**Overview of Bayesian Network**

Loc Nguyen

International Engineering and Technology Institute (IETI), Hong Kong

Anum Shafiq

Preston University, Islamabad, Pakistan

**Abstract**

|  |  |  |
| --- | --- | --- |
|  |  | (9.99) |

**1. Introduction**

This introduction section starts with a little bit discussion of Bayesian inference which is the base of both Bayesian network and inference in Bayesian network described later.

**Bayesian inference** [1], a form of statistical method, is responsible for collecting evidences to change the current belief in given hypothesis. The more evidences are observed, the higher degree of belief in hypothesis is. First, this belief was assigned by an initial probability or prior probability. Note, in classical statistical theory, the random variable’s probability is objective (physical) through trials. But, in Bayesian method, the probability of hypothesis is “personal” because its initial value is set subjectively by expert. When evidences were gathered enough, the hypothesis is considered trustworthy.

Bayesian inference is based on so-called Bayes’ rule or Bayes’ theorem [1] specified in equation 1.1 as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (1.1) |

Where,

* *H* is probability variable denoting a hypothesis existing before evidence.
* *D* is also probabilistic variable denoting an observed evidence. It is conventional that notations *d*, *D* and are used to denote evidence, evidences, evidence sample, data sample, sample, training data and corpus (another term for data sample). Data sample or evidence sample is defined as a set of data or a set of observations which is collected by an individual, a group of person, a computer software or a business process, which focuses on a particular analysis purpose [2]. The term “data sample” is derived from statistics; please read the book “Applied Statistics and Probability for Engineers” by authors Montgomery and Runger [3, p. 4] for more details about sample and statistics.
* *P*(*H*) is *prior probability* of hypothesis *H*. It reflects the degree of subjective belief in hypothesis *H*.
* *P*(*H|D*), conditional probability of *H* with given *D*, is called *posterior probability*. It tells us the changed belief in hypothesis when occurring evidence. Whether or not the hypothesis in Bayesian inference is considered trustworthy is determined based on the posterior probability. In general, posterior probability is cornerstone of Bayesian inference.
* *P*(*D|H*) is conditional probability of occurring evidence *D* when hypothesis *H* was given. In fact, likelihood ratio is *P*(*D|H*)/ *P*(*D*) but *P*(*D*) is constant value. So we can consider *P*(*D*|*H*) as *likelihood function* of *H* with fixed *D*. Please pay attention to conditional probability because it is mentioned over the whole research.
* *P*(*D*) is probability of occurring evidence *D* together all mutually exclusive cases of hypothesis. If *H* and *D* are discrete, then , otherwise with *H* and *D* being continuous, *f* denoting probability density function [3, p. 99]. Because of being sum of products of prior probability and likelihood function, *P*(*D*) is called *marginal probability*.

Note: *H*, *D* must be random variables [3, p. 53] according to theory of probability and statistics and *P*(.) *denotes random probability*.

Beside Bayes’ rule, there are three other rules such as additional rule, multiplication rule and total probability rule which are relevant to conditional probability. Given two random events (or random variables) *X* and *Y*, the additional rule [3, p. 33] and multiplication rule [3, p. 44] are expressed in equations 1.2 and 1.3, respectively as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (1.2) |

|  |  |  |
| --- | --- | --- |
|  |  | (1.3) |

Where notations and denote union operator and intersection operator in set theory [4]. Your attention please, when *X* and *Y* are numerical variables, notations and also denote operators “*or*” and “*and*” in theory logic [5, pp. 1-12]. If *X* and *Y* are mutually independent (mutually exclusive) then, and are often denoted as *X*+*Y* and *XY*, respectively and so, we have:

The probability *P*(*XY*)=*P*(*X*,*Y*) is often known as joint probability.

Given a complete set of mutually exclusive events *X*1, *X*2,…, *Xn* such that

The total probability rule [3, p. 44] is specified in equation 1.4 as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (1.4) |

If *X* and *Y* are continuous variables, the total probability rule is re-written in integral form as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (1.5) |

Note, *P*(*Y|X*) and *P*(*X*) are continuous functions known as probability density functions mentioned right after.

Please pay attention to Bayes’ rule (equation 1.1) and total probability rule (equations 1.4 and 1.5) because they are used frequently over the whole research.

**Bayesian network (BN)** [6, p. 40] [7, p. 1] is combination of graph theory and Bayesian inference. It having a set of nodes and a set of directed arcs is the directed acyclic graph (DAG); please pay attention to the terms “DAG” and “BN” because they are used over the whole research. Each node represents a random variable which can be an evidence or hypothesis in Bayesian inference. Each arc reveals the relationship among two nodes. If there is the arc from node *A* to *B*, we call “*A* causes *B*” or “*A* is parent of *B*”, in other words, *A* depends conditionally on *B*. Otherwise there is no arc between *A* and *B*, it asserts the conditional independence. Note, in BN context, terms: *node and variable are the same*.

A node has a local Conditional Probability Distribution (CPD) with attention that conditional probability distribution is often called shortly *probability distribution* or *distribution*. If variables are discrete, CPD is simplified as Conditional Probability Table (CPT). If variables are continuous, CPD is often called conditional Probability Density Function (PDF) which will be mentioned in section 4 – how to learn CPT from beta density function. PDF can be called *density function*, in brief. CPD is the general term for both CPT and PDF; there is convention that CPD, CPT and PDF indicate both probability and conditional probability. In general, each CPD, CPT or PDF specifies a random variable and is known as the *probability distribution* or *distribution* of such random variable.

Another representation of CPD is cumulative distribution function (CDF) [3, p. 64] [3, p. 102] but CDF and PDF have the same meaning and they share interchangeable property when PDF is derivative of CDF; in other words, CDF is integral of PDF. In practical statistics, PDF is used more commonly than CDF is used and so, PDF is mentioned over the whole research. Note, notation *P*(.) often denotes probability and it can be used to denote PDF but we prefer to use lower case letters such as *f* and *g* to denote PDF. Given a variable having PDF *f*, we often state that “such variable has distribution *f* or such variable has density function *f*”. Let *F*(*X*) and *f*(*X*) be CDF and PDF, respectively, equation 1.6 is the definition of CDF and PDF.

|  |  |  |
| --- | --- | --- |
|  |  | (1.6) |

Because this introduction section focuses on BN, please read [3, pp. 98-103] for more details about CDF and PDF.

Now please pay attention to the concept CPT because it occurs very frequently in the research; you can understand simply that CPT is essentially collection of discrete conditional probabilities of each node (variable). It is easy to infer that CPT is discrete form of PDF. When one node is conditionally dependent on another, there is a corresponding probability (in CPT or CPD) measuring the influence of causal node on this node. In case that node has no parent, its CPT *degenerates into prior probabilities*. This is the reason CPT is often identified with probabilities and conditional probabilities.

E.g., in figure 1.1, event “cloudy” is cause of event “rain” which in turn is cause of “grass is wet” [8]. So we have three causal relationships of: 1-cloudy to rain, 2- rain to wet grass, 3- sprinkler to wet grass. This model is expressed below by BN with four nodes and three arcs corresponding to four events and three relationships. Every node has two possible values True (1) and False (0) together its CPT.



**Figure 1.1.** Bayesian network (a classic example about wet grass)

Note that random variables *C*, *S*, *R*, and *W* denote phenomena or events such as cloudy, sprinkler, rain, and wet grass, respectively and the table next to each node expresses the CPT of such node. For instance, focusing on the CPT attached to node “Wet grass”, if it is rainy (*R*=1) and garden is sprinkled (*S*=1), it is almost certain that grass is wet (*W*=1). Such assertion can be represented mathematically by the condition probability of event “grass is wet” (*W*=1) given evident events “rain” (*R*=1) and “sprinkler” (*S*=1) is 0.99 as in the attached table, *P*(*W*=1|*R*=1,*S*=1) = 0.99. As seen, the conditional probability *P*(*W*=1|*R*=1,*S*=1) is an entry of the CPT attached to node “Wet grass”. In general, BN consists of two models such as qualitative model and quantitative model. Qualitative model is the structure as the graph shown in figure 1.1. Quantitative model includes parameters which are CPT (s) attached nodes in BN. Thus, CPT (s) as well as conditional probabilities are known as parameters of BN. Parameter learning and structure learning will be mentioned in sections 4 and 5.

Beside important subjects of BN such as parameter learning and structure learning, there is a more essential subject which is inference mechanism inside BN when the inference mechanism is a very powerful mathematical tool that BN provides us. Before studying inference mechanism in this wet grass example, we should know some advanced concepts of Bayesian network.

Suppose we use two notations *Xi* and *PA*(*Xi*) to indicate a node and a set of its parent, respectively. Let *X* be vector which was constituted of all *Xi*, *X* = (*X*1, *X*2,…, *Xn*). The **G**lobal **J**oint **P**robability **D**istribution (GJPD) *P*(*X*) being product of all local CPD (s) or CPT (s) is formulated as:

|  |  |  |
| --- | --- | --- |
|  |  | (1.7) |

Suppose is the subset of *PA(Xi*) such that *Xi* must depend conditionally and directly on every variable in *.* In other words, there is always an arc from each variable in to *Xi* and no intermediate node between them. Thus, equation 1.7 becomes:

|  |  |  |
| --- | --- | --- |
|  |  | (1.8) |

Note that *P*(*Xi*|) in equation 1.8 is the CPT of *Xi*. According to Bayesian rule, given evidence (random variables) , the posterior probability *P*(*Xi*|) of variable *Xi* is computed in equation 1.9 as below:

|  |  |  |
| --- | --- | --- |
|  |  | (1.9) |

Where *P*(*Xi*) is prior probability of random variable *Xi* and *P*(|*Xi*) is conditional probability of occurring given *Xi* and *P*() is probability of occurring together all mutually exclusive cases of *X*. From equations 1.8 and 1.9, we gain equation 1.10 as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (1.10) |

Where and are all possible values *X* = (*X*1, *X*2,…, *Xn*) with fixing (excluding) and fixing (excluding) , respectively. Note that evidence including at least one random variable *Xi* is a subset of *X* and the sign “\” denotes the subtraction (excluding) in set theory [4]. Please pay attention that the equation 1.10 is the base for inference inside Bayesian network, which is used over the whole research. Equations 1.9 and 1.10 are extensions of Bayes’ rule specified by equation 1.1. It is not easy to understand equation 1.10 and so, please see equations 1.12 and 1.13 which are advanced posterior probabilities applied into wet grass example in order to comprehend equation 1.10.

From figure 1.1 of wet grass example and according to equation 1.7, we have:

Applying equation 1.8, *P*(*S*|*C*)=*P*(*S*) due to no conditional independence assertion about variables *S* and *C*. Furthermore, because *S* is intermediate node between *C* and *W*, we should remove *C* from *P*(*W | C*, *R*, *S*), hence *P*(*W* | *C*, *R*, *S*)= *P*(*W* | *R*, *S*). In short, applying equation 1.8, we have equation 1.11 for determining global joint probability distribution of “wet grass” Bayesian network as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (1.11) |

Using **Bayesian inference**, we need to compute the posterior probability of each hypothesis node in network. In general, the computation based on Bayesian rule is known as the inference in BN.

Reviewing figure 1.1, suppose *W* becomes evidence variable which is observed as the fact that the grass is wet, so, *W* has value 1. There is request for answering the question: how to determine which cause (sprinkler or rain) is more possible for wet grass. Hence, we will calculate two posterior probabilities of *R* (=1) and *S* (=1) in condition *W* (=1). Such probabilities called *explanations* for *W* are simple forms of equation 1.10, expended by equations 1.12 and 1.13 as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (1.12) |

|  |  |  |
| --- | --- | --- |
|  |  | (1.13) |

Note that the numerator in the right side of equation 1.12 is the sum of possible probabilities over possible values of *C* and *S*. Concretely, we have an interpretation for the numerator as follows:

Applying equation 1.11 for global joint probability distribution of “wet grass” Bayesian network, we have:

It is easy to infer that there is the same interpretation for numerators and denominators in right sides of equations 1.12 and 1.13 and the previous equation 1.10 is also understood simply by this way when {*C*, *S*} = {*C*, *R*, *S*, *W*}\{*R*, *W*} and fixing {*R*, *W*}. In similar, we have:

In fact, equations 1.12 and 1.13 are expansions of equation 1.10. As a result, we have:

Obviously, the posterior probability of event “sprinkler” (*S*=1) is larger than the posterior probability of event “rain” (*R*=1) given evidence “wet grass” (*W*=1), which leads to conclusion that sprinkler is the most likely cause of wet grass. Now a short description of Bayesian is introduced. Next section will concern advanced concepts of Bayesian network.

**2. Advanced concepts**

**3. Inference**

**4. Parameter learning**

As a convention, uppercase letter such as *X*, *Y*, and *Z* denotes random variable whereas lowercase letters such as *x*, *y*, and *z* denote instances or values of random variables. According to Bayesian approach, parameters such as mean *μ* and variance *σ*2 of normal distribution and probability *p* of binominal distribution are random variables. These random variables are commonly denoted Θ, which are hypotheses according to equation 1.1. The prior distribution (prior probability) is denoted *P*(Θ | *ξ*) where *ξ* denotes background knowledge about Θ. Note that *ξ* is often parameter of the prior distribution and so it can be called hyper-parameter of prior distribution. For example, if Θ follows beta distribution, its prior distribution is:

Where *ξ* = (*a*, *b*) are two parameters of such prior (beta) distribution. Note that Γ(.) is gamma function:

For another example, if Θ is mean *μ* and it follows normal distribution, its prior distribution is:

Where *ξ* = (*μ*0, *σ*02) are mean and variance of such prior (normal) distribution. Given sample *D* = {*X*1, *X*2,…, *Xn*) consisting *n* observations (evidences) *Xi*. Equation 4.1 specifies posterior distribution (posterior probability) of Θ, according to Bayes’ rule [9, p. 6] [1].

|  |  |  |
| --- | --- | --- |
|  |  | (4.1) |

Where *P*(*D* | Θ) is the likelihood function of Θ and *P*(*D*) is the marginal probability of sample. Note, the equation 4.1 is written fully as follows:

However *P*(*D* | Θ, *ξ*) = *P*(*D* | Θ) and *P*(*D* | *ξ*) = *P*(*D*) because *D* is only dependent on Θ. The marginal probability *P*(*D*) is expectation of the likelihood function *P*(*D* | Θ) given prior probability *P*(Θ | *ξ*).

Equation 4.1 which defines posterior probability of parameter Θ is used to assess hypothesis Θ after surveying sample *D*. This is a so-called *Bayesian inference*.

Suppose there is a requirement of predicting possibility of a new observation *Xn*+1 given previous sample *D* with note that *Xn*+1 is independent from *D*. In other words, we need to calculate the probability *P*(*Xn*+1 | *D*) called posterior predictive probability. Equation 4.2 specifies the posterior predictive probability [9, p. 6] [1].

|  |  |  |
| --- | --- | --- |
|  |  | (4.2) |

According to equation 4.2, the posterior predictive probability *P*(*Xn*+1 | *D*) is expectation of the probability *P*(*Xn*+1 | Θ) given posterior distribution *P*(Θ | *D*). Equation 4.2 establishes a so-called *Bayesian prediction*. The probability *P*(*Xn*+1 | Θ, *ξ*) is always determined because it is probability of observation.

Given a sample *D* = {*X*1, *X*2,…, *Xn*), there is a requirement of estimating parameter (hypothesis) Θ according to Bayesian inference. Let denote a Bayesian estimate of Θ. The squared-error loss function is defined according to equation 4.3 [10, p. 717]:

|  |  |  |
| --- | --- | --- |
|  |  | (4.3) |

The mean of posterior distribution *P*(Θ | *D*, *ξ*) is a Bayesian estimate of Θ under squared-error loss function, according to equation 4.4 [10, p. 717]. In other words, such mean minimizes the squared-error loss function.

|  |  |  |
| --- | --- | --- |
|  |  | (4.4) |

The absolute loss function is defined according to equation 4.5 [10, p. 718]:

|  |  |  |
| --- | --- | --- |
|  |  | (4.5) |

The median of posterior distribution *P*(Θ | *D*, *ξ*) is a Bayesian estimate of Θ under absolute loss function, according to equation 4.6 [10, p. 718]. In other words, such median minimizes the absolute loss function.

|  |  |  |
| --- | --- | --- |
|  |  | (4.6) |

Therefore, equations 4.4 and 4.6 are two popular equations for *Bayesian parameter estimation*. If the posterior distribution *P*(Θ | *D*, *ξ*) is symmetric, these two equations produces the same estimate .

Now we survey a common case of Bayesian inference in which *D* is binominal sample. At that time which every *Xi* is binary random variable and the likelihood function *P*(*D* | Θ) becomes [9, p. 6]:

Where *h* and *t* are the numbers of *Xi* = 1 and *Xi* = 0, respectively. The notation “*h*” and “*t*” are abbreviations of “head” and “tail” when tossing a coin. The *h* and *t* are sufficient statistics of binomial sampling. Hence, Θ = *P*(*Xi* = 1) is the probability of *Xi* = 1. Equation 4.7 [9, p. 6], which is a special case of equation 4.1, specifies Bayesian inference (posterior probability of Θ) in case of binominal sampling.

|  |  |  |
| --- | --- | --- |
|  |  | (4.7) |

Derived from equation 4.2, the posterior predictive probability *P*(*Xn*+1 = 1 | *D*) in case of binomial sampling becomes [9, p. 6]:

Due to:

Equation 4.8, which is a variant of equation 4.2, specifies the posterior predictive probability *P*(*Xn*+1 = 1 | *D*) in case of binomial sampling [9, p. 6].

|  |  |  |
| --- | --- | --- |
|  |  | (4.8) |

Where *E*(Θ | *D*, *ξ*) denotes the expectation of Θ given the posterior probability *P*(Θ | *D*, *ξ*) specified by equation 4.7.

Suppose Θ is distributed according to beta distribution, its prior probability is specified by equation 4.9.

|  |  |  |
| --- | --- | --- |
|  |  | (4.9) |

Where *ξ* = (*a*, *b*) are two parameters of such prior beta distribution. The theoretical mean of beta distribution is:

Note that Γ(.) is gamma function:

The marginal probability *P*(*D*) is calculated as follows:

Where *N* = *h* + *t*. The posterior probability of Θ is re-calculated as follows:

Therefore, in case of binomial sampling, if the prior probability of Θ conforms beta distribution beta(Θ | *a*, *b*) then, the posterior probability of Θ conforms beta distribution beta(Θ | *a* + *h*, *b* + *t*), which is a beautiful result according to equation 4.10 [9, p. 7].

|  |  |  |
| --- | --- | --- |
|  |  | (4.10) |

Where *N* = *h* + *t*. Equation 4.10 is the special case of equation 4.7 in case of beta prior distribution. Equation 4.11 [9, p. 7] specifies the posterior predictive probability *P*(*Xn*+1 = 1 | *D*) in case of binomial sampling and prior beta distribution.

|  |  |  |
| --- | --- | --- |
|  |  | (4.11) |

|  |  |  |
| --- | --- | --- |
|  |  | (9.99) |

**5. Structure learning**

**6. Applications**

**7. Conclusions**

**References**

|  |  |
| --- | --- |
| [1] | Wikipedia, "Bayesian inference," Wikimedia Foundation, 2006. [Online]. Available: http://en.wikipedia.org/wiki/Bayesian\_inference. [Accessed 2007]. |
| [2] | Wikipedia, "Sample (statistics)," Wikimedia Foundation, 29 October 2014. [Online]. Available: http://en.wikipedia.org/wiki/Sample\_(statistics). [Accessed 31 October 2014]. |
| [3] | D. C. Montgomery and G. C. Runger, Applied Statistics and Probability for Engineers, 3rd Edition ed., New York, NY: John Wiley & Sons, Inc, 2003, p. 706. |
| [4] | Wikipedia, "Set (mathematics)," Wikimedia Foundation, 10 October 2014. [Online]. Available: http://en.wikipedia.org/wiki/Set\_(mathematics). [Accessed 11 October 2014]. |
| [5] | K. H. Rosen, Discrete Mathematics and Its Applications, 7nd Edition ed., M. Lange, Ed., McGraw-Hill Companies, 2012, p. 1071. |
| [6] | R. E. Neapolitan, Learning Bayesian Networks, Upper Saddle River, New Jersey: Prentice Hall, 2003, p. 674. |
| [7] | L. Nguyen, "Overview of Bayesian Network," Science Journal Publication, Warri, Delta State, Nigeria, 2013. |
| [8] | K. P. Murphy, "A Brief Introduction to Graphical Models and Bayesian Networks," University of British Columbia, 1998. [Online]. Available: http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html. [Accessed 2008]. |
| [9] | D. Heckerman, "A Tutorial on Learning With Bayesian Networks," Microsoft Research, Redmond, 1995. |
| [10] | R. E. Walpole, R. H. Myers, S. L. Myers and K. Ye, Probability & Statistics for Engineers & Scientists, 9th ed., D. Lynch, Ed., Boston, Massachusetts: Pearson Education, Inc., 2012, p. 816. |