BAYESIAN NETWORKS and STATISTICAL GENETICS :
Probabilistic Learning of Bayesian Networks
Timo Koski 2010-05-04

1 Introduction

By learning from data one often means the process of inferring a general law or principle from the observations of particular instances. The general law is a piece of knowledge about the mechanism of nature that generates the data. The learning can be done by use of 'MODELS', which serve as the language in which the constraints predicated on the data can be described. In this course the language is that of directed acyclic graphs (involving a language for causality) and the joint distributions recursively factorized along them. These notes give a more formal treatment of some parts of chapter 3 in (Jensen 2001). The presentation here is in many essential parts based on (Heckerman 1996, 1997).

2 Probabilistic Models with Conditional Independence

There is one type of learning that we will be concerned with: this is inferring, analysing and using a family of models indexed by parameters.

The first family of models to be studied is conditional independence. We shall consider two examples of this to start with.

2.1 Modeling and Learning for Tosses of a Thumbtack

2.1.1 The Model Family

The mathematics involved here is found in greater detail in (v. Mises and Geiringer 1964) but goes in fact back to P.S. Laplace.

We consider a sequence of flips of a thumbtack (Heckerman 1996, 1997). If we throw a thumbtack in the air, it will come to rest either on its point (0) or on its head (1). Suppose we flip the thumbtack n times (fixing n in advance), making sure that that the physical properties of the thumbtack and the conditions under which it is flipped remain stable over time. We let $\mathbf{x}$ denote the sequence of outcomes of the flips, $\mathbf{x} = x_{i_1} x_{i_2} \ldots x_{i_n}$, $x_{i_l} \in \{0, 1\}$. Let now Θ be a random variable (quantity), whose values are numbers, denoted by θ, between zero and one, $0 \leq \theta \leq 1$. These values θ correspond to the possible values of the *chance of obtaining heads* in tossing thumbtack .

<div align="center">MODEL FAMILY:</div>

CONDITIONED ON $\Theta = \theta$, THE SYMBOLS IN $\mathbf{x}$ ARE INDEPENDENT.

Or more completely: the symbols in $\mathbf{x}$ are outcomes of independent Bernoulli random variables with the parameter θ. Hence a model in the family is given by the probability assignment

$$P\left(\mathbf{x} \mid \Theta = \theta\right) = \prod_{l=1}^{n} \theta^{x_{i_l}} \cdot (1 - \theta)^{1 - x_{i_l}} =$$

$$\theta^{\sum_{l=1}^{n} x_{i_l}} \cdot (1 - \theta)^{n - \sum_{l=1}^{n} x_{i_l}} = \theta^k \cdot (1 - \theta)^{n-k} ,$$

if $\sum_{l=1}^{n} x_{i_l} = k$.

One *problem of inference* is to find the model (within a preestablished family) that is best in some sense given some observed data. In the thumbtack example we understand this as follows. We have observed n outcomes of flips of a thumbtack $\mathbf{x}$ and wish to determine which of the models in the family that best describes this set of flips.

2.1.2 The Posterior Density

To progress with this we express our uncertainty about Θ using a probability density function $f_\Theta(\theta)$, which is called the *prior*. Formally this means

$$f_\Theta(\theta) \geq 0, 0 \leq \theta \leq 1$$

and $f_\Theta(\theta) = 0$ elsewhere, and

$$\int_0^1 f_\Theta(\theta)\, d\theta = 1.$$

Also $P(a < \Theta \le b) = \int_a^b f_\Theta(\theta)\, d\theta$.

By an extension of Bayes' rule to continuous random variables we get the *posterior*

$$f_{\Theta|\mathbf{x}}(\theta \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid \Theta = \theta) \cdot f_\Theta(\theta)}{\int_0^1 P(\mathbf{x} \mid \Theta = \theta) \cdot f_\Theta(\theta)\, d\theta}, 0 \le \theta \le 1 \qquad (2.1)$$

and zero elsewhere. Due to the standardization $f_{\Theta|\mathbf{x}}(\theta \mid \mathbf{x})$ is another probability density for Θ.

The posterior $f_{\Theta|\mathbf{x}}(\theta \mid \mathbf{x})$ expresses our updated belief in the statement that θ is the true chance of obtaining heads given that we have observed $\mathbf{x}$.

One way to get further from here is to use an explicit form for $f_\Theta(\theta)$. There could be several choices, but some are at least analytically more advantageous. Let us consider the *uniform prior* given by

$$f_\Theta(\theta) = \begin{cases} 1 & 0 \le \theta \le 1 \\ 0 & \text{elsewhere.} \end{cases}$$

The uniform prior is often interpreted as a representation of complete ignorance (v. Mises and Geiringer 1964).

By an insertion we can calculate

$$\int_0^1 P(\mathbf{x} \mid \Theta = \theta) \cdot f_\Theta(\theta)\, d\theta = \int_0^1 \theta^k \cdot (1-\theta)^{n-k}\, d\theta = \frac{k!(n-k)!}{(n+1)!}$$

by the *Beta integral* found in many handbooks on integral calculus or as a special case of the Dirichlet integral, see (A.10), recapitulated in the Appendix. Then we have

$$f_{\Theta|\mathbf{x}}(\theta \mid \mathbf{x}) = \begin{cases} \frac{(n+1)!}{k!(n-k)!} \cdot \theta^k (1-\theta)^{n-k} & 0 \le \theta \le 1 \\ 0 & \text{elsewhere.} \end{cases} \qquad (2.2)$$

This is a *Beta density*, see appendix 5.3.

2.1.3 The Maximum Likelihood Estimate

To understand better the alluded properties of $f_{\Theta|\mathbf{x}}(\theta \mid \mathbf{x})$ we introduce the *maximum likelihood estimate* $\widehat{\theta}_{ML}$ *of* θ defined by

$$\widehat{\theta}_{ML} = \text{argmax}_{0 \leq \theta \leq 1} P(\mathbf{x} \mid \Theta = \theta) = \text{argmax}_{0 \leq \theta \leq 1} \theta^k \cdot (1-\theta)^{n-k}.$$

The rationale for this is that we *try to find the model within the family that gives the (training) sequence* $\mathbf{x}$ *the highest possible probability*. The probability $P(\mathbf{x} \mid \Theta = \theta)$ regarded as a function of θ is known as the *likelihood function*

$$L(\theta) = P(\mathbf{x} \mid \Theta = \theta).$$

The likelihood function $L(\theta)$ thus compares the plausibilities of different models for given $\mathbf{x}$.

A straightforward maximization of the likelihood function gives

$$\widehat{\theta}_{ML} = \frac{k}{n}. \tag{2.3}$$

By a Taylor expansion of $\log P(\mathbf{x} \mid \Theta = \theta)$ around $\widehat{\theta}_{ML}$ and by using the definition of the Fisher information $J(\theta)$, or for any probability distribution $f(x; \theta)$ with L values

$$J(\theta) = \sum_{i=1}^{L} f(x_i; \theta) \left(\frac{\partial}{\partial \theta} \log f(x_i; \theta) \right)^2, \tag{2.4}$$

we can show

$$f_{\Theta|\mathbf{x}}(\theta \mid \mathbf{x}) \approx e^{-\frac{1}{2} n J(\widehat{\theta}_{ML}) \cdot (\theta - \widehat{\theta}_{ML})^2} = e^{-\frac{1}{2} \frac{n}{\widehat{\theta}_{ML} \cdot (1-\widehat{\theta}_{ML})} \cdot (\theta - \widehat{\theta}_{ML})^2}. \tag{2.5}$$

We can empirically, say for $\mathbf{x}$ drawn from a pseudo random number generator, plot the posterior density (2.2) as a function of θ and observe the property (2.5), when the length of a string $\mathbf{x}$ increases. In fact this holds independently of the prior density. This behaviour is clearly present in a typical simulation.

2.1.4 The Personal Probability for the Outcome of the Next Toss

In the thumbtack model we may be concerned with

$$P(X_{n+1} = \text{head}|\mathbf{x}),$$

if X_{n+1} is a random variable modeling the next toss, given n flips of the thumbtack as recorded in $\mathbf{x}$. The evaluation of this probability is left as an exercise. The computation of $P\left(X_{n+1} = \text{head}|\mathbf{x}\right)$ has also been discussed from the pedagogical point of view as a potential item in high school curriculum of statistics in (Lindley 1970). Lindley sees here a natural transition from rules of probability to the rules of inference that should be easier at a high school level than the frequentist approach. Personal probability for expert systems is discussed in (Lindley 1987).

2.2 More on Modeling and Learning

2.2.1 The Model Family

Let $X_1, X_2, \ldots, X_n$ be independent random variables assuming values in

$$\mathcal{X} = \{x_1, \cdots, x_L\}$$

with the common distribution

$$\theta_l = P\left(X_i = x_l\right), l = 1, 2, \ldots, L.$$

Hence $\theta_1 + \theta_2 + \ldots + \theta_L = 1$. Let $\mathbf{x} = x_{i_1} x_{i_2} \ldots x_{i_n}$ be a string of symbols from $\mathcal{X}$ and let for $l = 1, 2, \ldots, L$

$$n_l = \text{ the number of times the symbol } x_l \text{ is found in } x_{i_1} x_{i_2} \ldots x_{i_n}.$$

We set

$$\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_L)$$

and consider $\underline{\Theta}$ as a random variable (element) that assumes values in the simplex

$$S_L = \{\underline{\theta} \mid \theta_1 + \theta_2 + \ldots + \theta_L = 1, \theta_l \geq 0, l = 1, \ldots, L\}.$$

THE MODEL FAMILY:

CONDITIONED ON $\underline{\Theta} = \underline{\theta}$, THE SYMBOLS IN $\mathbf{x}$ ARE INDEPENDENT.

Thus, as shown before,

$$P\left(\mathbf{x} \mid \underline{\theta}\right) = \theta_{i_1} \cdot \theta_{i_2} \cdots \theta_{i_n} = \theta_1^{n_1} \cdot \theta_2^{n_2} \cdots \theta_L^{n_L}.$$

Again we find a prior $\phi_{\underline{\Theta}}(\underline{\theta})$ for Θ. Let us consider the *Dirichlet prior* given by

$$\phi_{\underline{\Theta}}(\underline{\theta}) = \begin{cases} \frac{\Gamma(\alpha)}{\Gamma(\prod_{j=1}^{L} \alpha q_j)} \prod_{j=1}^{L} \theta_j^{\alpha q_j - 1} & \underline{\theta} \in S_L \\ 0 & \text{elsewhere,} \end{cases}$$

where *the hyperparameters* are $\alpha > 0$, $q_j \geq 0$, $\sum_{j=1}^{L} q_j = 1$, $\Gamma(z)$ is Euler's gamma function as given in the appendix. The prior $\phi_{\underline{\Theta}}$ is in (A.7) in the appendix given the symbol

$$Dir(\alpha q_1, \ldots, \alpha q_L).$$

By extension of Bayes' rule we get the *posterior*

$$\phi_{\underline{\Theta}|\mathbf{x}}(\underline{\theta}|\mathbf{x}; \underline{\alpha}) = \frac{P(\mathbf{x} \mid \underline{\Theta} = \underline{\theta}) \cdot \phi_{\underline{\Theta}}(\underline{\theta})}{\int_{S_L} P(\mathbf{x} \mid \underline{\Theta} = \underline{\theta}) \cdot \phi_{\underline{\Theta}}(\underline{\theta}) \, d\underline{\theta}}, \underline{\theta} \in S_L \qquad (2.6)$$

and zero elsewhere. Using the Dirichlet integral expounded in the appendix we get

Proposition 2.1 *The posterior density* $\phi_{\underline{\Theta}|\mathbf{x}}(\underline{\theta}|\mathbf{x}; \underline{\alpha})$ *is a Dirichlet density*

$$Dir(n_1 + \alpha q_1, \ldots, n_L + \alpha q_L)$$

or

$$\phi_{\underline{\Theta}|\mathbf{x}}(\underline{\theta}|\mathbf{x}; \underline{\alpha}) = \frac{\Gamma(n + \alpha)}{\prod_{i=1}^{L} \Gamma(\alpha q_i + n_i)} \prod_{i=1}^{L} \theta_i^{n_i + \alpha q_i - 1}. \qquad (2.7)$$

∎

This property says that the posterior density is in the same family of densities as the prior. Hence the prior is called *closed under sampling* or a *conjugate prior*.

2.2.2 Mean Posterior Estimate

One useful property of the Dirichlet density is that we can compute explicitly the expectation of any θ_i with respect to the posterior density. In fact this expectation is by (A.9) and (2.7)

$$\widehat{\theta}_i = \int_{S_L} \theta_i \phi(\theta_1, \ldots, \theta_L | \mathbf{x}; \underline{\alpha}) \, d\theta_1 \ldots d\theta_L = \frac{n_i + \alpha q_i}{n + \alpha}. \qquad (2.8)$$

This result can be seen as a *regularization* adding pseudocounts αq_i to the vector of observed counts $\underline{n}$ and then normalising so that $\sum_{i=1}^{L} \widehat{\theta}_i = 1$. If we have $n = 0$, the estimate is simply q_i.

The probabilty in (2.8) is known as a *rule of succession*. Wilson (1927) suggests for the thumbtack model a different rule of ssuccession

$$\widehat{\theta}^W = \frac{k + \alpha^2/2}{n + \alpha^2},\qquad(2.9)$$

and says that the value of α depends on 'our readiness to to gamble on the typicalness of our experience'.

2.2.3 Maximum Likelihood

The maximum likelihood estimate of $\underline{\theta}$ (a finite table of probabilities) is by a familiar principle given by

$$\widehat{\underline{\theta}}_{ML} = \text{argmax}_{\underline{\theta} \in S_L} P\left(\mathbf{x} \mid \underline{\theta}\right) = \text{argmax}_{\underline{\theta} \in S_L} \theta_1^{n_1} \cdot \theta_2^{n_2} \cdots \theta_L^{n_L}.$$

The presence of S_L imposes a constrained problem of maximization. We take the natural logarithm of $P\left(\mathbf{x} \mid \underline{\theta}\right)$, which gives us the *loglikelihood function*

$$l\left(\theta_1, \theta_2, \ldots, \theta_L\right) = \log P\left(\mathbf{x} \mid \underline{\theta}\right)$$

We may equivalently seek the maximum of $l\left(\theta_1, \theta_2, \ldots, \theta_L\right)$. Since the constraint $\theta_1 + \theta_2 + \ldots + \theta_L = 1$ must be met, we consider the new auxiliary function in $L - 1$ free variables

$$\widetilde{l}\left(\theta_1, \theta_2, \ldots, \theta_{L-1}\right) = l\left(\theta_1, \theta_2, \ldots, 1 - \left(\theta_1 + \theta_2 + \ldots + \theta_{L-1}\right)\right).$$

This gives

$$\widetilde{l}\left(\theta_1, \theta_2, \ldots, \theta_{L-1}\right) = n_1 \cdot \log \theta_1 + n_2 \cdot \log \theta_2 + \ldots + n_L \cdot \log\left(1 - \left(\theta_1 + \theta_2 + \ldots + \theta_{L-1}\right)\right).$$

Vi differentiate partially $\widetilde{l}\left(\theta_1, \theta_2, \ldots, \theta_{L-1}\right)$ with respect to $\theta_1, \theta_2, \ldots, \theta_{L-1}$ and set the partial derivatives equal to zero. This gives us the system of equations

$$\frac{\partial}{\partial \theta_1}\widetilde{l}\left(\theta_1, \theta_2, \ldots, \theta_{L-1}\right) = \frac{n_1}{\theta_1} - \frac{n_L}{1 - \left(\theta_1 + \theta_2 + \ldots + \theta_{L-1}\right)} = 0,$$

$$\vdots$$

$$\frac{\partial}{\partial \theta_{L-1}} \widetilde{l}\left(\theta_1, \theta_2, \ldots, \theta_{L-1}\right) = \frac{n_{L-1}}{\theta_{L-1}} - \frac{n_L}{1 - \left(\theta_1 + \theta_2 + \ldots + \theta_{L-1}\right)} = 0.$$

This leads to the equalities

$$\frac{n_1}{\theta_1} = \frac{n_2}{\theta_2} = \ldots = \frac{n_L}{1 - \left(\theta_1 + \theta_2 + \ldots + \theta_{L-1}\right)}.$$

Let us denote the common value of these ratios as λ so that

$$\theta_1 = \frac{n_1}{\lambda}, \theta_2 = \frac{n_2}{\lambda}, \ldots, \theta_L = \frac{n_L}{\lambda}.$$

We determine λ from the constraint $\theta_1 + \theta_2 + \ldots + \theta_L = 1$, which gives

$$1 = \theta_1 + \theta_2 + \ldots + \theta_L = \frac{n_1}{\lambda} + \frac{n_2}{\lambda} + \ldots + \frac{n_L}{\lambda}$$

or

$$\lambda = n_1 + n_2 + \ldots + n_L = n.$$

Hence we have obtained the solution to $\nabla \widetilde{l}\left(\theta_1, \theta_2, \ldots, \theta_{L-1}\right) = 0$ written in a componentwise form as

$$\widehat{\theta}_i = \frac{n_i}{n}, i = 1, \ldots, L.$$

Strictly taken we have yet to prove that this yields a maximum. For this we could check the matrix of second order partial derivatives of $\widetilde{l}$, (Khuri 1993 p. 283). There is a more instructive way to prove that the estimate found above actually gives the maximum. In fact the proof of the next proposition shows that it is not even necessary to differentiate to prove that we have found the maximum likelihood estimate.

Proposition 2.2 *The maximum likelihood estimate $\widehat{\theta}_{ML}$ of $\underline{\theta}$ is*

$$\widehat{\underline{\theta}}_{ML} = \left(\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_L}{n}\right).$$

Proof: Clearly the candidate solution $\widehat{\underline{\theta}}_{ML}$ belongs to S_L and is thus admissible. Since $P\left(\mathbf{x} \mid \underline{\theta}\right) = \prod_{i=1}^{L} \theta_i^{n_i}$, the following identity is evident

$$H\left(\widehat{\underline{\theta}}_{ML}\right) = -\frac{1}{n} \log P\left(\mathbf{x} \mid \widehat{\underline{\theta}}_{ML}\right), \tag{2.10}$$

8

where

$$H\left(\widehat{\underline{\theta}}_{ML}\right) = -\sum_{i=1}^{L}\widehat{\theta}_i \log \widehat{\theta}_i \qquad (2.11)$$

is the (empirical) Shannon's entropy in nats.

Next we take an arbitrary $\underline{\theta}$ in S_L. Then we have in view of (2.11) for an arbitrary $\underline{\theta}$ in S_L another evident identity

$$P\left(\mathbf{x} \mid \underline{\theta}\right) = \prod_{i=1}^{L}\theta_i^{n_i} = e^{-n\left(D\left(\widehat{\underline{\theta}}_{ML} \mid \underline{\theta}\right) + H\left(\widehat{\underline{\theta}}_{ML}\right)\right)}. \qquad (2.12)$$

Here we have used the Kullback distance between two discrete probability distributions defined as

$$D\left(f \mid g\right) = \sum_{x \in \mathcal{X}} f(x) \log \frac{f(x)}{g(x)}.$$

Thus from (2.10) and (2.12)

$$\frac{P\left(\mathbf{x} \mid \widehat{\underline{\theta}}_{ML}\right)}{P\left(\mathbf{x} \mid \underline{\theta}\right)} = e^{-nH\left(\widehat{\underline{\theta}}_{ML}\right)} \cdot e^{n\left(D\left(\widehat{\underline{\theta}}_{ML} \mid \underline{\theta}\right) + nH\left(\widehat{\underline{\theta}}_{ML}\right)\right)} =$$

$$= e^{nD\left(\widehat{\underline{\theta}}_{ML} \mid \underline{\theta}\right)} \geq 1,$$

where the last inequality follows due to the fact that $D\left(\widehat{\underline{\theta}}_{ML} \mid \underline{\theta}\right)$ is the Kullback distance, which is known to be nonnegative. Equality holds if and only if $\widehat{\underline{\theta}}_{ML} = \underline{\theta}$. Thus

$$P\left(\mathbf{x} \mid \widehat{\underline{\theta}}_{ML}\right) \geq P\left(\mathbf{x} \mid \underline{\theta}\right)$$

for every $\underline{\theta}$ in S_L and the assertion is proved. ∎

Here we may note that another way of referring to α in (2.8) is to talk about the *flattening constant* (Bender 1996, pp. 554 - 555). The flattening constant determines a linear interpolation between the maximum likelihood estimate $\frac{n_i}{n}$ and the prior estimate q_i. Hence α has the interpretation as the degree of confidence we distribute between the data and the prior.

9

2.3 General Summary

A formal Bayesian modeling articulates the information in a (training) sequence with evidence other than that of the (training) sequence. It is thought that there is always such evidence or that there is no such thing as the 'right analysis' if there is none. The evidence is assessed by judgement and is expressed in probability theory terms:

(1) a probability distribution specifies the probability of any sequence conditional on certain parameters;

(2) a prior expresses uncertainty about the parameters.

When (1) is combined with the training sequence we get the *likelihood function* of the sequence. The likelihood function is combined with (2) via Bayes' rule to produce a *posterior distribution* for the parameters of the model and this is the output of the formal Bayesian analysis.

3 Learning of Bayesian Networks from Complete Data

3.1 Notations

Let $\mathcal{G} = (V, E)$ be a directed acyclic graph with the set of nodes $V = \{1, \ldots, d\}$ and the edges E. Each edge (j, i) in E is a statement telling that X_j is influencing or is a direct cause of X_i. The absence of an edge indicates lack of direct influence. For each node there is a discrete random variable X_j the instantiations of which are designated by

$$x_j^i \in \mathcal{X}_j = \{x_j^1, \ldots, x_j^{k_j}\}.$$

Since the graph is acyclic and directed, the *structure* of the graph $\mathcal{G}$ is determined by the parent sets

$$(\Pi[1], \Pi[2], \ldots, \Pi[d]),$$

where $\Pi[j]$ is the set of parents of the node j. This is the same notation as $\mathrm{pa}(j)$.

Let π_j^l denote a *parent configuration*. The parent configuration is the set of instantiations assumed by the variables in the parent nodes so that

$$\pi_j^l \in \mathcal{X}_{\Pi[j]} = \times_{t \in \Pi[j]} \mathcal{X}_t,$$

and $\mathcal{X}_{\Pi[j]}$ is the notation for the set of all possible parent configurations at node j. The number of possible parent configurations at node j is

$$q_j = \prod_{l \in \Pi[j]} k_l,$$

where k_l is the number of symbols in $\mathcal{X}_l$.

The joint distribution of $(X_1, \ldots, X_d)$ is recursively factorized along $\mathcal{G}$ in the sense that

$$P\left(X_1 = x_1^{i_1}, \ldots, X_d = x_d^{i_d}\right) = \prod_{j=1}^{d} P\left(X_j = x_j^{i_j} | \pi_j^{l_j}\right), \tag{3.1}$$

We write now the table of conditional probabilities $P\left(X_j = x_j^i | \pi_j^l\right)$ using a special system of notations.

We set for $x_j^i \in \mathcal{X}_j$, $i = 1, \ldots, k_j$, $\pi_j^l \in \mathcal{X}_{\Pi[j]}$, $l = 1, \ldots, q_j$ and $j = 1, \ldots, d$

$$\theta_{jil} = P\left(X_j = x_j^i | \pi_j^l\right), \tag{3.2}$$

and

$$\theta^{j,l} = (\theta_{jil}; i = 1, \ldots, k_j), \tag{3.3}$$

and

$$\Theta = \left(\theta^{j,l}; j = 1, \ldots, d, l = 1, \ldots, q_j\right)$$

denotes the overall parameter consisting of the local child-parent parameters.

For convenience of expression we shall in the sequel write the joint distributions

$$P\left(X_1 = x_1^{i_1}, \ldots, X_d = x_d^{i_d}\right)$$

and other distributions by omitting notationally the random variables but including the graph and parameters by setting

$$\mathbf{x} = \left(x_1^{i_1}, \ldots, x_d^{i_d}\right),$$

and

$$P_\Theta\left(\mathbf{x} \mid \mathcal{G}\right) = P\left(X_1 = x_1^{i_1}, \ldots, X_d = x_d^{i_d}\right). \tag{3.4}$$

3.2 Parameter estimation

Suppose we are given a sample of cases $\mathbf{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ and we wish to find a Bayesian network on basis of this data. Note that each sample is multivariate,

$$\mathbf{x}^{(k)} = \left(x_{1,k}^{i_1}, \ldots, x_{d,k}^{i_d}\right)$$

corresponding to a configuration of states of the d variables in the graph. We assume that **no value** $x_{j,k}^{i_1}$ **is missing**, i.e., that **the data is complete**.

There are two aspects of a BN that in general need be learned:

(1) the structure of $\mathcal{G}$;

(2) the parameters Θ, given the graph structure $\mathcal{G}$.

We shall first discuss the latter problem, that of learning of the parameters given the structure of the graph.

Let now $n_k(x_j^i|\pi_j^l)$ be equal to one (1) if we see the case (x_j^i, π_j^l) (*a family configuration*) in the sample $\mathbf{x}_k$ and zero (0) otherwise. The joint probability of a case $\mathbf{x}^{(k)}$ can now be written invoking (3.1), (3.2) and (3.4) as

$$P_\Theta\left(\mathbf{x}^{(k)} \mid \mathcal{G}\right) = \prod_{j=1}^{d}\prod_{l=1}^{q_j}\prod_{i=1}^{k_j} [\theta_{jil}]^{n_k(x_j^i|\pi_j^l)} \tag{3.5}$$

If the cases in $\mathbf{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ are modeled as independent outcomes of the graph (variables), then we have

$$\prod_{k=1}^{n} P_\Theta\left(\mathbf{x}^{(k)} \mid \mathcal{G}\right) = \prod_{j=1}^{d}\prod_{l=1}^{q_j}\prod_{i=1}^{k_j}\prod_{k=1}^{n} [\theta_{jil}]^{n_k(x_j^i|\pi_j^l)}$$

$$= \prod_{j=1}^{d}\prod_{l=1}^{q_j}\prod_{i=1}^{k_j} [\theta_{jil}]^{\sum_{k=1}^{n} n_k(x_j^i|\pi_j^l)} \tag{3.6}$$

Let us set

$$n(x_j^i|\pi_j^l) = \sum_{k=1}^{n} n_k(x_j^i|\pi_j^l),$$

which is the number of times we see the family configuration (x_j^i, π_j^l) in $\mathbf{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$.

3.2.1 Maximum Likelihood

Thus we have

$$L(\Theta) = \prod_{k=1}^{n} P_\Theta\left(\mathbf{x}_k \mid \mathcal{G}\right) = \prod_{j=1}^{d}\prod_{l=1}^{q_j}\prod_{i=1}^{k_j} [\theta_{jil}]^{n(x_j^i|\pi_j^l)}. \tag{3.7}$$

For any node j and any parent configuration π_j^l there is the constraint

$$\sum_{i=1}^{k_j} \theta_{jil} = 1. \tag{3.8}$$

Thus, the likelihood $L(\Theta)$ in (3.7) is seen to factorize into local parent-child factors and additionally to $d \times q_j$ separate maximum likelihood estimations all of the basic form treated in proposition 2.2. Hence we have that

$$\widehat{\theta}_{\mathrm{ML}}^{j,l} = \left(\frac{n(x_j^i|\pi_j^l)}{n(\pi_j^l)}; i = 1, \ldots, k_j\right),$$

where

$$n(\pi_j^l) = \sum_{i=1}^{k_j} n_k(x_j^i|\pi_j^l) \tag{3.9}$$

is the frequency of the parent configuration π_j^l in $\mathbf{X}$. The maximum likelihood estimate of θ_{jil} is thus

$$\widehat{\theta}_{jil} = \frac{\text{frequency of the family configuration}}{\text{frequency of the parent configuration}}.$$

3.2.2 Bayesian Learning

The factorization properties of $\prod_{k=1}^{n} P_\Theta\left(\mathbf{x}_k \mid \mathcal{G}\right)$ can now be taken advantage in Bayesian learning. We assume that the parameters $\theta^{j,l}$ in (3.3) are independent random variables and have the Dirichlet density

$$Dir\left(\alpha_{1,j,l}, \ldots, \alpha_{k_j,j,l}\right).$$

Then we obtain from proposition 2.1 that the posterior density $\phi_{\underline{\Theta}|\mathbf{X}}\left(\theta^{j,l}|\mathbf{x}; \underline{\alpha}_{j,l}\right)$ is the Dirichlet density

$$Dir\left(n(x_j^1|\pi_j^l) + \alpha_{1,j,l}, \ldots, n(x_j^{k_j}|\pi_j^l) + \alpha_{k_j,j,l}\right).$$

In addition the mean posterior estimate must in view of (2.8) be given by

$$\widehat{\theta^{j,l}} = \int_{S_{j,l}} \theta_{jil}\phi_{\underline{\Theta}|\mathbf{x}}\left(\theta^{j,l}|\mathbf{x};\underline{\alpha}_{j,l}\right)d\theta^{j,l} = \frac{n(x_j^i|\pi_j^l) + \alpha_{i,j,l}}{n(\pi_j^l) + \sum_{i=1}^{k_j}\alpha_{i,j,l}}, \qquad (3.10)$$

where $S_{j,l}$ is the manifold of values of $\theta^{j,l}$ given in (3.3).

4 Cooper-Herskovitz Likelihood for the Graph Structure

Suppose now that the structure of a Bayesian network is to learned from data. Assume that the samples are complete, that the indiviudal cases are independent conditioned on Θ and that the prior distributions are $Dir\left(\alpha_{1,j,l}, \ldots, \alpha_{k_j,j,l}\right)$ for all nodes and parent configurations. Then we define

$$p\left(\mathbf{X}\mid\mathcal{G}\right) = \int_{\times_{j,l}S_{j,l}} \prod_{k=1}^{n} P_{\Theta}\left(\mathbf{x}_k\mid\mathcal{G}\right) \prod_{j=1}^{d}\prod_{l=1}^{q_j} d\phi\left(\theta_{j1l},\ldots,\theta_{jk_jl},\underline{\alpha}_{j,l}\right), \qquad (4.1)$$

where $\phi\left(\theta_{j1l},\ldots,\theta_{jdl},\underline{\alpha}_{j,l}\right)$ is another more complete way of writing the Dirichlet density $Dir\left(\alpha_{1,j,l},\ldots,\alpha_{k_j,j,l}\right)$.

Then it follows by a straightforward computation using the properties of the Dirichlet integral and the factorization in (3.7) that

$$p\left(\mathbf{X}\mid\mathcal{G}\right) = \prod_{j=1}^{d}\prod_{l=1}^{q_j} \frac{\Gamma\left(\sum_{i=1}^{k_j}\alpha_{i,j,l}\right)}{\Gamma\left(n(\pi_j^l) + \sum_{i=1}^{k_j}\alpha_{i,j,l}\right)} \prod_{i=1}^{k_j} \frac{\Gamma\left(n(x_j^i|\pi_j^l) + \alpha_{i,j,l}\right)}{\Gamma\left(\alpha_{i,j,l}\right)}, \qquad (4.2)$$

where $n(\pi_j^l)$ is given by (3.9). This is the *Cooper-Herskovitz likelihood for the graph structure* (Cooper and Herskovitz 1992). There is only a finite number of different DAG:s $\mathcal{G}_r$ with d nodes. Some expertise on domain knowledge, may, of course, reduce the number of graph structures to be considered in a special situation.

The Bayesian rule of selection of the structure of the Bayesian network using the cases in $\mathbf{X}$ is to use the graph that maximizes the posterior probability $p\left(\mathcal{G}_r\mid\mathbf{X}\right)$ or

$$\mathrm{argmax}_r p\left(\mathcal{G}_r\mid\mathbf{X}\right) = \frac{p\left(\mathbf{X}\mid\mathcal{G}_r\right)p\left(\mathcal{G}_r\right)}{p\left(\mathbf{X}\right)},$$

where $p\left(\mathbf{X}\right) = \sum_{r=1}^{N} p\left(\mathbf{X} \mid \mathcal{G}_r\right) p\left(\mathcal{G}_r\right)$ and $p\left(\mathcal{G}_r\right)$, $r = 1, \ldots, N$ is a prior distribution on the graphs. The task is known to be NP-hard, see (Chickering 1996). There do not seem to be many well known suggestions on the prior distribution (?), there are some references and points of view in (Buntine 1996).

5 Appendix: Some Formulas for Dirichlet Densities

5.1 Euler's gamma function

The *gamma* function $\Gamma(z)$ is defined for complex numbers z, whose real part is positive, by the definite integral

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx. \tag{A.1}$$

A special case, obtained by the substitution $x = u^2/2$ is

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}.$$

The recursion formula is

$$\Gamma(z) = (z-1)\Gamma(z-1). \tag{A.2}$$

Hence, if $z = n$, where n is a positive integer, we have the factorial

$$\Gamma(n) = (n-1)!. \tag{A.3}$$

5.2 The Dirichlet density

Let $S_L \subset R^k$ be the *simplex*

$$S_L = \left\{ (\theta_1, \ldots, \theta_L) \,|\, \theta_i \geq 0, i = 1, \ldots, L, \sum_{i=1}^L \theta_i = 1 \right\}. \tag{A.4}$$

Let for $\alpha_i > 0$

$$\phi(\theta_1, \ldots, \theta_L) = \begin{cases} \frac{\prod_{i=1}^L \theta_i^{\alpha_i - 1}}{Z}, & \text{if } \theta_1, \ldots, \theta_L \in S_L \\ 0 & \text{otherwise.} \end{cases} \tag{A.5}$$

Here

$$\frac{1}{Z} = \frac{\Gamma\left(\sum_{i=1}^L \alpha_i\right)}{\prod_{i=1}^L \Gamma(\alpha_i)}. \tag{A.6}$$

The density $\phi(\theta_1, \ldots, \theta_L)$ is called a *Dirichlet density*. We designate it symbolically by

$$Dir(\alpha_1, \ldots, \alpha_L). \tag{A.7}$$

If $\alpha_1 = \alpha_2 = \ldots = \alpha_L = \kappa$, then we talk about a *symmetric Dirichlet density*. For the proof of the fact that

$$\int_{S_L} \phi(\theta_1, \ldots, \theta_L) d\theta_1 \ldots d\theta_L = 1 \tag{A.8}$$

we refer to (Wilks 1962). This means also that

$$\int_{S_L} \prod_{i=1}^{L} \theta_i^{\alpha_i - 1} d\theta_1 \ldots d\theta_L = \frac{\prod_{i=1}^{L} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{L} \alpha_i\right)}. \tag{A.9}$$

(Gupta and Richards 1987) is a concise compendium of knowledge about the Dirichlet distributions.

5.3   Beta density

As a special case for $L = 2$ we obtain in (A.9) the *Beta integral*

$$\int_0^1 \theta^{\alpha_1 - 1}(1 - \theta)^{\alpha_2 - 1} d\theta = \frac{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)}. \tag{A.10}$$

Thus

$$f(\theta) = \begin{cases} \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2)} \theta^{\alpha_1 - 1} (1 - \theta)^{\alpha_2 - 1} & 0 \leq \theta \leq 1 \\ 0 & \text{elsewhere.} \end{cases} \tag{A.11}$$

is a probability density called the *Beta density* and denoted by

$$\mathcal{B}e(\theta; \alpha_1, \alpha_2).$$

Note the difference in the heuristic notation between Beta and Bernoulli $Be(p)$. If $\theta = (\theta_1, \ldots, \theta_L)$ is a random variable that assumes values in S_L in (A.4) and has the symmetric $Dir(\alpha, \ldots, \alpha)$ distribution, then the marginal density of any θ_i is given by

$$\theta_i \in \mathcal{B}e(\theta; \alpha, (L-1)\alpha). \tag{A.12}$$

17

6   References and further reading:

1 Journal articles and technical reports on Bayesian learning/machine learning and Dirichlet distribution:

- W. Buntine (1996): A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering*, 8, pp. 195−210.

- P. Cheeseman (1988): An Inquiry into Computer Understanding. *Computational Intelligence*, **4**, 58−66.

- D.M. Chickering (1996): Learning Bayesian Networks is NP-Complete. *Learning from Data. Artificial Intelligence and Statistics* V. D. Fisher and H-J. Lenz (editors), Springer-Verlag, New York, Berlin, Heidelberg, pp. 121−130.

- G. Cooper and E. Hershkovitz (1992): A Bayesian method for the induction of of probabilistic networks from data. *Machine Learning*, 9, pp. 309−347.

- J.M. Dickey (1983): Multiple Hypergeometric Functions: Probabilistic Interpretation and Statistical Uses. *Journal of the American Statistical Association*, 78, pp. 628−637.

- R.D. Gupta and D.St.P. Richards (1987): Multivariate Liouville Distributions. *Journal of Multivariate Analysis*, 23, pp. 232−256.

- D. Heckerman (1996): A Tutorial on Learning with Bayesian Networks. *Microsoft Research. Technical Report*, MSR-TR-95-06, Redmond, Washington.

- D. Heckerman (1997): Bayesian Networks for Data Mining. *Data Mining*, pp. 81−119.

- P. Kontkanen, P. Myllymäki, T. Silander, H. Tirri and P. Grünewald (2000): On predictive distributions and Bayesian networks. *Statistics and Computing*, 10, pp. 39−54.

- D.V. Lindley (1987): The Probability Approach to the Treatment of Uncertainty in Artificial Intelligence and Expert Systems. *Statistical Science*, 2, pp. 17−24.

- D.V. Lindley (1970): A non-frequentist view of probability and statistics. *The Teaching of Probability & Statistics*, L. Råde editor, Almqvist & Wicksell, Uppsala, pp. 209−222.

- J. Rissanen (1997): Stochastic complexity and learning. *Journal of Computer and System Sciences* , 55, pp. 89−95.

- H.V. Roberts (1965): Probabilistic Prediction. *Journal of the American Statistical Association*, 60, pp. 50−62.

- E.B. Wilson (1927): Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22, pp. 209−212.

2 Books:

- E.A. Bender (1996): *Mathematical Methods in Artificial Intelligence.* IEEE Computer Society Press, Los Alamitos, California.

- J.M. Bernardo and A.F.M. Smith (1994): *Bayesian Theory.* John Wiley and Sons, Chichester, New York, Brisbane, Toronto and Singapore.

- A.I. Khuri (1993): *Advanced Calculus with Applications in Statistics.* John Wiley and Sons, Inc. New York.

- R. v. Mises with H. Geiringer (1964): *Mathematical Theory of Probability and Statistics.* Academic Press, New York and London.

- F.V. Jensen (2001): *Bayesian Networks and Decision Graphs.* Springer Verlag.

- S.S. Wilks (1962): *Mathematical Statistics*, John Wiley and Sons, New York and London.