Chapter 10
Directed graphical models (Bayes nets)

10.1 Introduction

10.1.1 Chain rule

$$p(x_{1:V}) = p(x_1)p(x_2|x_1)p(x_3|x_{1:2})\cdots p(x_N|x_{1:V-1}) \tag{10.1}$$

10.1.2 Conditional independence

X and Y are **conditionally independent** given Z, denoted $X \perp Y|Z$, iff the conditional joint can be written as a product of conditional marginals, i.e.

$$X \perp Y|Z \iff p(X,Y|Z) = p(X|Z)p(Y|Z) \tag{10.2}$$

first order **Markov assumption**: the future is independent of the past given the present,

$$x_{t+1} \perp x_{1:t-1}|x_t \tag{10.3}$$

first-order **Markov chain**

$$p(x_{1:V}) = p(x_1)\prod_{t=2}^{V} p(x_t|x_{t-1}) \tag{10.4}$$

10.1.3 Graphical models

A **graphical model**(GM) is a way to represent a joint distribution by making CI assumptions. In particular, the nodes in the graph represent random variables, and the (lack of) edges represent CI assumptions.

There are several kinds of graphical model, depending on whether the graph is directed, undirected, or some combination of directed and undirected. In this chapter, we just study directed graphs. We consider undirected graphs in Chapter 19.

10.1.4 Directed graphical model

A **directed graphical model**or **DGM** is a GM whose graph is a DAG. These are more commonly known as **Bayesian networks**. However, there is nothing inherently Bayesian about Bayesian networks: they are just a way of defining probability distributions. These models are also called **belief networks**. The term belief here refers to subjective probability. Once again, there is nothing inherently subjective about the kinds of probability distributions represented by DGMs.

Ordered Markov property

$$x_s \perp x_{\text{pred}(s) \setminus \text{pa}(s)} \perp x_{\text{pa}(s)} \tag{10.5}$$

where $\text{pa}(s)$ are the parents of nodes, and $\text{pred}(s)$ are the predecessors of nodes in the DAG.

Markov chain on a DGM

$$p(x_{1:V}|G) = \prod_{t=1}^{V} p(x_t|x_{\text{pa}(t)}) \tag{10.6}$$

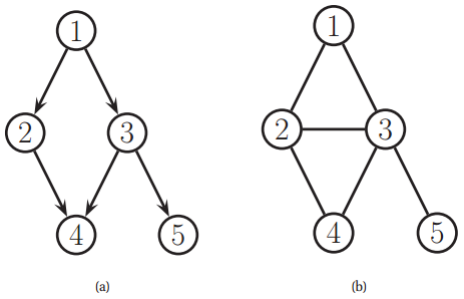

Fig. 10.1: (a) A simple DAG on 5 nodes, numbered in topological order. Node 1 is the root, nodes 4 and 5 are the leaves. (b) A simple undirected graph, with the following maximal cliques: 1,2,3, 2,3,4, 3,5.

10.2 Examples

10.2.1 Naive Bayes classifiers

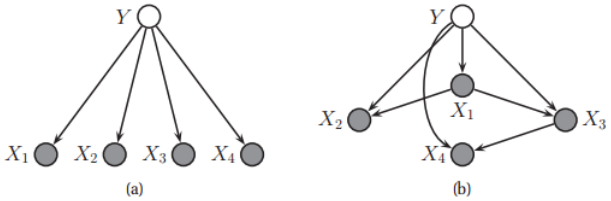

Fig. 10.2: (a) A naive Bayes classifier represented as a DGM. We assume there are $D = 4$ features, for simplicity. Shaded nodes are observed, unshaded nodes are hidden. (b) Tree-augmented naive Bayes classifier for $D = 4$ features. In general, the tree topology can change depending on the value of y.

10.2.2 Markov and hidden Markov models

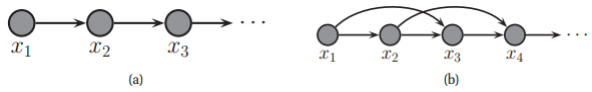

Fig. 10.3: A first and second order Markov chain.

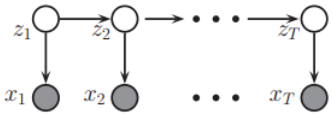

Fig. 10.4: A first-order HMM.

10.3 Inference

Suppose we have a set of correlated random variables with joint distribution $p(\boldsymbol{x}_{1:V}|\boldsymbol{\theta})$. Let us partition this vector into the **visible variables** $\boldsymbol{x}_v$, which are observed, and the **hidden variables**, $\boldsymbol{x}_h$, which are unobserved. Inference refers to computing the posterior distribution of the unknowns given the knowns:

$$p(\boldsymbol{x}_h|\boldsymbol{x}_v,\theta) = \frac{p(\boldsymbol{x}_h,\boldsymbol{x}_v|\boldsymbol{\theta})}{p(\boldsymbol{x}_v|\boldsymbol{\theta})} = \frac{p(\boldsymbol{x}_h,\boldsymbol{x}_v|\boldsymbol{\theta})}{\sum_{\boldsymbol{x}'_h} p(\boldsymbol{x}'_h,\boldsymbol{x}_v|\boldsymbol{\theta})} \quad (10.7)$$

Sometimes only some of the hidden variables are of interest to us. So let us partition the hidden variables into **query variables**, $\boldsymbol{x}_q$, whose value we wish to know, and the remaining **nuisance variables**, $\boldsymbol{x}_n$, which we are not interested in. We can compute what we are interested in by **marginalizing out** the nuisance variables:

$$p(\boldsymbol{x}_q|\boldsymbol{x}_v,\boldsymbol{\theta}) = \sum_{\boldsymbol{x}_n} p(\boldsymbol{x}_q,\boldsymbol{x}_n|\boldsymbol{x}_v,\boldsymbol{\theta}) \quad (10.8)$$

10.4 Learning

MAP estimate:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log p(\boldsymbol{x}_{i,v}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \quad (10.9)$$

10.4.1 Learning from complete data

If all the variables are fully observed in each case, so there is no missing data and there are no hidden variables, we say the data is **complete**. For a DGM with complete data, the likelihood is given by

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{\theta}) &= \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{\theta}) \\ &= \prod_{i=1}^{N}\prod_{t=1}^{V} p(\boldsymbol{x}_{it}|\boldsymbol{x}_{i,\mathrm{pa}(t)},\boldsymbol{\theta}_t) \\ &= \prod_{t=1}^{V} p(\mathcal{D}_t|\boldsymbol{\theta}_t) \end{aligned} \quad (10.10)$$

where $\mathcal{D}_t$ is the data associated with node t and its parents, i.e., the t'th family.

Now suppose that the prior factorizes as well:

$$p(\boldsymbol{\theta}) = \prod_{t=1}^{V} p(\boldsymbol{\theta}_t) \quad (10.11)$$

Then clearly the posterior also factorizes:

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = \prod_{t=1}^{V} p(\mathcal{D}_t|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t) \quad (10.12)$$

10.4.2 Learning with missing and/or latent variables

If we have missing data and/or hidden variables, the likelihood no longer factorizes, and indeed it is no longer convex, as we explain in detail in Section TODO. This means we will usually can only compute a locally optimal ML or MAP estimate. Bayesian inference of the parameters is even harder. We discuss suitable approximate inference techniques in later chapters.

10.5 Conditional independence properties of DGMs

10.5.1 d-separation and the Bayes Ball algorithm (global Markov properties)

1. P contains a chain

$$\begin{aligned} p(x,z|y) &= \frac{p(x,y,z)}{p(y)} = \frac{p(x)p(y|x)p(z|y)}{p(y)} \\ &= \frac{p(x,y)p(z|y)}{p(y)} = p(x|y)p(z|y) \end{aligned} \tag{10.13}$$

2. P contains a fork

$$p(x,z|y) = \frac{p(x,y,z)}{p(y)} = \frac{p(y)p(x|y)p(z|y)}{p(y)} = p(x|y)p(z|y) \tag{10.14}$$

3. P contains v-structure

$$p(x,z|y) = \frac{p(x,y,z)}{p(y)} = \frac{p(x)p(z)p(y|x,z)}{p(y)} \neq p(x|y)p(z|y) \tag{10.15}$$

10.5.2 Other Markov properties of DGMs

10.5.3 Markov blanket and full conditionals

$$mb(t) = ch(t) \cup pa(t) \cup copa(t) \tag{10.16}$$

10.5.4 Multinoulli Learning

Multinoulli Distribution

$$Cat(x|\mu) = \prod_{k=1}^{K} \mu_k^{x_k} \tag{10.17}$$

then from **??** and 10.17:

$$p(x|G,\theta) = \prod_{v=1}^{V} \prod_{c=1}^{C_v} \prod_{k=1}^{K} \theta_{vck}^{y_{vck}} \tag{10.18}$$

Likelihood

$$p(D|G,\theta) = \prod_{n=1}^{N} p(x_n|G,\theta) = \prod_{n=1}^{N} \prod_{v=1}^{V} \prod_{c=1}^{C_{nv}} \prod_{k=1}^{K} \theta_{vck}^{y_{nvck}} \tag{10.19}$$

where $y_{nv} = f(pa(x_{nv}))$, f(x) is a map from x to a vector, there is only one element in the vector is 1.

10.6 Influence (decision) diagrams *