METHODS FOR DISCRETIZING CONTINUOUS VARIABLES WITHIN THE FRAMEWORK OF BAYESIAN NETWORKS

Mihaela-Daciana Crăciun, Violeta Chiş and Cristina Băla

Abstract. Often a Bayesian network (BN) contains discrete and continuous random variables. Discretizing the continuous variables mean that if the possible values of the node are n ranges than the probability of each of these ranges is specified in the network. Many BN inference packages allow the user to specify the both continuous variables and discrete variables in the same network. We can sometimes obtain simpler and better inference results by representing the variables as discrete. One reason for this is that, if we discretize the variables, we do not need to assume any particular continuous probability density function. In this paper we will present two methods for discretizing continuous variables within the BN: Bracket Medians Method and Pearson-Tukey Method.

2000 *Mathematics Subject Classification*: 62E99, 62C10 / Subject Classification for Computer Science: 255

1. Introduction

A. Probability Spaces
Definitions:
Suppose we have a sample space containing n distinct elements: that is,

$$\Omega = \{e_1, e_2, ..., e_n\}$$

A function that assigns a real number P(E) to each event E is called a *probability function* on the set of subsets of if satisfies the following conditions:
1. $0 \leq P(e_i) \leq 1, for 1 \leq i \leq n$; 2. $P(e_1) + P(e_2) + ... + P(e_n) = 1$.
3. For each event that is not an elementary event, P(E) is the sum of the probabilities of the elementary events whose outcomes are in E.

The pair (Ω, P) is called *probability space*.
The most straightforward way to assign probabilities is to use the *Principle of Indifference*, which says that outcomes are to be equiprobable if we have no reason to expect one over the other. According to this principle, when there are n elementary events, each has probability equal to 1/n.

Let E and F be events such that $P(F) \neq 0$. Then the *conditional probability* of E given F, denoted $P(E|F)$, is given by:

$$P(E|F) = \frac{P(E \bigcap F)}{P(F)}$$

Theorem:
Let (Ω, P) be a probability space. Then:
1. $P(\Omega) = 1$; 2. $0 \leq P(E) \leq 1, for every E \subseteq \Omega$
3. $For every two subsets E and F of \Omega such that E \cap F = \emptyset$,

$$P(E \cup F) = P(E) + P(F)$$

where $\emptyset$ denotes the empty space.

B. Random variables
Definitions:
Given a probability space (Ω, P), a random variable X is a function whose domain is Ω.
The range of X is called the space of X.
We call $P(X = x)$ the *probability distribution* of the random variable X.

C. Bayesian Network - BN
Bayesian networks - BN consist of:
- a direct acyclic graph (DAG), whose edges represent relationships among random variables that are often (but not always) causal;
- the prior probability distribution of every variable that is a root in the DAG;
- the conditional probability distribution of every non-root variable given each set of values of its parents.

D. Variance and Covariance
Definitions:
Suppose we have a discrete numeric random variable X, whose space is

$$\{x_1, x_2, ..., x_n\}$$

Then the *variance* Var (X) is given by

$$Var(X) = E([X - E(X)]^2)$$

Suppose we have two discrete numeric random variables X and Y. Then the covariance Cov(X,Y) of X and Y is given by

$$Cov(X, Y) = E([X - E(X)][Y - E(Y)])$$

E. Disrectizing
Let be a BN that contains random variables that are discrete or continuous. For the continuous variable the possible values of the node are ranges and the probability of each of these ranges is specified in the network. This is called *discretizing* the continuous variables.

2.Methods for discretizing

A. Bracket Medians Method
In the Bracket Medians Method the mass in a continuous probability distribution function $F(x) = P(X \leq x)$ is divided into n equally spaced intervals. The method proceeds as follows. Typically we can use three, four ore five intervals. If we have more intervals, the computation is more accurate. Let be n=5 in this explanation.
1. Determine n equally spaced intervals in the interval [0, 1]. If n=5, the intervals are: [0, 0.2], [0.2, 0.4], [0.4, 0.6], [0.6, 0.8] and [0.8, 1].
2. Determine points x_1, x_2, x_3, x_4, x_5 and x_6 such that:

$$P(X \leq x_1) = 0.0, P(X \leq x_2) = 0.2$$
$$P(X \leq x_3) = 0.4, P(X \leq x_4) = 0.6$$
$$P(X \leq x_5) = 0.8, P(X \leq x_6) = 1.0$$

where the values on the right in these equalities are the endpoints of the five intervals.
3. For each interval $[x_i, x_{i+1}]$ compute the bracket median d_i, which is the value such that

$$P(xi \leq X \leq d_i) = P(d_i \leq X \leq x_{i+1}).$$

4. Define the discrete variable D with the following probabilities:

$$P(D = d_1) = 0.2, P(D = d_2) = 0.2$$
$$P(D = d_3) = 0.2, P(D = d_4) = 0.2, P(D = d_5) = 0.2$$

B. Pearson-Tukey Method
In the Pearson-Tukey Method the mass in a continuous probability distribution function $F(x) = P(X \leq x)$ is divided into three intervals. The method proceeds as follows:
1. Determine points x_1, x_2 and x_3 such that

$$P(X \leq x_1) = 0.05, P(X \leq x_2) = 0.50, P(X \leq x_3) = 0.95$$

2. Define the discrete variable D with the following probabilities:

$$P(D = x_1) = 0.185, P(D = x_2) = 0.63, P(D = x_3) = 0.185$$

3. Applying the discretizing methods

Let be the BN for detecting credit card fraud, see the Figure 1.

A. Bracket Medians Method
Suppose we have the normal distribution function given by
$> with(Statistics); X := RandomVariabile(Normal(\mu, \Omega)); PDF(X, x)$

$$\frac{1}{2}\frac{\sqrt{2}e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}}{\sqrt{\pi}\sigma}$$

where, $> Mean(X)$ represent μ and $> Variance(X)$ represent σ^2
and the cumulative distribution function for this density function is given by
$> with(Statistics); CumulativeDistributionFunction(Normal(\mu, \Omega), x)$

$$\frac{1}{2} - \frac{1}{2}erf(\frac{1}{2}\frac{\sqrt{2}(-x+\mu)}{\sigma})$$

$> CumulativeDistributionFunction(Normal(\mu, \Omega), x, numeric)$

This functions for $\mu = 50$ and $\sigma = 15$ are shown in Figure 2 and 3. This might be the distribution of age for some particular population.
$> with(Statistics); X := RandomVariabile(Normal(50, 15)); PDF(X, x)$

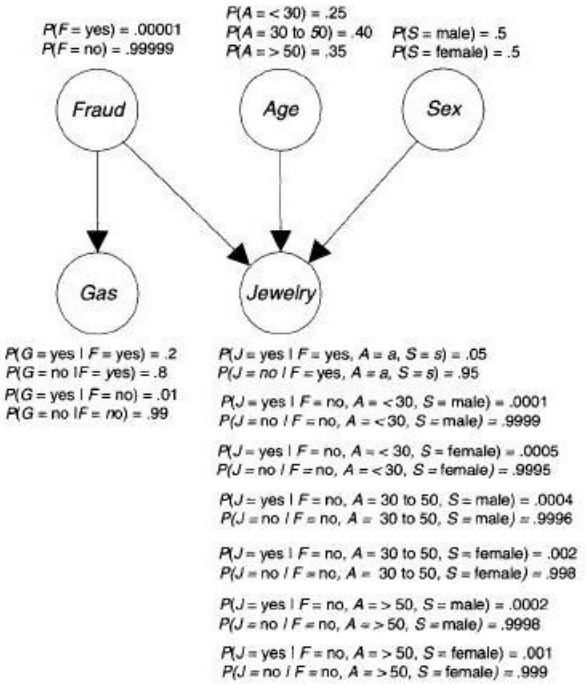


Figure 1: BN for detecting credit card fraud

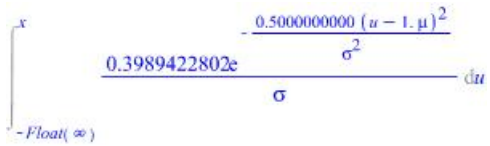
$$\frac{\frac{1}{30}\sqrt{2}e^{-\frac{1}{450}(x-50)^2}}{\sqrt{\pi}}$$

$> smartplot(PDF(X, x))$

$> CumulativeDistributionFunction(Normal(50, 15), x, numeric)$

$> smartplot(PDF(X, x))$

Next we use the Bracket Medians Method to discretize it into three ranges. Then n=3 and our four steps are as follows:
1. Since there is essentially no mass less then 0 and greater then 100, our three intervals are [0, .333], [.333, .666] and [.666, 1].
2. We need to find points x_1, x_2, x_3 and x_4 such that



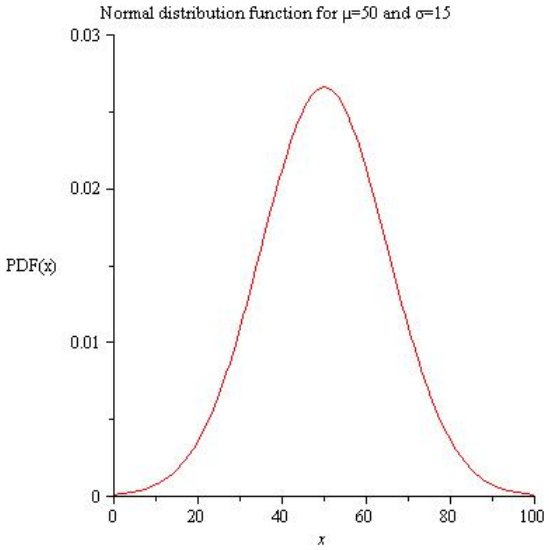


Figure 2: NDF with $\mu = 50$ and $\sigma = 15$

$$P(X \leq x_1) = 0.0, P(X \leq x_2) = 0.333, P(X \leq x_3) = 0.666, P(X \leq x_4) = 1.$$

Clearly, $x_1 = 0$ and $x_4 = 100$. To determine x_2 we need to determine

$$x_2 = F^{-1}(0.333)$$

Using Maple, we have:
$> T := Normal(50, 15)); X := RandomVariable(T); CDF(X, t)$

$$T := Normal(50, 15)$$
$$X := R6$$
$$\tfrac{1}{2} + \tfrac{1}{2} erf(\tfrac{1}{30} t\sqrt{2} - \tfrac{5}{3}\sqrt{2})$$

$> InverseSurvivalFunction(X, t)$

$$\int_{-Float(\infty)}^{x} 0.02659615201 e^{-0.002222222222\,(u-50.)^2}\,du$$

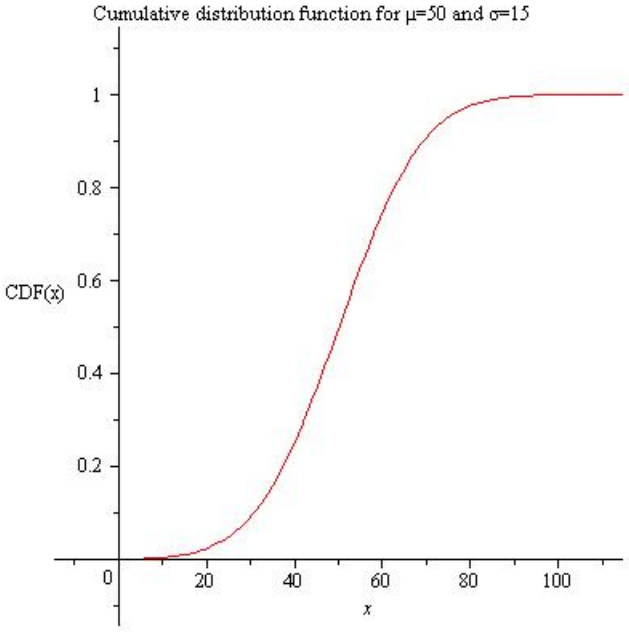


Figure 3: CDF with $\mu = 50$ and $\sigma = 15$

$$5(5\sqrt{2} + 3RootOf(erf(Z) - 1 + 2t))\sqrt{2}$$

$> InverseSurvivalFunction(X, 1 - 0.333)$

$$43.525336409240651$$

$$x_2 = 43.5$$

Similarly,
$> InverseSurvivalFunction(X, 1 - 0.333)$

$$56.433417561113032$$
$$x_3 = 56.4$$

3. Compute the bracket medians. We compute them using Maple by solving the following equations:

$> CumulativeDistributionFunction(Normal(50,15), d_1) = CumulativeDistributionFunction(Normal(50,15), 43.5) - CumulativeDistributionFunction(Normal(50,15), d_1)$

$$\frac{1}{2} + \frac{1}{2}erf(\frac{1}{30}d_l\sqrt{2} - \frac{5}{3}\sqrt{2}) = -0.1676136874$$

$$-\frac{1}{2}erf(\frac{1}{30}d_l\sqrt{2} - \frac{5}{3}\sqrt{2})$$

$solve([d_1])$

$$[[d_1 = 35.46022283]]$$

Solution is $d_1 = 35.5$

$> CumulativeDistributionFunction(Normal(50,15), d_2) - CumulativeDistributionFunction(Normal(50,15), 43.5) = CumulativeDistributionFunction(Normal(50,15), 56.4) - CumulativeDistributionFunction(Normal(50,15), d_2)$

$$0.1676136874 + \frac{1}{2}erf(\frac{1}{30}d_2\sqrt{2} - \frac{5}{3}\sqrt{2}) = 0.1651889337$$

$$-\frac{1}{2}erf(\frac{1}{30}d_2\sqrt{2} - \frac{5}{3}\sqrt{2})$$

$solve([d_2])$

$$[[d_2 = 49.95441526]]$$

Solution is $d_1 = 50.0$

$> CumulativeDistributionFunction(Normal(50,15), d_3) - CumulativeDistributionFunction(Normal(50,15), 56.4) = 1 - CumulativeDistributionFunction(Normal(50,15), d_3)$

$$-0.1651889337 + \frac{1}{2}erf(\frac{1}{30}d_3\sqrt{2} - \frac{5}{3}\sqrt{2}) = \frac{1}{2}$$

$$-\frac{1}{2}erf(\frac{1}{30}d_3\sqrt{2} - \frac{5}{3}\sqrt{2})$$

$solve([d_3])$

$$[[d_3 = 64.46702832]]$$

Solution is $d_3 = 64.5$

4. Finally, we set

$$P(D = 35.5) = 0.333, P(D = 50.0) = 0.333, P(D = 64.5) = 0.333$$

The variable D requires a numeric value if we need to perform computations using it. However, if the variable does not require a numeric value for computational purposes, we need to perform Step3 in the Bracket Medians Method. We just show ranges as the values of D. In the pervious example, we would set

$$P(D \leq 43.5) = 0.333,\ P(D = 43.5 \text{ to } 56.4) = 0.333, P(D \geq 56.4) = 0.333$$

This example is what we did for the node Age in the BN in Figure 1. In this case, if a data item's continuous value is between 0 and 43.5, we simply assign the data item that range.

B. Pearson-Tukey Method

Suppose we have the normal distribution [6] discussed by the Bracket Medians Method. Next, we apply the Pearson-Tukey Method to that distribution. 1. Using Maple, we have

$> T := Normal(50, 15)); X := RandomVariable(T); CDF(X, t)$

$$T := Normal(50, 15)$$
$$X := R$$
$$\tfrac{1}{2} + \tfrac{1}{2}erf(\tfrac{1}{30}t\sqrt{2} - \tfrac{5}{3}\sqrt{2})$$

$> InverseSurvivalFunction(X, t)$

$$5(5\sqrt{2} + 3RootOf(erf(Z) - 1 + 2t))\sqrt{2}$$

$> InverseSurvivalFunction(X, 1 - 0.05)$

$$25.327195595717995$$

Solution is $x_1 = 25.3$.

$> InverseSurvivalFunction(X, 0.50)$

50.

Solution is $x_2 = 50$.
$> InverseSurvivalFunction(X, 1 - 0.95)$

74.672804404281990

Solution is $x_3 = 74.7$.

2. We set

$$P(D = 25.3) = 0.185, P(D = 50.0) = 0.63, P(D = 74.7) = 0.185$$

To assign data items discrete values, we need to determine the range of values corresponding to each of the cutoff points. That is, we compute the following:

$> InverseSurvivalFunction(X, 1 - 0.185)$

36.552899539965821

$> InverseSurvivalFunction(X, 0.185)$

63.447100460034178

If data item's continues value is less the 36.6, we assign the data item the value 25.3; if the value is in [36.5, 63.4], we assign the value 50; and if the value is greater then 63.4, we assign the value 74.7.

If the variable does not required a numeric value for computational purposes, we need to perform Steps 1 and 2, but rather just determine the range of values corresponding to each of the cutoff points and just show ranges as the values of D. In our example, we would set

$$P(D \leq 36.6) = 0.185, P(D = 36.6 \text{ to } 63.4) = 0.63, P(D \geq 63.4) = 0.185$$

In this case if a data item's continuous value is between 0 and 36.6, we simply assign the data item that range.

4. Conclusion

We observe that, when we used the Pearson-Tukey Method, the middle discrete value represented numbers in the interval [36.6, 63.4], while when we used the Bracket Median Method, the middle discrete value represented numbers in the interval [43.5, 56.4]. The interval for the Pearson-Tukey Method is larger, meaning more numbers in the middle are treated as the same discrete value, and the other two discrete values represent values only in the tails.

References

[1] Berry, D.A., Statistics: A Bayesian Perspective, Wadsworth, Belmont, California, 1996

[2] Cooper, G.F., "The Computational Complexity of Probabilistic Inference Using Bayesian Belief Networks," Artificial Intelligence, Vol. 33, 1990

[3] Hogg, R.V., and A.T. Craig, Introduction to Mathematical Statistics, Macmillan, New York, 1972

[4] Lindley, D.V., Introduction to Probability and Statistics from a Bayesian Viewpoint, Cambridge University Press, London, 1985

[5] D. Heckermann and D. Geiger. Learning Bayesian networks: a unification for discrete and gaussian domains. In UAI '95, pp. 274-284. 1995

[6] Evans, Merran; Hastings, Nicholas; and Peacock, Brian. Statistical Distributions - 3rd ed. Hoboken: Wiley, 2000

[7] Stuart, Alan, and Ord, Keith. Kendall's Advanced Theory of Statistics. 6th ed. London: Edward Arnold, 1998. Vol. 1: Distribution Theory.

Mihaela-Daciana Craciun, Violeta Chis, Cristina Bala
Department of Mathematics and Computer Science
"Aurel Vlaicu" University of Arad
Str.Elena Dragoi, No.2, 310330 - Arad, Complex Universitar M
email: *mihaeladacianacraciun@yahoo.com, viochis@yahoo.com, crisbala@yahoo.com*