Clasificación

Clasificadores Bayesianos

Clasificadores Bayesianos

- Modela relaciones probabilisticas entre el conjunto de atributos y el atributo clase
- Probabilidad condicional: probabilidad de que una variable aleatoria pueda tomar un valor particular dado el valor de otra variable aleatoria

$$P(Y=y \mid X=x)$$

se refiere a la probabillidad que la variable Y puede tomar el valor de y dado que la variable X toma el valor de x

Teorema de Bayes

- Las probabilidades condicionales de X y Y estan relacionadas:

  $$P(X,Y) = P(Y|X)\,P(X) = P(X|Y)\,P(Y)$$

- **Teorema de Bayes**

  $$P(Y|X) = P(X|Y) \cdot P(Y) \,/\, P(X)$$

- Ejercicio

  Football: 2 equipos. Equipo 0 gana el 65%, equipo 1 gana 35%. De los juegos ganados por el equipo 0, el 30% son jugados en la cancha del equipo 1. El 75%, de las victorias del equipo 1 son ganados cuando juegan en casa. Si el equipo 1 es local el siguiente juego, cual equipo es el favorito a ganar?

Ejemplo

- X variable aleatoria que representa el equipo local
- Y variable aleatoria que representa el ganador
- Probabilidad que equipo 0 gane: $P(Y=0) = 0.65$
- Probabilidad que equipo 1 gane: $P(Y=1) = 0.35$
- Probabilidad de que el equipo 1 juegue como local gane:

$$P(X=1|Y=1) = 0.75$$

- Probabilidad de que el equipo 1 juegue como local y equipo 0 gane:

$$P(X=1|Y=0) = 0.3$$

Ejemplo

- Objetivo

  $P(Y=1|X=1)$ probabilidad condicional de que el equipo 1 gane el siguiente juego estando como local, y comparar con $P(Y=0|X=1)$

- Usando Bayes

  $P(Y=1|X=1) = P(X=1|Y=1)\, P(Y=1) / P(X=1)$ Ley de probabilidad total

  $= P(X=1|Y=1)\, P(Y=1) / P(X=1,Y=1)+P(X=1,Y=0)$

  $= P(X=1|Y=1)\, P(Y=1) / P(X=1|Y=1)P(Y=1) + P(X=1|Y=0)P(Y=0)$

  $= 0.75x0.35/(0.75x0.35 + 0.3x0.65) = 0.5738$

- $P(Y=0|X=1) = 1 - P(Y=1|X=1) = 0.4262$

Equipo1 tiene mas oportunidad de ganar

Clasificador Bayesiano

- X conjunto de atributos
- Y clase
- X y Y tratadas como variables aleatorias y la relación probabilística entre ellas es

$$P(Y|X)$$

Probabilidad posterior para Y

En el entrenamiento las probabilidades posteriores por cada combinación de X y Y son obtenidas

Clasificador Bayesiano

- El problema puede ser formalizado usando probabilidades a-posteriori

  P(Y|X) = probabilidad que el ejemplo $X=<x_1,\ldots,x_k>$ sea de la clase Y.
- Conjunto de test X' puede ser clasificado encontrando la clase Y' que maximice la probabilidad posterior P(Y'|X')

Problema de tomar una decisión

Dada las condiciones del clima, es posible jugar tennis?

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

Estimando probabilidades posteriores

- Teorema de Bayes: $P(Y|X) = P(X|Y) \cdot P(Y) / P(X)$
- $P(X)$ es constante por todas las clases (puede ser ignorado)
- $P(Y)$ = (prior probablities): $P(Y_i) = s_i/s$
- Y tal que $P(Y|X)$ es maxima=
  Y tal que $P(X|Y) \cdot P(Y)$ is maxima
- Problema: computo de $P(X|Y)$ no es disponible!

Naive Bayes Clasifier y Bayesian belief Network

Naïve Bayesian Classification

- Supuesto "Naïve": independencia de atributos

  $$P(x_1,\ldots,x_k|Y) = P(x_1|Y)\cdot\ldots\cdot P(x_k|Y)$$

- Si i-esimo atributo es categórico: $P(x_i|Y)$ es estimado como la frequencia relativa de ejemplos que tienen valor x_i como i-esimo atributo en clase Y:
  - $P(x_i|Y) = \dfrac{\text{\# de ejemplos en clase Y w/ i-th atributo} = x_i}{\text{\# de ejemplos en clase Y}}$

Naïve Bayesian Classification

- Si i-esimo atributo es continuo:
  $P(x_i|Y_j)$ es estimado con la función de densidad Gauss
  - compute mean ($\mu_{j,i}$) and stand. Deviation ($\sigma_{j,i}$) for EACH attribute (*i*) using data from *j*-th class (C_j) only
  - Densities:

$$P(x_i \mid Y_j) = \frac{1}{\sqrt{2\pi}\sigma_{j,i}} e^{-\frac{(x_i - \mu_{j,i})^2}{2\sigma_{j,i}^2}}$$

Probabilidades condicionales por clase $P(x_i|Y_j)$

- Computo de probabilidades(4 atributos, 2 clases)
- 14 ejemplos (9 positivos, 5 negativos)

Outlook	P	N		Humidity	P	N
sunny	2/9	3/5		high	3/9	4/5
overcast	4/9	0		normal	6/9	1/5
rain	3/9	2/5				
Tempreature				Windy		
hot	2/9	2/5		true	3/9	3/5
mild	4/9	2/5		false	6/9	2/5
cool	3/9	1/5				

Ejemplo:

- Nuevo ejemplo X = <rain, hot, high, false>

  $$P(Y|X) = P(X|Y) \cdot P(Y)$$

- $P(X|p) \cdot P(p) =$
  $P(rain|p) \cdot P(hot|p) \cdot P(high|p) \cdot P(false|p) \cdot P(p) =$
  $3/9 \cdot 2/9 \cdot 3/9 \cdot 6/9 \cdot 9/14 = 0.010582$
- $P(X|n) \cdot P(n) =$
  $P(rain|n) \cdot P(hot|n) \cdot P(high|n) \cdot P(false|n) \cdot P(n) =$
  $2/5 \cdot 2/5 \cdot 4/5 \cdot 2/5 \cdot 5/14 = 0.018286$

- Ejemplo X is clasificado en clase n (No jugar)

Supuesto de independencia

- … computacion posible
- … optimo clasificador cuando el supuesto se satisface
- … pero raro en la realidad, en la mayoria los atributos son correlacionados
- Intentos de manejar esta limitación:
  - Redes Bayesianas, combinan el razonamiento bayesiano con causal relationships (relaciones casuales) entre atributos
  - Árboles de decisión, analiza un atributo a la vez, considerando los mas importantes atributos primero

Bayesian Belief Networks
Redes Bayesianas

- Modelar la probabilidad condicional de clases

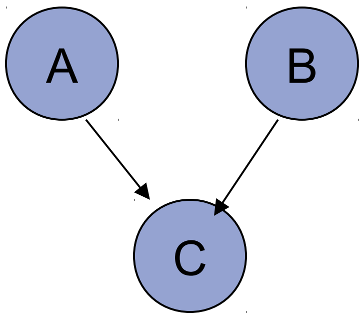
P(X|Y) sin el supuesto de independencia

- Permite especificar que par de atributos son condicionalmente independientes

1. Representación y construcción del modelo
2. Inferencia sobre el modelo

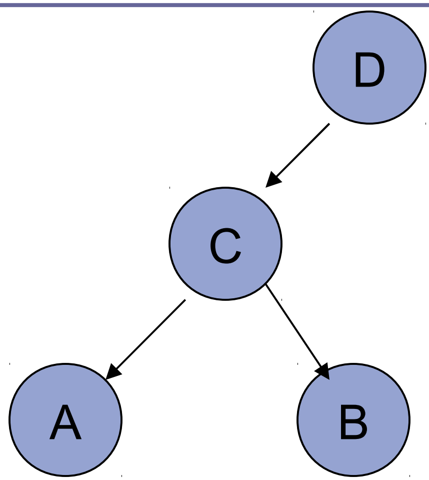
Representación del modelo

- Representación grafica de las relaciones probabilísticas entre el conjunto de variables aleatorias.
  - Grafo dirigido aciclico (representa las relaciones dependientes entre variables)
  - Tabla de probabilidades (asociando cada nodo con sus nodos padres)



A y B variables independientes. Cada una tiene
Influencia en la variable C
A y B son padres de C
C es hijo de A y B

Representación del modelo (2)



Path directo

D es ancestro de B

A es descendente de D

B no es descendente de A

D no es descendente de A

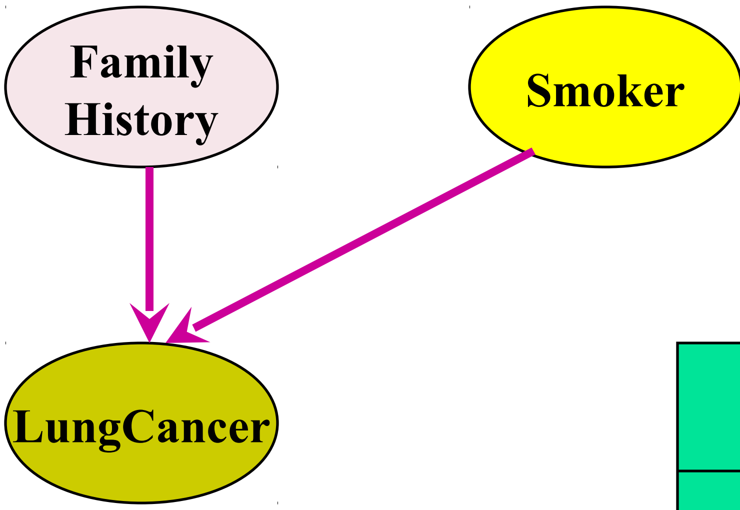
Un nodo en una red bayesiana es condicionalmente independiente de sus no descendientes, si sus padres son conocidos

Representación del modelo (3)

- Tabla de probabilidades
  - Si el nodo X no tiene padres, la tabla contiene la probabilidad a priori P(X)
  - Si el nodo X tiene solo un padre, Y, entonces la tabla contiene la probabilidad condicional P(X|Y)
  - Si el nodo X tiene varios padres $\{Y_1, Y_2, \ldots, Y_{k,}\}$ entonces la tabla contiene la probabilidad condicional $P(X| Y_1, Y_2, \ldots, Y_k)$

Representación del modelo (4)

Valores binarios

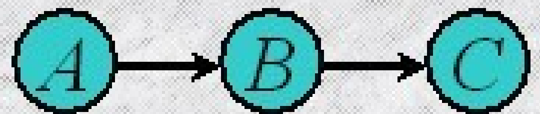


	(FH, S)	(FH, ~S)	(~FH, S)	(~FH, ~S)
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

Construcción del modelo

1. Crear la estructura de la red
   - Algoritmos para generar la topología (garantizar no hay ciclos)
1. Estimar las probabilidades en tablas asociadas a cada nodo

Basic Inference



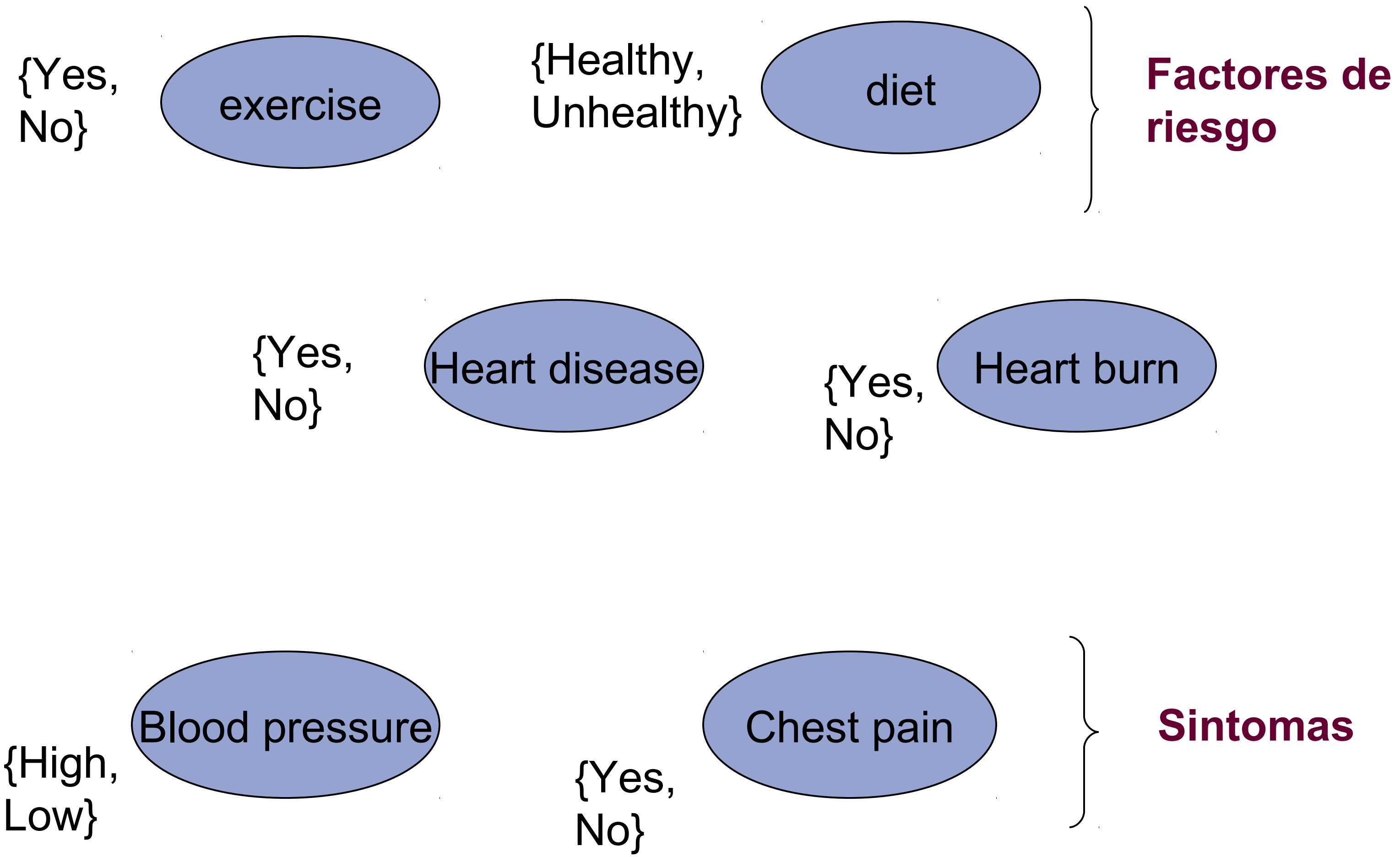
$$P(b) = \sum_{a} P(a, b) = \sum_{a} P(b \mid a)\, P(a)$$

$$P(c) = \sum_{b} P(c \mid b)\, P(b)$$

$$P(c) = \sum_{b,a} P(a, b, c) = \sum_{b,a} P(c \mid b)\, P(b \mid a)\, P(a)$$

$$= \sum_{b} P(c \mid b) \underbrace{\sum_{a} P(b \mid a)\, P(a)}_{P(b)}$$

Ejemplo: Modelar pacientes con enfermedad del corazón o problemas de gastritis



Variables binarias

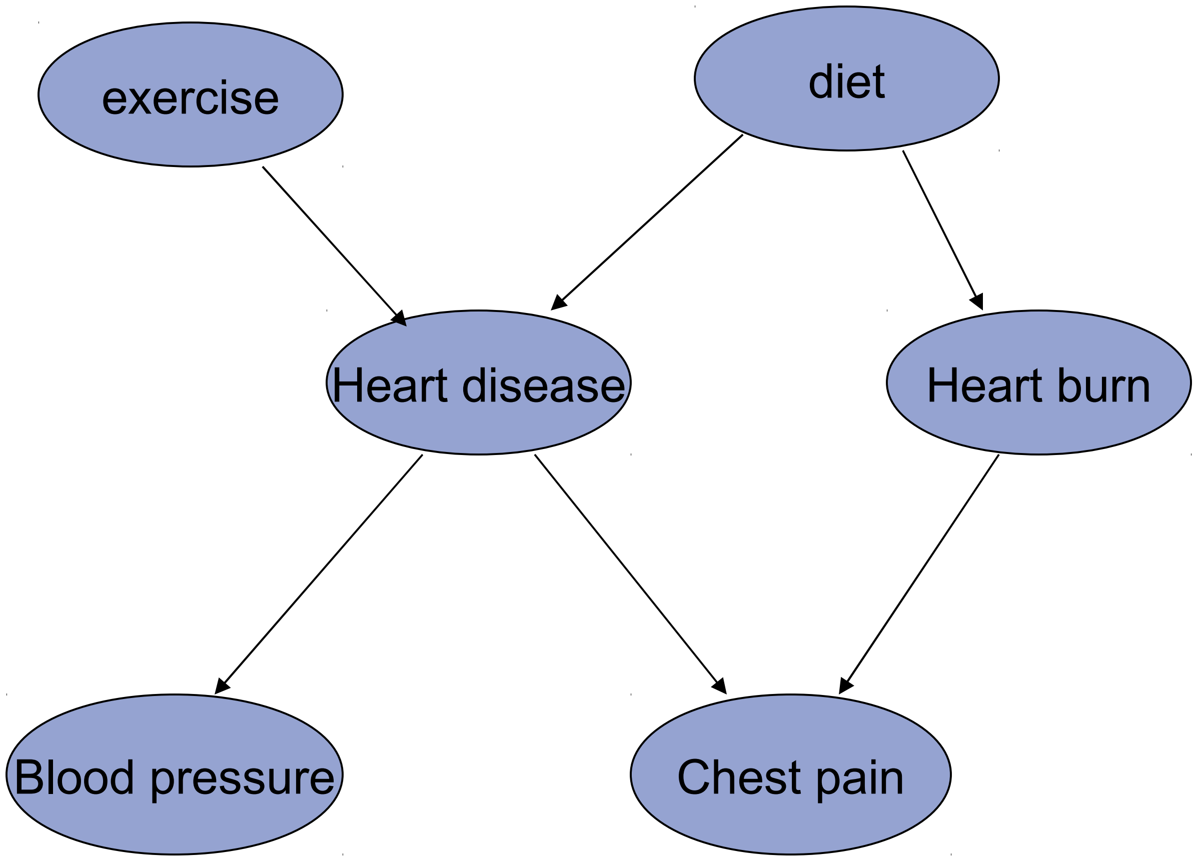
Ejemplo: generando la topología de la red

- Variables ordenadas (E,D,HD,Hb,CP,BP)

P(D|E)	P(D)
P(HD|E, D)	P(HD|E, D) No se puede simplificar
P(Hb|HD,E,D)	P(Hb,D)
P(CP|Hb,HD,E,D)	P(CP|Hb,HD)
P(BP|CP,Hb,HD,E,D)	P(BP|HD)

- Basados en las probabilidades condicionales se crean los arcos entre nodos: **(E,HD),(D,HD),(D,Hb),(HD,CP),(Hb,CP) y (HD,BP)**

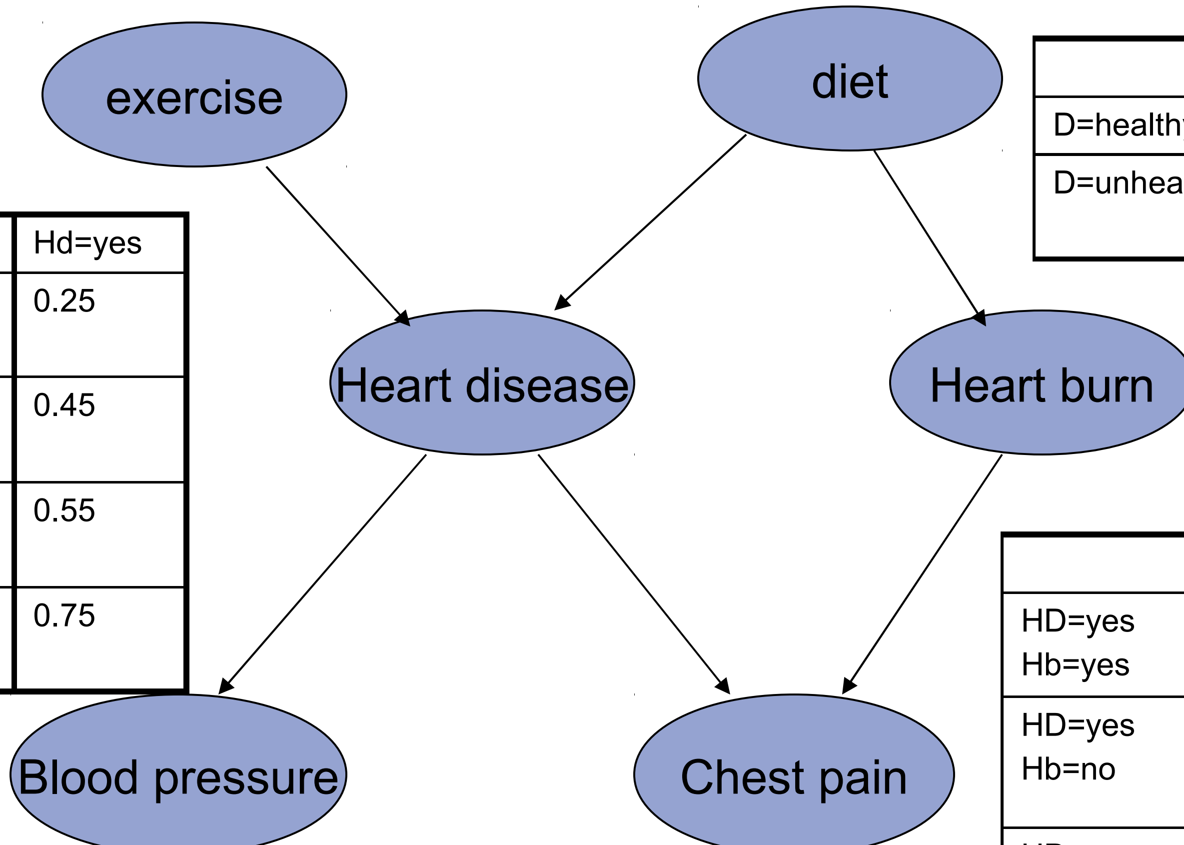
Ejemplo: Estructura de red



Ejemplo: tabla de probabilidades asociada a cada nodo

E=yes
0.7

D=Healthy
0.25



	Hb=yes
D=healthy	0.2
D=unhealthy	0.85

	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

	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

	Bp=high
HD=yes	0.85
HD=no	0.2

Ejemplo: Inferencia

- Diagnosticar cuando una persona esta enferma del corazón.
- Diagnostico puede ser hecho desde diferentes escenarios:
  1. Sin información previa
  2. Alta presión (High Blood preassure)
  3. Alta presión, dieta saludable (Healthy diet) y ejercicio regular (regular exercise)

1. Sin información previa

- Se puede determinar computando las probabilidades a priori:

  P(HD=yes) y P(HD=no)

  Supongamos:

  α Є {yes, no} valores de **exercise**

  β Є {healthy, unhealthy} valores de **diet**

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

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

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

$$= 0.49$$

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

El paciente tiene ligeramente mas probabilidad de no tener la enfermedad

$$P(HD = no) = 1 - P(HD = yes) = 0.51$$

2. Alta presión (High Blood preassure)

- Diagnostico comparando las probabilidades posteriores
  P(HD=yes | BP=high) vs P(HD=no | BP=high)
- Se debe computar P(BP=high)

{yes, no} valores de **Heart disease**

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

$$= 0.85 \times 0.49 + 0.2 \times 0.51 = 0.5185$$

Calculado con las probabilidades apriori

	Bp=high
HD=yes	0.85
HD=no	0.2

- La probabilidad posterior de que la persona tiene enfermedad del corazón es:

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

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

$$P(HD = no \mid BP = high) = 1 - 0.8033 = 0.1967$$

Cuando el paciente tiene presión alta incrementa El riesgo de sufrir Enfermedad del corazón

3.Alta presión, dieta saludable (Healthy diet) y ejercicio regular (regular exercise)

- Tarea para próxima clase!

Características

- Modelo grafico
- Construir la red puede ser costoso. Sin embargo, una vez construida la red, adicionar una nueva variable es directo
- Trabajan bien con datos perdidos (sumando o integrando las probabilidades)
- El modelo es un poco robusto a overfitting