CONDUCTING INFERENCE USING BAYESIAN NETWORKS IN R

AUGUST RAMBERG-GOMEZ, LAURA SCHMIDT-SCHWEDA, ANNA TALMAN

Introduction

Background

Bayesian Networks (BN) are a graphical method of representing the conditional dependencies between variables in a data set. A BN is a directed and a-cyclic graph (DAG) $G = (V, E)$ with vertex set $v \in V$ and directed edge set $e \in E$. Vertices represent variables where as edges represent the conditional dependencies between variables. They are a powerful form of modelling that can both represent and store the conditional probabilities between variables in multi-variate data sets. They are extremely useful when dealing with multivariate data that may be unwieldy to model using other methods of analysis. The primary difficulty in constructing BN's lies in determining the structure of the network. The structure can be determined purely algorithmically if sufficient data is available, with the help of experts if relationships between variables are known, or by a combination of the two methods. One of the strong suits of BN's is their ability to exploit expert knowledge about known relationships between variables into the construction of a network when there is a lack of data. When there is sufficient data to determine the structure of the network algorithmically, the problem of determining the optimal structure from the space of all possible structures is NP-hard. This means that there is no fast or optimal way to determine if a given structure of a BN determined from data is optimal. There are, however, highly effective algorithms depending on whether a data set is discrete or continuous. We will explore, by simulation, the implementation of one such algorithms for use with discrete data. Following this, for our second data set, we will also construct a BN purely from expert knowledge of known relationships.

For our analysis we will create two networks using the different construction approaches; a simple discrete network using simulated data and a more complex model that incorporates expert knowledge on relationships between variables. To perform these different analyses we will be using different combinations of algorithms and data sets. Since the problem of learning the structure of a BN from data is NP-hard we will be using a single greedy, score-based algorithm called the hill climbing algorithm to learn the structure of our network. These greedy algorithms are computationally more efficient in finding optimal network constructions. Although they do not always find the optimal network, they generally find network constructions that are very good. These algorithms operate by selecting the best immediate solution until they have constructed the entire structure. A version of the the hill climbing algorithm is able to scale to construct networks with thousands of variables. We selected the hill climbing algorithm since it is included in the "bnlearn" package and because it has been suggested to be an extremely effective algorithm for discrete data sets (Beretta *et al*, 2018).

Date: 26 February 2020.

Research Question: How can we use R to create a Bayesian Network and the accompanying probability distribution by hand? Furthermore, can we simulate a data set using the relationships we created by hand and then from this data algorithmically learn our constructed network?

Bayes Theorem

The probability distributions in BN's are determined using Bayes theorem. Bayes theorem is an elaboration of conditional probability. Below we have equations for standard conditional probability for two events, A and B.

This equation gives the probability that an event A occurs given that an event B has already occurred.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

This equation gives the probability that an event B occurs given that an event A has already occurred.

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

Remark: $P(A \cap B) = P(B \cap A) = P(B) \times P(A)$

Bayes theorem elaborates on standard conditional probability by using the remark above to combine the two equations algebraically. By some simple substitutions, we can arrive at the theorem:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

The theorem provides a more nuanced view of conditional probability that can be continuously updated given new information. It also provides a more complete picture of the probability of an event occurring given prior knowledge. In other words, the formula is very effective at incorporating evidence. We can conceptualize this as computing the probability of a hypothesis given new evidence. Using the conditional probability of hypotheses and evidence allows for the use of expert opinion in the event that a data set is incomplete. We can re-write Bayes theorem below using hypothesis and evidence:

$$P(Hypothesis|Evidence) = \frac{P(Evidence|Hypothesis) \times P(Hypothesis)}{P(Evidence)}$$

Methods

To summarize our methods briefly, we will be using steps outlined previously by data scientist Marco Scutari. Following these outlined steps we will be able to gain an understanding of how Bayesian networks function within R. The first step in our process is to manually build a BN, and then to simulate a data set, with the same conditional probabilities, and to compare the resulting BN with the original BN. We chose to simulate our own data because it provides the ability to manipulate it in order to investigate and understand the construction process of BN's.

The simulated discrete data set to be used with the hill-climbing algorithm and our experiment is explained below. We will be modeling the conditional dependencies between variables in a classic game theory children's experiment. In this experiment children of various ages and gender are placed in a room, either alone or with a partner. The child(ren) are then given instructions and have a single marshmallow placed in-front of them. They have the choice to eat the marshmallow right away, or wait an undetermined length of time in order to receive a second marshmallow. The variables that will affect whether or not the treat will be eaten right away are the gender, and age of the participant, as well as whether or not they have companion in the experiment with them.

With our experiment outlined, we will build a BN structure using semi-arbitrarily determined conditional probabilities between variables. The attached R script follows through the steps of building the network, visualizing it, and finally inputting the conditional probabilities. This will give us a fully constructed BN. We will simulate a data set with a sample size of $n = 100$. Then, we will then use algorithms to learn the BN structure of this data set and see if it is similar to the BN we created by hand. They should be the same as this is a test of concept. From these two networks, we will be able to infer how likely a child is to eat their marshmallow before the time is up based on our explanatory variables.

Results

This section will show the conditional probabilities of each variable affecting the outcome. These results will also include inferences made with our BN. We will also show the accuracy of the network and whether or not it presented us with the probabilities of the manually constructed network.

Discussion

This section will discuss the relevance of the conclusions of our model, potential further applications of these network structures, and limitations of our BN.

Reflection

In this section we will reflect on the process, including which parts were more or less challenging and why.

References

Beretta, S., Castelli, M., Gonçalves, I., Henriques, R., Ramazzotti, D. (2018). Learning the structure of Bayesian Networks: A quantitative assessment of the effect of different algorithmic schemes. Complexity, 2018.

(R)eliability. “Heart Disease Prediction From Patient Data in R.” R, 28 Sept. 2019, www.r-bloggers.com/heart-disease-prediction-from-patient-data-in-r/.b

Scutari, Marco. “Bayesian Network Structure Learning.” Bnlearn, 2019, www.bnlearn.com/.

Appendix

Please see our new attached schedule. Some troubles that we have come across along the way have included understanding how to incorporate and simulate Bayes Theorem into our data set. This particular challenge has taken more time than initially expected and has resulted in a shift of our schedule.

Saturday February 22	Understand Bayes Theorem and how we can simulate it in our data
Sunday February 23	Successfully simulate first data set, attempt simulation on a second more complicated data set
Monday February 24	Incorporate feedback, summarize results and discussion
Tuesday February 25	Finish paper editing, further annotate code, start presentation, **Final Paper due 4:00**
Wednesday February 26	**Presentation**