Data Mining
Classification: Alternative Techniques

Bayesian Belief Network

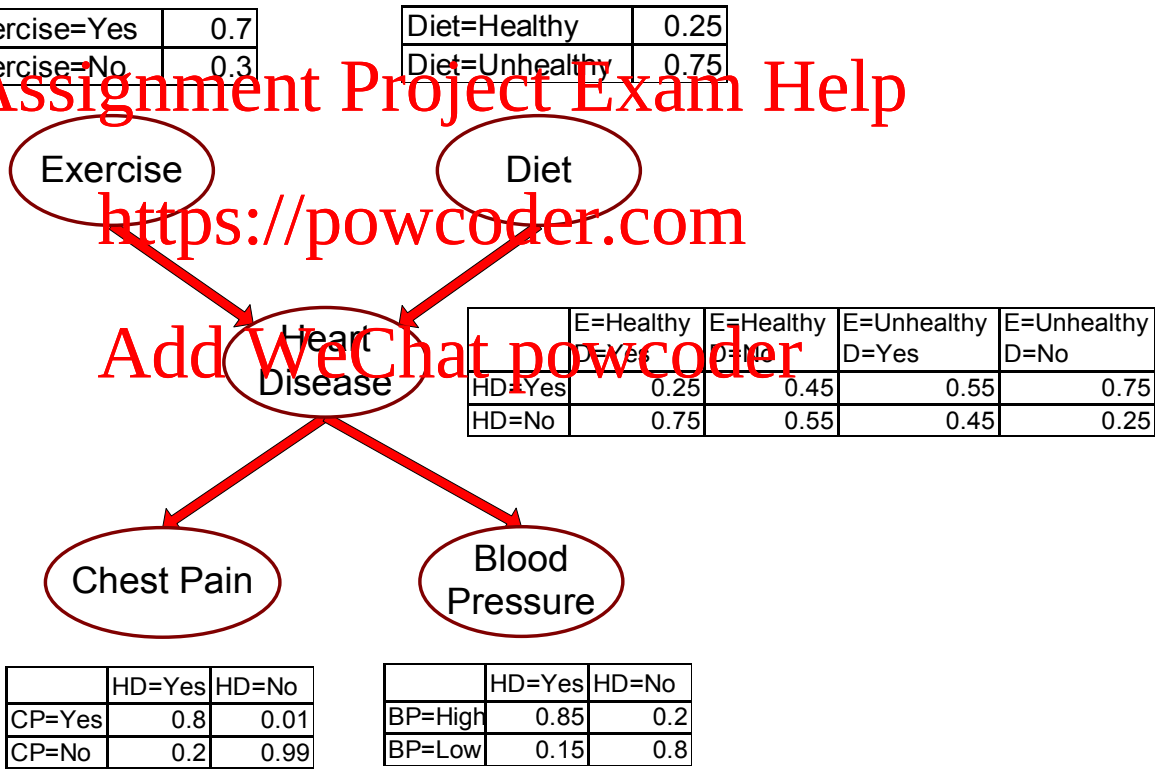
Introduction to Data Mining

by

Tan, Steinbach, Karpatne, Kumar

A motivation example for Bayes belief network

A man exercises regularly and eats health. He recently complained about chest pain. But the blood pressure is normal. The doctor would like to know the chance of him developing the heart disease.



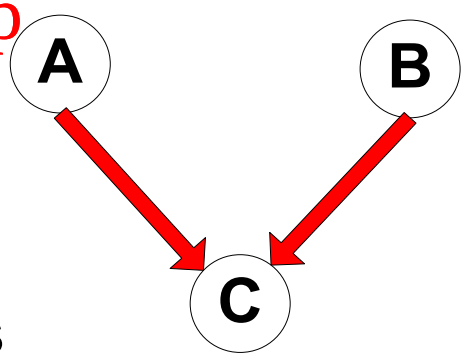
	E=Healthy D=Yes	E=Healthy D=No	E=Unhealthy D=Yes	E=Unhealthy D=No
HD=Yes	0.25	0.45	0.55	0.75
HD=No	0.75	0.55	0.45	0.25

	HD=Yes	HD=No
CP=Yes	0.8	0.01
CP=No	0.2	0.99

	HD=Yes	HD=No
BP=High	0.85	0.2
BP=Low	0.15	0.8

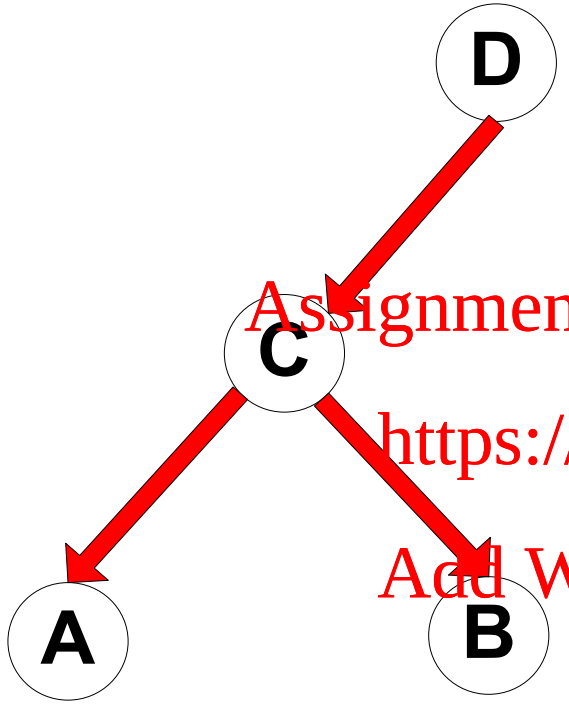
Bayesian Belief Networks

- Provides graphical representation of probabilistic relationships among a set of random variables
- Consists of:
  - A directed acyclic graph (DAG)
    - Node corresponds to a variable
    - Arc corresponds to dependence relationship between a pair of variables



  - A probability table associating each node to its immediate parent

Conditional Independence



D is parent of C

A is child of C

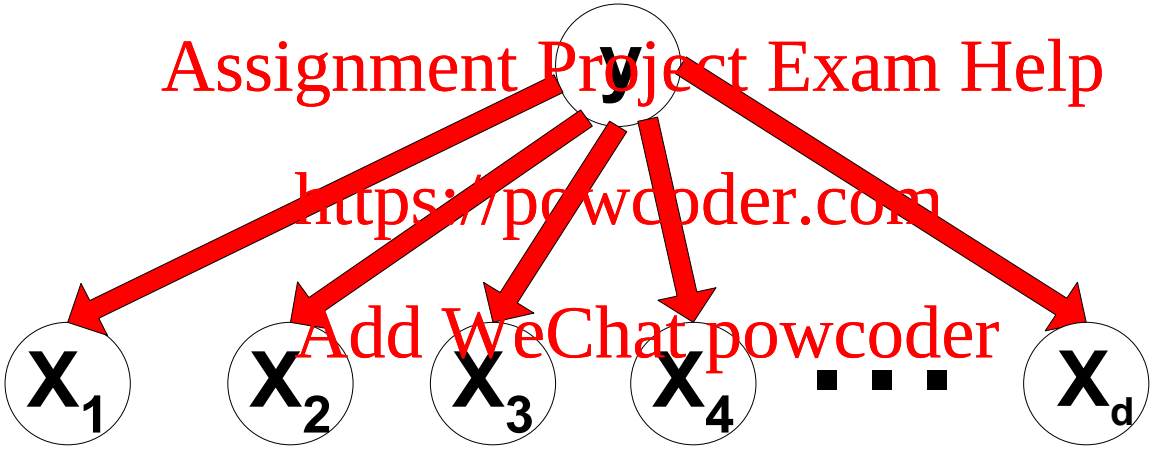
B is descendant of D

D is ancestor of A

- A node in a Bayesian network is conditionally independent of all of its nondescendants, if its parents are known

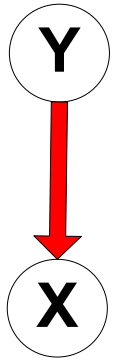
Conditional Independence

- Naïve Bayes assumption:



Probability Tables

- If X does not have any parents, table contains prior probability P(X)



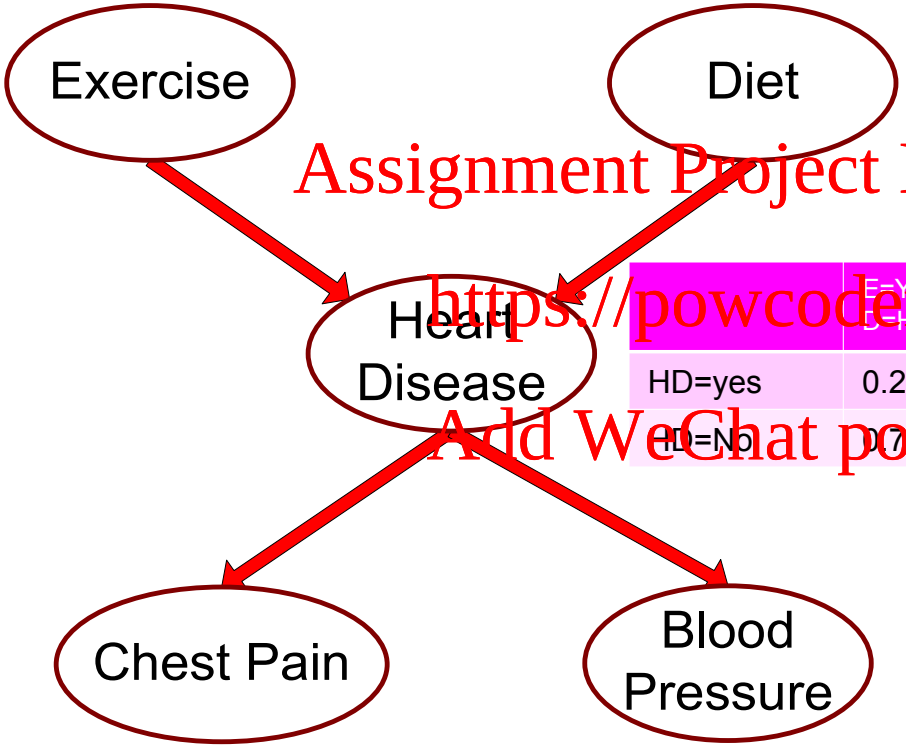
- If X has only one parent (Y), table contains conditional probability P(X|Y)

- If X has multiple parents (Y_1, Y_2,…, Y_k), table contains conditional probability P(X|Y_1, Y_2,…, Y_k)

Example of Bayesian Belief Network

Exercise=Yes	0.7
Exercise=No	0.3

Diet=Healthy	0.25
Diet=Unhealthy	0.75

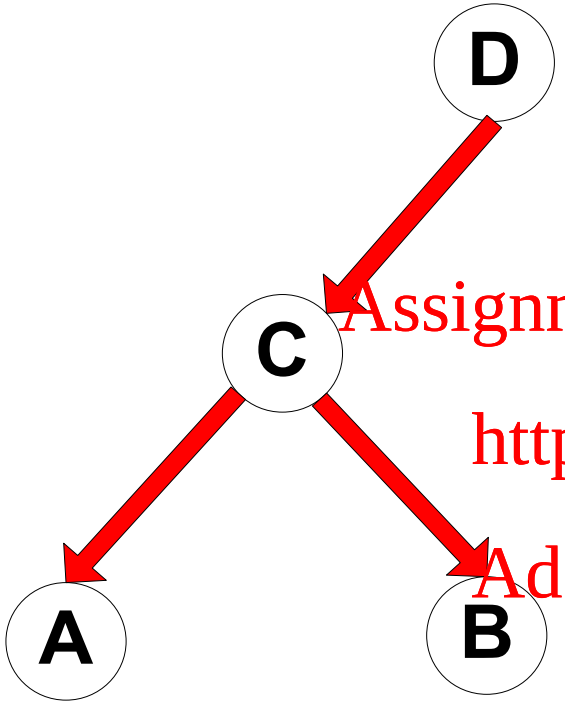


	E=Yes D=Healthy	E=Yes D=Unhealthy	E=No D=healthy	E=No D=unhealthy
HD=yes	0.25	0.45	0.55	0.75
HD=No	0.75	0.55	0.45	0.25

	HD=Yes	HD=No
CP=Yes	0.8	0.01
CP=No	0.2	0.99

	HD=Yes	HD=No
BP=High	0.85	0.2
BP=Low	0.15	0.8

Conditional Independence



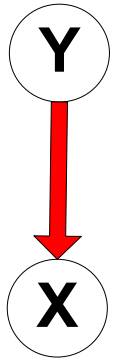
The following statements, true or false?

1. D is conditionally independent of C
2. C is conditionally independent of D
3. A is conditionally independent of C
4. B is conditionally independent of C
5. C is conditionally independent of A
6. C is conditionally independent of B
7. A is conditionally independent of B
8. A is conditionally independent of D
9. B is conditionally independent of D

- A node in a Bayesian network is conditionally independent of all of its nondescendants, if its parents are known

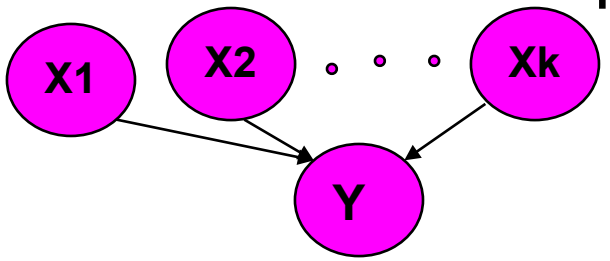
Probability Tables

- If X does not have any parents, table contains prior probability P(X)



- If X has only one parent (Y), table contains conditional probability P(X|Y)

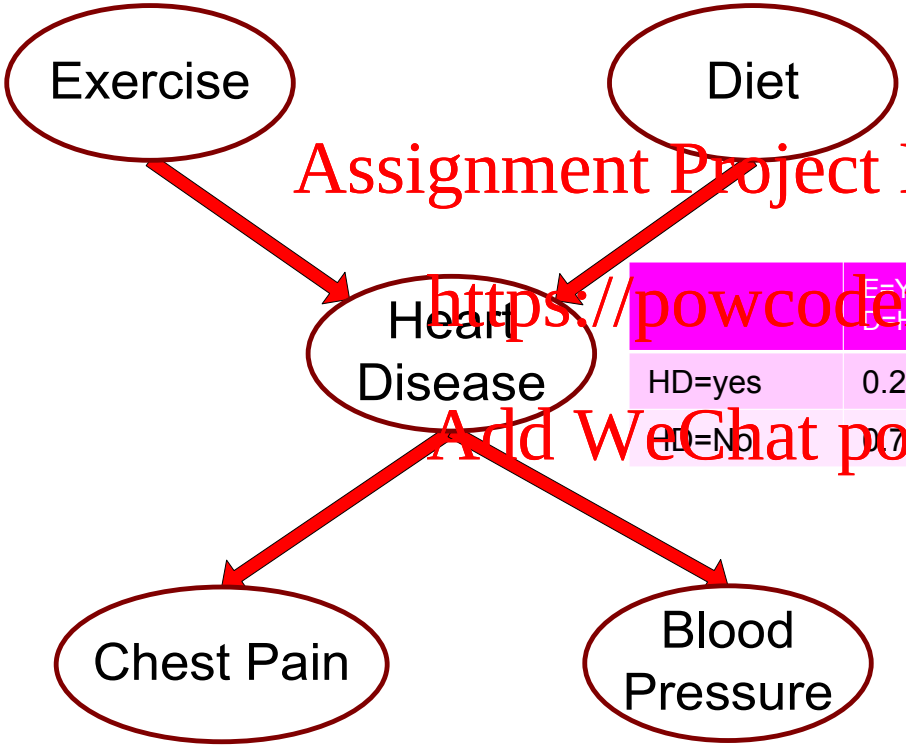
- If Y has multiple parents (X_1, X_2,…, X_k), table contains conditional probability P(Y|X_1, X_2,…, X_k)



Example of Bayesian Belief Network

Exercise=Yes	0.7
Exercise=No	0.3

Diet=Healthy	0.25
Diet=Unhealthy	0.75



	E=Yes D=Healthy	E=Yes D=Unhealthy	E=No D=healthy	E=No D=unhealthy
HD=yes	0.25	0.45	0.55	0.75
HD=No	0.75	0.55	0.45	0.25

	HD=Yes	HD=No
CP=Yes	0.8	0.01
CP=No	0.2	0.99

	HD=Yes	HD=No
BP=High	0.85	0.2
BP=Low	0.15	0.8

Use domain knowledge to construct this network; Causal → effect variables

In-class exercise (to submit)

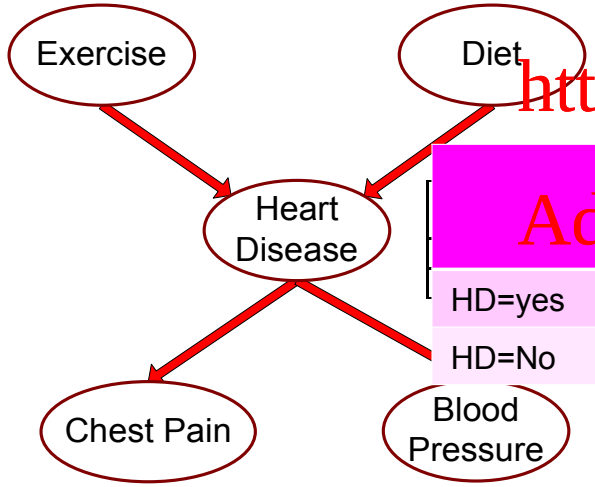
- Based on your knowledge, construct a Bayes belief network for COVID-19 using the following attributes:
  - Chest pain, difficulty of breathing, sore throat, fever, sneeze, coughing, high blood pressure, obesity, age, gender, smoking.
  - Draw the figure and upload the file.

Classification using Bayes Belief Network

- Q1: Can the following equations be simplified?

Exercise=Yes	0.7
Exercise=No	0.3

Diet=Healthy	0.25
Diet=Unhealthy	0.75



	HD=Yes	HD=No
CP=Yes	0.8	0.01
CP=No	0.2	0.99

	HD=Yes	HD=No
BP=High	0.85	0.2
BP=Low	0.15	0.8

P (Heart disease| Exercise, Diet)

P (Chest pain | Heart disease, Exercise, Diet) = P (Chest Pain | Heart Disease)

	E=Yes D=Healthy	E=Yes D=Unhealthy	E=No D=healthy	E=No D=unhealthy
HD=yes	0.25	0.45	0.55	0.75
HD=No	0.75	0.55	0.45	0.25

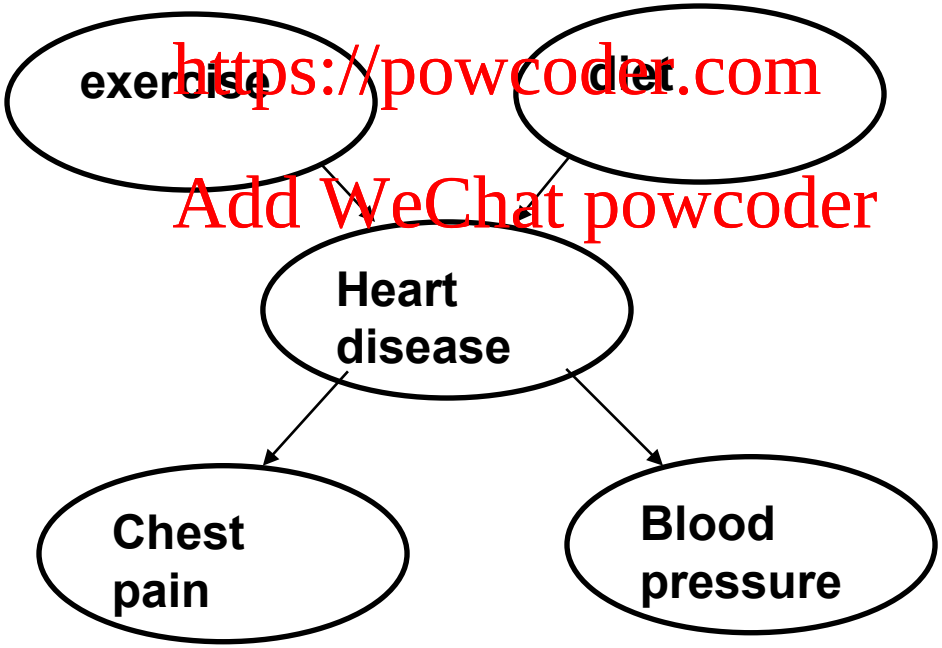
P (Blood Pressure |Heart disease, Exercise, Diet)= P(Blood Pressure| Heart disease)

P (Heart disease | Exercise, Diet, Chest Pain, Blood Pressure) = ? (cannot be simplified)

Classification using Bayes Belief Network

- Q2: Without any prior information, compute whether a person is likely to have heart disease

P(heart disease=yes)



Classification using Bayes Belief Network

- Q2: Using the given probabilities, compute whether a person is likely to have heart disease

<u>P(heart disease=yes)</u>

Apply marginal probability

$P(A)=P(AB_1)+P(AB_2)+\ldots+P(AB_k)$

B_1 to B_k are mutually independent;

B_1 to B_k are all possible outcomes of B

P(heart disease=Yes)

=P(HD=yes,E=Yes,D=healthy)

+P(HD=yes,E=Yes,D=unhealthy)

+P(HD=yes,E=No,D=healthy)

+P(HD=Yes,E=No,D=unhealthy)

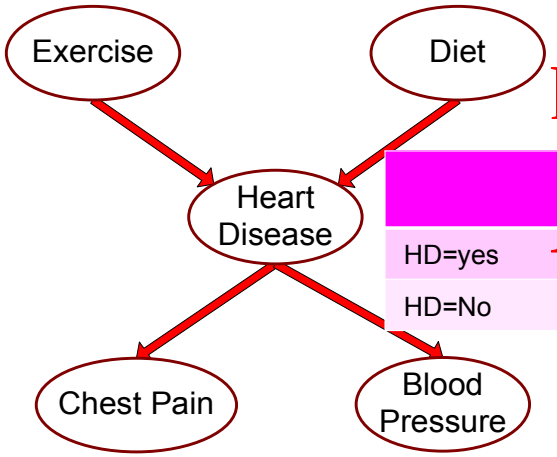
=0.25*0.7*0.25+0.45*0.7*0.75+0.55*0.3*0.25+0.75*0.3*0.75=0.49

Classification using Bayes Belief Network

- Q3: If the person has high blood pressure, what is P(Heart disease = Yes| Blood Pressure = high)

Exercise=Yes	0.7
Exercise=No	0.3

Diet=Healthy	0.25
Diet=Unhealthy	0.75



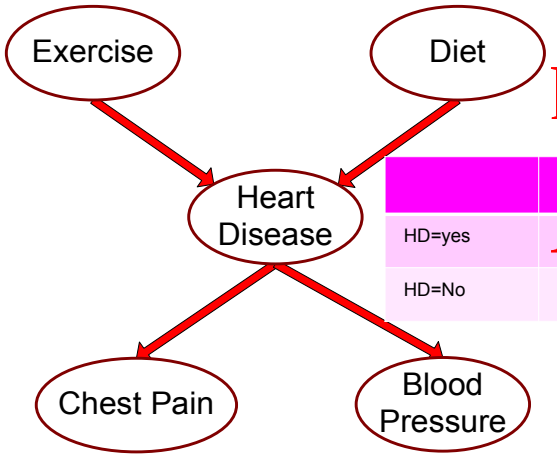
	E=Yes D=Healthy	E=Yes D=Unhealthy	E=No D=healthy	E=No D=unhealthy
HD=yes	0.25	0.45	0.55	0.75
HD=No	0.75	0.55	0.45	0.25

	HD=Yes	HD=No
CP=Yes	0.8	0.01
CP=No	0.2	0.99

	HD=Yes	HD=No
BP=High	0.85	0.2
BP=Low	0.15	0.8

Classification using Bayes Belief Network

- Q3: If the person has high blood pressure, what is P(Heart disease = Yes| Blood Pressure = high)



Exercise=Yes	0.7
Exercise=No	0.3

Diet=Healthy	0.25
Diet=Unhealthy	0.75

	HD=Yes	HD=No
CP=Yes	0.8	0.01
CP=No	0.2	0.99

	HD=Yes	HD=No
BP=High	0.85	0.2
BP=Low	0.15	0.8

P (HD=yes| BP=high)
=P(BP=high | HD=yes)P(HD=yes) /
P(BP=high)
Only unknown: P(BP=high)
Apply marginal probability again:

	E=Yes D=Healthy	E=Yes D=Unhealthy	E=No D=healthy	E=No D=unhealthy
HD=yes	0.25	0.45	0.55	0.75
HD=No	0.75	0.55	0.45	0.25

P(BP=high)=P(BP=high, HD=yes)+P(BP=high, HD=no)
=P(BP=high|HD=yes)P(HD=yes)+P(BP=high|HD=no)P(HD=no)=0.5158

Result=0.8033

Marginal probability is needed in order to get the accurate probability, rather than just "classification"

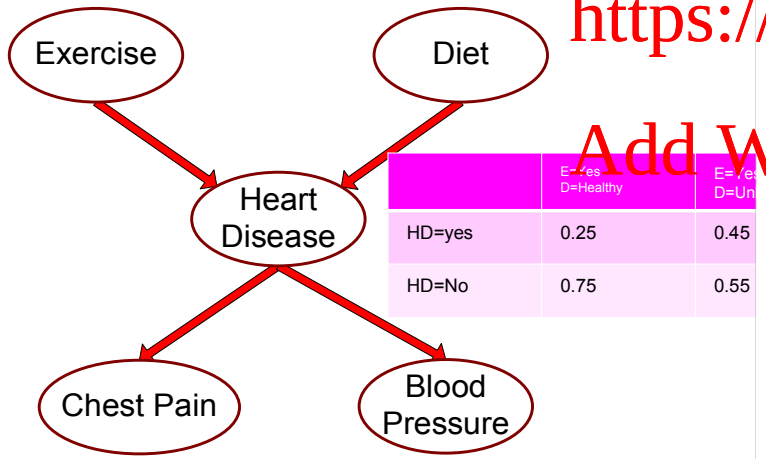
Classification using Bayes Belief Network

- Q4: If the person has high blood pressure, healthy diet, and regular exercise, what is the chance that he/she has heart disease?

P(Heart disease = Yes| Blood Pressure = high, Diet = healthy, Exercise=Yes)

Exercise=Yes	0.7
Exercise=No	0.3

Diet=Healthy	0.25
Diet=Unhealthy	0.75



	E=Yes D=Healthy	E=Yes D=Unhealthy
HD=yes	0.25	0.45
HD=No	0.75	0.55

	HD=Yes	HD=No
CP=Yes	0.8	0.01
CP=No	0.2	0.99

	HD=Yes	HD=No
BP=High	0.85	0.2
BP=Low	0.15	0.8

P(HD=yes| BP=high, D=Yes, E=Yes)
=P(HD=yes,BP=high,D=Yes,E=yes) /
P(BP=high,D=Yes, E=Yes)
= P(BP=high|HD=yes) P(HD=yes|D=Yes, E=Yes)P(D=Yes, E=Yes) /
P(BP=high, D=Yes, E=Yes)

How to compute P(BP=high, D=Yes, E=Yes) ?

Marginal probability again:
Above = P(BP=high,D=Y, E=Y, HD=Yes)+
P(BP=high,D=Y, E=Y, HD=No)

Result: 0.5862

P(HD=yes, BP=high, D=Yes (healthy), E=Yes)=P(BP=high|HD=Yes,D=Yes,E=Yes)P(HD=Yes,D=Yes,E=Yes)
=P(BP=high|HD=Yes)P(HD=Yes,D=Yes,E=Yes)…

Classification using Bayes Belief Network

Compute:
1.P(B=good, F=empty, G=empty, S=yes)

2.Given that the battery is bad, compute the probability that the car can still start.

In-class exercise (no need to submit)

Classification using Bayes Belief Network

In-class exercise

Compute:
1.P(B=good, F=empty, G=empty, S=yes)
This is a joint probability. You can apply chain rule and then add simplifications based on the network (conditional independence). If you are familiar with this process, you can also directly decompose this into product of given conditional probabilities and priors.
P(B=good,F=empty,G=empty,S=yes)=P(S=yes|B=good,F=empty)P(G=empty|B=good,F=empty)P(B=good)P(F=empty)
=(1-0.8)*0.8*(1-0.1)*0.2.

Note: ”Start” is conditionally independent of Gauge given Batter and Fuel

More details:
P(B=good, F=empty, G=empty, S=yes)
=P(S=yes|B=good, F=empty, G=empty)
P(B=good, F=empty, G=empty)
=P(S=yes|B=good, F=empty)
P(G=empty|B=good,F=empty)P(B=good,F=empty)
=….

Classification using Bayes Belief Network

In-class exercise

2. Given that the battery is bad, compute the probability that the car can still start.

First, write down the equation in probability: P(S=yes|B=bad)
Node S (start) has two parents: Battery and Fuel. Thus, you need to add Fuel to this equation, using **marginalization.** To avoid any mistakes, it is safe to convert this into joint probability and then sum over "Fuel".
P(S=yes|B=bad)=P(S=yes,B=bad)/ P(B=bad)

P(S=yes,B=bad)=
P(S=yes,B=bad,F=empty)+
P(S=yes,B=bad,F=not empty)
=P(S=yes|
B=bad,F=empty)P(B=bad)P(F=e
mpty)+P(S=yes|B=bad,F=not
empty)P(B=bad)P(F=not empty)
=0+0.1**0.1*0.8