1. Write a program to construct a Bayesian network considering medical data. Use this model to demonstrate the diagnosis of heart patients using standard Heart Disease Data Set. You can use Java/Python ML library classes/API

## Theory

A Bayesian network is a directed acyclic graph in which each edge corresponds to a conditional dependency, and each node corresponds to a unique random variable.

Bayesian network consists of two major parts: a directed acyclic graph and a set of conditional probability distributions

* + The directed acyclic graph is a set of random variables represented by nodes.
  + The conditional probability distribution of a node (random variable) is defined for every possible outcome of the preceding causal node(s).

For illustration, consider the following example. Suppose we attempt to turn on our computer, but the computer does not start (observation/evidence). We would like to know which of the possible causes of computer failure is more likely. In this simplified illustration, we assume only two possible causes of this misfortune: electricity failure and computer malfunction.

The corresponding directed acyclic graph is depicted in below figure.

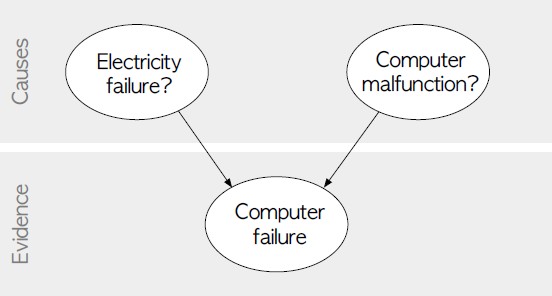


Fig: Directed acyclic graph representing two independent possible causes of a computer failure.

The goal is to calculate the posterior conditional probability distribution of each of the possible unobserved causes given the observed evidence, i.e. P [Cause | Evidence].

## Data Set:

**Title:** Heart Disease Databases

The Cleveland database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "Heartdisease" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4.

Database: 0 1 2 3 4 Total

Cleveland: 164 55 36 35 13 303

# Attribute Information:

1. age: age in years
2. sex: sex (1 = male; 0 = female)
3. cp: chest pain type
   * Value 1: typical angina
   * Value 2: atypical angina
   * Value 3: non-anginal pain
   * Value 4: asymptomatic
4. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
5. chol: serum cholestoral in mg/dl
6. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7. restecg: resting electrocardiographic results
   * Value 0: normal
   * Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
   * Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. thalach: maximum heart rate achieved
9. exang: exercise induced angina (1 = yes; 0 = no)
10. oldpeak = ST depression induced by exercise relative to rest
11. slope: the slope of the peak exercise ST segment
    * Value 1: upsloping
    * Value 2: flat
    * Value 3: downsloping
12. ca = number of major vessels (0-3) colored by flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
14. Heartdisease: It is integer valued from 0 (no presence) to 4. Diagnosis of heart disease (angiographic disease status)

# Some instance from the dataset:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | Heartdisease |
| 63 | 1 | 1 | 145 | 233 | 1 | 2 | 150 | 0 | 2.3 | 3 | 0 | 6 | 0 |
| 67 | 1 | 4 | 160 | 286 | 0 | 2 | 108 | 1 | 1.5 | 2 | 3 | 3 | 2 |
| 67 | 1 | 4 | 120 | 229 | 0 | 2 | 129 | 1 | 2.6 | 2 | 2 | 7 | 1 |
| 41 | 0 | 2 | 130 | 204 | 0 | 2 | 172 | 0 | 1.4 | 1 | 0 | 3 | 0 |
| 62 | 0 | 4 | 140 | 268 | 0 | 2 | 160 | 0 | 3.6 | 3 | 2 | 3 | 3 |
| 60 | 1 | 4 | 130 | 206 | 0 | 2 | 132 | 1 | 2.4 | 2 | 2 | 7 | 4 |

## Program:

import numpy as np import csv

import pandas as pd

from pgmpy.models import BayesianModel

from pgmpy.estimators import MaximumLikelihoodEstimator from pgmpy.inference import VariableElimination

#read Cleveland Heart Disease data heartDisease = pd.read\_csv('heart.csv')

heartDisease = heartDisease.replace('?',np.nan)

#display the data print('Few examples from the dataset are given below') print(heartDisease.head())

#Model Bayesian Network Model=BayesianModel([('age','trestbps'),('age','fbs'),

('sex','trestbps'),('exang','trestbps'),('trestbps','heartdise

ase'),('fbs','heartdisease'),('heartdisease','restecg'),

('heartdisease','thalach'),('heartdisease','chol')])

#Learning CPDs using Maximum Likelihood Estimators print('\n Learning CPD using Maximum likelihood estimators') model.fit(heartDisease,estimator=MaximumLikelihoodEstimator)

# Inferencing with Bayesian Network print('\n Inferencing with Bayesian Network:') HeartDisease\_infer = VariableElimination(model)

#computing the Probability of HeartDisease given Age print('\n 1. Probability of HeartDisease given Age=30') q=HeartDisease\_infer.query(variables=['heartdisease'],evidence

={'age':28})

print(q['heartdisease'])

#computing the Probability of HeartDisease given cholesterol print('\n 2. Probability of HeartDisease given cholesterol=100') q=HeartDisease\_infer.query(variables=['heartdisease'],evidence

={'chol':100})

print(q['heartdisease'])

## Output:

Few examples from the dataset are given below

age sex cp trestbps ...slope ca thal heartdisease 0 63 1 1 145 ... 3 0 6 0

1 67 1 4 160 ... 2 3 3 2

2 67 1 4 120 ... 2 2 7 1

3 37 1 3 130 ... 3 0 3 0

4 41 0 2 130 ... 1 0 3 0

[5 rows x 14 columns]

Learning CPD using Maximum likelihood estimators Inferencing with Bayesian Network:

1. Probability of HeartDisease given Age=28

╒════════════════╤═════════════════════╕

│ heartdisease │ phi(heartdisease) │

╞════════════════╪═════════════════════╡

│ heartdisease\_0 │ 0.6791 │

├────────────────┼─────────────────────┤

│ heartdisease\_1 │ 0.1212 │

├────────────────┼─────────────────────┤

│ heartdisease\_2 │ 0.0810 │

├────────────────┼─────────────────────┤

│ heartdisease\_3 │ 0.0939 │

├────────────────┼─────────────────────┤

│ heartdisease\_4 │ 0.0247 │

╘════════════════╧═════════════════════╛

1. Probability of HeartDisease given cholesterol=100

╒════════════════╤═════════════════════╕

│ heartdisease │ phi(heartdisease) │

╞════════════════╪═════════════════════╡

│ heartdisease\_0 │ 0.5400 │

├────────────────┼─────────────────────┤

│ heartdisease\_1 │ 0.1533 │

├────────────────┼─────────────────────┤

│ heartdisease\_2 │ 0.1303 │

├────────────────┼─────────────────────┤

│ heartdisease\_3 │ 0.1259 │

├────────────────┼─────────────────────┤

│ heartdisease\_4 │ 0.0506 │

╘════════════════╧═════════════════════╛