Probabilistic Graphical Models

Dylan Dijk

1 Introduction

Our goal is to represent the joint distribution over some set of random variables in the most compact way possible.

We can measure the "compactness" of a representation of a joint distribution, by counting the number of *independent parameters* that are required.

Independent parameters, are parameters whose values are not determined by others.

For example, if we have n random variables, $X_1, \dots X_n$, that represent n independent coin tosses, we can factorise the joint distribution in the following way:

$$p(x_1, \dots x_n) = \prod_i p(x_i) \tag{1}$$

These are binary random variables; therefore, we can determine the probability for any of the 2^n possibilities, $(x_1, \dots x_n)$, using n parameters. If we did not factorise the joint distribution, into the n probabilities, then we would require $2^n - 1$ independent parameters. Because, for every combination of possibilities, $(x_1, \dots x_n)$, we would need to denote the probability apart from one as we would be able to evaluate it using that the probabilities must sum to one.

2 Bayesian Network Representation

We first give an informal high-level definition of a Bayesian network (BN), and we then provide formal definitions afterwards.

A **Bayesian network** is a pair $(p, \mathcal{G})$. Where p is a probability distribution over a set of random variables, and $\mathcal{G}$ is a DAG. The DAG $\mathcal{G}$ can be viewed/constructed in two different ways:

1. Representing a set of **conditional independence statements** that are held by p.
2. A way to present the **parameterisation** of the joint distribution p.

It then turns out, that when constructing $\mathcal{G}$ using either perspective, the properties induced from the graph in the alternate perspective still hold. For example, if we start from a valid parameterisation of p, then use this to create a DAG. The set of independence statements induced from this DAG will hold for p.

We now will define both of these perspectives formally, and start with giving some core definitions related to graphs.

Definition 1. Graph terminology

DAG A directed graph $\mathcal{G} = (E, V)$, that has no directed cycles.

Parent/Children If there exists a directed **edge** from node $v_1 \in V$ to $v_2 \in V$ then v_1 is a parent of v_2, and v_2 is a child of v_1.

Descendant If there exists a directed **path** from $v_1 \in V$ to $v_2 \in V$ then v_2 is a descendant of v_1. The children of a node is therefore a subset of its descendants.

We now look at the first perspective (1.) of a Bayesian network structure $\mathcal{G}$. Below we provide a definition for how a DAG, where each node represents a random variable, encodes conditional independence statements.

Definition 2. BN structure - encoding conditional indpenedence statements

A **Bayesian network structure** $\mathcal{G}$ is a DAG whose nodes represent random variables $X_1, ..., X_n$. $\mathcal{G}$ encodes the following set of conditional local independence assumptions, called the **local independencies**, and denoted by $\mathcal{I}_{\mathcal{L}}(\mathcal{G})$:

$$(X_i \perp NonDescendants^{\mathcal{G}}(X_i) | Parents^{\mathcal{G}}(X_i)) \quad \forall X_i \in V^{\mathcal{G}} \tag{2}$$

Local independencies are sometimes referred to as **local Markov independencies**.

The independence statements encoded by a graph, and how it relates to the distribution is an important point; therefore, we formalise it here now. If we let $\mathcal{I}(p)$ denote the set of independence statements that hold in p, we say that $\mathcal{G}$ is an **I-Map** for p if $\mathcal{I}_{\mathcal{L}}(\mathcal{G}) \subseteq \mathcal{I}(p)$. In other words, if $\mathcal{G}$ is an I-Map for p then any independence that $\mathcal{G}$ asserts must hold in p.

We now look at the second perspective (2.), and provide a definition for howshow how $\mathcal{G}$ can be used to present the parameterisation of the joint distribution.

Definition 3. BN structure - Factorisation of joint probability over $\mathcal{G}$

Let $\mathcal{G}$ be a BN structure over the variables $X_1, \ldots, X_n$. We say that a distribution p, for these random variables, factorises according to $\mathcal{G}$ if p can be expressed as a product:

$$p(X_1, ..., X_n) = \prod_{i=1}^{n} p(X_i \mid Parents^{\mathcal{G}}(X_i)) \tag{3}$$

We give an example in the Appendix (A.1.2), that shows how we can use the factorisation of a distribution presented by a graph to decompose a conditional distribution.

Theorem 1. Structure $\mathcal{G}$ represents both conditional independence statements and factorisation of p

Let $\mathcal{G}$ be a BN structure for random variables $X_1, \ldots, X_n$, and let p be a joint distribution over the same space.

$$\mathcal{G} \text{ is an I-Map for } p \iff p \text{ factorises according to } \mathcal{G}$$

We now finally give the formal definition of a Bayesian Network.

Definition 4. Bayesian network

A Bayesian network is a pair $\mathcal{B} = (\mathcal{G}, p)$ where p factorizes over $\mathcal{G}$, and where p is specified as a set of CPDs associated with $\mathcal{G}$'s nodes.

So far, we know that if we have a Bayesian Network, then the joint probability density p must satisfy the **local independencies** of $\mathcal{G}$. By the ($\Leftarrow$) direction of Theorem 1.
But are there other additional independencies that can be defined from the graph $\mathcal{G}$ that also hold for p?

Definition 5. d-separated nodes

Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ be three sets of nodes in $\mathcal{G}$. We say that $\mathbf{X}$ and $\mathbf{Y}$ are d-separated given $\mathbf{Z}$, if in the skeleton of the moralised ancestral graph of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ there exists no paths from the node set $\mathbf{X}$ to the node set $\mathbf{Y}$ that does not include a node from the node set $\mathbf{Z}$.

Or more concisely: $\mathbf{Z}$ separates $\mathbf{X}$ and $\mathbf{Y}$ in the skeleton of the moralised ancestral graph of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$.

Definition 6. Global Markov Independencies

We name the set of independence statements, $\mathcal{I}(\mathcal{G})$, that are induced by the **d-separation theorem**, the global Markov independencies.
What we mean by induced is that if, $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$, then we say $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$

The local Markov property is implied by the construction of a Bayesian network, but it is also a special case of the d-separation theorem.

We have given some "extra" independence statements encoded by $\mathcal{G}$, but have not said whether these actually hold for the distribution that factorises over $\mathcal{G}$. The next theorem gives us exactly that, independencies in $\mathcal{I}(\mathcal{G})$ are those that are guaranteed for every Bayesian network with $\mathcal{G}$ as its graph.

Theorem 2. The d-separation Theorem

If a distribution p factorises according to $\mathcal{G}$ then $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(p)$

However, the converse of Theorem 2 does not hold in general, an example is shown in the Appendix (A.1.1). In other words, for a BN the independence statements induced by the d-separation theorem do not in general contain all of the independence statements of its joint distribution.

In the above definitions and theorems, we have shown if p factorises over $\mathcal{G}$ we can derive a set of independence statements. The question now is, when given a distribution p to find a graph $\mathcal{G}$ whose set of independence statements are as close as possible to the independencies in p. This is covered further in chapter 3.4 of Probabilistic Models by Daphne Koller and Nir Friedman.

The ideal case would be to have a graph that encodes all of the independence statements of the distribution. A graph $\mathcal{K}$ is called a perfect map (P-map) for a set of independencies $\mathcal{I}$ if $\mathcal{I}(\mathcal{K}) = \mathcal{I}$

3 Undirected Graphical Models

BN structures rely on a choice of direction, which can make it difficult to encode certain sets of independence statements. Using an undirected graph could make it easier.

Markov networks allow us to represent independence statements without selecting a direction to the influence. **A Markov network structure is an undirected graph**, where the nodes represent the random variables.

In Bayesian networks the directions in the structure told us how to factorise p, in terms of conditional probabilities, which are not symmetric. In a MN structure, as it is undirected, the parameterisation cannot be directed. We therefore use **factors** instead of conditional probabilities.

Definition 7. Factor

For a set of random variables $\mathbf{X}$, a factor is defined as a function from the values that the random variables can take to $\mathbb{R}$. The set of variables, $\mathbf{X}$, that the factor is defined for, is called the scope of the factor.

Factors generalise probability distributions, for example a joint distribution over $\mathbf{X}$ is a factor over $\mathbf{X}$. We can write the chain rule for conditional probabilities as a product of factors. Usually for Markov networks we use factors that map to the positive real numbers $\mathbb{R}^+$.

The product of factors is defined in a specific way, so that taking the products makes sense in terms of the random variables interacting.

Definition 8. Gibbs distribution

A distribution p_ϕ is a Gibbs distribution over the set of random variables $\mathbf{X} = \{X_1, \ldots, X_n\}$ parameterized by a set of **factors** $\boldsymbol{\phi} = \{\phi_1(\mathbf{X}_1), \ldots, \phi_K(\mathbf{X}_k)\}$, with $\mathbf{X}_i \subseteq \mathbf{X}$, if it is defined as:

$p_{\boldsymbol{\phi}}(X_1, ..., X_n) = \frac{1}{Z}\tilde{p}(X_1, ..., X_n)$,

where $\tilde{p}(X_1, ..., X_n) = \phi_1(\mathbf{X}_1) \times \phi_2(\mathbf{X}_2) \times \cdots \times, \phi_K(\mathbf{X}_k)$ is an unnormalised measure and Z is the normalising constant.

We can think of each of the individual factors as contributing to the joint distribution. And, if we have a factor that includes two random variables X and Y then we are introducing an interaction between them. Therefore, it makes sense that when we define MN structures later that all of the rvs in the scope of a factor should have an edge between them.

Just as we did in Bayesian networks, we want to express the parameterisation of a distribution using a graph structure, but we use an undirected graph $\mathcal{H}$. Analogous to Definition 3, we have:

Definition 9. Markov network factorisation

A Gibbs distribution p_ϕ, with $\phi = \{\phi(\mathbf{D}_1), \ldots, \phi(\mathbf{D}_K)\}$ factorises over a Markov network $\mathcal{H}$, if each $\mathbf{D}_k$ is a complete subgraph/clique of $\mathcal{H}$.

A clique is subgraph where every node is connected to each other.

So given a Gibbs distribution p_ϕ, we can then give a Markov structure $\mathcal{H}$ so that the distribution factorises over it. Now, every complete subgraph is a subset of a maximal clique. So we can reduce the number of factors in the parameterisation by having factors only for maximal cliques. This parameterisation will then still factorise over $\mathcal{H}$. I give an example showing this in the Appendix (A.2.1).

In the Appendix (A.2.5) we give an example of a distribution that factorises over an MN structure. In addition, in the Appendix we define and give an example of a pairwise Markov network (A.2.2).

Definition 10. Global Indedependencies for Markov networks

If a set of nodes $\mathbf{Z}$ separates $\mathbf{X}$ and $\mathbf{Y}$ in $\mathcal{H}$, then $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$. Define the global independencies as:

$$\mathcal{I}(\mathcal{G}) = \{(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}) : \mathrm{sep}_{\mathcal{H}}(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})\} \tag{4}$$

In the Appendix (A.2.4) we give an example of going from a set of conditional independence statements to an MN structure.

Theorem 3. Structure $\mathcal{H}$ represents conditional independence statements and factorisation of p

Let $\mathcal{H}$ be a MN structure for random variables $X_1, \ldots, X_n$, and let p be a distribution over the same space.

$$\begin{aligned} \mathcal{H} \text{ is an I-Map for } p &\Leftarrow p \text{ is a Gibbs distribution that factorises over } \mathcal{H} && (5) \\ \mathcal{H} \text{ is an I-Map for positive distribution } p &\Rightarrow p \text{ is a Gibbs distribution that factorises over } \mathcal{H} && (6) \end{aligned}$$

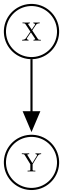
The direction (6) is known as the Hammersley-Clifford theorem. A positive distribution is one in which the probability assigned to measurable sets, apart from the empty set, is strictly greater than zero.

A Appendix

A.1 Bayesian Networks (BN)

A.1.1 Example showing that $\mathcal{I}(p) \not\subset \mathcal{I}(\mathcal{G})$

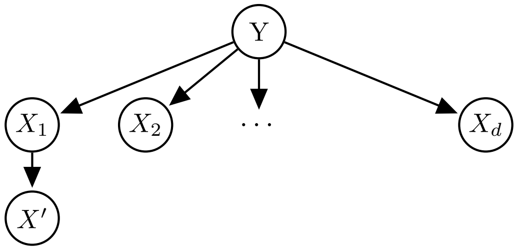
Suppose that we have two random variables Y and X that are independent. A possible Bayesian structure is:



We clearly cannot deduce that X and Y are independent using the d-separation theorem. The skeleton of the moralised ancestral graph will be the undirected version of the graph above and therefore X and Y will not be separated.

A.1.2 Redundant Feature in Classification Task

If we have the random variables $X_1, \dots, X_d$ and Y with an associated Bayesian network with $\mathcal{G}$ as given below:



We can write the conditional distribution for Y given the other variables as:

$$
\begin{aligned}
p(Y \mid X', X_1, \dots, X_d) &= \frac{p(Y, X', X_1, \dots, X_d)}{p(X', X_1, \dots, X_d)} && (7)\\
&= \frac{p(Y)p(X' \mid X_1) \prod_{i=1}^{d} p(X_i \mid Y)}{p(X', X_1, \dots, X_d)} && (8)\\
&\propto p(Y) \prod_{i=1}^{d} p(X_i \mid Y) && (9)
\end{aligned}
$$

We use that p factorises over $\mathcal{G}$ to get (8).

Now if we wanted to perform classification for the outcome variable Y, using the rest of the random variables as features, we would look for:

$$\underset{y}{\operatorname{argmax}}\ p(Y = y \mid X', X_1, \ldots, X_d) \tag{10}$$

And we can see from (9), that we can ignore X' completely to complete this classification task.

A.2 Markov Networks (MN)

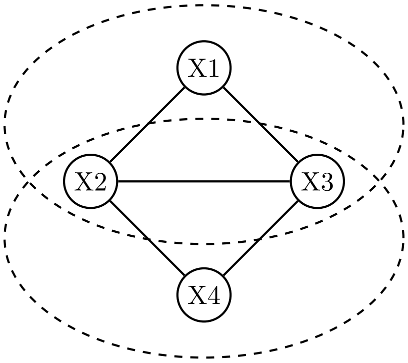
A.2.1 Different clique potentials

Factors that parameterize a Markov network are often called clique potentials.

Lets say we have a Gibbs distribution:

$$p_{\Phi}(X_1, ..., X_4) = \frac{1}{Z}\phi_1(X_1, X_2, X_3) \cdot \phi_1(X_2, X_3, X_4) \tag{11}$$

A MN structure $\mathcal{H}$, such that p_{Φ} factorises over it is given by:



Every complete subgraph is a subset of some (maximal) clique, so in general we can reduce the number of factors in our parameterisation by allowing factors only for maximal cliques.

So in this example we can just provide a factor for the maximal clique over the nodes $X_1, \ldots, X_4$, then $\mathcal{H}$ will still be a valid MN structure.

We can get this single factor $\phi(X_1, X_2, X_3, X_4)$ by taking the factor product of ϕ_1 and ϕ_2.

Definition 11. Factor Product

Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ be three disjoint sets of variables, and let $\phi_1(\mathbf{X}, \mathbf{Y}), \phi_2(\mathbf{Y}, \mathbf{Z})$ be two factors. The factor product $\phi_1 \times \phi_2$ is defined to be a factor:

$$\psi(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \phi_1(\mathbf{X}, \mathbf{Y}) \cdot \phi_1(\mathbf{Y}, \mathbf{Z}) \tag{12}$$

Although we can do this to reduce the number of factors used to parameterise a MN structure, this obscures the structure that is present in the original set of factors.

This example shows that starting from a MN structure, there are options for how many factors we want to use to parameterise it. Or in other words how “fine” or “coarse” we want the structure to be.

A.2.2 Pairwise Markov Network

Related to the previous section (A.2.1), a subclass of Markov networks are pairwise Markov networks. These are distributions that can be parameterised by factors that are over at most two random variables.

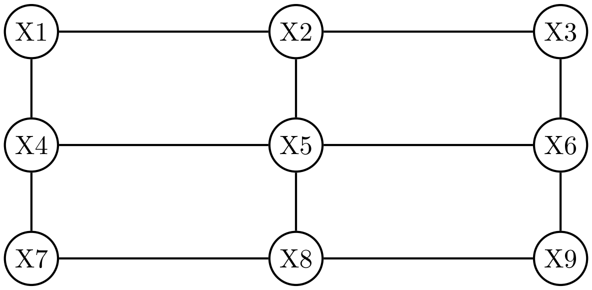
A common example of a pairwise Markov network, is one that can be represented by a grid.

Lets say we had a distribution that could be parameterised in the following way:

$$p_{\boldsymbol{\phi}}(\mathbf{X}) = \frac{1}{Z} \prod_{(i,j)\in\epsilon} \phi_{i,j}(X_i, X_j) \tag{13}$$

$$\text{Where} \quad \epsilon = \{(i,j) : |i-j| \in \{1,3\},\ i<j\} \tag{14}$$

A $\mathcal{H}$ structure that $p_{\boldsymbol{\phi}}$ will factorise over will be given by:



A.2.3 Graphical Modelling - Senators Pairwise Markov Network

Problem description:

- 100 senators
- Dataset of rollcall votes for each senator. Vote is either yes or no.
- Want to analyse dependencies between senators
- $x^{(i)} \in \{-1,1\}$ is a vote for senator i.

We model the joint distribution $p(x^{(1)},\ldots,x^{(100)})$ as a type of pairwise Markov network. Precisely, we assume all factors are over **pairs** of variables.

Therefore we have that:

$$p(x^{(1)}, \dots, x^{(100)}) \propto \prod_{u,v \in E} g'_{u,v}(x^{(u)}, x^{(v)}) \tag{15}$$

For some edge set E, which determines which pairs of random variables have a corresponding factor.

Now we make some further modelling assumptions, and assume that

$$g'_{u,v} := \exp(w_{u,v} x^{(u)} x^{(v)}) \tag{16}$$

For all u, v, we can now write:

$$p(x^{(1)}, \dots, x^{(100)}; W) \propto \exp(\sum_{u=1}^{100} \sum_{v>u} w_{u,v} x^{(u)} x^{(v)}) \tag{17}$$

Now if we want to carry out inference, but for an individual senator we can look at the conditionals.

$$\begin{align} p(x^{(1)} \mid x^{(2)} \dots, x^{(100)}; W) &\propto \exp(\sum_{u=1}^{100} \sum_{v>u} w_{u,v} x^{(u)} x^{(v)}) \tag{18} \\ &\propto \exp(\sum_{v=2}^{100} w_{1,v} x^{(1)} x^{(v)}) \tag{19} \end{align}$$

The normalising constant for the conditional distribution will be given by:

$$\sum_{x^{(1)} \in \{-1,1\}} \exp(\sum_{v=2}^{100} w_{1,v} x^{(1)} x^{(v)}) \tag{20}$$

Therefore we get that conditional probability distribution is given by:

$$\begin{align} p(x^{(1)} \mid x^{(2)} \dots, x^{(100)}; W) &= \frac{\exp(\sum_{v=2}^{100} w_{1,v} x^{(1)} x^{(v)})}{\exp(\sum_{v=2}^{100} w_{1,v} x^{(v)}) + \exp(\sum_{v=2}^{100} -w_{1,v} x^{(v)})} \tag{21} \\ &= \sigma(x^{(1)} \cdot (2 \sum_{v=2}^{100} w_{1,v} x^{(v)})) \tag{22} \end{align}$$

Where σ is the sigmoid function.

Now, if we observe a dataset $D := \{\underline{\mathbf{x}}_i\}_{i=1}^{n}$, where $\underline{\mathbf{x}}_i$ is the votes of the 100 senators on bill i.

The likelihood is given by:

$$p(\mathbf{x}^{(1)} \mid \mathbf{x}^{(2)} \ldots, \mathbf{x}^{(100)}; W) = \prod_{i=1}^{n} \sigma(x_i^{(1)} \cdot (2 \sum_{v=2}^{100} w_{1,v} x_i^{(v)})) \tag{23}$$

Therefore, this is equivalent to logistic regression.

A.2.4 Conditional independence statements to MN structure

Suppose we have the set of random variables: {Math, ML, SM1, Python}, and a probability distribution that can be factorised as:

$p(\text{Math, ML, SM1, Python}) \propto \phi_1(\text{Math, SM1}) \times \phi_2(\text{SM1, ML, Python})$

You can show that if we have three disjoint subsets $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ with $\mathcal{X} = \mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z}$

$$p(\mathcal{X}) = \phi_1(\mathbf{X}, \mathbf{Z}) \times \phi_2(\mathbf{Z}, \mathbf{Y}) \iff \mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z} \tag{24}$$

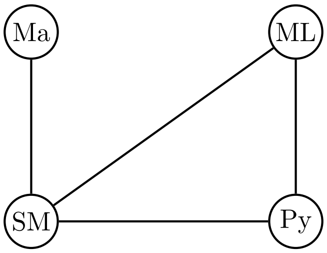
Therefore from our factorization we have the following conditional independence statement:

$$\text{Math} \perp \text{Python, ML} \mid \text{SM1} \tag{25}$$

Then using **decomposition** and **weak union** we have the following conditional independence statements:

$$\text{Math} \perp \text{Python, ML} \mid \text{SM1} \tag{26}$$
$$\text{Math} \perp \text{ML} \mid \text{SM1} \tag{27}$$
$$\text{Math} \perp \text{Python} \mid \text{SM1} \tag{28}$$
$$\text{Math} \perp \text{ML} \mid \text{SM1, Python} \tag{29}$$
$$\text{Math} \perp \text{Python} \mid \text{SM1, ML} \tag{30}$$

The following undirected graph, encodes all of these statements.

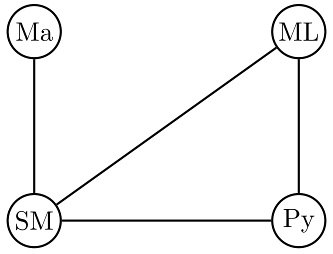


A.2.5 Factorisation of distribution to MN structure

Now if we use the factorisation of the joint probability distribution over the random variables {Math, ML, SM1, Python}

$p(\text{Math, ML, SM1, Python}) \propto \phi_1(\text{Math, SM1}) \times \phi_2(\text{ML, SM1, Python})$

To create an undirected graph, such that the distribution factorises over it, we get:



Can see that the graph is identical to the graph generated to encode the independence statements. This is an example, showing Theorem 3 in action. We first constructed the graph so that it encoded the independence statements, and then we created a graph so that the distribution factorises over it.