COMP7/8118 M50

Data Mining

Bayesian Classifier

Xiaofei Zhang

Slides compiled from Jiawei Han and Raymond C.W. Wong's work

THE UNIVERSITY OF MEMPHIS

Bayes Classifier

- A probabilistic framework for solving classification problems
- **A, C** random variables
- Joint probability: **Pr(A=a,C=c)**
- Conditional probability: **Pr(C=c | A=a)**
- Relationship between joint and conditional probability distributions

$$\Pr(C, A) = \Pr(C \mid A) \times \Pr(A) = \Pr(A \mid C) \times \Pr(C)$$

- **Bayes Theorem**:

$$P(C \mid A) = \frac{P(A \mid C) P(C)}{P(A)}$$

Bayesian Classifiers

- How to classify the new record X = (‘Yes’, ‘Single’, 80K)

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Find the class with the highest probability given the vector values.

Maximum Aposteriori Probability estimate:

- Find the value c for class C that maximizes P(C=c| X)

How do we estimate P(C|X) for the different values of C?

- We want to estimate P(C=Yes| X)
- and P(C=No| X)

Bayesian Classifiers

- In order for probabilities to be well defined:
  - Consider each attribute and the class label as random variables
  - Probabilities are determined from the data

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Evade C
Event space: {Yes, No}
$P(C) = (0.3, 0.7)$

Refund A_1
Event space: {Yes, No}
$P(A_1) = (0.3,0.7)$

Marital Status A_2
Event space: {Single, Married, Divorced}
$P(A_2) = (0.4,0.4,0.2)$

Taxable Income A_3
Event space: R
$P(A_3) \sim Normal(\mu,\sigma^2)$
$\mu = 104$:sample mean, $\sigma^2=1874$:sample var

Bayesian Classifiers

- Approach:
  - compute the posterior probability $P(C \mid A_1, A_2, \ldots, A_n)$ using the Bayes theorem

$$\boxed{P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}}$$

  - Maximizing $P(C \mid A_1, A_2, \ldots, A_n)$ is equivalent to maximizing $P(A_1, A_2, \ldots, A_n \mid C)\ P(C)$
    - The value $P(A_1, \ldots, A_n)$ is the same for all values of C.

- How to estimate $P(A_1, A_2, \ldots, A_n \mid C)$?

Naïve Bayes Classifier

- Assume conditional independence among attributes A_i when class C is given:
  - $P(A_1, A_2, \ldots, A_n|C) = P(A_1|C)\, P(A_2|C) \cdots P(A_n|C)$
  - We can estimate $P(Ai|\, C)$ from the data.
  - New point $X = (A_1 = \alpha_1, \ldots A_n = \alpha_n)$ is classified to class c if

$$P(C = c|X) \sim P(C = c) \prod_i P(A_i = \alpha_i|c)$$

is maximum over all possible values of C.

Example

- Record
  $X = (Refund = Yes, Status = Single, Income = 80K)$
- For the class C = 'Evade', we want to compute:
  $P(C = Yes|X)$ and $P(C = No|X)$
- We compute:
  - $P(C = Yes|X) = P(C = Yes)*P(Refund = Yes\ |C = Yes)$
    $*P(Status = Single\ |C = Yes)$
    $*P(Income = 80K\ |C = Yes)$
  - $P(C = No|X) = P(C = No)*P(Refund = Yes\ |C = No)$
    $*P(Status = Single\ |C = No)$
    $*P(Income = 80K\ |C = No)$

How to Estimate Probabilities from Data?

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Class Prior Probability:

$$P(C = c) = \frac{N_c}{N}$$

N_c: Number of records with class c

N = Number of records

P(C = No) = 7/10

P(C = Yes) = 3/10

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Discrete attributes:

$$P(A_i = a | C = c) = \frac{N_{a,c}}{N_c}$$

$N_{a,c}$: number of instances having attribute $A_i = a$ and belong to class c

N_c: number of instances of class c

How to Estimate Probabilities from Data?

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Discrete attributes:

$$P(A_i = a|C = c) = \frac{N_{a,c}}{N_c}$$

$N_{a,c}$: number of instances having attribute $A_i = a$ and belong to class c

N_c: number of instances of class c

P(Refund = Yes|No) = 3/7

How to Estimate Probabilities from Data?

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Discrete attributes:

$$P(A_i = a | C = c) = \frac{N_{a,c}}{N_c}$$

$N_{a,c}$: number of instances having attribute $A_i = a$ and belong to class c

N_c: number of instances of class c

P(Refund = Yes|Yes) = 0

How to Estimate Probabilities from Data?

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Discrete attributes:

$$P(A_i = a|C = c) = \frac{N_{a,c}}{N_c}$$

$N_{a,c}$: number of instances having attribute $A_i = a$ and belong to class c

N_c: number of instances of class c

P(Status=Single|No) = 2/7

How to Estimate Probabilities from Data?

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Discrete attributes:

$$P(A_i = a | C = c) = \frac{N_{a,c}}{N_c}$$

$N_{a,c}$: number of instances having attribute $A_i = a$ and belong to class c

N_c: number of instances of class c

P(Status=Single|Yes) = 2/3

How to Estimate Probabilities from Data?

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

- Normal distribution:

$$P(A_i = a \mid c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(a-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

  - One for each (a, c) pair

- For Class=No
  - sample mean $\mu = 110$
  - sample variance $\sigma^2 = 2975$
- For Income = 80

$$P(Income = 80 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(80-110)^2}{2(2975)}} = 0.0062$$

How to Estimate Probabilities from Data?

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Normal distribution:

$$P(A_i = a \mid c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(a-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

  - One for each (a, c) pair

- For Class=Yes
  - sample mean $\mu = 90$
  - sample variance $\sigma^2 = 2975$

- For Income = 80

$$P(Income = 80 \mid Yes) = \frac{1}{\sqrt{2\pi}(5)} e^{-\frac{(80-90)^2}{2(25)}} = 0.01$$

Example

- Record

  X = (Refund = Yes, Status = Single, Income =80K)

- We compute:
  - P(C = Yes|X) = P(C = Yes)*P(Refund = Yes |C = Yes)
    *P(Status = Single |C = Yes)
    *P(Income =80K |C= Yes)
    = 3/10* 0 * 2/3 * 0.01 = 0
  - P(C = No|X) = P(C = No)*P(Refund = Yes |C = No)
    *P(Status = Single |C = No)
    *P(Income =80K |C= No)
    = 7/10 * 3/7 * 2/7 * 0.0062 = 0.0005

Example of Naïve Bayes Classifier

- Creating a Naïve Bayes Classifier, essentially means to compute counts:

Total number of records: N = 10

Class No:
Number of records: 7
Attribute Refund:
- Yes: 3
- No: 4

Attribute Marital Status:
- Single: 2
- Divorced: 1
- Married: 4

Attribute Income:
- mean: 110
- variance: 2975

Class Yes:
Number of records: 3
Attribute Refund:
- Yes: 0
- No: 3

Attribute Marital Status:
- Single: 2
- Divorced: 1
- Married: 0

Attribute Income:
- mean: 90
- variance: 25

Example of Naïve Bayes Classifier

Given a Test Record:

X = (Refund = Yes, Status = Single, Income =80K)

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1

P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

- P(X|Class=No) = P(Refund=Yes|Class=No) × P(Married| Class=No) × P(Income=120K| Class=No) = 3/7 * 2/7 * 0.0062 = 0.00075

- P(X|Class=Yes) = P(Refund=No| Class=Yes) × P(Married| Class=Yes) × P(Income=120K| Class=Yes) = 0 * 2/3 * 0.01 = 0

- P(No) = 0.3, P(Yes) = 0.7

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)

=> Class = No

Naïve Bayes Classifier

- If one of the conditional probabilities is zero, then the entire expression becomes zero

- Laplace Smoothing:

$$P(A_i = a | C = c) = \frac{N_{ac} + 1}{N_c + N_i}$$

  - N_i: number of attribute values for attribute A_i

Example of Naïve Bayes Classifier

Given a Test Record:

With Laplace Smoothing

X = (Refund = Yes, Status = Single, Income =80K)

naive Bayes Classifier:

P(Refund=Yes|No) = 4/9
P(Refund=No|No) = 5/9
P(Refund=Yes|Yes) = 1/5
P(Refund=No|Yes) = 4/5

P(Marital Status=Single|No) = 3/10
P(Marital Status=Divorced|No)=2/10
P(Marital Status=Married|No) = 5/10
P(Marital Status=Single|Yes) = 3/6
P(Marital Status=Divorced|Yes)=2/6
P(Marital Status=Married|Yes) = 1/6

For taxable income:

If class=No: sample mean=110
sample variance=2975

If class=Yes: sample mean=90
sample variance=25

- P(X|Class=No) = P(Refund=No|Class=No) $\times$ P(Married| Class=No) $\times$ P(Income=120K| Class=No) = 4/9 $\times$ 3/10 $\times$ 0.0062 = 0.00082
- P(X|Class=Yes) = P(Refund=No| Class=Yes) $\times$ P(Married| Class=Yes) $\times$ P(Income=120K| Class=Yes) = 1/5 $\times$ 3/6 $\times$ 0.01 = 0.001

- P(No) = 0.7, P(Yes) = 0.3
- P(X|No)P(No) = 0.0005
- P(X|Yes)P(Yes) = 0.0003

=> Class = No

Implementation details

- Computing the conditional probabilities involves multiplication of many very small numbers
  - Numbers get very close to zero, and there is a danger of numeric instability
- We can deal with this by computing the logarithm of the conditional probability

$$
\begin{aligned}
\log P(C|A) &\sim \log P(A|C) + \log P(C) \\
&= \sum_i \log P(A_i|C) + \log P(C)
\end{aligned}
$$

Naïve Bayes for Text Classification

- Naïve Bayes is commonly used for text classification
- For a document with k terms $d = (t_1, \ldots, t_k)$

$$P(c|d) = P(c)P(d|c) = P(c)\prod_{t_i \in d} P(t_i|c)$$

Fraction of documents in c

- $P(t_i|c)$= Fraction of terms from all documents in c that are t_i.

Number of times t_i appears in all documents in c

$$P(t_i|c) = \frac{N_{ic} + 1}{N_c + T}$$

Laplace Smoothing

Number of unique words (vocabulary size)

Total number of terms in all documents in c

- Easy to implement and works relatively well
- Limitation: Hard to incorporate additional features (beyond words).
    - E.g., number of adjectives used.

Example

News titles for Politics and Sports

	Politics	Sports
documents	"Obama meets Merkel" "Obama elected again" "Merkel visits Greece again"	"OSFP European basketball champion" "Miami NBA basketball champion" "Greece basketball coach?"
	P(p) = 0.5	P(s) = 0.5
terms Vocabulary size: 14	obama:2, meets:1, merkel:2, elected:1, again:2, visits:1, greece:1	OSFP:1, european:1, basketball:3, champion:2, miami:1, nba:1, greece:1, coach:1
	Total terms: 10	Total terms: 11

New title: X = "Obama likes basketball"

P(Politics|X) ~ P(p)*P(obama|p)*P(likes|p)*P(basketball|p)
= 0.5 * 3/(10+14) *1/(10+14) * 1/(10+14) = 0.000108

P(Sports|X) ~ P(s)*P(obama|s)*P(likes|s)*P(basketball|s)
= 0.5 * 1/(11+14) *1/(11+14) * 4/(11+14) = 0.000128

Naïve Bayes (Summary)

- Robust to isolated noise points
- Handle missing values by ignoring the instance during probability estimate calculations
- Robust to irrelevant attributes
- Independence assumption may not hold for some attributes
    - Use other techniques such as Bayesian Belief Networks (BBN)
- Naïve Bayes can produce a probability estimate, but it is usually a very biased one
    - Logistic Regression is better for obtaining probabilities.

Bayesian Belief Network

- Naïve Bayes Classifier
  - Independent Assumption
- Bayesian Belief Network
  - Do not have independent assumption

Bayesian Belief Network

Yes/No | Healthy/ Unhealthy | Yes/No | High/Low | Yes/No | Yes/No

Exercise	Diet	Heartburn	Blood Pressure	Chest Pain	Heart Disease
Yes	Healthy	No	High	Yes	No
No	Unhealthy	Yes	Low	Yes	No
No	Healthy	Yes	High	No	Yes
…	…	…	…	…	…

Some attributes are dependent on other attributes.

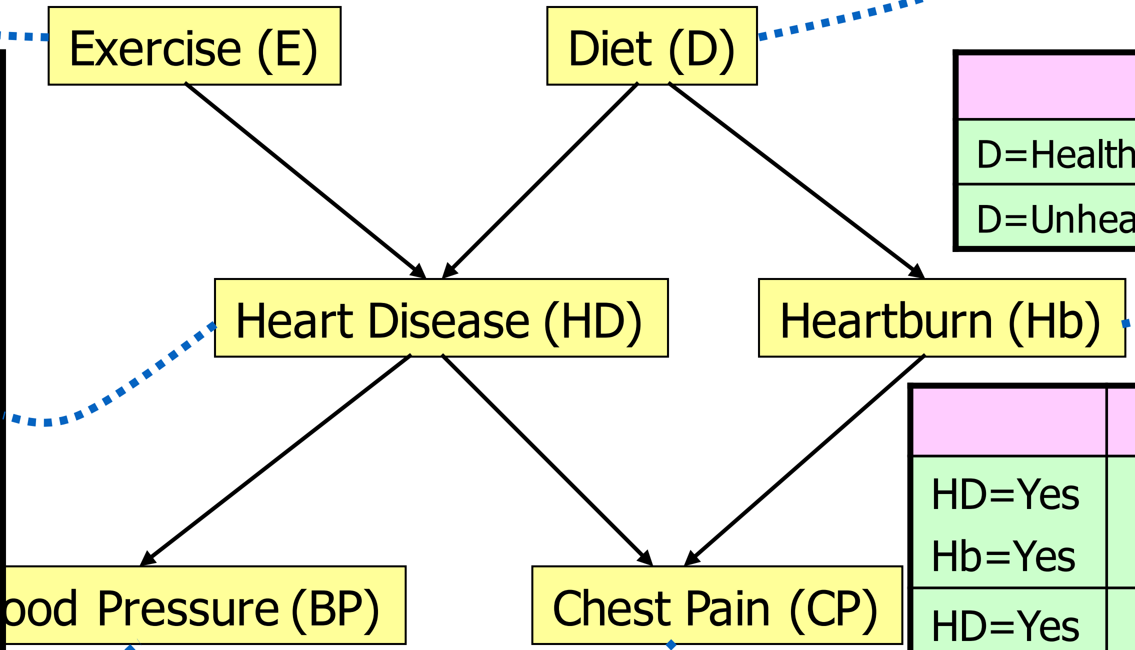
e.g., doing exercises may reduce the probability of suffering from Heart Disease

$$\begin{tikzcd} \text{Exercise (E)} \arrow[d] \\ \text{Heart Disease} \end{tikzcd}$$

Bayesian Belief Network

E = Yes
0.7

D = Healthy
0.25



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

	Hb=Yes
D=Healthy	0.85
D=Unhealthy	0.2

	BP=High
HD=Yes	0.85
HD=No	0.2

	CP=Yes
HD=Yes Hb=Yes	0.8
HD=Yes Hb=No	0.6
HD=No Hb=Yes	0.4
HD=No Hb=No	0.1

Bayesian Belief Network

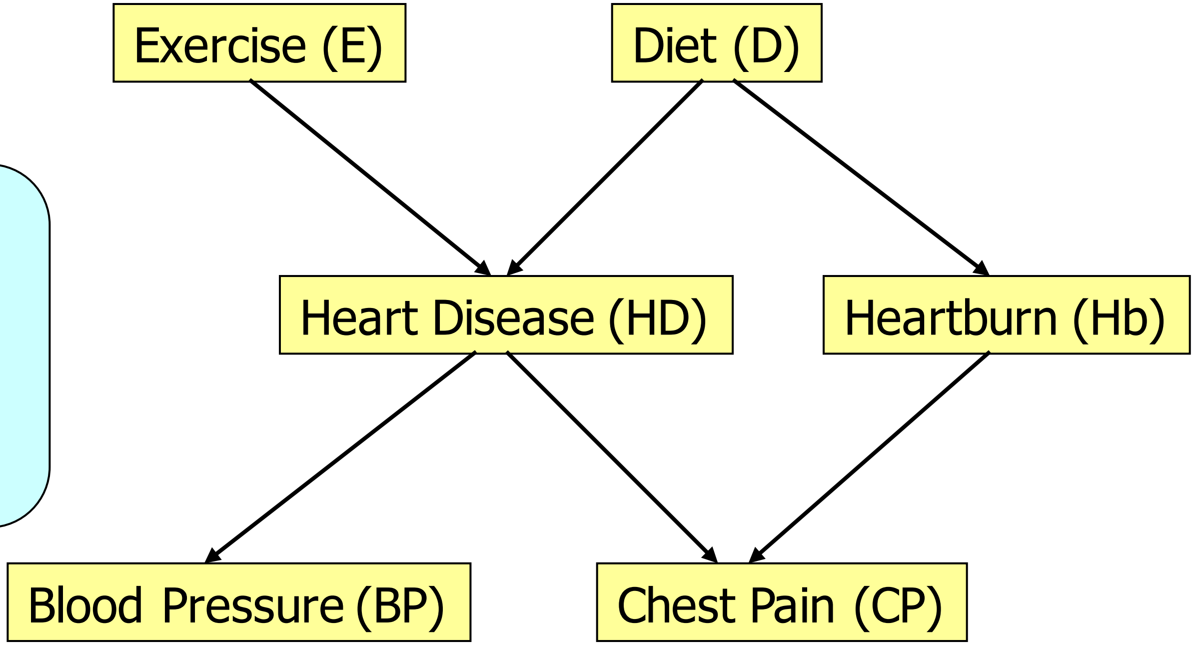
Let X, Y, Z be three random variables.
X is said to be **conditionally independent** of Y given Z if P(X | Y, Z) = P(X | Z)

Lemma:
If X is conditionally independent of Y given Z, P(X, Y | Z) = P(X | Z) x P(Y | Z)

Bayesian Belief Network

Let X, Y, Z be three random variables. X is said to be **conditionally independent** of Y given Z if P(X | Y, Z) = P(X | Z)

Property: A node is **conditionally independent** of its non-descendants if its parents are known.



e.g., P(BP = High | HD = Yes, D = Healthy) = P(BP = High | HD = Yes)

“BP = High” is **conditionally independent** of “D = Healthy” given “HD = Yes”

e.g., P(BP = High | HD = Yes, CP=Yes) = P(BP = High | HD = Yes)

“BP = High” is **conditionally independent** of “CP = Yes” given “HD = Yes”

Bayesian Belief Network

Exercise	Diet	Heartburn	Blood Pressure	Chest Pain	Heart Disease
Yes/No	Healthy/ Unhealthy	Yes/No	High/Low	Yes/No	Yes/No
Yes	Healthy	No	High	Yes	No
No	Unhealthy	Yes	Low	Yes	No
No	Healthy	Yes	High	No	Yes
…	…	…	…	…	…

Suppose there is a new person and I want to know whether he is likely to have Heart Disease.

Exercise	Diet	Heartburn	Blood Pressure	Chest Pain	Heart Disease
?	?	?	?	?	?

Exercise	Diet	Heartburn	Blood Pressure	Chest Pain	Heart Disease
?	?	?	High	?	?

Exercise	Diet	Heartburn	Blood Pressure	Chest Pain	Heart Disease
Yes	Healthy	?	High	?	?

Bayesian Belief Network

Suppose there is a new person and I want to know whether he is likely to have Heart Disease.

Exercise	Diet	Heartburn	Blood Pressure	Chest Pain	Heart Disease
?	?	?	?	?	?

$$P(HD = Yes) = \sum_{x \in \{Yes, No\}} \sum_{y \in \{Healthy, Unhealthy\}} P(HD=Yes|E=x, D=y) \times P(E=x, D=y)$$

$$= \sum_{x \in \{Yes, No\}} \sum_{y \in \{Healthy, Unhealthy\}} P(HD=Yes|E=x, D=y) \times P(E=x) \times P(D=y)$$

$$= 0.25 \times 0.7 \times 0.25 + 0.45 \times 0.7 \times 0.75 + 0.55 \times 0.3 \times 0.25 + 0.75 \times 0.3 \times 0.75$$

$$= 0.49$$

$$P(HD = No) = 1- P(HD = Yes)$$

$$= 1- 0.49$$

$$= 0.51$$

Bayesian Belief Network

Suppose there is a new person and I want to know whether he is likely to have Heart Disease.

Exercise	Diet	Heartburn	Blood Pressure	Chest Pain	Heart Disease
?	?	?	High	?	?

$$P(BP = High) = \sum_{x \in \{Yes,\ No\}} P(BP = High|HD=x) \times P(HD = x)$$

$$= 0.85x0.49 + 0.2x0.51$$

$$= 0.5185$$

$$P(HD = Yes|BP = High) = \frac{P(BP = High|HD=Yes) \times P(HD = Yes)}{P(BP = High)}$$

$$= \frac{0.85 \times 0.49}{0.5185}$$

$$= 0.8033$$

$$P(HD = No|BP = High) = 1 - P(HD = Yes|BP = High)$$

$$= 1 - 0.8033$$

$$= 0.1967$$

Bayesian Belief Network

Suppose there is a new person and I want to know whether he is likely to have Heart Disease.

Exercise	Diet	Heartburn	Blood Pressure	Chest Pain	Heart Disease
Yes	Healthy	?	High	?	?

$$P(HD = Yes \mid BP = High, D = Healthy, E = Yes)$$

$$= \frac{P(BP = High \mid HD = Yes, D = Healthy, E = Yes)}{P(BP = High \mid D = Healthy, E = Yes)} \times P(HD = Yes \mid D = Healthy, E = Yes)$$

$$= \frac{P(BP = High \mid HD = Yes)\, P(HD = Yes \mid D = Healthy, E = Yes)}{\sum_{x \in \{Yes, No\}} P(BP{=}High \mid HD{=}x)\ P(HD{=}x \mid D{=}Healthy, E{=}Yes)}$$

$$= \frac{0.85 \mathrm{x} 0.25}{0.85 \mathrm{x} 0.25 + 0.2 \mathrm{x} 0.75} = 0.5862$$

$$P(HD = No \mid BP = High, D = Healthy, E = Yes)$$

$$= 1 - P(HD = Yes \mid BP = High, D = Healthy, E = Yes)$$

$$= 1 - 0.5862 = 0.4138$$