BNs to represent ML models

- Machine learning research papers frequently use Bayesian networks to graphically represent machine learning models.
- They represent the data-generating process.
- Here's an example from NeurIPS 2019 [BS19].


## Differentially private Bayesian linear regression

### 3.1 Privacy mechanism

Using the Laplace mechanism to release the noisy sufficient statistics z results in the model shown in Figure 1. This is the same model used in non-private linear regression except for the introduction of $z$, which requires the exact sufficient statistics s to have finite sensitivity. A standard assumption in literature [Awan and Slavkovic, 2018, Sheffet, 2017, Wang, 2018, Zhang et al. 2012] is to assume x and $y$ have known a priori lower and upper bounds, $\left(a_{\mathbf{x}}, b_{\mathbf{x}}\right)$ and ( $\left.a_{y}, b_{y}\right)$, with bound widths $w_{\mathbf{x}}=b_{\mathbf{x}}-a_{\mathbf{x}}$ (assuming, for simplicity, equal bounds for all covariate dimensions) and $w_{y}=b_{y}-$ $a_{y}$, respectively. We can then reason about the worst case influence of an individual on each component of $\mathrm{s}=\left[X^{T} X, X^{T} \mathbf{y}, \mathbf{y}^{T} \mathbf{y}\right]$, recalling that $\mathbf{s}=\sum_{i} t\left(\mathbf{x}^{(i)}, y^{(i)}\right)$, so that $\left[\Delta_{\left(X^{T} X\right)_{j k}}, \Delta_{(X y)_{j}}, \Delta_{y^{2}}\right]=\left[w_{\mathbf{x}}^{2}, w_{\mathbf{x}} w_{y}, w_{y}^{2}\right]$. The number of unique element $\left.s^{2}\right]$ in s is $[d(d+1) / 2, d, 1]$, so $\Delta_{\mathbf{s}}=w_{\mathbf{x}}^{2} d(d+$


Figure 1: Private regression model. $1) / 2+w_{\mathbf{x}} w_{n} d+w_{n}^{2}$. The noisy sufficient statistics fit for public release are

## Another example

- An example from a paper on 'causal representation learning' [ $\mathrm{LM} \mathrm{M}+23]$


Figure 2: A representation of our assumptions. Observed variables are shown in gray ( $X^{\tau}$ and $R^{\tau}$ ) and latent variables in white. Optional causal edges are shown as dashed lines. A latent causal variable $C_{i}^{t}$ has as parents a subset of the causal factors at the previous time step $C^{t-1}=\left\{C_{1}^{t-1}, \ldots, C_{K}^{t-1}\right\}$, and its latent binary interaction variable $I_{i}^{t}$. The interaction variables are determined by an observed regime variable $R^{t}$ and potentially by the variables from the previous time step $C^{t-1}$ (e.g., in a collision). The regime variable can be a dynamical process over time as well, for example, by depending on the previous time step. The observation $X^{\tau}$ is a high-dimensional entangled representation of all causal variables $C^{\tau}$ at time step $\tau$.

## Naive Bayes

- In a naive Bayes model for classification [Bisō variables $\mathbf{x}=\left(x_{1}, \ldots x_{D}\right)$ are assumed independent conditional on the class variable $\mathbf{z}$ :

$$
\begin{equation*}
P(\mathbf{x}, \mathbf{z})=P(\mathbf{z}) P(\mathbf{x} \mid \mathbf{z})=P(\mathbf{z}) \prod_{i=1}^{D} P\left(x_{i} \mid \mathbf{z}\right) \tag{5}
\end{equation*}
$$

- Let's have a look at a naive Bayes model. [Muriz3; p. 163].
- And a latent variable model [Mūr2̄3', p. 159].


## Hierarchical Linear Regression

Here's a nice example of using Bayesian networks to represent different approaches to a linear regression problem where there is extra 'structure'.

## Standard regression (abbreviated)

parameter

$$
\theta
$$

observations

$$
P(\theta, y)=P(\theta) \prod_{i=1}^{k} P(y ; \mid \theta)
$$

## Separate regressions (abbreviated)

parameters

${ }^{\theta_{k}}$
observations
$y_{1}$
$y_{2}$
$y_{k}$

$$
P(\theta, y)=\prod_{i=1}^{k} P\left(y_{i} \mid \theta_{i}\right) P\left(\theta_{i}\right)
$$

## Hierarchical rearession (abbreviated)

model
$\mu, \sigma^{2}$
parameters

${ }^{\theta_{k}}$
observations $\boldsymbol{Y}_{1}$
$y_{2}$
$y_{k}$

$$
P\left(\theta, y, \mu, \sigma^{2}\right)=P\left(\mu, \sigma^{2}\right) \prod_{i}^{k} P\left(y_{i} \mid \theta_{i}\right) P\left(\theta_{i} \mid \mu, \sigma^{2}\right)
$$

## Conditional independence

- A random variable $x$ is independent of another random variable $y$ conditional on a set of random variables $S$ if and only if:

$$
\begin{equation*}
P(x, y \mid S)=P(x \mid S) P(y \mid S) \tag{6}
\end{equation*}
$$

Equivalently:

$$
\begin{equation*}
P(x \mid S)=P(x \mid y, S) \tag{7}
\end{equation*}
$$

- The DAG for a BN encodes conditional independence relations.



## Conditional independence

- A random variable $x$ is independent of another random variable $y$ conditional on a set of random variables $S$ if and only if:

$$
\begin{equation*}
P(x, y \mid S)=P(x \mid S) P(y \mid S) \tag{6}
\end{equation*}
$$

Equivalently:

$$
\begin{equation*}
P(x \mid S)=P(x \mid y, S) \tag{7}
\end{equation*}
$$

- The DAG for a BN encodes conditional independence relations.
- Some of the following slides are modified versions of those made available by David Barber,
- who has written a great (freely available) book on Bayesian machine learning [Bar-12]


## Independence $\perp$ in Bayesian Networks - Part I

All Bayesian networks with three nodes and two links:


- In (a), (b) and (c), $A$ and $B$ are conditionally independent given $C$.
(a) $p(A, B \mid C)=\frac{p(A, B, C)}{p(C)}=\frac{p(A \mid C) p(B \mid C) p(C)}{p(C)}=p(A \mid C) p(B \mid C)$
(b) $p(A, B \mid C)=\frac{p(A) p(C \mid A) p(B \mid C)}{p(C)}=\frac{p(A, C) p(B \mid C)}{p(C)}=p(A \mid C) p(B \mid C)$
(c) $p(A, B \mid C)=\frac{p(A \mid C) p(C \mid B) p(B)}{p(C)}=\frac{p(A \mid C) p(B, C)}{p(C)}=p(A \mid C) p(B \mid C)$
- In (d) the variables $A, B$ are conditionally dependent given $C$, $p(A, B \mid C) \propto p(A, B, C)=p(C \mid A, B) p(A) p(B)$.


## Independence $\perp$ in Bayesian Networks - Exercises



- Show that in (d), we have $A \perp B$.
- For each of (a), (b) and (c), assume that each variable is binary, and find parameters so that $A \not \subset B$


## Paths and colliders

$$
p(A, B, C, D, E)=p(A) p(B) p(C \mid A, B) p(D \mid C) p(E \mid B, C)
$$



- A node is a collider on some path if both arrows point into it on that path.
- $C$ is a collider on the path $(A, C, B)$ but is not a collider on the path $(A, C, E)$ or on any of the following paths: $(A, C, E, B),(D, C, B)$ or ( $D, C, E$ ).


## $d$-separation

- If all paths from node $x$ to node $y$ are blocked given nodes $S$ then $x$ and $y$ are $d$-separated by $S$.
- A path is blocked by $S$ if at least one of the following is the case:

1. there is a collider on the path that is not in $S$ and none of its descendants are in $S$
2. there is a non-collider on the path that is in $S$.

- If $x$ and $y$ are $d$-separated by $S$ then $x \perp y \mid S$ for any probability distribution which factorises according to the DAG.
- Let's do some d-separation exercises.


## Checking for $d$-separation



A path is blocked by $S$ if at least one of the following is the case:

1. there is a collider on the path that is not in $S$ and none of its descendants are in $S$
2. there is a non-collider on the path that is in $S$.

## Hierarchical regression revisited

model
$\mu, \sigma^{2}$
parameters

${ }^{\theta_{k}}$
observations $\boldsymbol{Y}_{1}$
$y_{2}$
$y_{k}$

$$
P\left(\theta, y, \mu, \sigma^{2}\right)=P\left(\mu, \sigma^{2}\right) \prod_{i}^{k} P\left(y_{i} \mid \theta_{i}\right) P\left(\theta_{i} \mid \mu, \sigma^{2}\right)
$$

David Barber.
Bayesian Reasoning and Machine Learning.
Cambridge University Press, 2012.
Christopher M. Bishop.
Pattern Recognition and Machine Learning.
Springer, 2006.
Rarrett Bernstein and Daniel R Sheldon.
Differentially private Bayesian linear regression.
In Advances in Neural Information Processing Systems, pages
525-535, 2019.
Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves.
BISCUIT: Causal representation learning from binary interactions.
In Proc. UAl23, 2023.
Kevin P. Murphy.
Probabilistic Machine Learning: Advanced Topics.
MIT Press, 2023.

