

# Data Mining

## Classification: Alternative Techniques

---

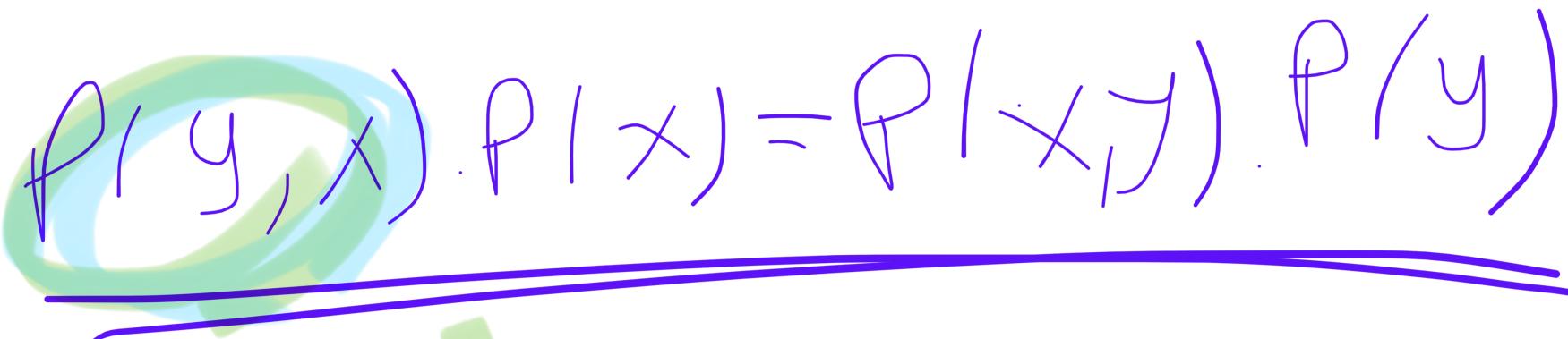
---

### Bayesian Classifiers

Introduction to Data Mining, 2<sup>nd</sup> Edition

by

Tan, Steinbach, Karpatne, Kumar

$$P(y|x) \cdot P(x) = P(x|y) \cdot P(y)$$


# Bayes Classifier

Let  $X$  and  $Y$  be a pair of random variables.

- Sınıflandırma problemlerini çözmek için olasılıksal bir yaklaşım joint probability,  $P(X = x, Y = y)$ ,  $X$  değişkeninin  $x$  değerini alması ve  $Y$  değişkeninin  $y$  değerini alması olasılığını ifade eder.
- Conditional Probability:

Koşullu olasılık (conditional probability), başka bir rastgele değişkenin sonucunun bilindiği göz önüne alındığında, rastgele bir değişkenin belirli bir değeri olması olasılığıdır. Örneğin, koşullu olasılık  $P(Y = y | X = x)$ ,  $X$  değişkeninin  $x$  değerine sahip olduğu gözlendiğinde,  $Y$  değişkeninin  $y$  değerini alma olasılığını ifade eder.

$$P(Y | X) = \frac{P(X, Y)}{P(X)}$$

$$P(X | Y) = \frac{P(X, Y)}{P(Y)}$$

- Bayes theorem:

Sınıfların önceki bilgilerini (prior knowledge) verilerden toplanan yeni kanıtlarla (new evidence gathered from data) birleştirmek için istatistiksel bir ilke

$$P(Y | X) = \frac{P(X | Y)P(Y)}{P(X)}$$

152 Nell  
D. I. . .

# Example of Bayes Theorem

19/0

- Verilen:

- Bir doktor, menenjitin% 50 oranında boyun tutulmasına (*stiff neck*) neden olduğunu bilir.
- Menenjit olan herhangi bir hastanın önsel olasılığı  $1 / 50.000$ 'dir (*Prior probability*)
- Boyun tutulması olan herhangi bir hastanın önsel olasılığı  $1 / 20$ 'dir (*Prior probability*)

Önsel olasılık(*Prior probability*), Bayesci İstatistikte gözlemlere atıf yapmadan önce değerlendirdilen özellikle öznel olabilen olasılıktır. Tecrübeye dayalı olasılık olarak da adlandırılır.

- Bir hastanın boynu tutulmuşsa menenjit olma olasılığı nedir?

$$P(M | S) = \frac{P(S | M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

# Using Bayes Theorem for Classification

- Her bir niteliği ve sınıf etiketini rastgele değişkenler olarak düşünün
- Öznitelikleri ( $X_1, X_2, \dots, X_d$ ) olan bir kayıt verildiğinde
  - Amaç,  $Y$  sınıfını tahmin etmektir
  - Tam olarak şu ifadeyi maksimize eden  $Y$  değerini bulmak istiyoruz :  $P(Y|X_1, X_2, \dots, X_d)$
- Doğrudan verilerden  $P(Y|X_1, X_2, \dots, X_d)$  tahmin edebilir miyiz?

Bu koşullu olasılık, önsel olasılık (**prior probability**)  $P(Y)$  'nin aksine,  $Y$  için sonsal olasılık (**Posterior probability**) olarak da bilinir.

# Example Data

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120K)$$

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

- Can we estimate  
 $P(\text{Evade} = \text{Yes} | X)$  and  $P(\text{Evade} = \text{No} | X)$ ?

In the following we will replace

Evade = Yes by Yes, and

Evade = No by No

$$P(y, x) = \underbrace{P(x, y)}_{P(x)} \times P(y)$$

$$P(x)$$

# Using Bayes Theorem for Classification

- Approach:

- compute posterior probability  $P(Y | X_1, X_2, \dots, X_d)$  using the Bayes theorem

$$P(Y | X_1 X_2 \dots X_n) = \frac{P(X_1 X_2 \dots X_d | Y) P(Y)}{P(X_1 X_2 \dots X_d)}$$

- Maximum a-posteriori*: Choose  $Y$  that maximizes

$$P(Y | X_1, X_2, \dots, X_d)$$

En büyük artçıl (*Maximum a-posteriori*)

- Equivalent to choosing value of  $Y$  that maximizes  $P(X_1, X_2, \dots, X_d | Y) P(Y)$

- How to estimate  $P(X_1, X_2, \dots, X_d | Y)$ ?

# Example Data

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

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

## Using Bayes Theorem:

- $P(\text{Yes} | X) = \frac{P(X | \text{Yes})P(\text{Yes})}{P(X)}$
- $P(\text{No} | X) = \frac{P(X | \text{No})P(\text{No})}{P(X)}$
- How to estimate  $P(X | \text{Yes})$  and  $P(X | \text{No})$ ?

# Naïve Bayes Classifier

Af! /

- Sınıf verildiğinde,  $X_i$  nitelikleri arasında bağımlılık olmadığını (*independence*) varsayıncı:
  - $P(X_1, X_2, \dots, X_d | Y_j) = P(X_1 | Y_j) P(X_2 | Y_j) \dots P(X_d | Y_j)$
  - Artık eğitim verilerinden tüm  $X_i$  ve  $Y_j$  kombinasyonları için  $P(X_i | Y_j)$  tahmin edebiliriz
  - $P(Y_j) \prod P(X_i | Y_j)$ , maksimum ise yeni nokta  $Y_j$  olarak sınıflandırılır.

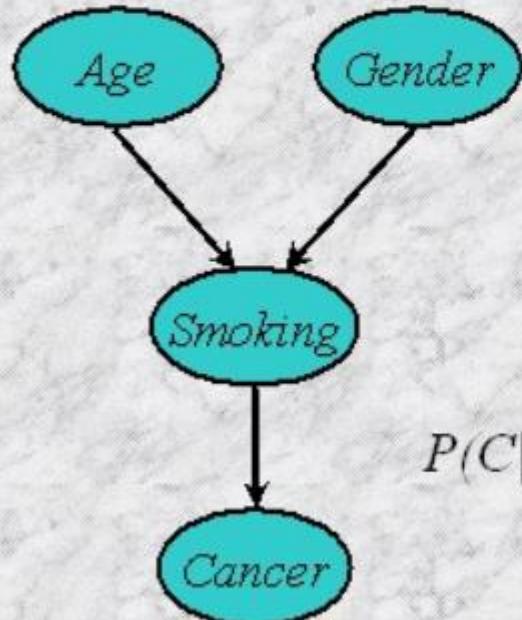
Kosullu Boşluk MS(2) / K

# Conditional Independence

- |  $P(X|YZ) = P(X|Z)$  ise  $X$  and  $Y$  koşullu olarak bağımsızdır (**conditionally independent**)
- | Örnek: Kol uzunluğu ve okuma becerileri
- | Çocuklar, yetişkinlere kıyasla daha kısa kol uzunluğuna ve sınırlı okuma becerisine sahiptir.
  - Yaş sabitse, kol uzunluğu ile okuma becerileri arasında belirgin bir ilişki yok
  - Kol uzunluğu ve okuma becerileri, yaşa göre koşullu olarak bağımsızdır

# Conditional Independence

## Conditional Independence



*Cancer* is independent  
of *Age* and *Gender*  
given *Smoking*.

$$P(C|A,G,S) = P(C|S) \quad C \perp A,G \mid S$$

# Naïve Bayes on Example Data

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120K)$$

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

Ayastis digimiza

$$P(X, \text{Yes}) = \frac{P(\text{Yes}, X) \cdot P(X)}{P(\text{Yes})}$$

- $P(X | \text{Yes}) =$

$$P(\text{Refund} = \text{No} | \text{Yes}) \times$$

$$P(\text{Divorced} | \text{Yes}) \times$$

$$P(\text{Income} = 120K | \text{Yes})$$

$$= P(R \text{ No}, Y_{\text{Yes}}) * P(D \text{ Divorced}, Y_{\text{Yes}}) * P(I \text{ Income} = 120K, Y_{\text{Yes}})$$

- $P(X | \text{No}) =$

$$P(\text{Refund} = \text{No} | \text{No}) \times$$

$$P(\text{Divorced} | \text{No}) \times$$

$$P(\text{Income} = 120K | \text{No})$$

$$P(X, N) = P(R \text{ No}, N) * P(D, N) * P(I \text{ Income} = 120K, N)$$

# Estimate Probabilities from Data

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

| Class:  $P(Y) = N_c/N$

- e.g.,  $P(\text{No}) = 7/10$ ,  
 $P(\text{Yes}) = 3/10$

| For categorical attributes:

$$P(X_i | Y_k) = |X_{ik}| / N_{c_k}$$

helps

NP

where  $|X_{ik}|$  is number of instances having attribute value  $X_i$  and belonging to class  $Y_k$

- Examples:

$$P(\text{Status}=\text{Married} | \text{No}) = 4/7$$

$$P(\text{Refund}=\text{Yes} | \text{Yes}) = 0$$

And so on |

$0/3 = 0$

# Estimate Probabilities from Data

~~Sudzki Page 81'~~

| For continuous attributes:

- Discretization: Partition the range into bins:
  - ◆ Replace continuous value with bin value  $k$
  - Attribute changed from continuous to ordinal
- Probability density estimation:
  - ◆ Assume attribute follows a normal distribution
  - ◆ Use data to estimate parameters of distribution (e.g., mean and standard deviation)
  - ◆ Once probability distribution is known, use it to estimate the conditional probability  $P(X_i|Y)$

# 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

Sample: deg ✓

$$P(\text{Income} = 120 \mid \text{No}) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

| Normal distribution:

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

- One for each  $(X_i, Y_i)$  pair

| For (Income, Class=No):

- If Class=No

• sample mean = 110

• sample variance = 2975

# Example of Naïve Bayes Classifier

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

Naïve Bayes Classifier:

$$P(\text{Refund} = \text{Yes} | \text{No}) = 3/7$$

$$\rightarrow 4 / 7 = \frac{4}{7}$$

$$P(\text{Refund} = \text{No} | \text{No}) = 4/7$$

$$\rightarrow 3 / 7 = \frac{3}{7}$$

$$P(\text{Refund} = \text{Yes} | \text{Yes}) = 0$$

$$P(\text{Refund} = \text{No} | \text{Yes}) = 1$$

$$P(\text{Marital Status} = \text{Single} | \text{No}) = 2/7$$

$$P(\text{Marital Status} = \text{Divorced} | \text{No}) = 1/7$$

$$P(\text{Marital Status} = \text{Married} | \text{No}) = 4/7$$

$$P(\text{Marital Status} = \text{Single} | \text{Yes}) = 2/3$$

$$P(\text{Marital Status} = \text{Divorced} | \text{Yes}) = 1/3$$

$$P(\text{Marital Status} = \text{Married} | \text{Yes}) = 0$$

- $P(X | \text{No}) = P(\text{Refund} = \text{No} | \text{No}) \times P(\text{Divorced} | \text{No}) \times P(\text{Income} = 120\text{K} | \text{No}) = 4/7 \times 1/7 \times 0.0072 = 0.0006$

$$P(X | \text{Yes}) = P(\text{Refund} = \text{No} | \text{Yes}) \times P(\text{Divorced} | \text{Yes}) \times P(\text{Income} = 120\text{K} | \text{Yes}) = 1 \times 1/3 \times 1.2 \times 10^{-9} = 4 \times 10^{-10}$$

For Taxable Income:

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

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

$$P(\text{No}) = 7/10, \quad P(\text{Yes}) = 3/10$$

Since  $P(X|\text{No})P(\text{No}) > P(X|\text{Yes})P(\text{Yes})$

Therefore  $P(\text{No}|X) > P(\text{Yes}|X)$   
 $\Rightarrow \text{Class} = \text{No}$

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

# Example of Naïve Bayes Classifier

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

Naïve Bayes Classifier:

$$P(\text{Refund} = \text{Yes} | \text{No}) = 3/7$$

$$P(\text{Refund} = \text{No} | \text{No}) = 4/7$$

$$P(\text{Refund} = \text{Yes} | \text{Yes}) = 0$$

$$P(\text{Refund} = \text{No} | \text{Yes}) = 1$$

$$P(\text{Marital Status} = \text{Single} | \text{No}) = 2/7$$

$$P(\text{Marital Status} = \text{Divorced} | \text{No}) = 1/7$$

$$P(\text{Marital Status} = \text{Married} | \text{No}) = 4/7$$

$$P(\text{Marital Status} = \text{Single} | \text{Yes}) = 2/3$$

$$P(\text{Marital Status} = \text{Divorced} | \text{Yes}) = 1/3$$

$$P(\text{Marital Status} = \text{Married} | \text{Yes}) = 0$$

For Taxable Income:

If class = No: sample mean = 110

sample variance = 2975

If class = Yes: sample mean = 90

sample variance = 25

- $P(\text{Yes}) = 3/10$

$$P(\text{No}) = 7/10$$

Using Bayes Theorem:

$$P(\text{Yes} | X) = \frac{P(X | \text{Yes})P(\text{Yes})}{P(X)}$$

$$P(\text{No} | X) = \frac{P(X | \text{No})P(\text{No})}{P(X)}$$

- $P(\text{Yes} | \text{Divorced}) = 1/3 \times 3/10 / P(\text{Divorced})$

$$P(\text{No} | \text{Divorced}) = 1/7 \times 7/10 / P(\text{Divorced})$$

- $P(\text{Yes} | \text{Refund} = \text{No}, \text{Divorced}) = 1 \times 1/3 \times 3/10 / P(\text{Divorced, Refund} = \text{No})$

$$P(\text{No} | \text{Refund} = \text{No}, \text{Divorced}) = 4/7 \times 1/7 \times 7/10 / P(\text{Divorced, Refund} = \text{No})$$

Tid	Refund	Marital Status	Taxable Income	Evaade
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

# Issues with Naïve Bayes Classifier

## Naïve Bayes Classifier:

$$P(\text{Refund} = \text{Yes} | \text{No}) = 3/7$$

$$P(\text{Refund} = \text{No} | \text{No}) = 4/7$$

$$P(\text{Refund} = \text{Yes} | \text{Yes}) = 0$$

$$P(\text{Refund} = \text{No} | \text{Yes}) = 1$$

$$P(\text{Marital Status} = \text{Single} | \text{No}) = 2/7$$

$$P(\text{Marital Status} = \text{Divorced} | \text{No}) = 1/7$$

$$P(\text{Marital Status} = \text{Married} | \text{No}) = 4/7$$

$$P(\text{Marital Status} = \text{Single} | \text{Yes}) = 2/3$$

$$P(\text{Marital Status} = \text{Divorced} | \text{Yes}) = 1/3$$

$$P(\text{Marital Status} = \text{Married} | \text{Yes}) = 0$$

- $P(\text{Yes}) = 3/10$

- $P(\text{No}) = 7/10$

Using Bayes Theorem:

$$P(\text{Yes} | X) = \frac{P(X | \text{Yes})P(\text{Yes})}{P(X)}$$

$$P(\text{No} | X) = \frac{P(X | \text{No})P(\text{No})}{P(X)}$$

- $P(\text{Yes} | \text{Married}) = 0 \times 3/10 / P(\text{Married})$

- $P(\text{No} | \text{Married}) = 4/7 \times 7/10 / P(\text{Married})$

For Taxable Income:

If class = No: sample mean = 110

sample variance = 2975

If class = Yes: sample mean = 90

sample variance = 25

Tid	Refund	Marital Status	Taxable Income	Evaade
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

# Issues with Naïve Bayes Classifier

7	Yes	Divorced	220K	No
---	-----	----------	------	----

Consider the table with Tid = 7 deleted

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
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

## Naïve Bayes Classifier:

$$P(\text{Refund} = \text{Yes} | \text{No}) = 2/6$$

$$P(\text{Refund} = \text{No} | \text{No}) = 4/6$$

$$\xrightarrow{\text{P(Refund = Yes | Yes) = 0}}$$

$$P(\text{Refund} = \text{No} | \text{Yes}) = 1$$

$$P(\text{Marital Status} = \text{Single} | \text{No}) = 2/6$$

$$\xrightarrow{\text{P(Marital Status = Divorced | No) = 0}}$$

$$P(\text{Marital Status} = \text{Married} | \text{No}) = 4/6$$

$$P(\text{Marital Status} = \text{Single} | \text{Yes}) = 2/3$$

$$P(\text{Marital Status} = \text{Divorced} | \text{Yes}) = 1/3$$

$$P(\text{Marital Status} = \text{Married} | \text{Yes}) = 0/3$$

For Taxable Income:

If class = No: sample mean = 91

sample variance = 685

If class = Yes: sample mean = 90

sample variance = 25

Given  $X = (\text{Refund} = \text{Yes}, \text{Divorced}, 120\text{K})$

$$P(X | \text{No}) = 2/6 \times 0 \times 0.0083 = 0$$

$$P(X | \text{Yes}) = 0 \times 1/3 \times 1.2 \times 10^{-9} = 0$$

**Naïve Bayes will not be able to  
classify X as Yes or No!**

# Issues with Naïve Bayes Classifier

- Eğer koşullu olasılıklardan biri sıfırsa, tüm ifade sıfır olur
- Basit kesirlerden başka koşullu olasılık tahminlerini kullanma ihtiyacı
- Probability estimation:

$$\text{Original : } P(A_i | C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace: } P(A_i | C) = \frac{N_{ic} + 1}{N_c + c}$$

$$\text{m - estimate: } P(A_i | C) = \frac{N_{ic} + mp}{N_c + m}$$

c: sınıf sayısı

p: sınıfın önceki  
olasılığı(prior  
probability)

m: parameter

$N_c$ : sınıftaki örnek sayısı

$N_{ic}$ : c sınıfında  $A_i$   
öznitelik değerine sahip  
örneklerin sayısı

# Issues with Naïve Bayes Classifier

Consider the table with Tid = 7 deleted

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
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

$$\text{Original : } P(A_i | C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace: } P(A_i | C) = \frac{N_{ic} + 1}{N_c + c}$$

$$\text{m - estimate: } P(A_i | C) = \frac{N_{ic} + mp}{N_c + m}$$

## Naïve Bayes Classifier:

$$P(\text{Refund} = \text{Yes} | \text{No}) = 2/6$$

$$P(\text{Refund} = \text{No} | \text{No}) = 4/6$$

$$P(\text{Refund} = \text{Yes} | \text{Yes}) = 0$$

$$P(\text{Refund} = \text{No} | \text{Yes}) = 1$$

$$P(\text{Marital Status} = \text{Single} | \text{No}) = 2/6$$

$$P(\text{Marital Status} = \text{Divorced} | \text{No}) = 0$$

$$P(\text{Marital Status} = \text{Married} | \text{No}) = 4/6$$

$$P(\text{Marital Status} = \text{Single} | \text{Yes}) = 2/3$$

$$P(\text{Marital Status} = \text{Divorced} | \text{Yes}) = 1/3$$

$$P(\text{Marital Status} = \text{Married} | \text{Yes}) = 0/3$$

koşullu olasılık  $P(\text{Status} = \text{Married} | \text{Yes}) = 0$  çünkü bu sınıfı yönelik eğitim kayıtlarının hiçbir ilgili öznitelik değerine sahip değildir.  $m = 3$  and  $p = 1/3$  ile **m-estimate** yaklaşımı kullanıldığında, koşullu olasılık artık sıfır değildir:

$$P(\text{Marital Status} = \text{Married} | \text{Yes}) = (0+3 \times 1/3)/(3 + 3) = 1/6.$$

# Example of Naïve Bayes Classifier

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

M: mammals

N: non-mammals

$$P(\text{GiveBirth}=\text{Yes} \mid \text{Mammals}) = 6/7$$

$$P(\text{GiveBirth}=\text{Yes} \mid \text{Non-mammals}) = 1/13$$

$$P(\text{CanFly}=\text{No} \mid \text{Mammals}) = 6/7$$

$$P(\text{CanFly}=\text{No} \mid \text{Non-mammals}) = 10/13$$

$$P(\text{LiveInWater}=\text{Yes} \mid \text{Mammals}) = 2/7$$

$$P(\text{LiveInWater}=\text{Yes} \mid \text{Non-mammals}) = 3/13$$

$$P(\text{HaveLegs}=\text{No} \mid \text{Mammals}) = 2/7$$

$$P(\text{HaveLegs}=\text{No} \mid \text{Non-mammals}) = 4/13$$

$$m = 7 \\ n = 13 \\ 20$$

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

# Example of Naïve Bayes Classifier

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

A: attributes

M: mammals

N: non-mammals

$$P(A | M) = \frac{1}{7} \times \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} = 0.06$$

$$P(A | N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A | M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A | N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

$$\begin{aligned} P(A|M)P(M) &> P(A|N)P(N) \\ \Rightarrow & \text{Mammals} \end{aligned}$$

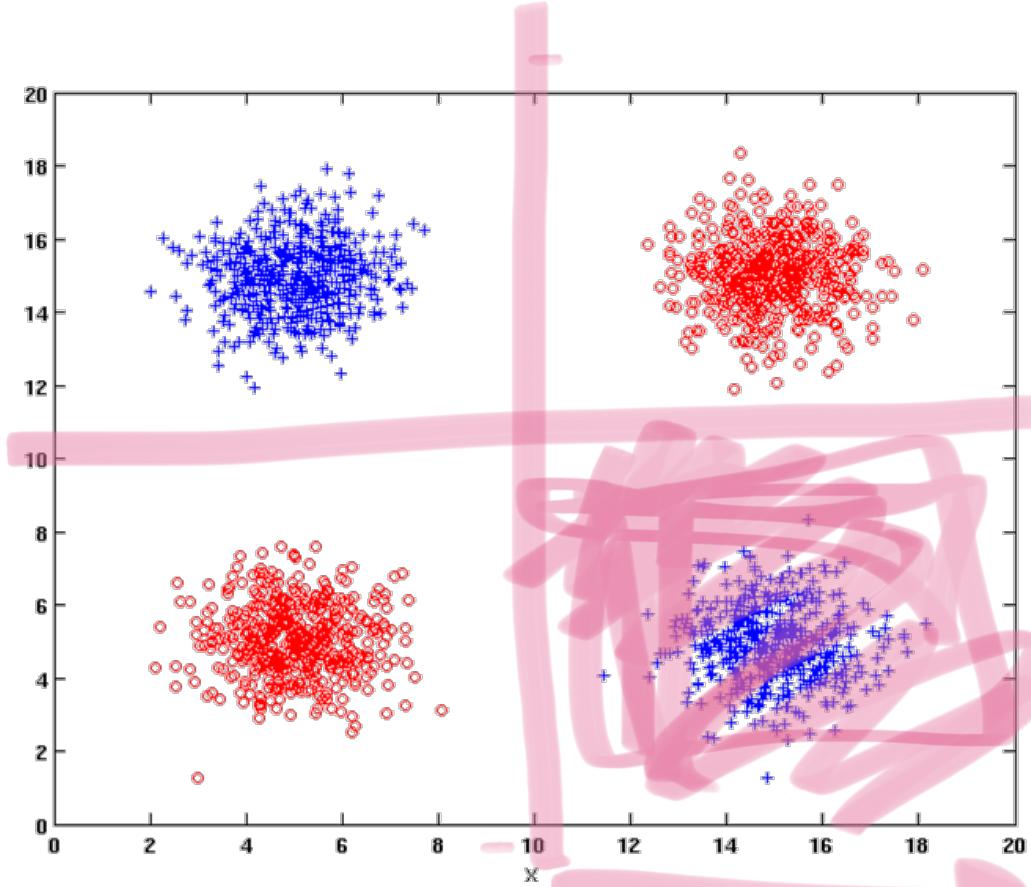
Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

# Naïve Bayes (Summary)

- | İzole gürültü noktalarına karşı sağlam
- | Olasılık tahmini hesaplamaları sırasında ilgili örneği yok sayarak eksik değerlerle başa çıkabilir
- | Alakasız özniteliklere karşı sağlam
- | Bağımsızlık varsayıımı (*independent assumption*) bazı özellikler için geçerli olmayıabilir
  - Bayesian Belief Networks (BBN) gibi diğer teknikleri kullanın

# Naïve Bayes

- How does Naïve Bayes perform on the following dataset?

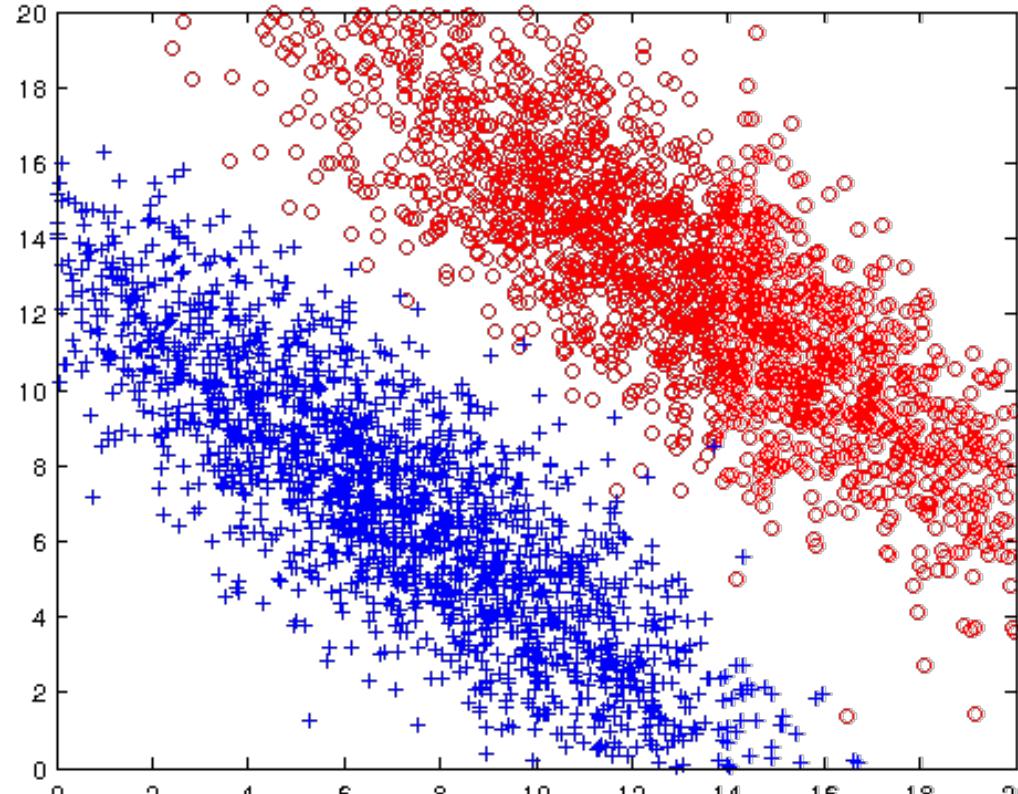


Özniteliklerin koşullu bağımsızlığı ihlal edilmiştir

# Naïve Bayes

---

- How does Naïve Bayes perform on the following dataset?



Naïve Bayes eğik karar sınırları oluşturabilir

# Naïve Bayes

---

- How does Naïve Bayes perform on the following dataset?

$Y = 1$	1	1	1	0
$Y = 2$	0	1	0	0
$Y = 3$	0	0	1	1
$Y = 4$	0	0	1	1
	$X = 1$	$X = 2$	$X = 3$	$X = 4$

Özniteliklerin koşullu bağımsızlığı ihlal edilmiştir

# Bayesian Belief Networks

---

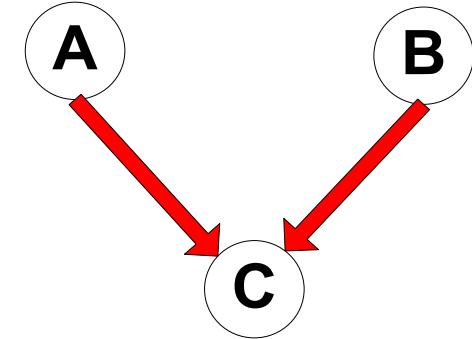
- Naive Bayes sınıflandırıcıları tarafından yapılan koşullu bağımsızlık varsayıımı (**conditional independence assumption**) **çok katı** görünebilir,
  - bilhassa özniteliklerin bir şekilde ilişkili olduğu sınıflandırma problemleri için.
- Sınıf-koşullu olasılıkların (class-conditional probabilities)  $P(\mathbf{X}|\mathbf{Y})$  modellenmesi için **daha esnek bir yaklaşım** sağlar.
- Sınıfa göre tüm özniteliklerin koşullu olarak bağımsız olmasını **zorunlu kılmak** yerine,
  - bu yaklaşım, hangi öznitelik çiftlerinin koşullu olarak bağımsız olduğunu belirlememizi sağlar.

# Bayesian Belief Networks

| Bir dizi rastgele değişken arasındaki olasılık ilişkilerinin grafiksel gösterimini sağlar

| Şunlardan oluşur:

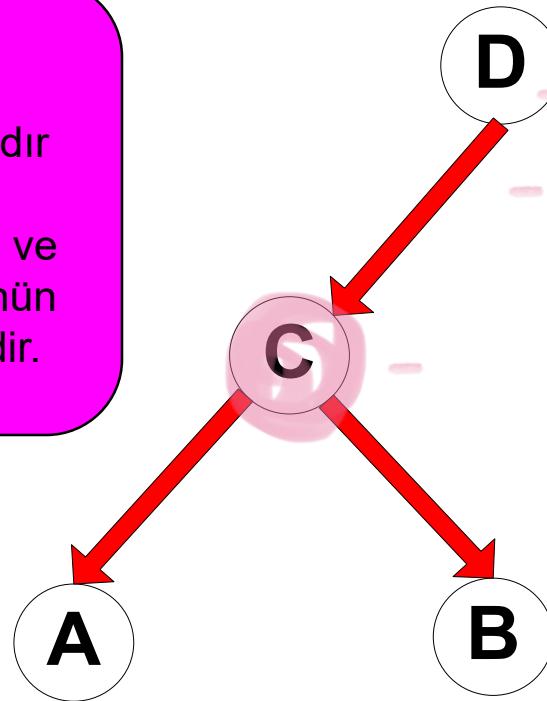
- A directed acyclic graph (dag)
  - ◆ Düğüm bir değişkene karşılık gelir
  - ◆ Yay ise, bir çift değişken arasındaki bağımlılık ilişkisine karşılık gelir
- Her düğümü en yakın üst ögesi (*immediate parent*) ile ilişkilendiren bir olasılık tablosu



A ve B'nin bağımsız değişkenler (**independent variables**) olduğu ve her birinin üçüncü bir değişken olan C üzerinde doğrudan bir etkisi olduğu A, B ve C olmak üzere üç rastgele değişkeni düşünün.

# Conditional Independence

C verildiğinde A, hem B'den hem de D'den koşullu olarak bağımsızdır (**conditionally independent**) çünkü B ve D düğümleri A düşümünün soyundan gelmemektedir.



D is parent of C

A is child of C

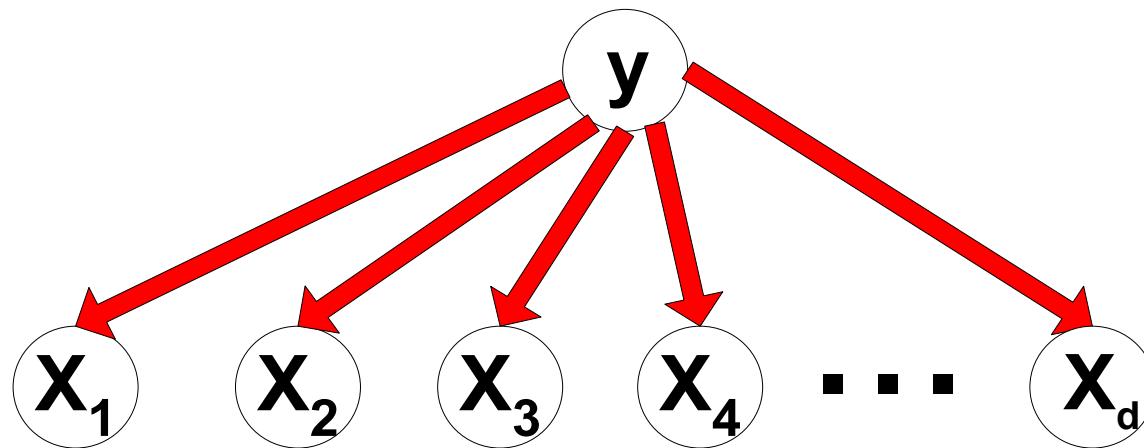
B is descendant of D

D is ancestor of A

- | Bayes ağındaki bir düğüm, ebeveynleri biliniyorsa, onun soyundan gelmeyen (**nondescendants**) diğer tüm düğümlerden koşullu olarak bağımsızdır.

# Conditional Independence

- Naïve Bayes assumption:



Bir naïve Bayes sınıflandırıcı tarafından yapılan koşullu bağımsızlık varsayıımı, yukarıda gösterildiği gibi bir Bayes ağı kullanılarak da temsil edilebilir, burada  $y$  hedef sınıfıdır ve  $\{X_1, X_2, \dots, X_d\}$  öznitelik kümesidir.

# Probability Tables

Olasılık Tablosu

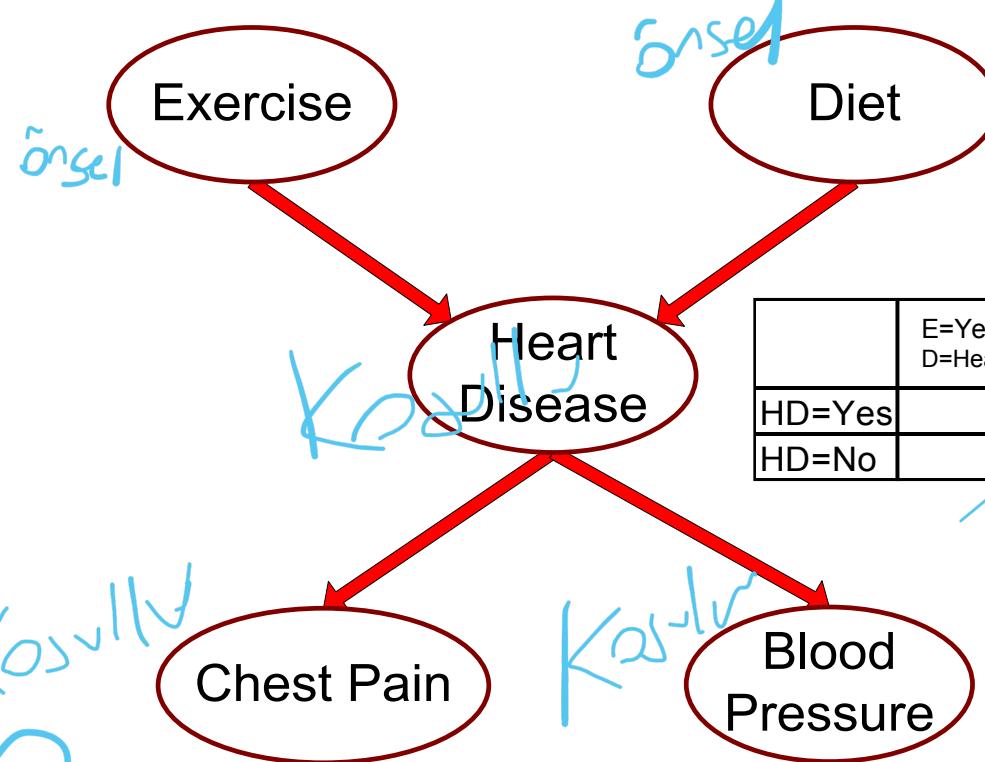
- X'in herhangi bir ebeveyni (üst ögesi) yoksa, tablo önceki olasılık  $P(X)$  içerir
- X'in yalnızca bir ebeveyni (Y) varsa, tablo koşullu olasılık  $P(X|Y)$  içerir
- X'in birden çok ebeveyni ( $Y_1, Y_2, \dots, Y_k$ ) varsa, tablo koşullu olasılık  $P(X|Y_1, Y_2, \dots, Y_k)$  içerir



# Example of Bayesian Belief Network

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

Kalp rahatsızlığı (HD) için ebeveyn düğümler (**parent nodes**), egzersiz (E) ve diyet (D) gibi bu rahatsızlığı etkileyebilecek risk faktörlerine karşılık gelir.

	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

Kalp rahatsızlığı için çocuk düğümler (**child nodes**) hastalıkın şu semptomlarına karşılık gelir: göğüs ağrısı (CP) ve yüksek tansiyon (BP) gibi hastalık.

Risk faktörleriyle ilişkili düğümler yalnızca önsel olasılıkları (**prior probabilities**) içerirken, kalp rahatsızlığı düğümleri ve bunlara karşılık gelen semptomları koşullu olasılıkları içerir.

# Example of Inferencing using BBN

Bayes inonç Aşağıda kılavuz Gikarın örneğ;

- Given:  $X = (E=\text{No}, D=\text{Yes}, CP=\text{Yes}, BP=\text{High})$ 
  - Compute  $P(HD|E,D,CP,BP)$ ?
- $P(HD=\text{Yes} | E=\text{No}, D=\text{Yes}) = 0.55$   
 $P(CP=\text{Yes} | HD=\text{Yes}) = 0.8$   
 $P(BP=\text{High} | HD=\text{Yes}) = 0.85$ 
  - $P(HD=\text{Yes} | E=\text{No}, D=\text{Yes}, CP=\text{Yes}, BP=\text{High})$   
 $\approx 0.55 \times 0.8 \times 0.85 = 0.374$
- $P(HD=\text{No} | E=\text{No}, D=\text{Yes}) = 0.45$   
 $P(CP=\text{Yes} | HD=\text{No}) = 0.01$   
 $P(BP=\text{High} | HD=\text{No}) = 0.2$ 
  - $P(HD=\text{No} | E=\text{No}, D=\text{Yes}, CP=\text{Yes}, BP=\text{High})$   
 $\approx 0.45 \times 0.01 \times 0.2 = 0.0009$

Classify X  
as Yes

# Example of Inferencing using BBN

## Case 1: No Prior Information

*Herhangi bir önsel bilgi olmadan, birisinin kalp krizi riski taşıyıp taşımadığına dair olasılık hesabı yapabiliyoruz.*

Without any prior information, we can determine whether the person is likely to have heart disease by computing the prior probabilities  $P(\text{HD} = \text{Yes})$  and  $P(\text{HD} = \text{No})$ . To simplify the notation, let  $\alpha \in \{\text{Yes}, \text{No}\}$  denote the binary values of Exercise and  $\beta \in \{\text{Healthy}, \text{Unhealthy}\}$  denote the binary values of Diet.

$$\begin{aligned} P(\text{HD} = \text{Yes}) &= \sum_{\alpha} \sum_{\beta} P(\text{HD} = \text{Yes} | E = \alpha, D = \beta) P(E = \alpha, D = \beta) \\ &= \sum_{\alpha} \sum_{\beta} P(\text{HD} = \text{Yes} | E = \alpha, D = \beta) P(E = \alpha) P(D = \beta) \\ &= 0.25 \times 0.7 \times 0.25 + 0.45 \times 0.7 \times 0.75 + 0.55 \times 0.3 \times 0.25 \\ &\quad + 0.75 \times 0.3 \times 0.75 \\ &= 0.49. \end{aligned}$$

	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

Since  $P(\text{HD} = \text{no}) = 1 - P(\text{HD} = \text{yes}) = 0.51$ , the person has a slightly higher chance of not getting the disease.

*%51 ile kişi, kalp krizi riski olmama şansı hafif de olsa daha yüksektir.*

# Example of Inferencing using BBN

## Case 2: High Blood Pressure

	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

If the person has high blood pressure, we can make a diagnosis about heart disease by comparing the posterior probabilities,  $P(\text{HD} = \text{Yes} | \text{BP} = \text{High})$  against  $P(\text{HD} = \text{No} | \text{BP} = \text{High})$ . To do this, we must compute  $P(\text{BP} = \text{High})$ :

$$\begin{aligned} P(\text{BP} = \text{High}) &= \sum_{\gamma} P(\text{BP} = \text{High} | \text{HD} = \gamma)P(\text{HD} = \gamma) \\ &= 0.85 \times 0.49 + 0.2 \times 0.51 = 0.5185. \end{aligned}$$

where  $\gamma \in \{\text{Yes}, \text{No}\}$ . Therefore, the posterior probability the person has heart disease is

$$\begin{aligned} P(\text{HD} = \text{Yes} | \text{BP} = \text{High}) &= \frac{P(\text{BP} = \text{High} | \text{HD} = \text{Yes})P(\text{HD} = \text{Yes})}{P(\text{BP} = \text{High})} \\ &= \frac{0.85 \times 0.49}{0.5185} = 0.8033. \end{aligned}$$

Similarly,  $P(\text{HD} = \text{No} | \text{BP} = \text{High}) = 1 - 0.8033 = 0.1967$ . Therefore, when a person has high blood pressure, it increases the risk of heart disease.

# Characteristics of BBN

---

- BBN, grafiksel bir model kullanarak belirli bir alanın (*domain*) önceki bilgilerini yakalamak için bir yaklaşım sağlar.
- Ağı oluşturmak **zaman alıcı** olabilir ve **büyük miktarda çaba** gerektirir.
  - Bununla birlikte, ağıın yapısı belirlendikten sonra, yeni bir değişken eklemek oldukça basittir.
- Bayes ağları, eksik verilerle (**incomplete data**) başa çıkmak için çok uygundur.
  - Eksik özniteliklere sahip örnekler, özniteliğin tüm olası değerleri üzerinden olasılıklar toplanarak veya bütünleştirilerek ele alınabilir.
- Veriler olasılıkla önsel bilgilerle birleştirildiğinden, yöntem modelin ezberlemesine karşı oldukça sağlamdır. (**robust to model overfitting**.)