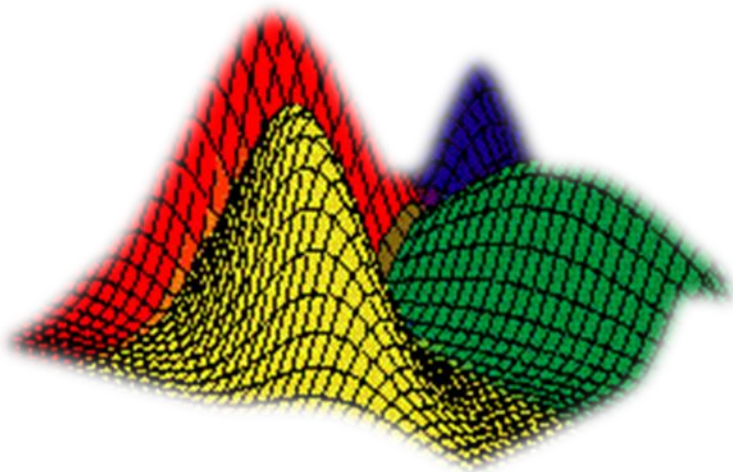
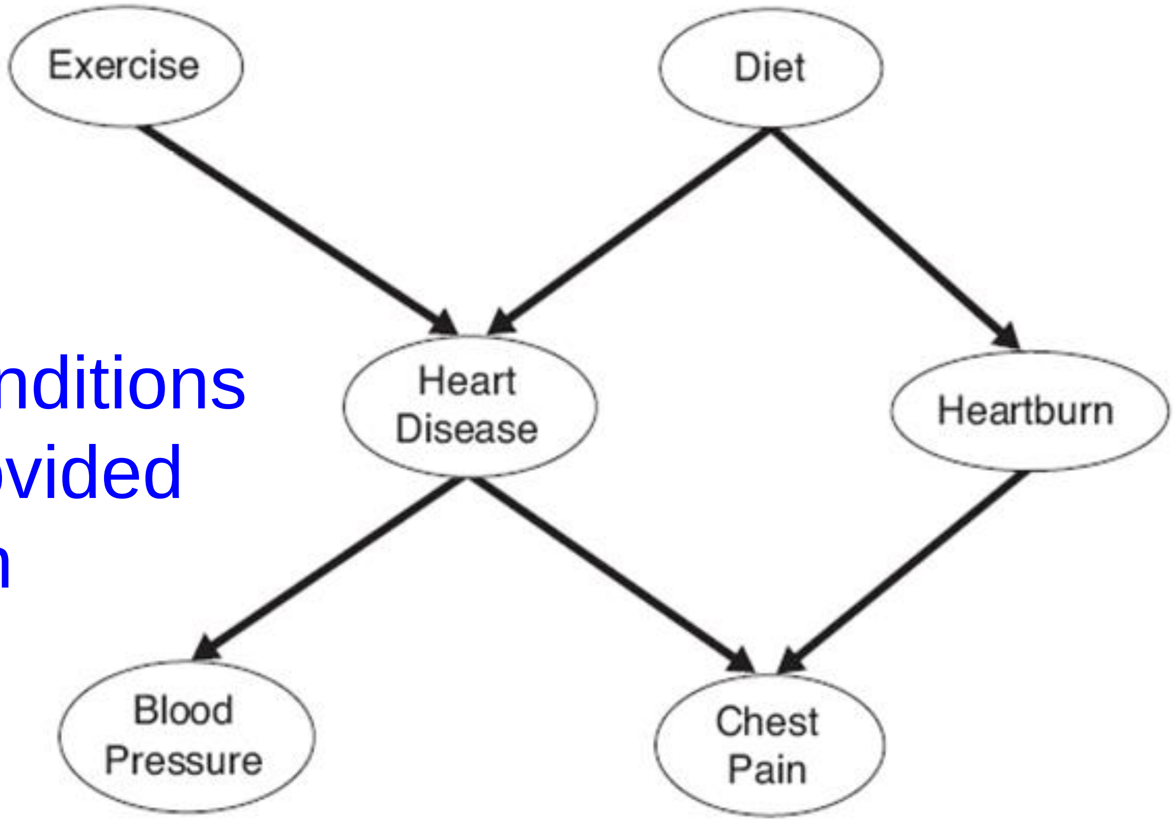


CSE 473
Pattern Recognition

Bayesian Classifier and its Variants

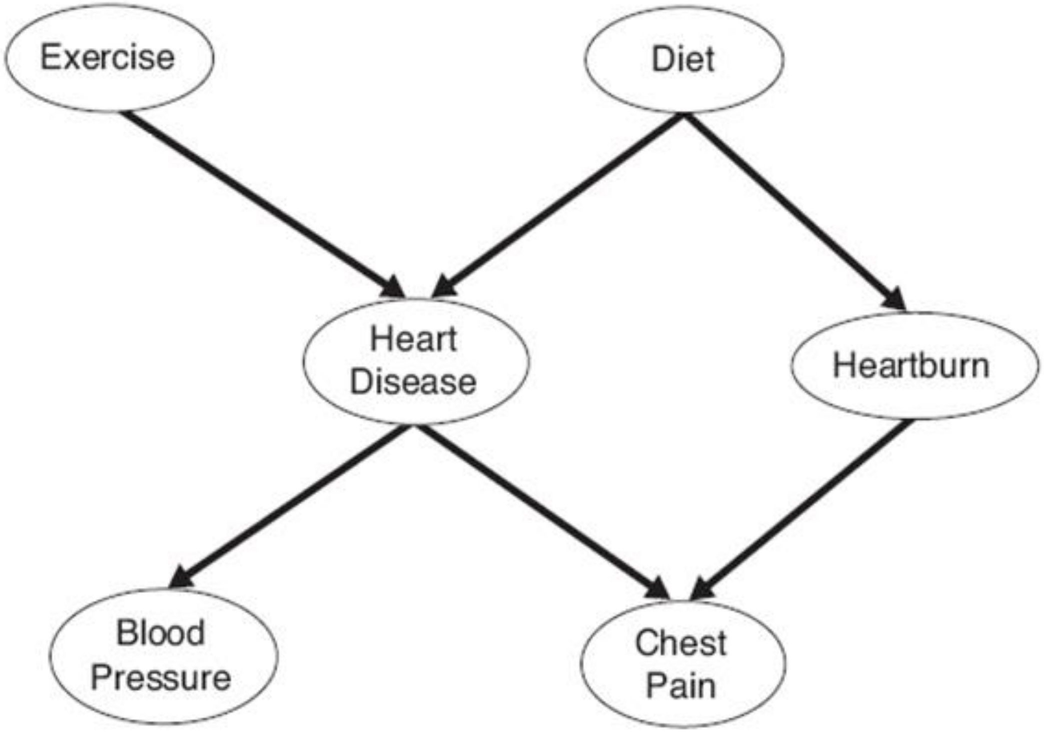
We will study this graph



Non-descendant conditions can be removed provided all parents are given

We can show:

- $P(D|E)=P(D)$
- $P(Hb|HD, E, D, CP) = P(Hb|D, CP)$
- $P(CP|Hb, HD, E, D) = P(CP|Hb, HD)$
- $P(BP|CP, Hb, HD, E, D) = P(BP|HD)$
- However, $P(HD|E,D)$ cannot be simplified



Exercise:

- *P (CP|HD, BP, E, D)*= No simplification

- BBN Model Building

$T = \{X_1, X_2, X_3, \cdots, X_d\}$ Set of ordered variables

for $j = 1$ to d do

$X_{T(j)} = j$th highest order variable

$\pi(X_{T(j)}) = \{X_{T(1)}, X_{T(2)}, X_{T(3)}, \cdots, X_{T(j-1)}\}$: preceding variables

remove non-dependent variables

create links between $X_{T(j)}$ and remaining $\pi(X_{T(j)})$

We will study this graph

$T = \{X_1, X_2, X_3, \cdots, X_d\}$ Set of ordered variables

for $j = 1$ to d do

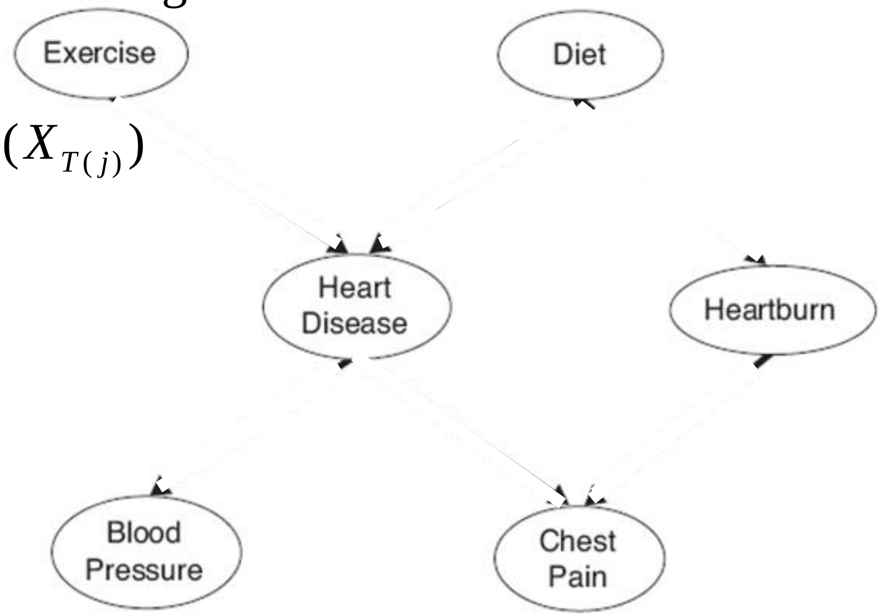
$X_{T(j)} = j$th highest order variable

$\pi(X_{T(j)}) = \{X_{T(1)}, X_{T(2)}, X_{T(3)}, \cdots, X_{T(j-1)}\}$: preceding variables

remove non-dependent variables

create links between $X_{T(j)}$ and remaining $\pi(X_{T(j)})$

Order: *E, D, HD, Hb, CP, BP*



We will study this graph

$T = \{X_1, X_2, X_3, \cdots, X_d\}$ Set of ordered variables

for $j = 1$ to d do

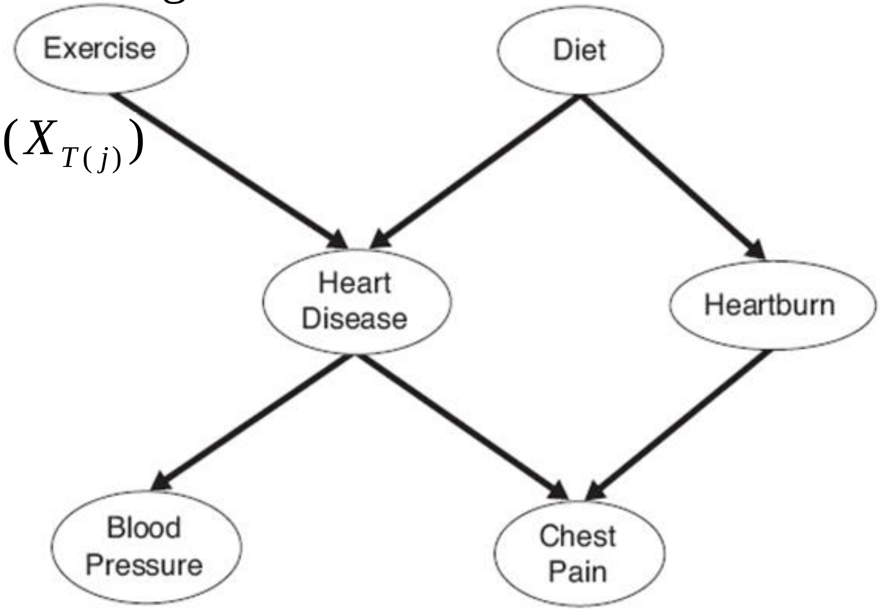
$X_{T(j)} =$ jth highest order variable

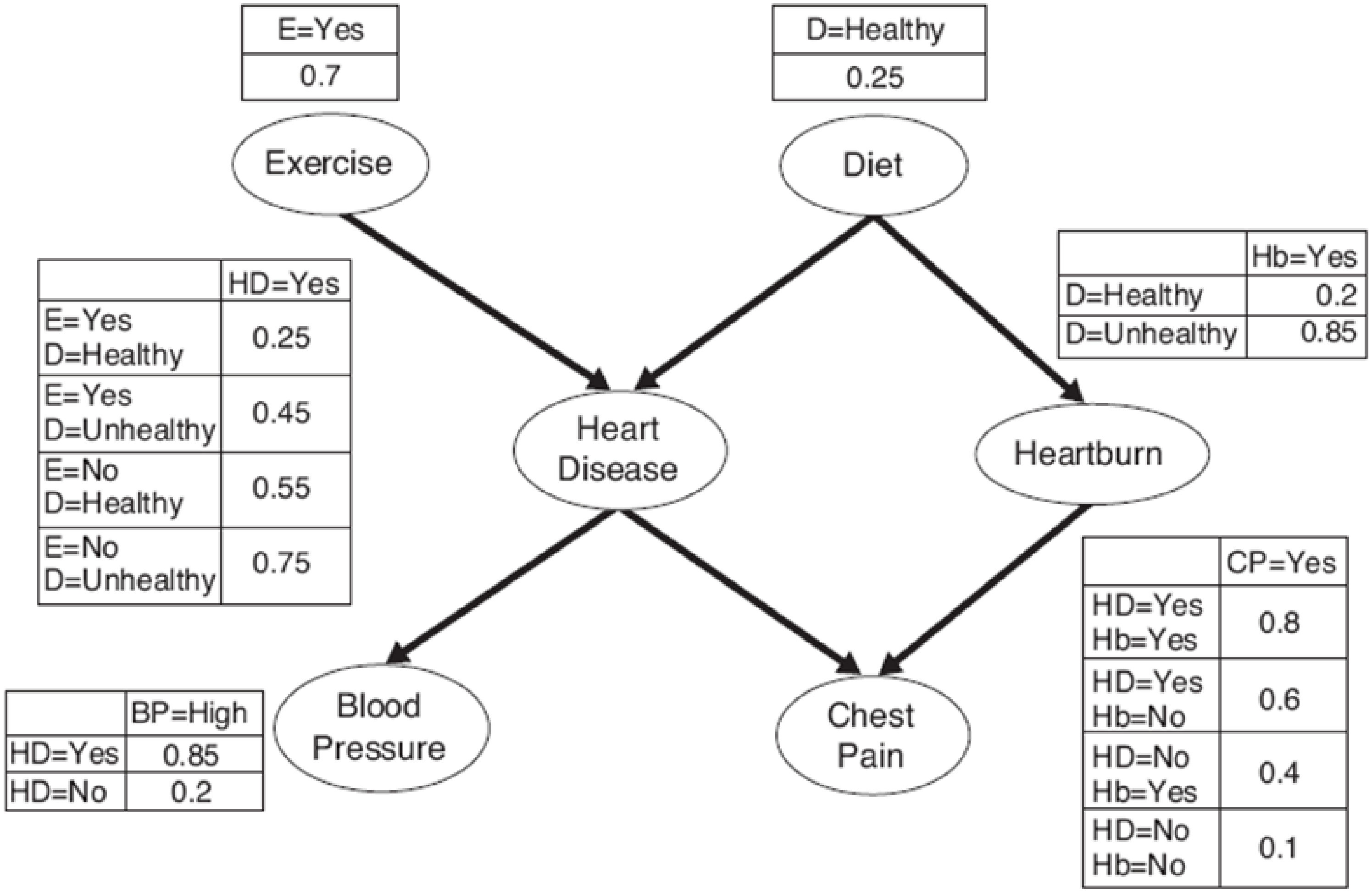
$\pi(X_{T(j)}) = \{X_{T(1)}, X_{T(2)}, X_{T(3)}, \cdots, X_{T(j-1)}\}$: preceding variables

remove non-dependent variables

create links between $X_{T(j)}$ and remaining $\pi(X_{T(j)})$

Order: *E, D, HD, Hb, CP, BP*





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

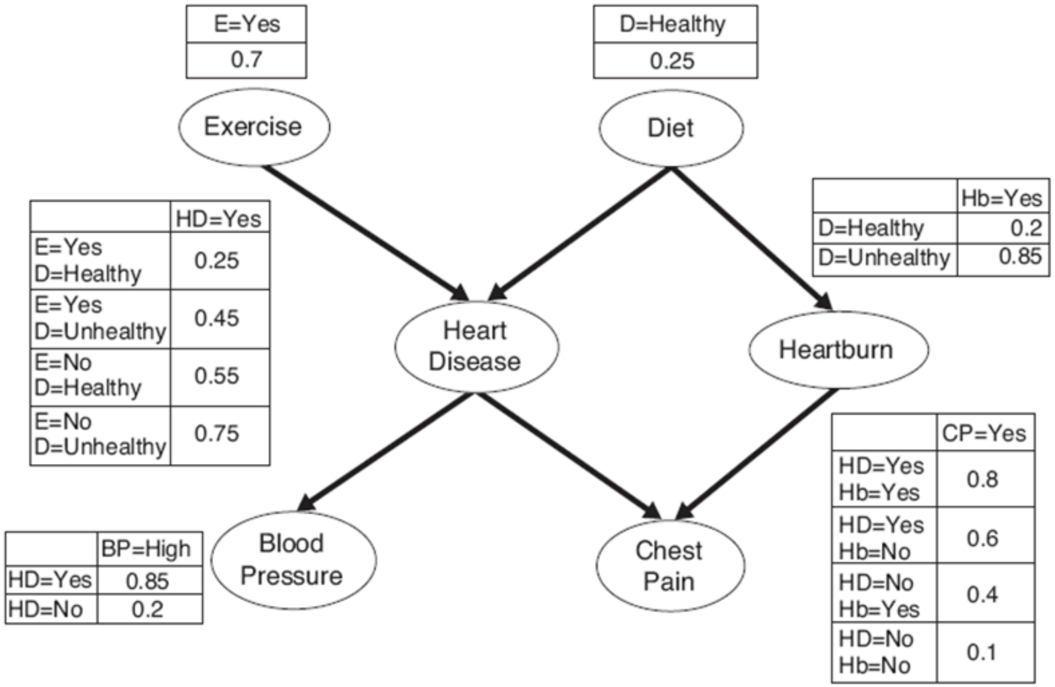
	Hb=Yes
D=Healthy	0.2
D=Unhealthy	0.85

	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

Calculate P(HD=yes) ?

Calculate P(HD=yes) ?



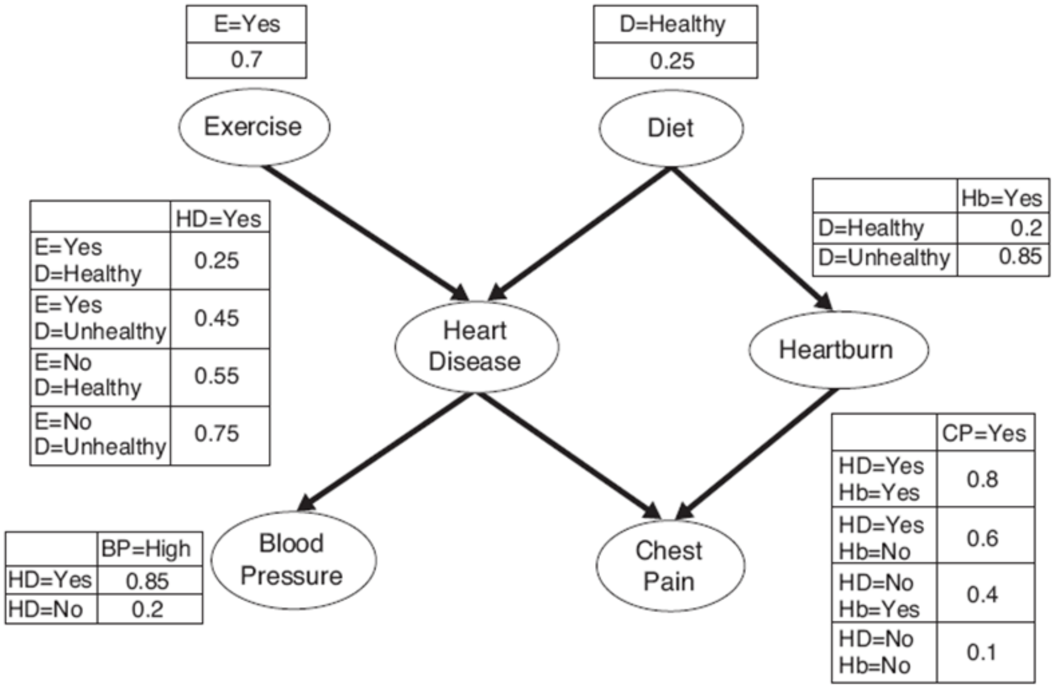
$$P(HD = Yes) = \sum_{\alpha}\sum_{\beta} P(HD = yes \mid E = \alpha, D = \beta) P(E = \alpha, D = \beta)$$

where,

α = Set of Values of Exercise(E) = {Yes, No}

β = Set of Values of Diet(D) = {Healthy, Not Healthy}

Calculate P(HD=yes) ?

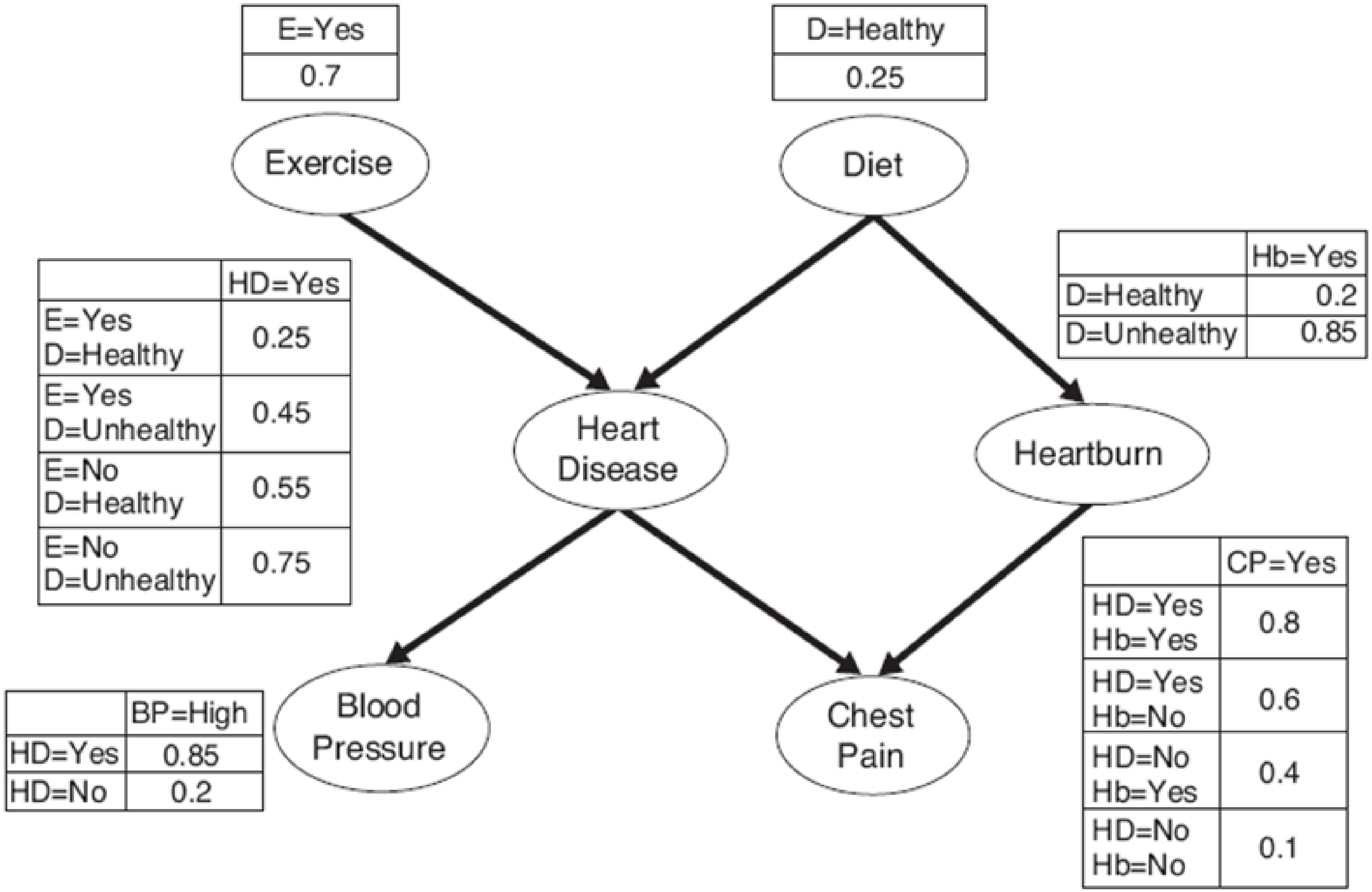


$$P(HD = Yes) = \sum_{\alpha}\sum_{\beta} P(HD = yes \mid E = \alpha, D = \beta) P(E = \alpha, D = \beta)$$

$$= \sum_{\alpha}\sum_{\beta} P(HD = yes \mid E = \alpha, D = \beta) P(E = \alpha) P(D = \beta)$$

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

$$= 0.49$$



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

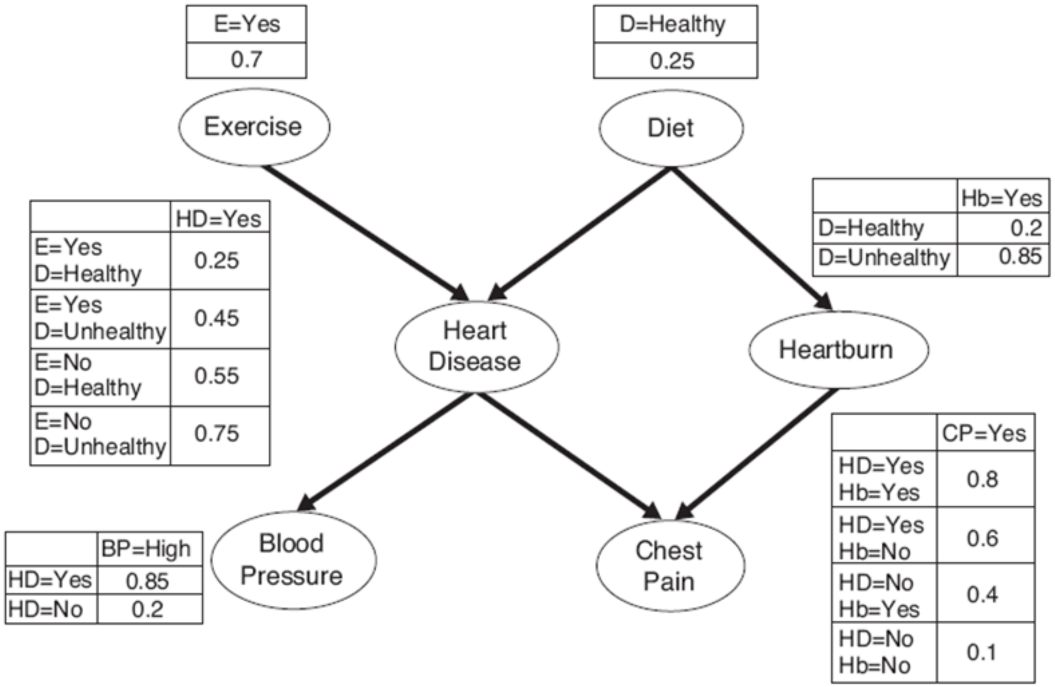
	Hb=Yes
D=Healthy	0.2
D=Unhealthy	0.85

	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

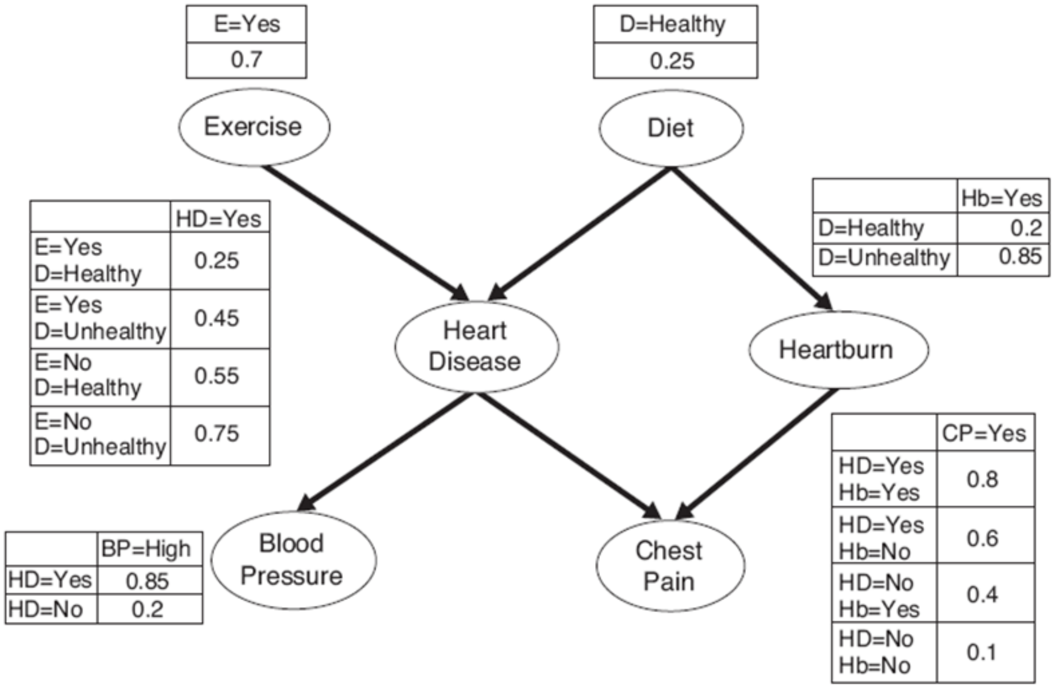
Calculate: P(HD=yes|BP=High)?

Calculate P(HD=yes|BP=High)



$P(HD = yes \mid BP = High)$ can be written as $\dfrac{P(BP = High \mid HD = yes)P(HD = yes)}{P(BP = High)}$

Calculate P(HD=yes|BP=High)

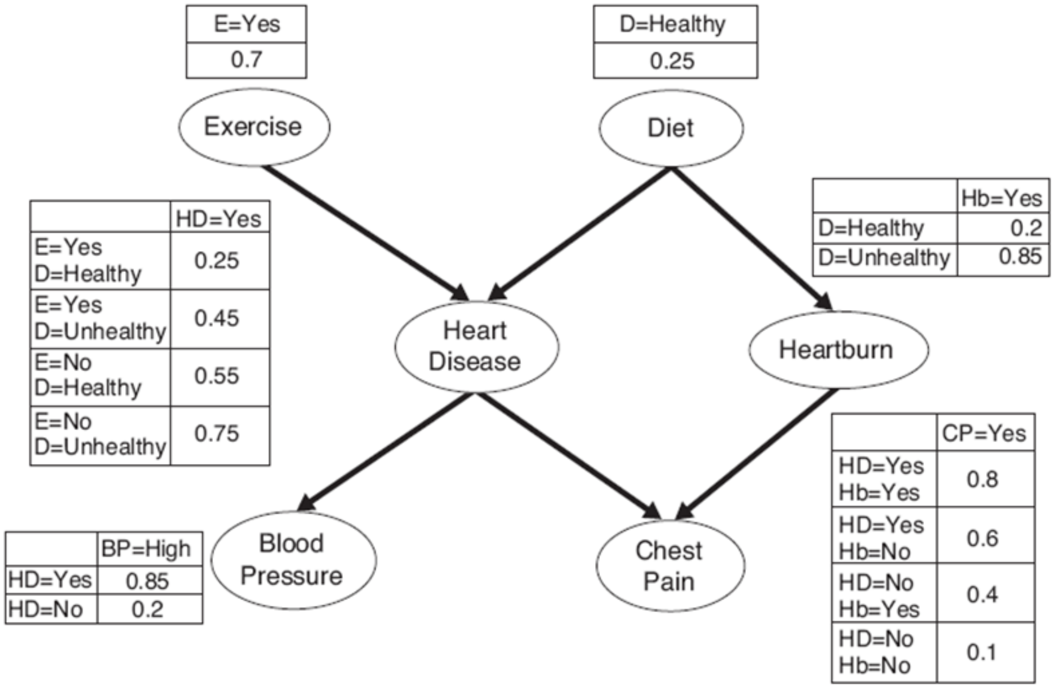


$$P(BP = High) = \sum_{\gamma} P(BP = high \mid HD = \gamma) P(HD = \gamma)$$

where,

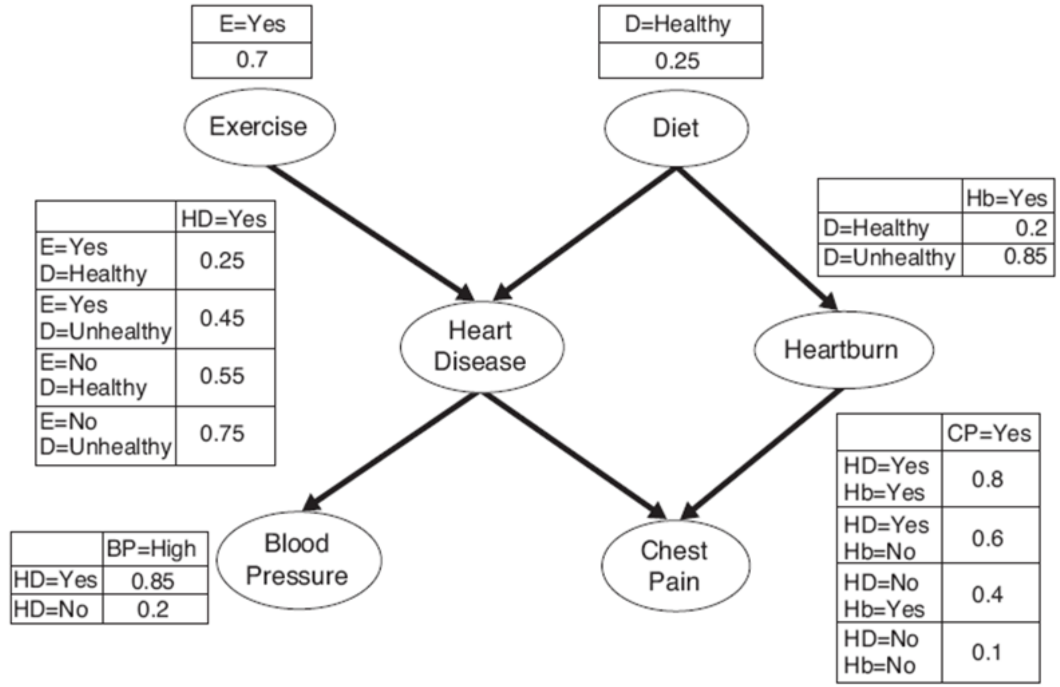
γ = Set of Values of Heart Disease (HD) = {Yes, No}

Calculate P(HD=yes|BP=High)



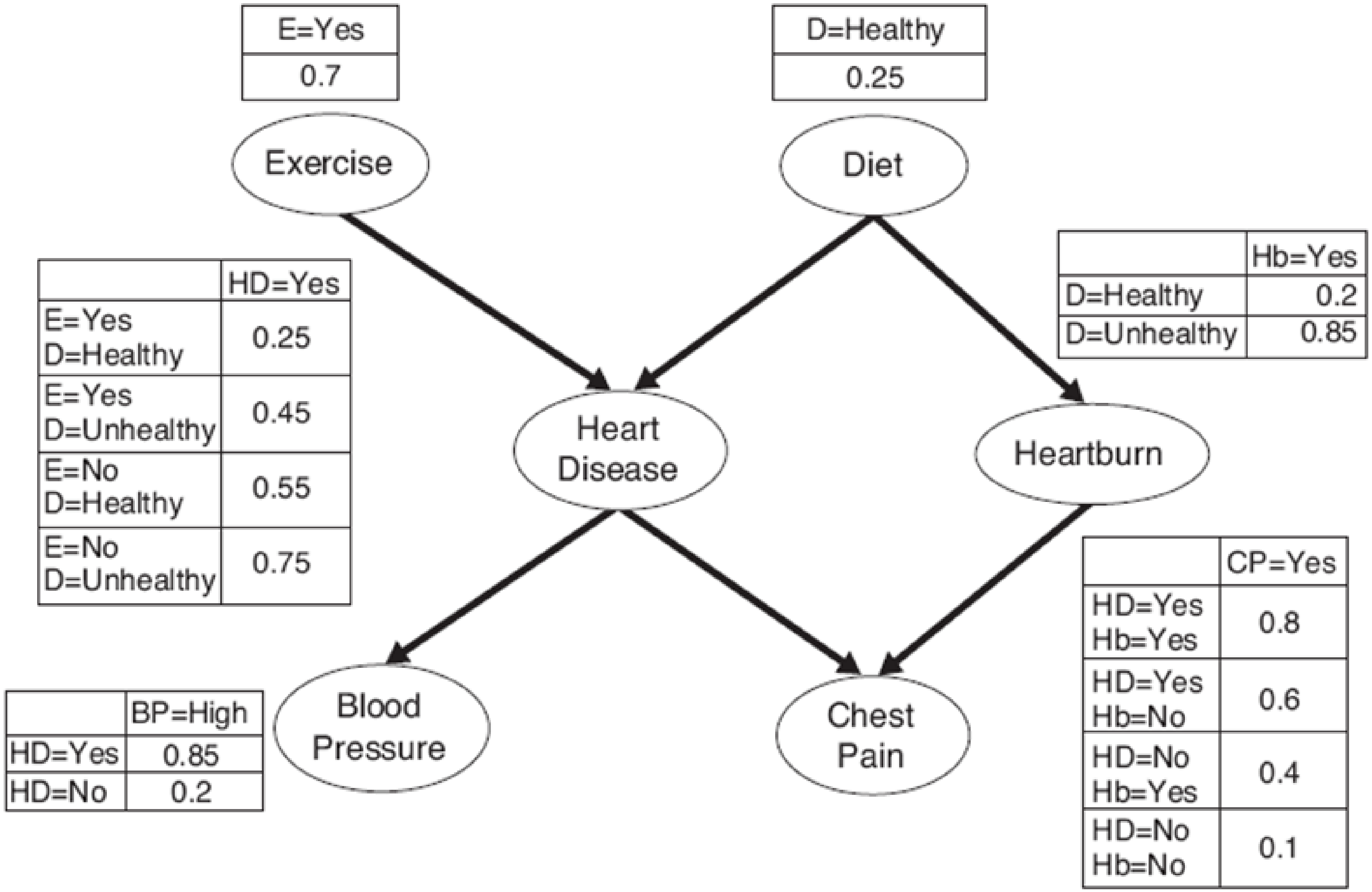
$$P(BP = High) = \sum_{\gamma} P(BP = high \mid HD = \gamma) P(HD = \gamma)$$
$$= 0.85 \times 0.49 + 0.2 \times 0.51 = 0.5185$$

Calculate P(HD=yes|BP=High)



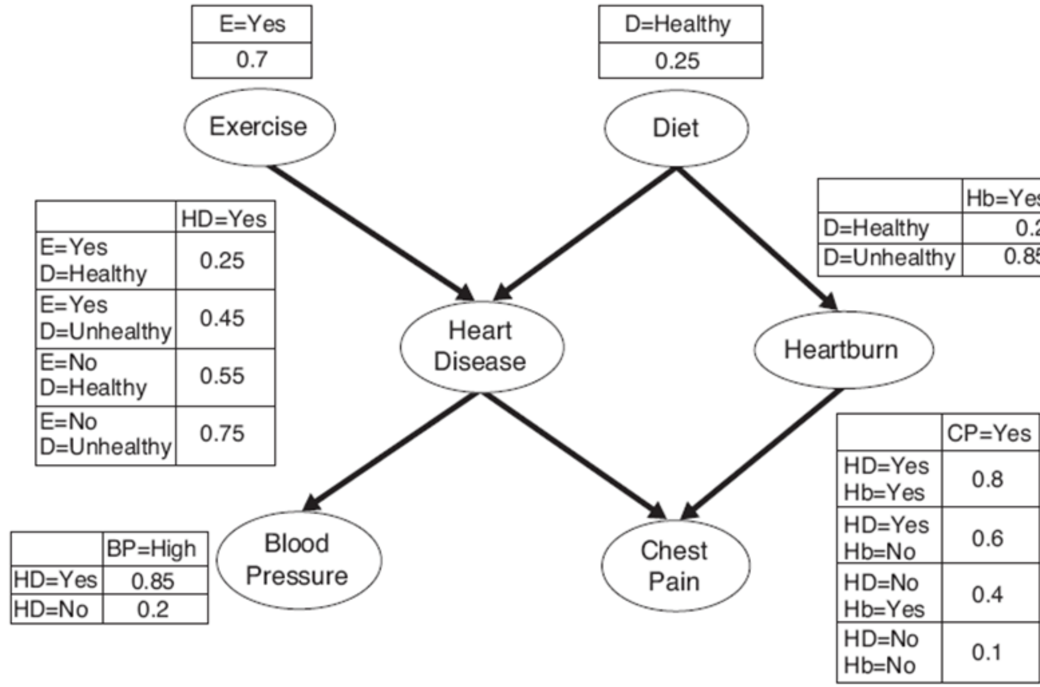
$$P(HD = yes \mid BP = High) = \frac{P(BP = High \mid HD = yes)P(HD = yes)}{P(BP = High)}$$

$$= \frac{0.85 \times 0.49}{0.5185} = 0.8033$$



Calculate P(HD=yes|BP=high, D=Healthy, E=yes)?

Calculate P(HD=yes|BP=high, D=Healthy, E=yes) ?



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

How is this formula true?

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

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

Let $$P(X \mid Y) = \frac{P(Y \mid X)}{P(Y)} \times P(X)$$

How is this formula true?

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

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

Let $$P(X \mid Y) = \frac{P(Y \mid X)}{P(Y)} \times P(X)$$

Now add *Z* and *W* as condition $$P(X \mid Y, Z, W) = \frac{P(Y \mid X, Z, W)}{P(Y \mid Z, W)} \times P(X \mid Z, W)$$

How is this formula true?

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

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

Let $$P(X \mid Y) = \frac{P(Y \mid X)}{P(Y)} \times P(X)$$

Now add *Z* and *W* as condition $$P(X \mid Y, Z, W) = \frac{P(Y \mid X, Z, W)}{P(Y \mid Z, W)} \times P(X \mid Z, W)$$

Similarly,

$$P(HD = yes \mid BP = high) = \frac{P(BP = high \mid HD = yes)}{P(BP = high)} \times P(HD = yes)$$

How is this formula true?

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

Let $$P(X \mid Y) = \frac{P(Y \mid X)}{P(Y)} \times P(X)$$

Now add *Z* and *W* as condition $$P(X \mid Y, Z, W) = \frac{P(Y \mid X, Z, W)}{P(Y \mid Z, W)} \times P(X \mid Z, W)$$

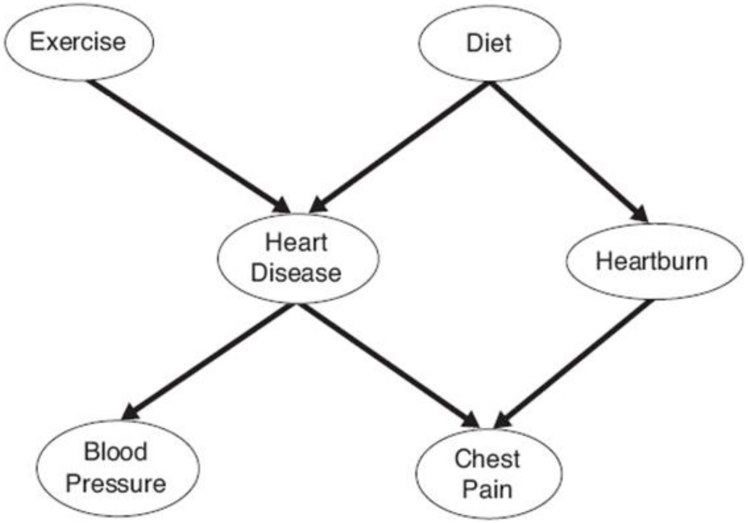
Similarly,

$$P(HD = yes \mid BP = high) = \frac{P(BP = high \mid HD = yes)}{P(BP = high)} \times P(HD = yes)$$

Now add conditions *D= Healthy* and *E= Yes* to above formula

Similarly,

$$P(BP = high \mid D = Healthy, E = Yes)$$
$$= \sum_{\gamma} P(BP = high \mid HD = \gamma, D = Healthy, E = Yes) \times P(HD = \gamma \mid D = Healthy, E = Yes)$$

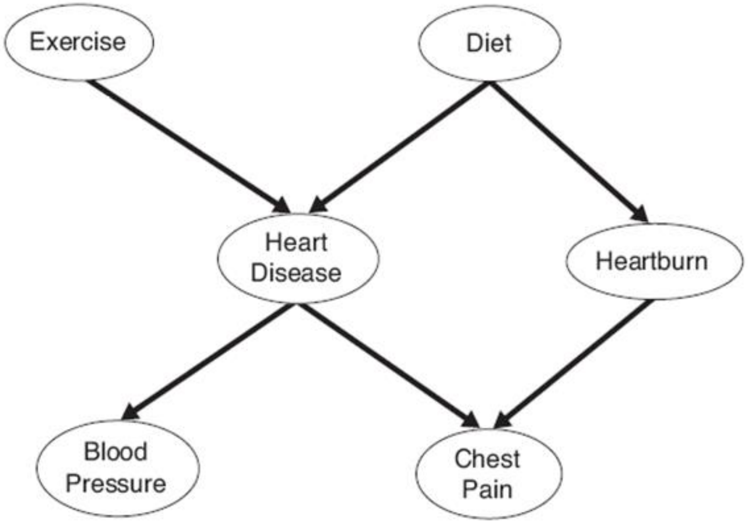


Similarly,

$$P(BP = high \mid D = Healthy, E = Yes)$$
$$= \sum_{\gamma} P(BP = high \mid HD = \gamma, D = Healthy, E = Yes) \times P(HD = \gamma \mid D = Healthy, E = Yes)$$

Proof:

$$P(BP = high) = \sum_{\gamma} P(BP = high \mid HD = \gamma) \times P(HD = \gamma)$$

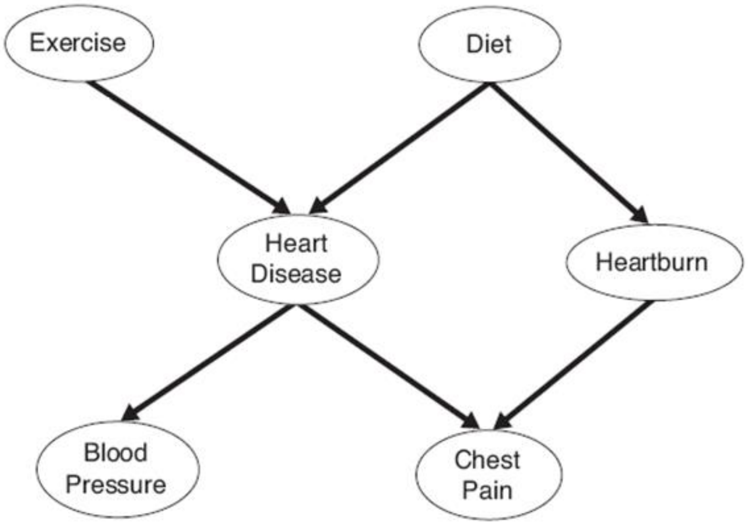


Similarly,

$$P(BP = high \mid D = Healthy, E = Yes)$$
$$= \sum_{\gamma} P(BP = high \mid HD = \gamma, D = Healthy, E = Yes) \times P(HD = \gamma \mid D = Healthy, E = Yes)$$

Proof:

$$P(BP = high) = \sum_{\gamma} P(BP = high \mid HD = \gamma) \times P(HD = \gamma)$$



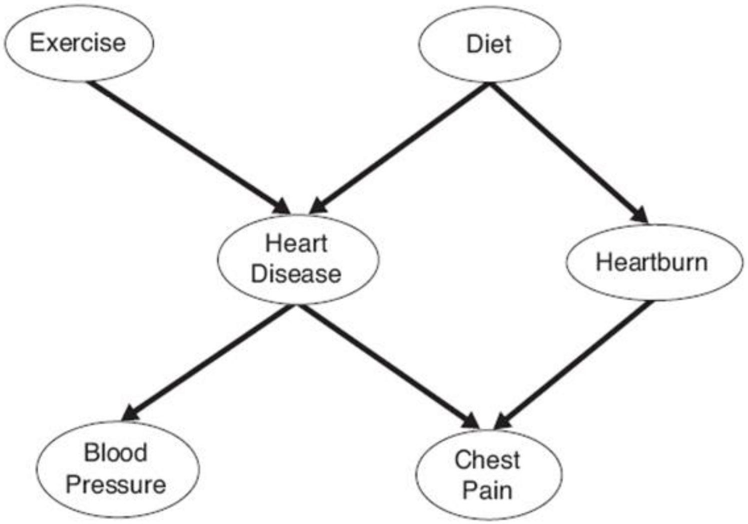
Adding conditions *D= Healthy* and *E= Yes*

we get,

$$P(BP = high \mid D = Healthy, E = Yes)$$
$$= \sum_{\gamma} P(BP = high \mid HD = \gamma, D = Healthy, E = Yes) \times P(HD = \gamma \mid D = Healthy, E = Yes)$$

Similarly,

$$P(BP = high \mid D = Healthy, E = Yes)$$
$$= \sum_{\gamma} P(BP = high \mid HD = \gamma, D = Healthy, E = Yes) \times P(HD = \gamma \mid D = Healthy, E = Yes)$$



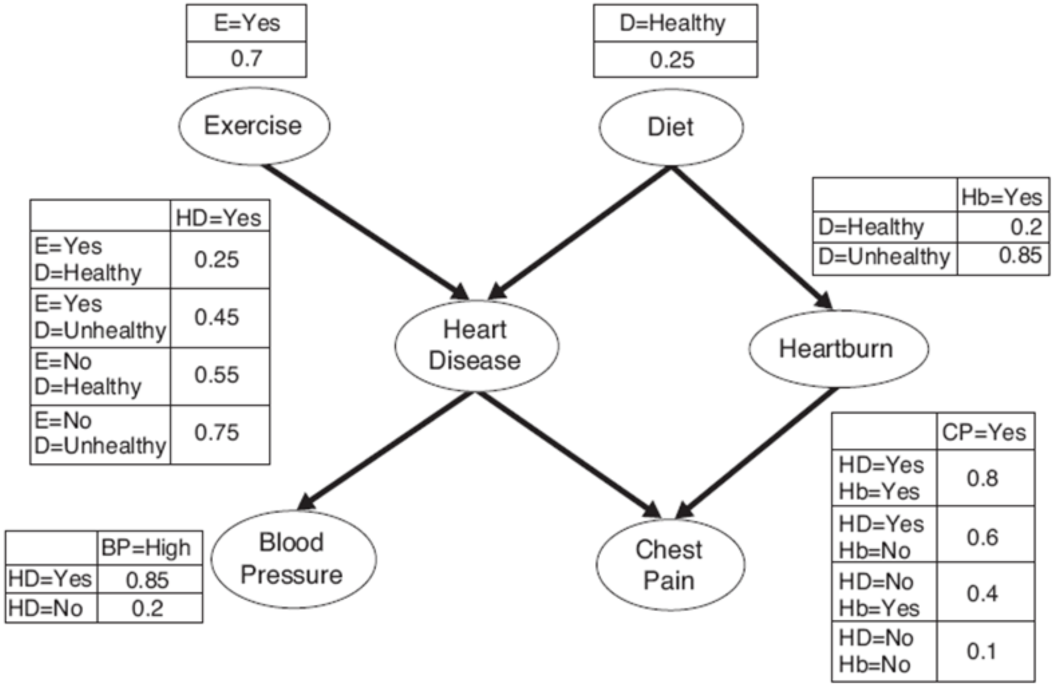
Proof:

$$P(BP = high) = \sum_{\gamma} P(BP = high \mid HD = \gamma) \times P(HD = \gamma)$$

Adding conditions *D= Healthy* and *E= Yes*

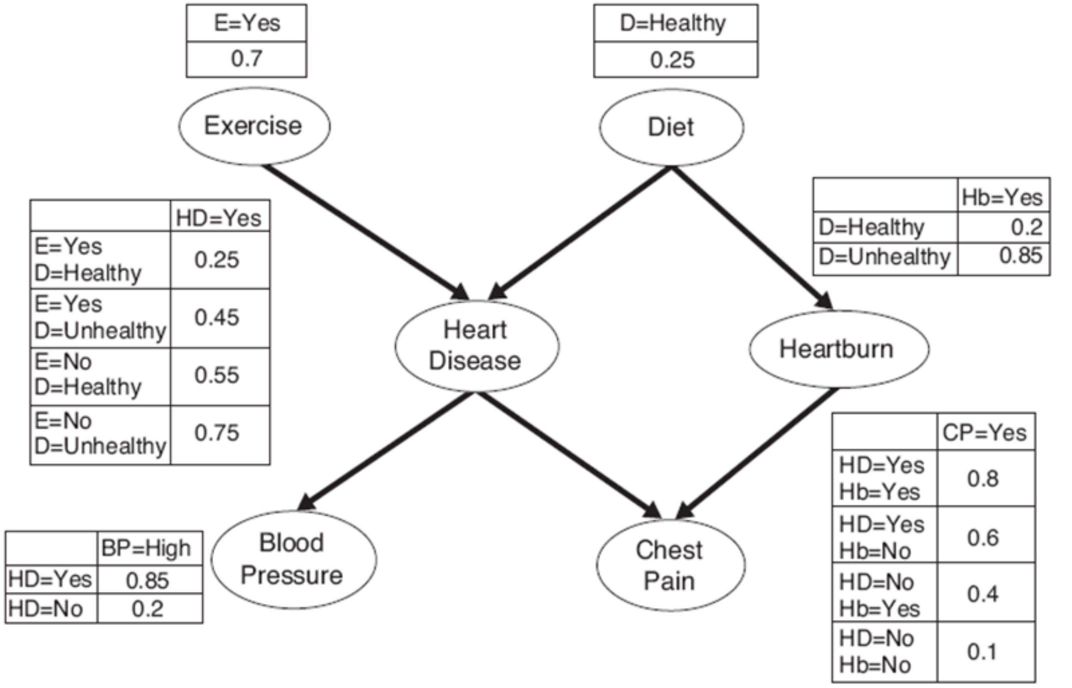
we get,

$$P(BP = high \mid D = Healthy, E = Yes)$$
$$= \sum_{\gamma} P(BP = high \mid HD = \gamma, D = Healthy, E = Yes) \times P(HD = \gamma \mid D = Healthy, E = Yes)$$
$$= \sum_{\gamma} P(BP = high \mid HD = \gamma) \times P(HD = \gamma \mid D = Healthy, E = Yes)$$



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

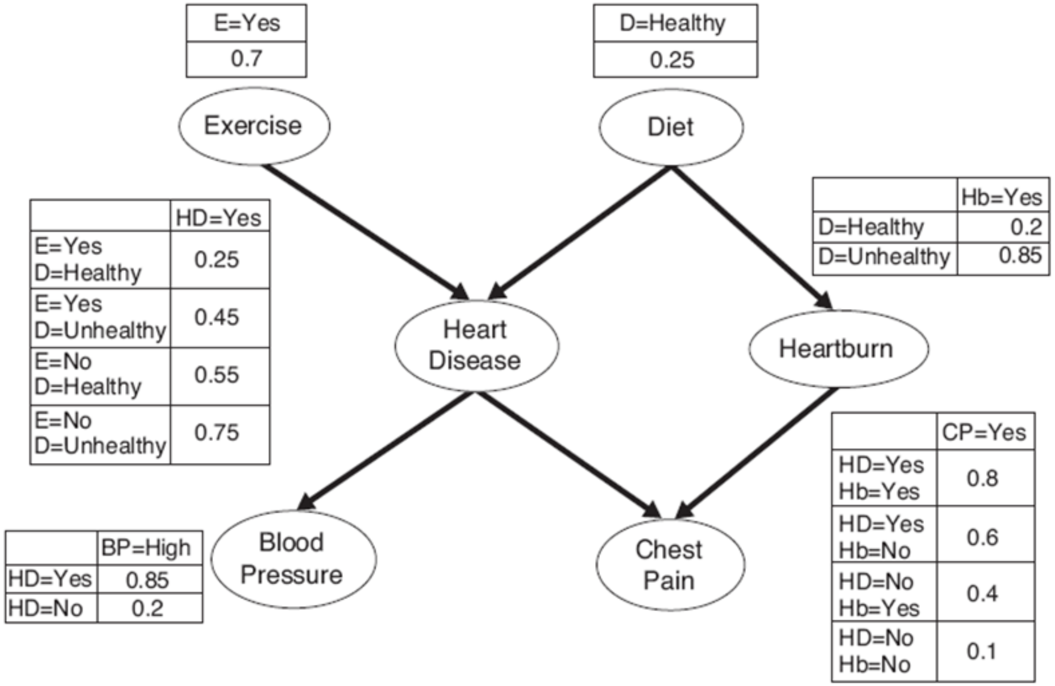
$$= \frac{P(BP = high \mid HD = yes, D = Healthy, E = Yes)}{P(BP = high \mid D = Healthy, E = Yes)} \times P(HD = yes \mid D = Healthy, E = Yes)$$



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

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

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

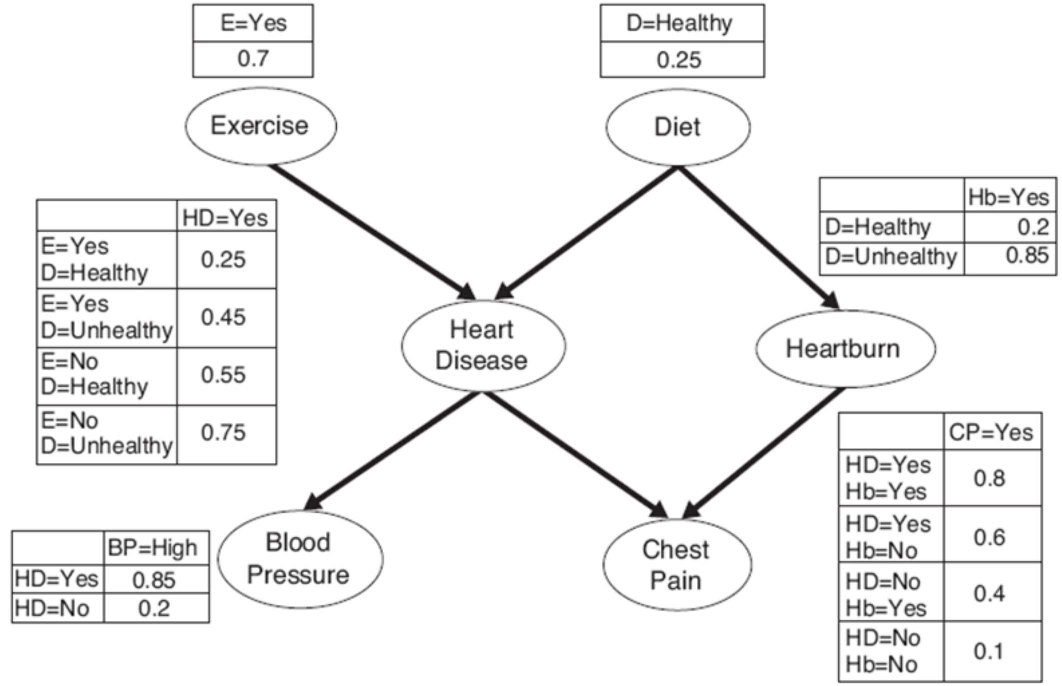


$$P(HD = yes \mid BP = high, D = Healthy, E = Yes)$$

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

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

$$= \frac{P(BP = high \mid HD = yes)}{\sum_{\gamma} P(BP = high \mid HD = \gamma) P(HD = \gamma \mid D = Healthy, E = Yes)} \times P(HD = yes \mid D = Healthy, E = Yes)$$



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

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

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

$$= \frac{P(BP = high \mid HD = yes)}{\sum_{\gamma} P(BP = high \mid HD = \gamma) P(HD = \gamma \mid D = Healthy, E = Yes)} \times P(HD = yes \mid D = Healthy, E = Yes)$$

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

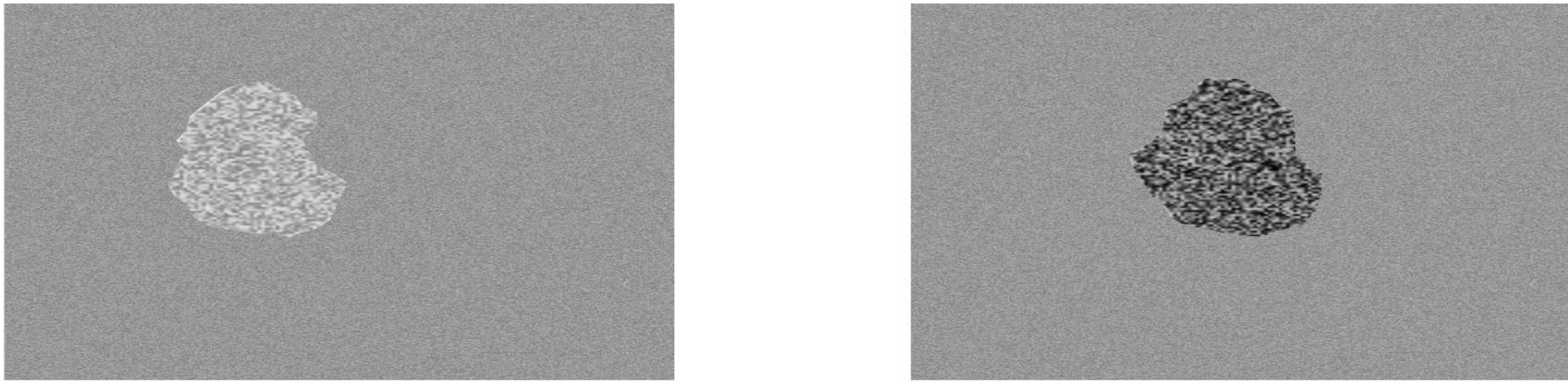
Review of Bayesian Classifier and its variants

- underlying probability densities were known
- training sample are used to estimate the probabilities

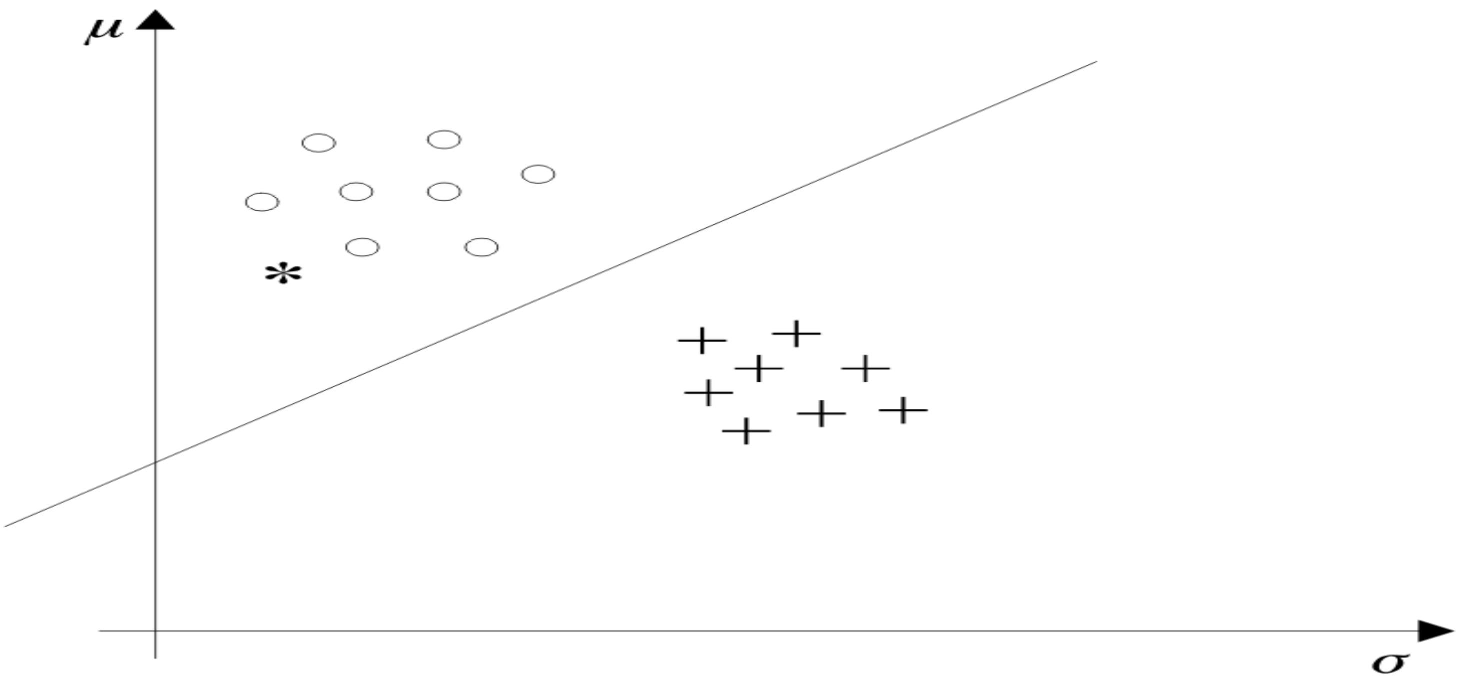
Linear Classifier: Introduction

- Classifies linearly separable patterns
- Assume proper forms for the discriminant functions
- may not be optimal
- very simple to use

Recall from Lecture 1



(a) (b)



Linear discriminant functions and decisions surfaces

- Definition

Let a pattern vector $\mathbf{x} = \{x_1, x_2, x_3, \ldots,\}$

a weight vector $\mathbf{w} = \{w_1, w_2, w_3, \ldots,\}$

A discriminant function :

$$g(\mathbf{x})=x_1w_1+x_2w_2+x_3w_3+\ldots$$

OR

$$g(\mathbf{x}) = w^t x + w_0 \qquad (1)$$

where w is the weight vector and w_0 the bias

Linear discriminant functions and decisions surfaces

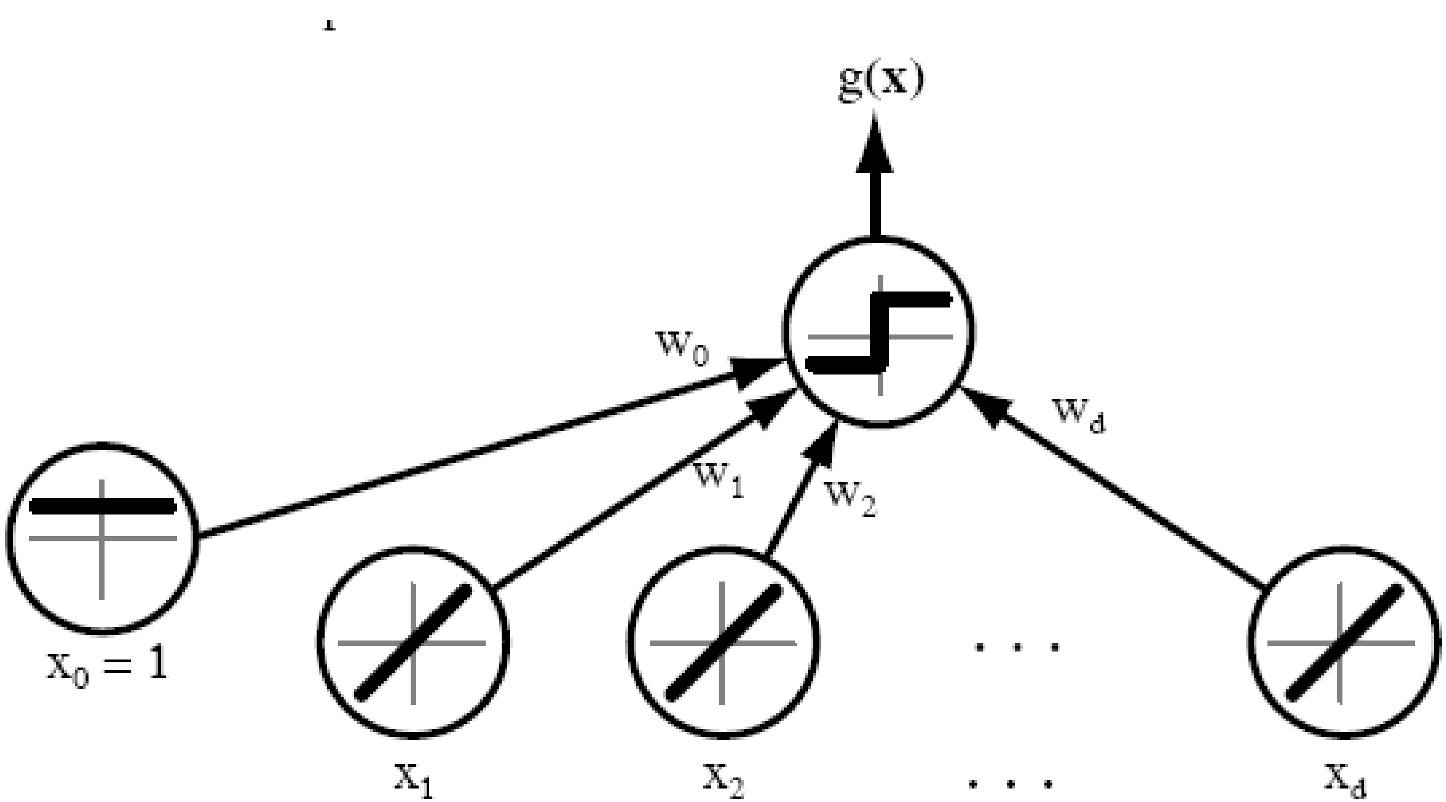
- Classify a new pattern **x** as follows

  Decide class ω_1 *if* $g(\mathrm{x}) > 0$

  and class ω_2 if $g(\mathrm{x}) < 0$

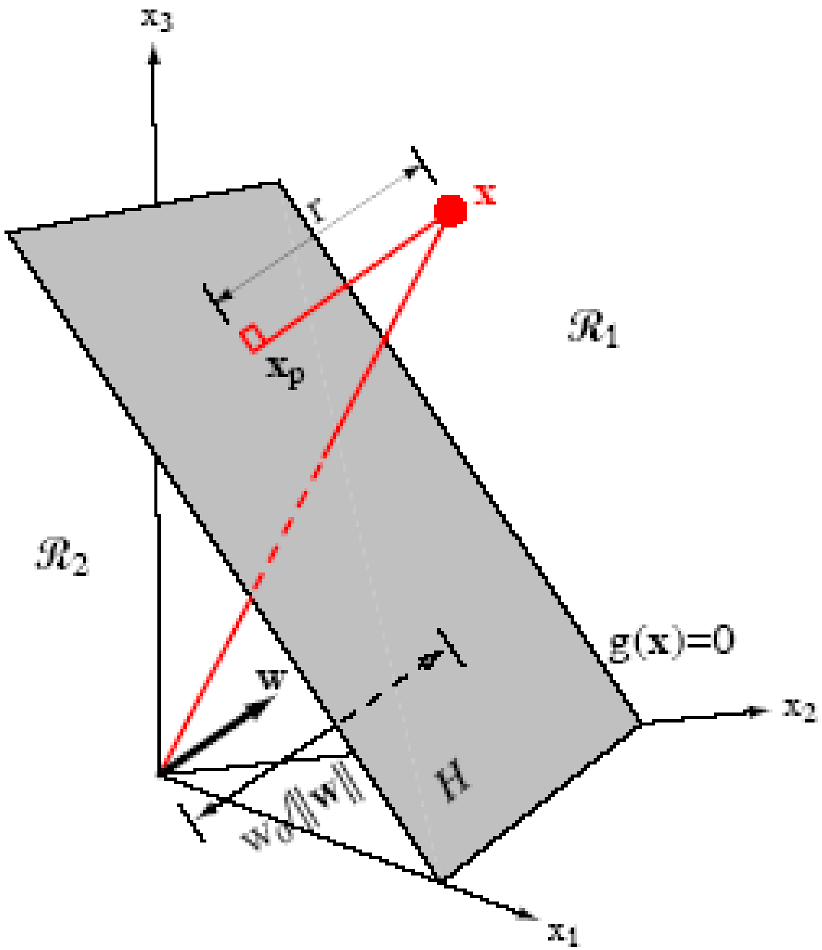
  If $g(x) = 0 \Rightarrow x$ is assigned to either class

Linear discriminant functions and decisions surfaces

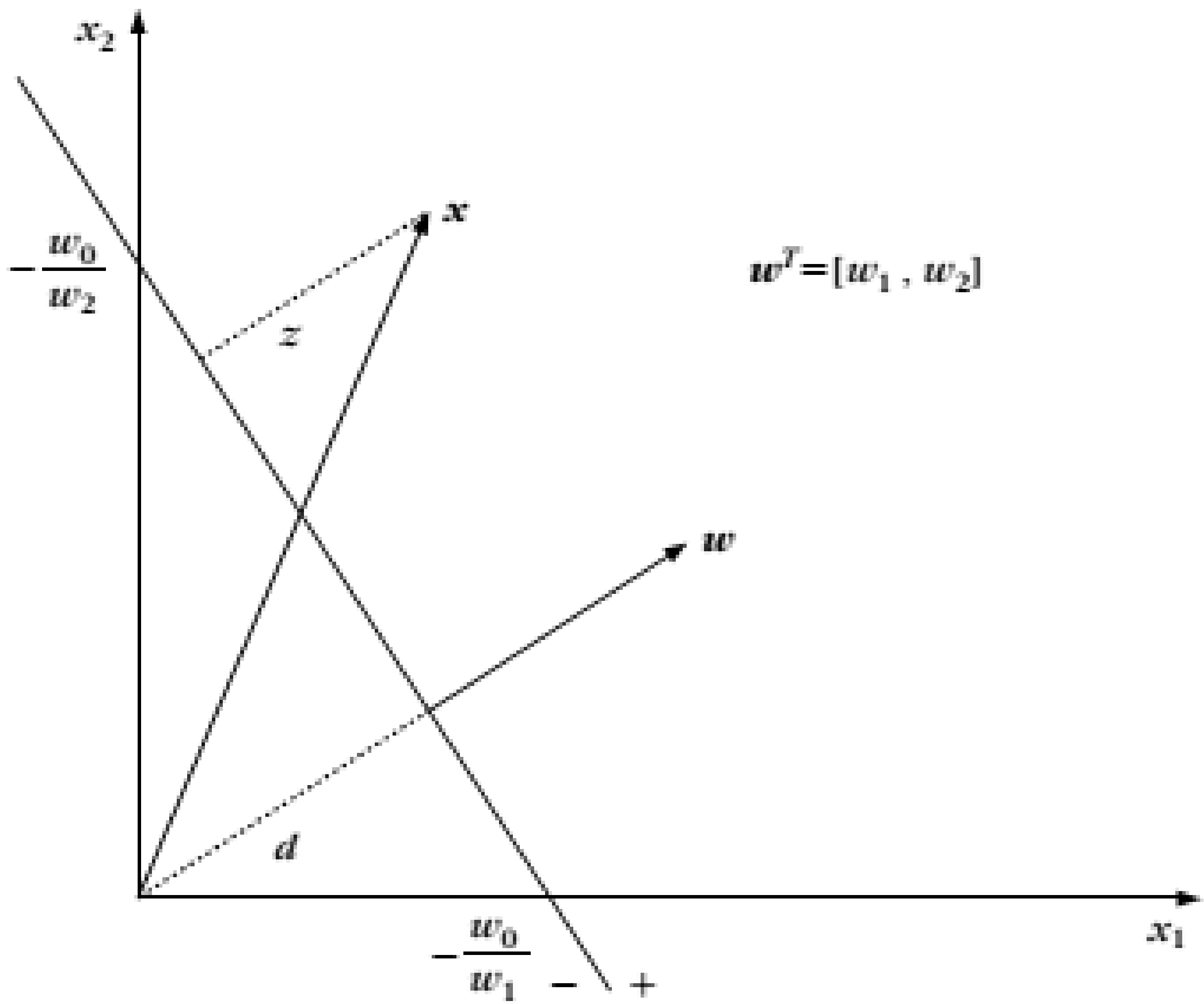


– The equation $g(x) = 0$ is the decision surface that separates patterns

– When $g(x)$ is linear, the decision surface is a hyperplane



x3
r
x
x
p
ℛ1
ℛ2
g(x)=0
w
x2
H
w0/||w||
x1



x_2
$-\frac{w_0}{w_2}$
x
$w^T = [w_1, w_2]$
z
w
d
x_1
$-\frac{w_0}{w_1}$
$-$
$+$

A little bit mathematics

- The Problem: Consider a two class task with ω_1, ω_2
  - $g(\underline{x}) = \underline{w}^T \underline{x} + w_0 = 0 =$
    $w_1 x_1 + w_2 x_2 + \ldots + w_l x_l + w_0$
  - Assume $\underline{x}_1, \underline{x}_2$ on the decision hyperplane:
    $0 = \underline{w}^T \underline{x}_1 + w_0 = \underline{w}^T \underline{x}_2 + w_0 \Rightarrow$
    $\underline{w}^T (\underline{x}_1 - \underline{x}_2) = 0 \ \ \forall \underline{x}_1, \underline{x}_2$

- Hence:

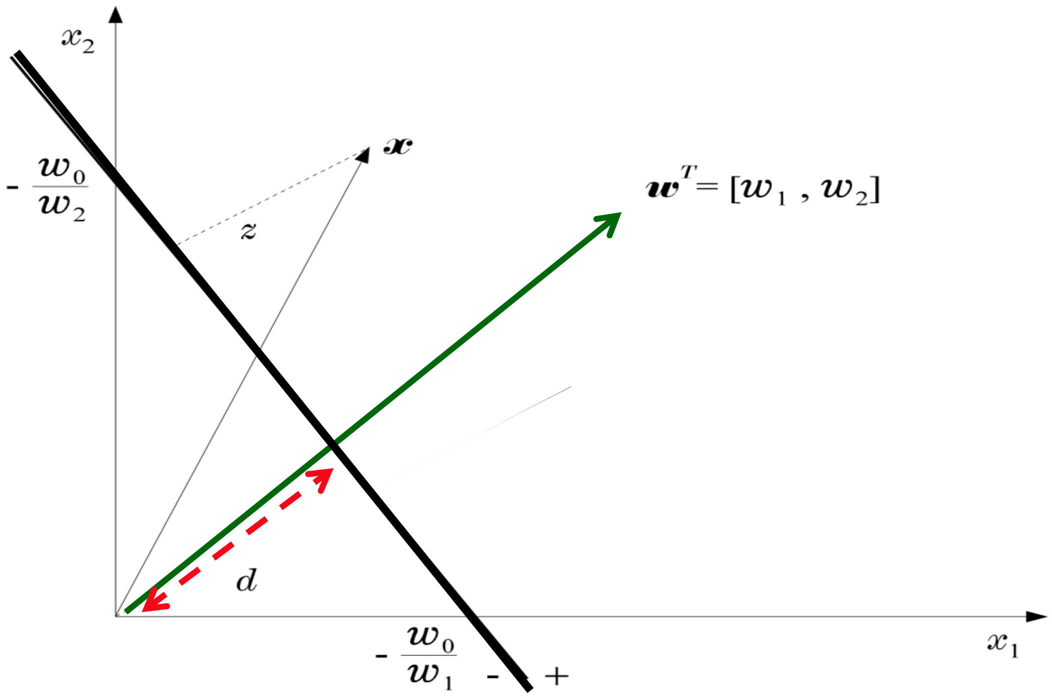
$$\boxed{\underline{w} \perp \text{on the hyperplane}}$$

$$g(\underline{x}) = \underline{w}^T \underline{x} + w_0 = 0$$

- Hence:

$$\boxed{\underline{w} \perp \text{on the hyperplane}}$$

$$g(\underline{x}) = \underline{w}^T \underline{x} + w_0 = 0$$



$$d = \frac{|w_0|}{\sqrt{w_1^2 + w_2^2}}, \quad z = \frac{|g(\underline{x})|}{\sqrt{w_1^2 + w_2^2}}$$

- The Perceptron Algorithm

  – Assume linearly separable classes, i.e.,

$$\exists \ \underline{w}^* : \ \underline{w}^{*T} \underline{x} > 0 \ \ \forall \underline{x} \in \omega_1$$

$$\underline{w}^{*T} \underline{x} < 0 \ \ \forall \underline{x} \in \omega_2$$

- The Perceptron Algorithm

  - Assume linearly separable classes, i.e.,

$$\exists\ \underline{w}^*:\ \underline{w}^{*T}\ \underline{x} > 0\ \ \forall \underline{x} \in \omega_1$$
$$\underline{w}^{*T}\ \underline{x} < 0\ \ \forall \underline{x} \in \omega_2$$

  - The case $\underline{w}^{*T}\ \underline{x} + w_0^*$
    falls under the above formulation, since

    - $\underline{w}' \equiv \begin{bmatrix} \underline{w}^* \\ w_0^* \end{bmatrix},\ \underline{x}' = \begin{bmatrix} \underline{x} \\ 1 \end{bmatrix}$

    - $\underline{w}^{*T}\ \underline{x} + w_0^* = \underline{w}'^T\ \underline{x}' = 0$

– Our goal: Compute a solution, i.e., a hyperplane $\underline{w}$, so that

$$\underline{w}^T \underline{x} (><) 0 \quad \underline{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

- The steps
  - Define a cost function to be minimized
  - Choose an algorithm to minimize the cost function
  - The minimum corresponds to a solution

– The Cost Function

$$J(\underline{w}) = \sum_{\underline{x} \in Y} (\delta_x \underline{w}^T \underline{x})$$

- Where Y is the subset of the vectors wrongly classified by $\underline{w}$.

- $$\delta_x = -1 \text{ if } \underline{x} \in Y \text{ and } \underline{x} \in \omega_1$$
  $$\delta_x = +1 \text{ if } \underline{x} \in Y \text{ and } \underline{x} \in \omega_2$$

– The Cost Function

$$J(\underline{w}) = \sum_{\underline{x} \in Y} (\delta_x \underline{w}^T \underline{x})$$

- Where Y is the subset of the vectors wrongly classified by $\underline{w}$.
- when Y=(empty set) a solution is achieved and

$$J(\underline{w}) = 0$$

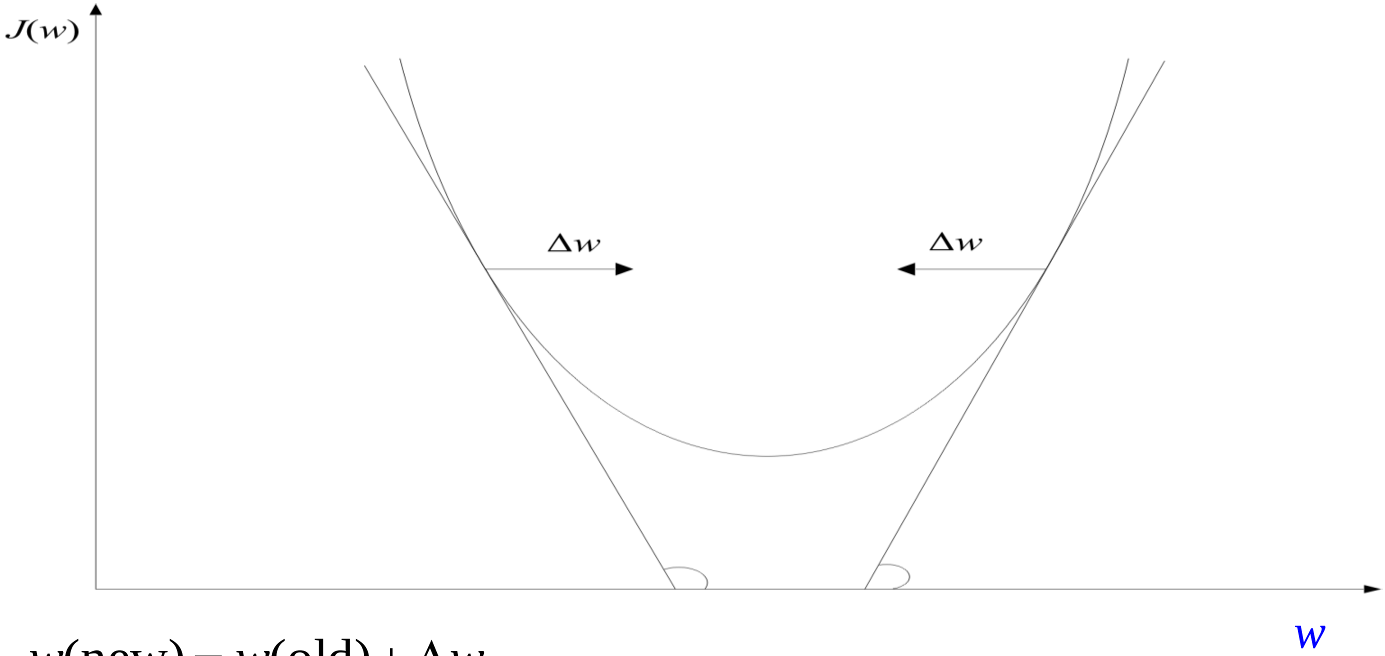
otherwise

$$J(\underline{w}) \geq 0$$

- $J(\underline{w})$ is piecewise linear (WHY?)

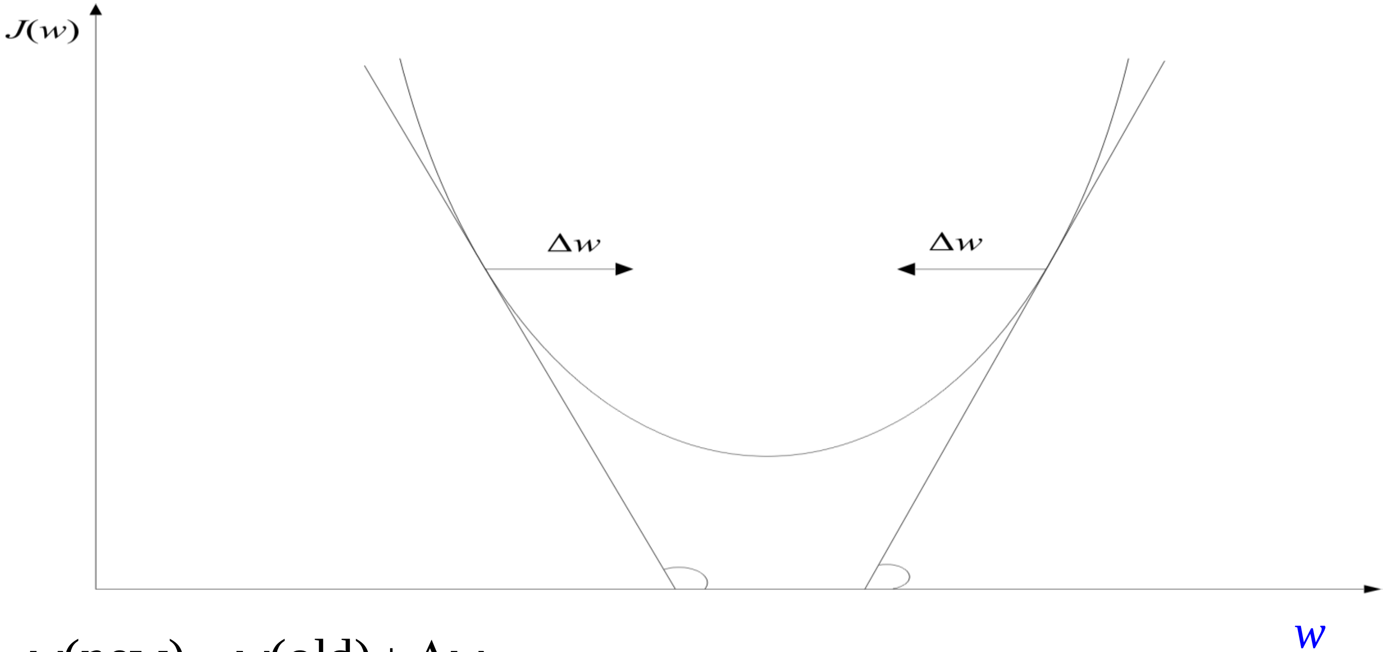
– The Algorithm

- The philosophy of the gradient descent is adopted.



$$\underline{w}(\text{new}) = \underline{w}(\text{old}) + \Delta\underline{w}$$

$$\Delta\underline{w} = -\mu \frac{\partial J(\underline{w})}{\partial \underline{w}} \Big| \underline{w} = \underline{w}(\text{old})$$



$$\underline{w}(\text{new}) = \underline{w}(\text{old}) + \Delta \underline{w}$$

$$\Delta \underline{w} = -\mu \frac{\partial J(\underline{w})}{\partial \underline{w}} \Big|_{\underline{w} = \underline{w}(\text{old})}$$

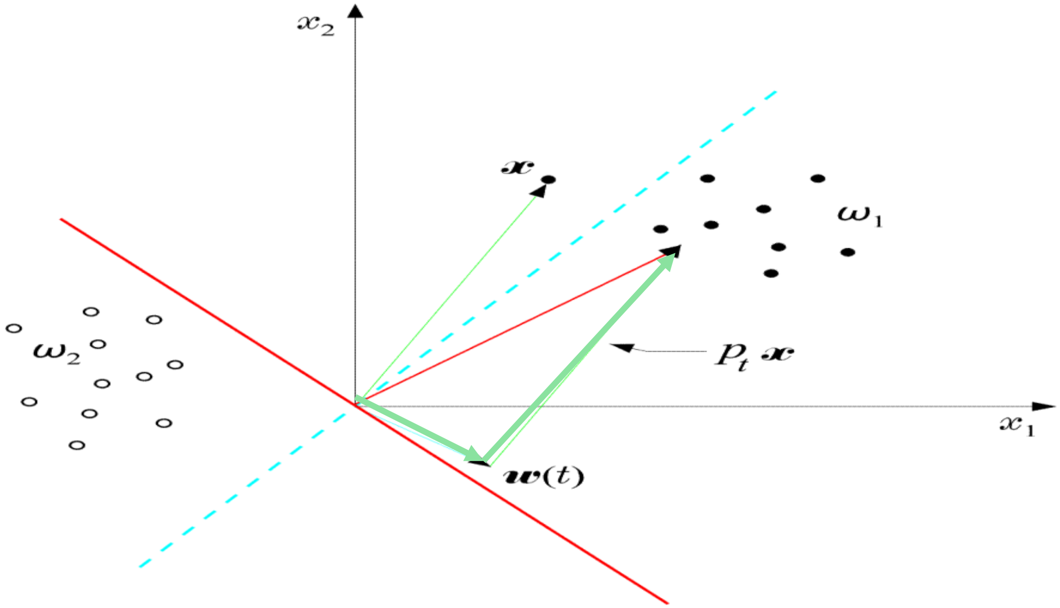
- Wherever valid

$$\frac{\partial J(\underline{w})}{\partial \underline{w}} = \frac{\partial}{\partial \underline{w}} \left(\sum_{\underline{x} \in Y} \delta_x \underline{w}^T \underline{x} \right) = \sum_{\underline{x} \in Y} \delta_x \underline{x}$$

- $$\boxed{\underline{w}(t+1) = \underline{w}(t) - \rho_t \sum_{\underline{x} \in Y} \delta_x \underline{x}}$$

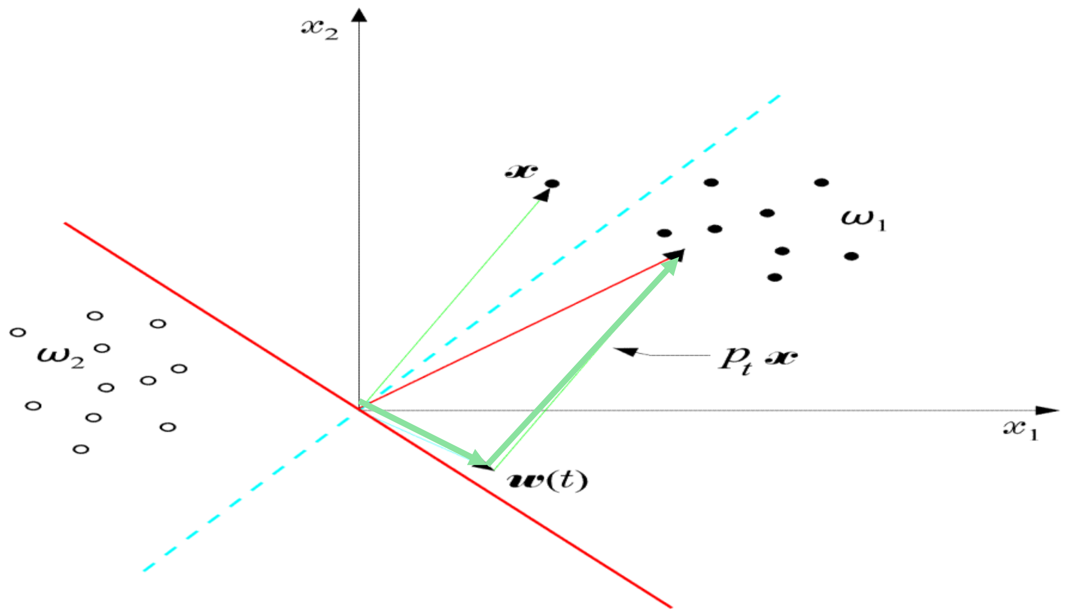
This is the celebrated Perceptron Algorithm

– An example:



$$\underline{w}(t+1) = \underline{w}(t) - \rho_t \delta_x \underline{x}$$
$$= \underline{w}(t) + \rho_t \underline{x} \quad (\delta_x = -1)$$

– An example:



$$\underline{w}(t+1) = \underline{w}(t) - \rho_t \delta_x \underline{x}$$
$$= \underline{w}(t) + \rho_t \underline{x} \quad (\delta_x = -1)$$

– The perceptron algorithm **converges** in a **finite** number of iteration steps to a solution **if patterns are linearly separable**

– Example: At some stage t the perceptron algorithm results in

$$w_1 = 1,\ w_2 = 1,\ w_0 = -0.5$$

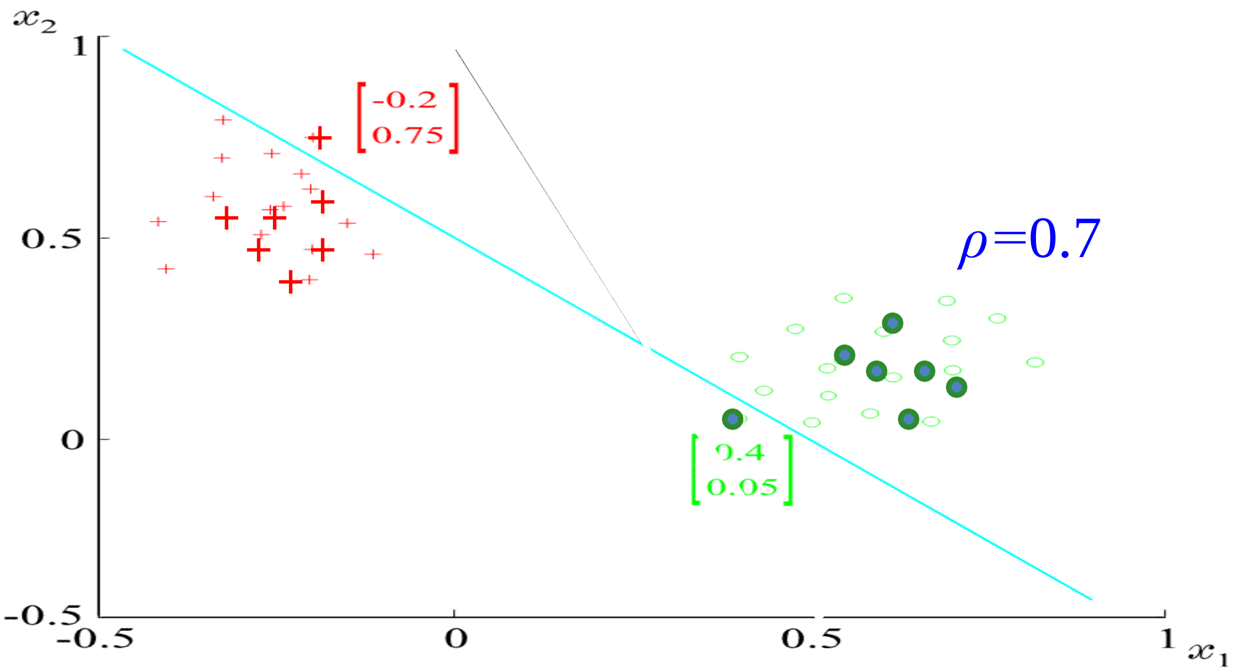
$$x_1 + x_2 - 0.5 = 0$$

– Example: At some stage t the perceptron algorithm results in

$$w_1 = 1,\ w_2 = 1,\ w_0 = -0.5$$

$$x_1 + x_2 - 0.5 = 0$$

The corresponding hyperplane is

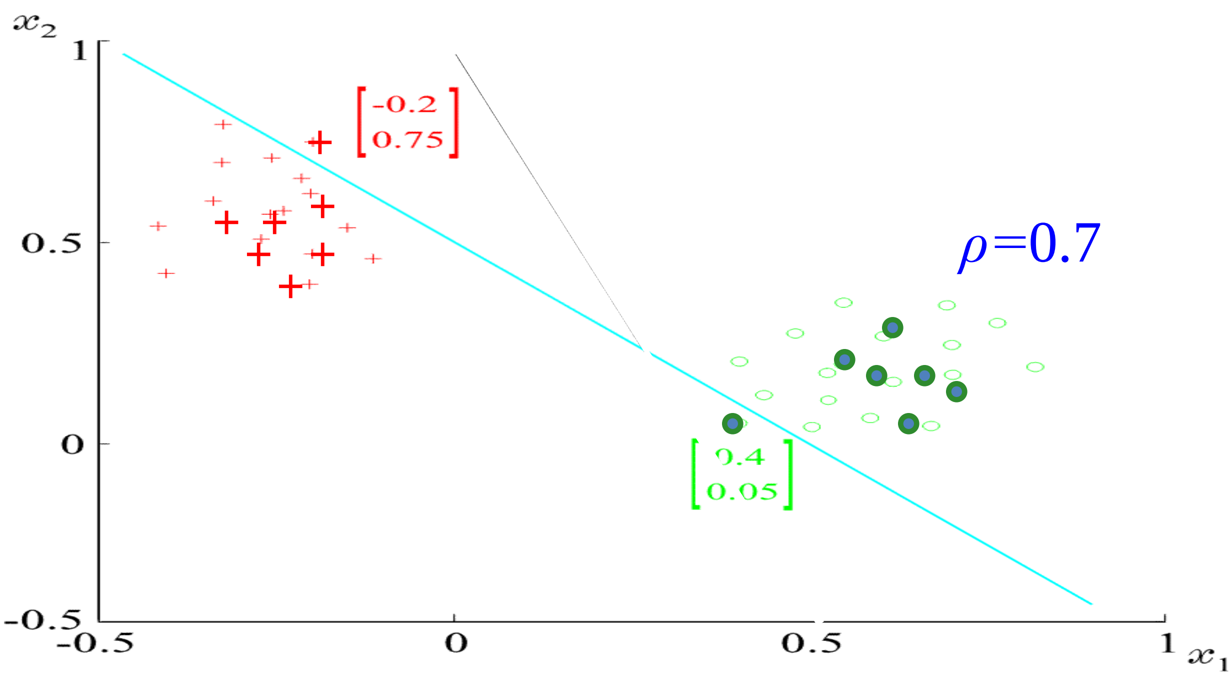


– Example: At some stage t the perceptron algorithm results in

$$w_1 = 1,\ w_2 = 1,\ w_0 = -0.5$$

$$x_1 + x_2 - 0.5 = 0$$

The corresponding hyperplane is



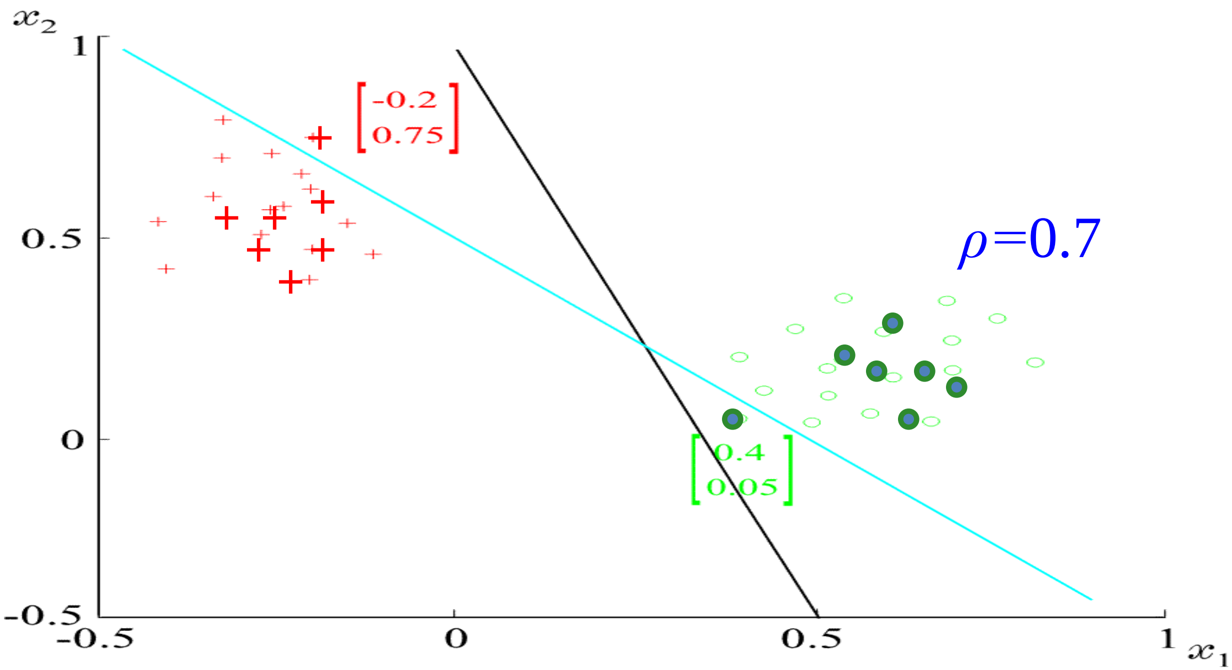
$$\underline{w}(t+1) = \begin{bmatrix} 1 \\ 1 \\ -0.5 \end{bmatrix} - 0.7(-1)\begin{bmatrix} 0.4 \\ 0.05 \\ 1 \end{bmatrix} - 0.7(+1)\begin{bmatrix} -0.2 \\ 0.75 \\ 1 \end{bmatrix} = \begin{bmatrix} 1.42 \\ 0.51 \\ -0.5 \end{bmatrix}$$

– Example: At some stage t the perceptron algorithm results in

$$w_1 = 1,\ w_2 = 1,\ w_0 = -0.5$$

$$x_1 + x_2 - 0.5 = 0$$

The corresponding hyperplane is



$$\underline{w}(t+1) = \begin{bmatrix} 1 \\ 1 \\ -0.5 \end{bmatrix} - 0.7(-1)\begin{bmatrix} 0.4 \\ 0.05 \\ 1 \end{bmatrix} - 0.7(+1)\begin{bmatrix} -0.2 \\ 0.75 \\ 1 \end{bmatrix} = \begin{bmatrix} 1.42 \\ 0.51 \\ -0.5 \end{bmatrix}$$