School of Engineering and Applied Science (SEAS), Ahmedabad University

Probabilistic Graphical Models (CSE 516)

Project Report

Group Name:
Team Members:

- Name of Team Member-1(Manav Yagnik AU2040040)
- Name of Team Member-2(Vrutik Barava AU2040236)
- Name of Team Member-3(Nand Patel AU2040183)
- Name of Team Member-4(Jay Patel AU2040014)

Summary

Weather plays an important role in every living species. Weather determines the convenience of doing any work or task. Weather directly affects the social and economic lifestyle of people around it. we are trying to predict tomorrow's rain probability using the bayesian network, which is based on weather forecasting. For that, we are using 10 years of daily Australian weather data. In this paper, we try to make two different networks and gonna compare the accuracy between networks, and try to understand the bayesian network concept from them. Try to conclude How it bayesian network help.

Till this point, we created one small bayesian network and tried to run it. So, here some doubt happened we can solve it now. And also found some information about topic which can help us to understand it better. In next step we want to do different changes in model and try to analyze it for next two week.

I. Introduction

A. Background

Weather plays an important role to do task or any kind of activity. It is very obvious that if you want to do any kind of activity then you need to plan it accordingly. Planning also needs an assurance about weather condition. So we needed to know about the future condition of weather. To solve this problem, we can use Probabilistic Graphical Models. Bayesian Belief Networks can get easily implemented to determine tomorrows rain. It is an efficient, compact and intuitive knowledge representation for handling uncertainty. Structure of a graphical model that de fines a set of dependence and independence statements over a set of random variables that represent the entire network. The main object of these networks is trying to understand the structure of causality relations.

Mathematically Belief network should contain variable $X = X_1, X_2, \ldots.., X_N$ of a Directed Acyclic Graph and probabilities can be calculated by the formula $(X_1, \ldots.., X_N) = \prod_{i=1}^{N} \mathrm{P}(X_i/\mathrm{Parents}(X_i))$

Where Parents(X_i) are the parents of X_i in a network.

B. Motivation

Many of our social and economic system depend on weather forecasts. Weather includes the interconnection of physical variables such as wind direction, speed, air pressure, temperature, humidity, date and location. Also due to global warming it becomes more complicated than before due to frequent climate changes. We can use Machine Learning algorithm on the data sets but Use of Graphical model is better option because by using this kind of algorithms or structure we can reduce the size of datasets and reduce the complexity of the network. Also use of Bayesian Network will help us to inferences in an easy manner.

Here Parents(X_i) are the parents of Xi in a network. Further we can calculate the joint distribution and conditional probabilities indicated by the network. Bayesian belief network can be used as representation tool for the decision making under uncertainty.

C. Contribution

We have used data set of Australia it contains past 10 years of daily weather observation from multiple Australian weather stations. We have tried to implement Bayesian Network and predicated the tomorrows rain.

II. Methodology

A. Analytical Overview/Product Details/

Data and it's preprocessing
We are using Australian data to predict tomorrow's rainfall. This data has a total of 23 variables. We are using only a few variables to predict the rain.

Taken parameter preprocessing:-

1) Season: - In this variable, we are converting the date data into the season. This season is helping to predict the rainfall of the specific season of the day. This variable data of different ranges are converted into three different states named 1, 2, and 3 as high, moderate, and low. Range:converted form date into season 1, season 2, season 3

2) Mintmp: - This is the minimum temperature over the allover day. This variable data of different ranges are converted into three different states named 1, 2, and 3 as high, moderate, and low. Range: mintmp$\leq$15 convert into 1 mintmp$>$ 30 and mintmp$<$50 convert into 2 mintmp$\geq$ 15 and mintmp $\leq$30 convert into 3

3) Maxtmp: - This is the maximum temperature over the allover day. This variable data of different ranges are converted into three different states named 1, 2, and 3 as high, moderate, and low. Range: Maxtmp$\leq$15 convert into 1 Maxtmp$>$ 30 and Maxtmp¡50 convert into 2 Maxtmp$\geq$15 and Maxtmp$\leq$30 convert into 3

4) Sunshine: - The number of hours of bright sunshine in the day. This variable data of different ranges are converted into three different states named 1, 2, and 3 as high, moderate, and low. Range: Sunshine$\leq$5 convert into 1 Sunshine$>$ 5 and Sunshine$<$10 convert into 2 Sunshine$\geq$10 convert into 3

5) Humidity: - It indicates the moisture level in the air. In this variable, we are taking the average humidity at 9 am and humidity at 3 pm. This variable data of different ranges are converted into three different states named 1 and 2 high and low. Range: Humidity$\leq$70 convert into 1 Humidity $>$ 70 convert into 2

6) Pressure: - Atmospheric pressure (hPa) reduced to mean sea level In this variable, we are taking the average pressure at 9 am and pressure at 3 pm. This variable data of different ranges are converted into three different states named 1 and 2 high and low. Range: Pressure$\leq$1010 convert into 1 Pressure$>$ 1010 convert into 2

7) Cloud: - It is defined as the fraction of the sky obscured by clouds at a particular time In this variable, we are taking the average Clouds at 9 am and clouds at 3 pm. This variable data of different ranges are converted into three different states named 1 and 2 high and low. Range: Cloud$\leq$5 convert into 1 Cloud$>$ 5convert into 2

8) Wind speed: - Wind speed (km/hr) averaged over 10 minutes prior to 9 am and 3 pm. This variable data of different ranges are converted into three different states named 1 and 2 high and low. Range: Wind speed$\leq$70 convert into 1 Wind speed$>$ 70 convert into 2

9) Evaporation: - It is the process by which a liquid turns into a gas. This variable data of different ranges are converted into three different states named 1and 2 high and low. Range: Evaporation$\leq$20 convert into 1 Evaporation$>$ 20 convert into 2

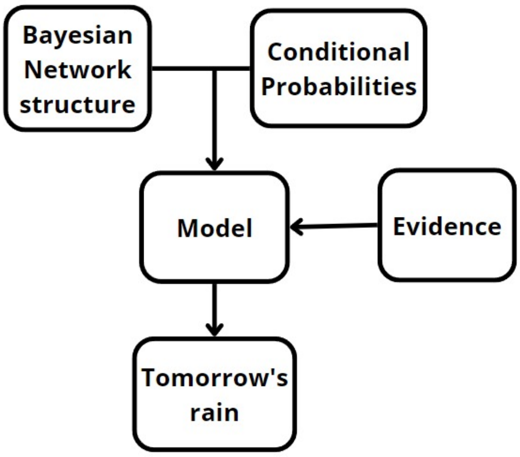
10) Rainfall: - The amount of rainfall recorded for the day in mm. This variable data of different ranges are converted into three different states named 1 and 2 high and low. Range: Rainfall$\leq$10 convert into 1 Rainfall$>$ 10 convert into 2

11) Rain Tommorow: - Describe the value Yes if rain is falling or not if rain is not falling. This variable data of different ranges are converted into three different states named 1and 2 Yes and No.

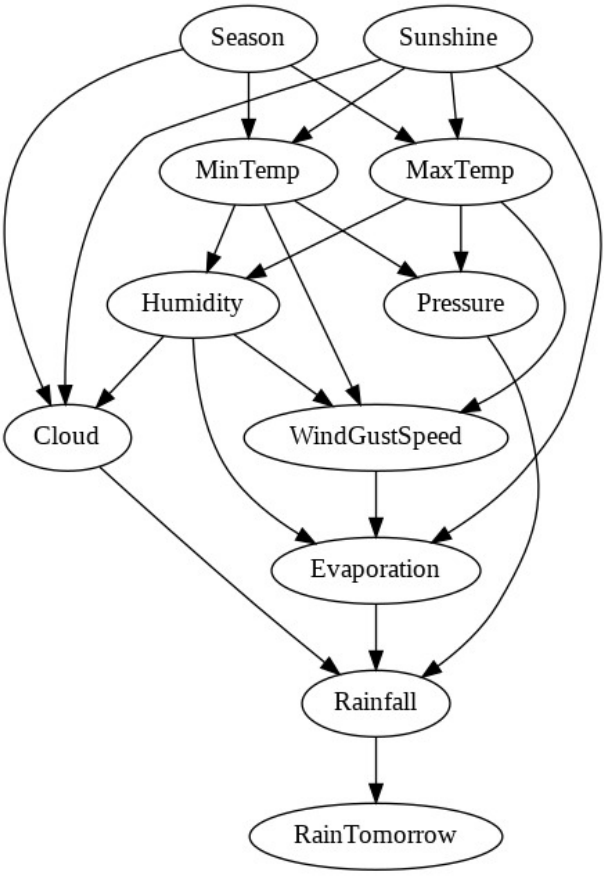
B. Mathematical Analysis/Working of the Product/Framework

In this paper, we make two Bayesian networks one is a Manually constructed bayesian network and another one is an Automatically generated bayesian network

We are using 11 variables are used to predict tomorrow's rain. With the help of these 11 nodes, we make our Bayesian network. We identify relevant nodes and the structural dependencies between them. Bayesian network is used to represent these dependencies/independencies. Tomorrow's rain is directly or indirectly dependent on present-day atmospheric phenomena. We are using atmospheric parameters which are known to cause present and tomorrow's rainfall. Here we are taking the season, sunshine, minimum and maximum temperature, humidity, pressure, cloud, average wind speed, evaporation, and if rain falls then the total amount of rain. All these variables are taken as nodes in the bayesian network and all nodes are taken as discrete values. All the data of the variable is converted into high, moderate, low, or high and low.



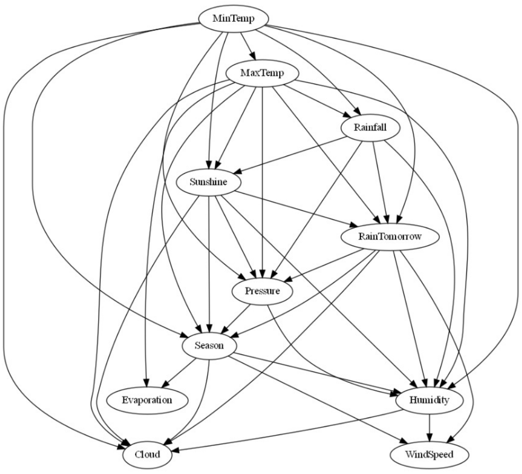
Manual constructed bayesian network



Working

In this model we are doing only parameter learning. This model work in two stages, training and testing. In training, the stage model is trained through parameter learning. By doing parameter learning we got a conditional probability table that helps to find the posterior probability of the given query. After the parameter learning, we do variable elimination the finding of the posterior probability for the given query. In the query, we are finding tomorrow's rain.

Automatically generated bayesian network



Working

In this model, we are doing both structured learning and parameter learning. This model works in two stages, training and testing. In the training process, we are adding the database to the model to learn the bayesian structure and conditional probability. Training includes structure learning using the Hill Climbing K2 algorithm and parameter learning using the Bayesian estimator which uses Bayesian Dirichlet equivalent uniform(BDeu).data of all variables is given during the training stage of the model. In testing, the model is an evaluation based on its performance. During the testing mode, for finding tomorrow's rain we are passing some queries with the evidence nodes, and the model predicts the value based on inferences. The predicted value is compared to the actual data to evaluate the performance.

Parameter learning: -

Parameter learning is the process of calculating the conditional probabilities between nodes. For this model, we are using a Bayesian estimator which uses a Bayesian Dirichlet equivalent uniform(BDeu). The bayesian estimator maximizes the posterior probability.

Structure learning: -

Structure learning is the process of finding the Bayesian network structure or Directed acyclic graph. We are using the Hill Clim K2 algorithm for finding the dependencies between different nodes. The K2 algorithm is a greedy search algorithm for finding the Bayesian network. It assumes all nodes have no parents and nodes are ordered. All the orders are decided manually.

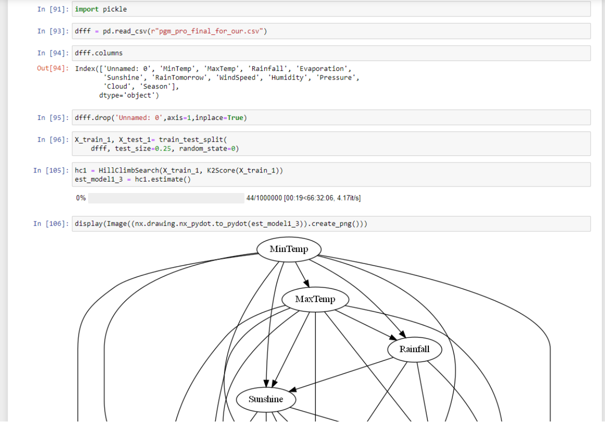
Hill Climb

It is a local search Algorithm, so it has the knowledge for training data sets. It works for a greedy approach, so it will work until it gets the best moves. If the state is better than the current state, then it is a new state; the state refers to nodes or variables. Also, you cannot do backtracking. The changing of state will occur until it gets stable. The use of hill climbing reduces the space complexity of the problem. It is also known as a Heuristic search. Due to the greedy concept of the Hill Climbing Algorithm, there are heavy chances of getting local maximum, plateau maximum, etc. The main reason for doing Hill climb is to define the best structure for the bayesian network.

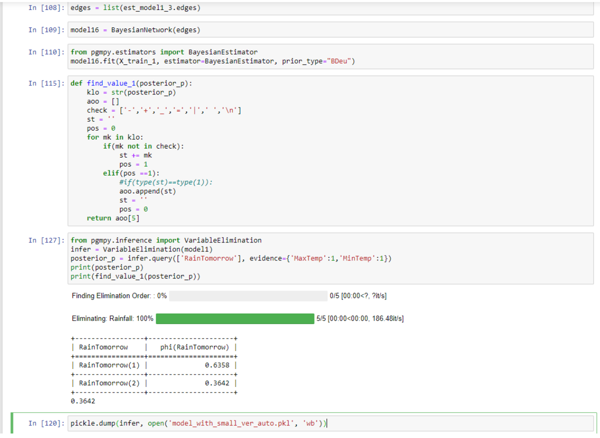
Variable elimination algorithm.

It is an efficient method for deriving inferences rather than joint distributions. Bayesian network. Final conditional probability can be calculated in two ways, one with solving bayes network with joint distribution or if we have found some query given some condition then we can directly apply the variable elimination method. In order to do Variable elimination we need to do in some order or in the form of a clique tree. For our bayesian network we calculated rain tomorrow with evidence Evaporation, WindGustSpeed.

III. Codes(only main document not supporting files with comments on each line)



```
In [91]:  import pickle

In [93]:  dfff = pd.read_csv(r"pgm_pro_final_for_our.csv")

In [94]:  dfff.columns

Out[94]:  Index(['Unnamed: 0', 'MinTemp', 'MaxTemp', 'Rainfall', 'Evaporation',
                 'Sunshine', 'RainTomorrow', 'WindSpeed', 'Humidity', 'Pressure',
                 'Cloud', 'Season'],
                dtype='object')

In [95]:  dfff.drop('Unnamed: 0',axis=1,inplace=True)

In [96]:  X_train_1, X_test_1= train_test_split(
              dfff, test_size=0.25, random_state=0)

In [105]: hc1 = HillClimbSearch(X_train_1, K2Score(X_train_1))
          est_model1_3 = hc1.estimate()

          0%                    44/1000000 [00:19<66:32:06, 4.17it/s]

In [106]: display(Image((nx.drawing.nx_pydot.to_pydot(est_model1_3)).create_png()))
```



```
In [108]: edges = list(est_model1_3.edges)

In [109]: model16 = BayesianNetwork(edges)

In [110]: from pgmpy.estimators import BayesianEstimator
          model16.fit(X_train_1, estimator=BayesianEstimator, prior_type="BDeu")

In [115]: def find_value_1(posterior_p):
              klo = str(posterior_p)
              aoo = []
              check = ['-','+','_','=','|',' ','\n']
              st = ''
              pos = 0
              for mk in klo:
                  if(mk not in check):
                      st += mk
                      pos = 1
                  elif(pos ==1):
                      #if(type(st)==type(1)):
                      aoo.append(st)
                      st = ''
                      pos = 0
              return aoo[5]

In [127]: from pgmpy.inference import VariableElimination
          infer = VariableElimination(model1)
          posterior_p = infer.query(['RainTomorrow'], evidence={'MaxTemp':1,'MinTemp':1})
          print(posterior_p)
          print(find_value_1(posterior_p))

          Finding Elimination Order: : 0%          0/5 [00:00<?, ?it/s]

          Eliminating: Rainfall: 100%          5/5 [00:00<00:00, 186.48it/s]

          +-----------------+---------------------+
          | RainTomorrow    |   phi(RainTomorrow) |
          +=================+=====================+
          | RainTomorrow(1) |              0.6358 |
          +-----------------+---------------------+
          | RainTomorrow(2) |              0.3642 |
          +-----------------+---------------------+
          0.3642

In [120]: pickle.dump(infer, open('model_with_small_ver_auto.pkl', 'wb'))
```

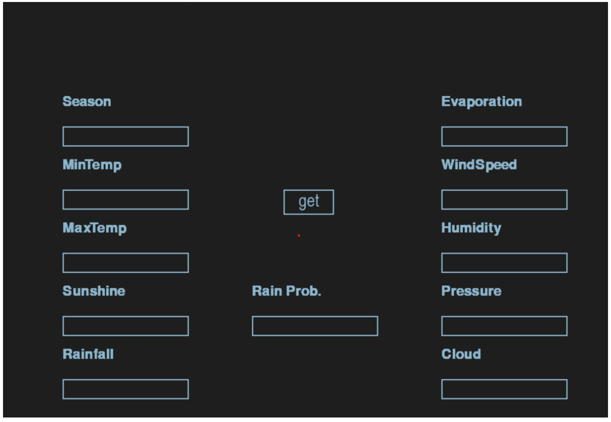
IV. Results and Inferences

we have creacted pygame gui for showing model and get evidence from user and print output which is probability of tomorrow.

We apply structure learning, parameter learning and variable elimination on both models. Top-down reasoning is performed for prediction of Rainfall in this model. Evidence is given to parent nodes and rainfall at different stations is found. Inputs and outputs are in the form of discretized values.

The Bayesian network structure (Bs) is obtained by the Hill climb K2 algorithm for tomorrow's rainfall. The Hill Climb K2 algorithm is very sensitive to the order of nodes given. A wrong order may lead to a false network. A wrong order may lead to a false network.

The second component of the Bayesian network, Condition probabilities (Bp) were computed using Bayesian estimators. Model is tested for different combinations of atmospheric parameters. While testing the both models we get different accuracy of the both models. Manual constructed bayesian network accuracy is 0.7872 And Automatically generated bayesian network accuracy is 0.7916



V. Conclusion

The presented approach to predicting rainfall is significantly capable of the expensive equipment and computer resources. This approach gives less accuracy than highly advanced techniques. Here we are using 11 nodes but increasing the number of nodes can increase the accuracy of the prediction of rainfall. Also, we can increase the atmospheric parameter to increase the accuracy. This study can be used for better water resources management. Further there are scope of improvements, Is rain very high low or moderate that can be determined. For that we need more parameters in data related to it. Currently we have only use hill climb and k-2 as scoring or network structure modifying method. We can also use RSMAX2 (General 2-phase Restricted Maximization) and H2PC (Hybrid HPC) for doing automatic learning for modelling the outcome.

VI. Team Activity Learning and Work distribution

Learned the concept of bayesian network, structure learning, parameter learning various scoring or structure-forming algorithms. Research in all fields to know the concept of the bayesian network.

Vrutik and Manav did research, PPT video editing, and report writing Jay and Nand did Bayesian network models, coding, and report writing.

VII. References

https://towardsdatascience.com/introduction-to-bayesian-belief-networks-c012e3f59f1b

Dataset: https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package

Khabarov, S Shilkina, M Vasiliev, N. (2021). Precipitation forecast based on the Bayesian Network. IOP Conference Series: Earth and Environmental Science. 806. 012016. 10.1088/1755-1315/806/1/012016.

https://en.wikipedia.org/wiki/Climate_of_Australia

A. Nandar, "Bayesian network probability model for weather prediction," 2009 International Conference on the Current Trends in Information Technology (CTIT), 2009, pp. 1-5, doi: 10.1109/CTIT.2009.5423132.