Loan prediction using Bayesian network

Sultan Baimukhanov
School of Engineering and Digital Sciences
Nazarbayev University
Nur-Sultan, Kazakhstan
sultan.baimukhanov@nu.edu.kz

I. Introduction

Artificial intelligence has become popular nowadays with a machine and deep learning implementation in data management. However AI has a probabilistic direction that uses widespread modeling in prediction like Bayesian networks (BN) or model Markov networks (MN). I implemented Bayesian and Markov networks to choose a database to create models and find inference output at the end of the project. The following report consists of the following stage describing motivation (importance of topic), a problem considering the project's purpose, methodology investigation of the chosen dataset and considering completing projects, solution implementation of MN and BN, comparison of the result of two methods and discussion of the evaluation.

II. Motivation

The primary motivation for this project was to attempt to implement a BN and MN network in the chosen dataset and consider the received result (inference). The project aims to prepare a model for Loan prediction agree probability. Also, Loan companies or banks can use the dataset and project overall in real life. For example, in Kazakhstan, the government can use it to prepare a model of the application from citizens that can take a credit of 4 percent to buy a car by filling out the form.

III. Problem

The current issue can lead society to misunderstand the importance of taking and returning money to the loaner. Today, taking out a loan has become a more straightforward process than in previous decades that requires certain conditions to take a loan from a bank. However, people can take loans from another financial organization (not a bank) that does not have a rule and does not look at income or ability to pay and return loans without a problem. Therefore the project can learn by using modeling and database to analyze all necessary factors and give answers of agreement or denied on the loan request.

IV. Solution

I have implemented BN and MN to the project. First, I have prepared data for BN modeling. Second, I have installed all necessary libraries (pgmpy.models import BayesianModel) that allow me to create models. The Bayesian network uses directed graphical models where nodes and edges are connected as a family (parents and children), while Markov networks use undirected graph and neighbors methods. During the execution project in the Bayesian network, I created a model. Next, I find probabilities for each node by using functions for one node, 2 nodes, and 5 nodes. Further, I find CPD and check the model state. The model state allows me to generate local independencies. As a result, I used variable elimination to get the inference. The initial process with the dataset was the same in the Markov network, but implementing the MN model requires different methods. In MN, I first add the list map and convert it to int and an array after finding an occurrence between nodes because, in MN, all nodes are neighbors. Next, I add nodes and edges to model l. Modeling gives the ability to find discrete factors and multiply them. These allow me to find local interdependencies and marginalize them.

V. Methodology

The database was from $Kaggle$, an open-source site with many different databases. I have used the $Kaggle$ site to find exciting datasets. However, I have met challenges such as too large size and too many columns and rows. As a result, I chose the Loan prediction dataset. The project preparation has been investigated using different methods such as Logistic regression, Support vector machine, and KNN implemented in the loan database. The data preparation was divided into two steps: process and exploration. The process consists of analyzing, finding empty rows, replacing them with numbers, and preparing data for modeling, exploring needs to see the impact of different nodes on loan status to imagine and draw the first (draft) version of modeling.

VI. Results

The result that I achieved after BN implementation shows that the following methods are appropriate to use with this data. According to our model, the perfect man who agrees on a loan should have a degree, work in the city, have a good credit history(good history), and be married. According to achievement results, BN can use it in real life to determine the status of requesting a person for a loan. MN allows me to model and consider how neighbors can communicate and impact each other. However, in comparison with MN, BN is more relevant to the project despite the logical reason for modeling.

VII. Evaluation

I have used MLE and SVC to estimate accuracy of inference. MLE is usually used with BN and SVC with MN. The

process of training accuracy was similar in some steps, both methods used Xtrain and Xtest, ytrain and ytest. The difference in SVC is that I used SVC kernel while in MLE estimators but both of them fit the model in the end. BN shows 78

VIII. Conclusion

During this project, I learned the basics of MN and BN. Also, the chosen dataset was used only with Logistic regression and SVM. Compared with BN and MN, last can give better accuracy because it communicates with each node as a neighbor. At the same time, in BN, I have created a model use logical reasoning and can be implemented in real life, as I mentioned in motivation.