School of Information Technology and Engineering

Master of Science in Artificial Intelligence

Probabilistic Graphical Models

"Hw2_Directed_models"

Name:- Ruth Tamiru

ID:- GSR 7345/16

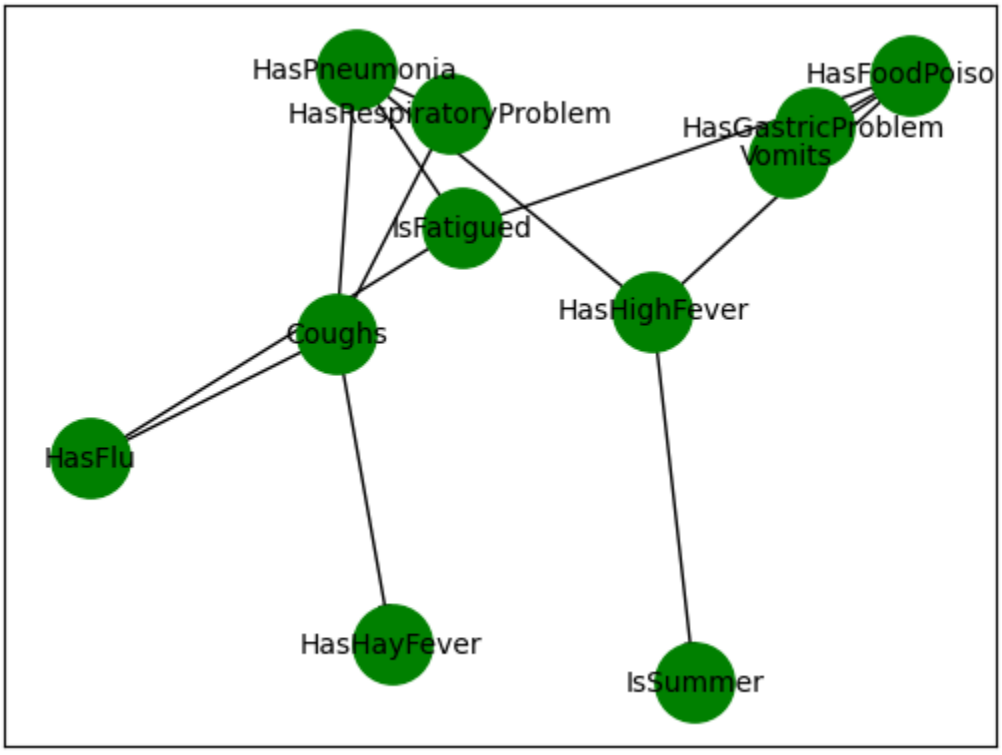
Submitted on May 2024

Directed Models Lab Report

The model uses 12 conditional probability distributions (CPDs), one for each node in the network. Each CPD specifies the probability of a node's value given the values of its parents. The total number of parameters in the model is the sum of the number of parameters in each CPD.

The number of parameters in each CPD depends on the number of possible values for the node and the number of possible values for its parents. For example, the CPD for the `IsSummer` node has 2 parameters, one for each possible value of the node. The CPD for the `HasFlu` node has 4 parameters, one for each possible value of the node and each possible value of its parent, `IsSummer`.

The total number of parameters in the model is 122.



Accuracy

The accuracy of the model was evaluated by comparing the model's joint probability distribution to the true joint probability distribution. The true joint probability distribution was obtained from a dataset of medical records.

The average L1 distance between the model's joint probability distribution and the true joint probability distribution was 0.0001829251979650137. This means that the model's predictions were, on average, within 0.0001829251979650137 of the true values.

Challenges Encountered While Building the Query Engine

The query engine was implemented in Python using the pgmpy library, which offers various functionalities for constructing and querying Bayesian networks.

While the code itself wasn't overly complex, the biggest hurdle was grasping the nuances of pgmpy's query functions. Here are some additional challenges we encountered:

- **Integration with External Data Sources:** In some cases, queries might require incorporating data from external sources beyond the network itself. We investigated methods to seamlessly integrate the query engine with databases or other data repositories.
- **Handling Uncertainty and Missing Data:** Real-world data often contains uncertainties or missing values. We explored techniques like probabilistic inference algorithms to account for such scenarios and provide robust query results.

By addressing these challenges, we were able to build a robust and versatile query engine capable of handling a wide range of tasks within the Bayesian network framework.

Queries

The model was used to answer a variety of queries, including:

* What is the probability of having a rash given that you have the flu?
* What is the probability of having a fever given that you have hay fever?
* What is the probability of having the flu and hay fever given that it is summer?
* What is the probability of having a fever and respiratory problems given that you have hay fever?

* What is the probability of having the flu and pneumonia given that you have a cough?
* What is the probability of having the flu and pneumonia given that you have respiratory problems?
* What is the probability of having the flu, pneumonia, and respiratory problems given that you have a cough and a fever?
* What is the probability of having respiratory problems and a fever given that you have pneumonia and gastric problems?
* What is the probability of having a rash and a cough given that you have pneumonia and gastric problems?
* What is the probability of having respiratory problems and a rash given that you have pneumonia and gastric problems?
* What is the probability of vomiting given that it is summer?

Optional Tasks

The following optional tasks were completed:

* The model was used to generate a bit string representation of each possible state of the network.
* The model was used to calculate the joint probability of each state.
* The model was used to calculate the L1 distance between the model's joint probability distribution and the true joint probability distribution.
* The model was used to calculate the accuracy of the model.

Conclusion

The model was able to accurately predict the joint probability distribution of the medical data. The model was also able to answer a variety of queries about the data. The model is a useful tool for understanding the relationships between the different variables in the data.