Modelos para el aprendizaje automático

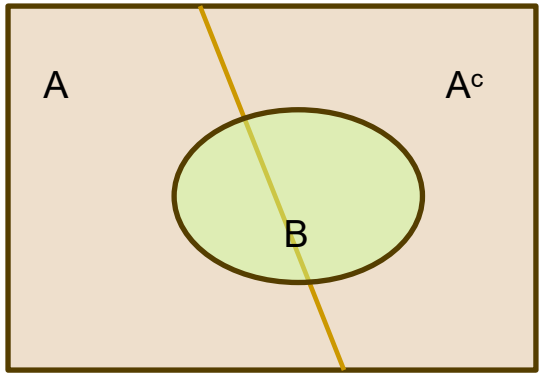
Silvia N. Pérez

Introducción a las Redes Bayesianas

Bibliografía:

• *Learning Bayesian Networks*. Neapolitan, R. (2004). Prentice Hall
• *Bayesian Networks in DM.* Heckerman, D. *DM&KD*. 1,79-119. (1997)
• Redes Bayesianas. Sangüesa i Solé. UOC.

Intro: Fórmula de Bayes



$$P(A/B) = \frac{P(B/A)P(A)}{P(B/A)P(A) + P(B/A^c)P(A^c)}$$

Esto dice que a partir de información de la ocurrencia de B si se dió o no A, podemos construir información acerca de la ocurrencia de A sabiendo que ocurrió B.

Este es el concepto básico para la inferencia que se usa en redes bayesianas.

Método bayesiano para estimación de parámetros

Sea θ un parámetro a estimar a partir de una muestra $x_1, x_2, \ldots, x_n$ proveniente de una distribución con densidad

$f(x \mid \theta)$.

Sea $\pi(\theta)$ la densidad *a priori* θ.

$\pi^*(\theta \mid x_1, x_2, \ldots, x_n)$ es la distribución *a posteriori* que resulta

$$\pi^*(\theta / muestra) \propto L(\theta / x_1, x_2, ..x_n).\pi(\theta)$$

Comparando:

- **Estimación de máxima verosimilitud**
  - Provee justificación a muchos métodos de estimación 'intuitivos'
  - MLE tiene buenas propiedades, se conoce la distribución asintótica
- **Estimación Bayesiana**
  - Incorpora conocimiento o información previa
  - Requiere manejo computacional para hallar distribuciones posterioris (por ejemplo, MCMC: Metrópolis-Hasting, Gibbs, etc.)

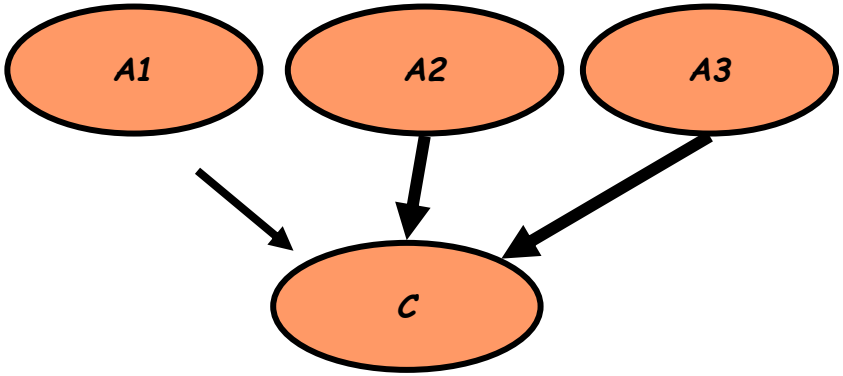
Redes Bayesianas (BN)

- Son estructuras gráficas-simbólicas para representar relaciones probabilísticas entre variables.
- Permiten definir modelos para diagnóstico (hallar P de la causa dados los síntomas) ó predicción (estando presente la causa, hallar P de los síntomas).
- Los nodos pueden se fuente de información u objeto de predicción, según sea la evidencia disponible.
- Se representan con un grafo DAG donde los nodos representan variables aleatorias y los arcos representan dependencias entre ellas.

Modelos gráficos probabilísticos

Grafo DAG (*Directed Acyclic Graph*): cuando no existen caminos dirigidos hacia un mismo nodo (o sea no puede volverse al mismo nodo).

Ejemplo: 4 variables dicotómicas relacionadas por el sig grafo:



Conocer estructuras de dependencia/indep nos permite reducir la información necesaria.

Algunas definiciones previas

- Dado un conjunto de nodos *N,* un *grafo dirigido* en N es un conjunto de arcos (pares ordenados) definidos sobre los elementos de *N.*
- *Camino dirigido*: secuencia ordenada de nodos
- X es padre de Y sii existe arco de X →Y. En este caso Y es hijo de X.

pa(X) = conjunto de padres de X

- X es *antepasado o ascendiente* de Z si existe un camino dirigido de X hasta Z. En ese caso Z es *descendiente* de X.
- X e Y son independientes condicionadas a Z si

$$P(X,Y / Z) = P(X / Z)\, P(Y / Z)$$

Notación: $X \perp Y \mid Z$

Definición

Una **Red Bayesiana** es un modelo gráfico probabilístico dado por

- un grafo DAG donde los nodos representan las variables y
- una distribución (conjunta) de probabilidades para las variables

tal que se verifica la **Condición de Markov:**

Cada nodo es independiente de sus no descendientes (nd) dado el conjunto de sus padres. Esto es:

$$X \perp nd(X) \mid pa(X)$$

Equivalentemente:

$$X \perp \{nd(X) - pa(X)\} \mid pa(X)$$

Distribución conjunta en una RB

Si un DAG con una función de probabilidad conjunta satisface la condición de Markov (esto es, se tiene una BN) entonces vale:

$$P(x_1, x_{2,}..x_n) = \prod_{i=1}^{n} P(x_i/pa(x_i))$$

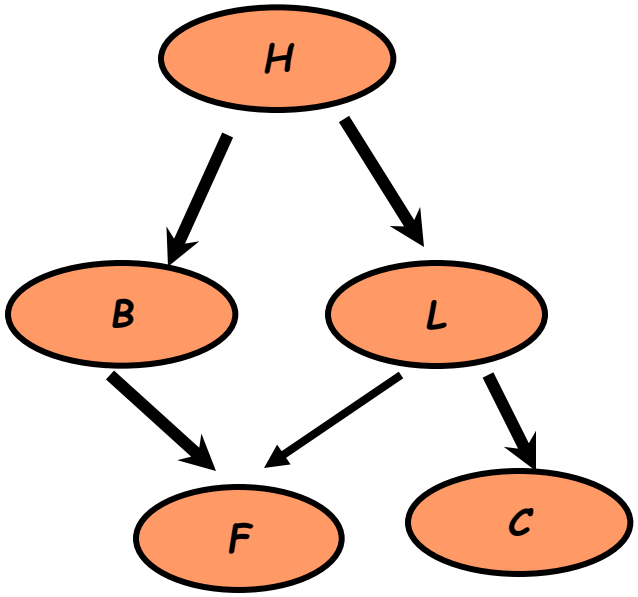
Donde para los nodos raíz, *P(X/pa(X)) = P(X)*

¡Esto permite escribir la distribución conjunta en una red bayesiana!

Obs: para variables discretas o Normales vale también la recíproca: si la conjunta se factoriza según esta expresión, entonces se cumple la condición de Markov.

Ejemplo de BN (Neapolitan, pag 4):

Se propone la siguiente estructura de BN para las variables:



H: historia de fumador?

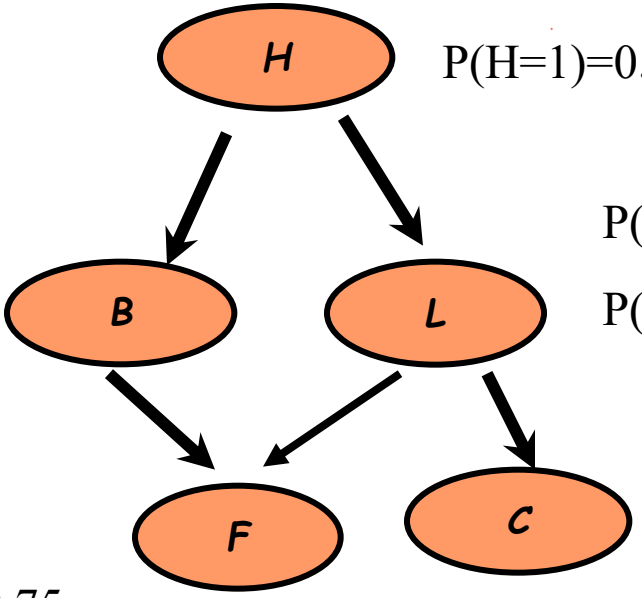
B: tiene bronquitis?

L: tiene cáncer de pulmón?

F: tiene fatiga pulmonar?

C: RX de torax positivo?

Si conocemos los parámetros (la distribución de las variables) y la estructura, tenemos la distribución conjunta



P(H=1)=0.2

P(B=1/H=1)=0.25

P(B=1/H=0)=0.05

P(L=1/H=1)=0.03

P(L=1/H=0)=0.00005

P(C=1/L=1)=0.6

P(C=1/L=0)=0.02

P(F=1/B=1,L=1)=0.75

P(F=1/B=1,L=0)=0.10

P(F=1/B=0,L=1)=0. 5

P(F=1/B=0,L=0)=0.05

Para dar la distribución necesitamos 11 parámetros

$$P(H,B,L,F,C) = P(H)P(B|H)P(L|H)P(C|L)P(F|B,L)$$

Cuestiones a resolver:

I. Determinar la estructura de la red (aprendizaje estructural)

II. Determinar la distribución conjunta de las variables involucradas (aprendizaje paramétrico)

III. Una vez determinada la red, hacer inferencia para alguna variable conocidas las demás (propagación de la evidencia).

Aprendizaje paramétrico

Se tienen las siguientes formas de estimar los parámetros para el caso de variables discretas:

$$\hat{\theta}_{x_i|pa_i} = \frac{N(x_i, pa_i)}{N(pa_i)}$$

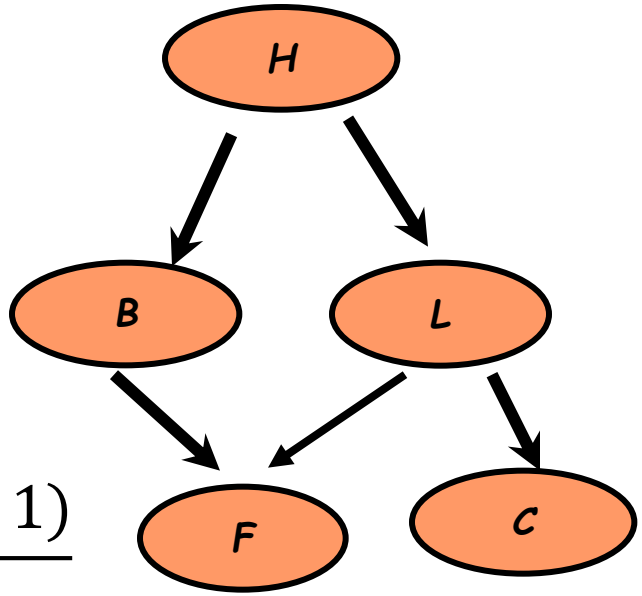
MLE

$$\tilde{\theta}_{x_i|pa_i} = \frac{\alpha(x_i, pa_i) + N(x_i, pa_i)}{\alpha(pa_i) + N(pa_i)}$$

Bayesiano

Ambas son asintóticamente equivalentes

Ejemplo



En este caso

$$P(\widehat{H = 1}) = \frac{\#casos\ (H = 1)}{\#total}$$

$$P(L = \widehat{1/H} = 1) = \frac{\#casos\ (L = 1 \wedge H = 1)}{\#casos\ (H = 1)}$$

$$P(F = \widehat{1/b \wedge H}) = \frac{\#casos\ (F = 1 \wedge b \wedge H)}{\#casos\ (b \wedge h)}$$

$etc\ldots$

Aprendizaje estructural

El primer paso para ajustar una RB es especificar su estructura. Esto puede hacerse basado en rutinas automáticas ó en juicio de expertos.

El aprendizaje automático consiste en inducir, a partir de los datos, una estructura. Esto es, determinar las relaciones de dependencia e independencia entre las variables.

Algunos paquetes de R permiten ajustar la estructura de la red (bnlearn).

Redes Bayesianas para clasificación

Son ampliamente usadas por varias ventajas:

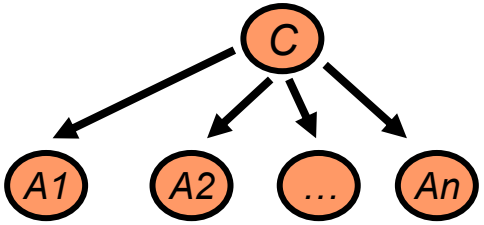
- fáciles de construír y entender
- la inferencia es sencilla
- son robustas a atributos irrelevantes

Clasificación con Redes Bayesianas

Supongamos un conjunto de variables C, A_1, A_2, …, A_n las cuales pueden pensarse como

C: variable de clasificación

A_i: atributos



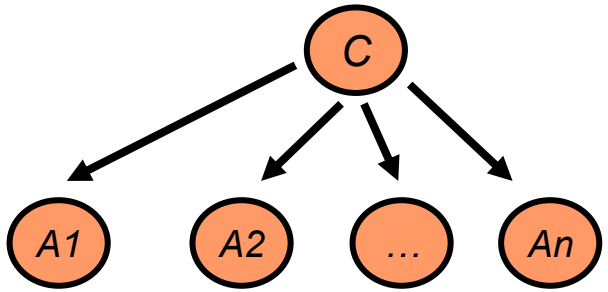
¿Cómo predecir C a partir de los atributos?

- buscar el valor de C que maximice P(A_1, A_2,., A_n/C) → E. M.V
- buscar el valor de C que maximice las probabilidades a posteriori P(C / A_1, A_2, …, A_n), o equivalentemente,

  Maximizar P(A_1, A_2, …, A_n /C)*P(C) → E. Bayesiana

Clasificador Naive Bayes

Asume independencia entre los atributos, condicionados a C, y una estructura dada por:



Entonces

$$P(A_1,..|C_i) = \prod_{k=1}^{n} P(a_k|C_i) = P(a_1|C_i) \times P(a_2|C_i) \times ... \times P(a_n|C_i)$$

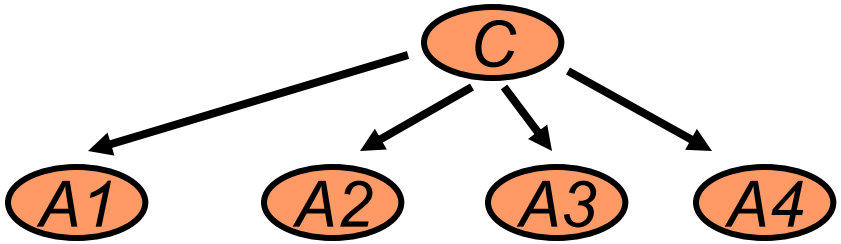
Con lo que

$$P(C_i | A_1 A_2 \ldots A_n) \propto \prod_{k=1}^{n} P(a_k|C_i) P(C_i)$$

La clasificación se hace maximizando esta última expresión.

Ejemplo Naïve Bayes:

Para una variable clasificatoria con *n=4* atributos binarios:



Necesitamos conocer :

P(Ai =0 /C=0) y P(Ai =0 /C=1) para cada i=1..4

y

P(C=0) *(la distribución a priori de C) y P(C=1)*

Con eso, las probabilidades a posteriori son

$P(C=1|A1=0,A2=0,..) \propto P(A1=0/C=1)\ldots P(A4=0/C=1)P(C=1)$

Cómo estimar probabilidades desde los datos?

- Para la distribución *a priori*:

$$P(C=c) = N_c/N$$

- Para la condicional de cada atributo:

$$P(A_i \mid C) = |A_i|/ N_c$$

  - donde $|A_{ic}|$ es el nº de casos A_i observados en la clase C
  - N_c es el nº de casos C observados

Alternativas para estimar los parámetros:

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

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

Donde

Nc=cantidad de datos de la clase condicionante

c= # categorías de Ai (por ejemplo, c=2 para atributo binario)

Ejemplo1: clasificando con Naive Bayes

Nombre	parición?	vuela?	vive en agua?	tiene piernas?	clase
human	si	no	no	si	mamífero
python	no	no	no	no	no-mamífero
salmon	no	no	si	no	no-mamífero
whale	si	no	si	no	mamífero
frog	no	no	a veces	si	no-mamífero
komodo	no	no	no	si	no-mamífero
bat	si	si	no	si	mamífero
pigeon	no	si	no	si	no-mamífero
cat	si	no	no	si	mamífero
leopard shark	si	no	si	no	no-mamífero
turtle	no	no	a veces	si	no-mamífero
penguin	no	no	a veces	si	no-mamífero
porcupine	si	no	no	si	mamífero
eel	no	no	si	no	no-mamífero
salamander	no	no	a veces	si	no-mamífero
gila monster	no	no	no	si	no-mamífero
platypus	no	no	no	si	mamífero
owl	no	si	no	si	no-mamífero
dolphin	si	no	si	no	mamífero
eagle	no	si	no	si	no-mamífero

caso	parición?	vuela?	vive en agua?	tiene piernas?	clase
	si	no	si	no	??????

Ejemplo1: clasificando con Naive Bayes

A: atributos observados

M: mamífero

N: no-mamífero

caso	parición?	vuela?	vive en agua?	tiene piernas?	clase
	si	no	si	no	??????

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

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

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

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

P(A|M)P(M) > P(A|N)P(N)

=> Se lo clasifica como mamífero

Cómo sería clasificar con MV???

Ejemplo1: (cont)

Si se quita de la tabla la fila que corresponde a “bat”, el murciélago, y los atributos observados fueran justamente los de este, se tendría que

$$P(A|M) = \frac{5}{6} \times \frac{0}{6} \times \frac{1}{6} \times \frac{1}{6} = 0$$

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

$$P(A|M)P(M) = 0 \times \frac{6}{19} = 0$$

$$P(A|N)P(N) = 0.004 \times \frac{13}{19}$$

P(A|M)P(M) < P(A|N)P(N)

⇒Se lo clasifica como

NO mamífero!!!!

El problema es que hay pocos (ningún) dato en algún nivel de atributos. Sug: usar Laplace

Propiedades de Naïve Bayes

Pros:

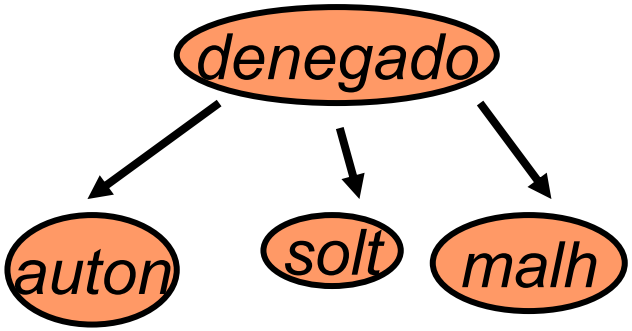
- Reduce significativamente el número de parámetros respecto a otras RB.
- Es robusto a puntos aislados.
- Es robusto a atributos irrelevantes.

Contras:

- El supuesto de independencia puede ser muy fuerte en algunos casos.
- Para datos esparsos, la probabilidad a posteriori es nula (como en el ejemplo anterior), y no permite tomar en cuenta las probabilidades de los otros atributos observados, produciendo a veces mala clasificación.
- El cálculo de la red es NP-hard, por lo que es muy difícil y posiblemente costoso.

Vamos a R

Ejemplo: hipoteca con Naive Bayes



Conjunta:

P(Y=1, M=yes, A=yes, S=yes)=

=P(M=yes|Y=1)P(A=yes|Y=1)P(S=yes|Y=1)P(Y=1) =

= 0.24 * 0.165 * 0.505 * 0.119

(idem cada conjunto de valores posibles de las variables)

A-priori probabilities:

Y

0 1

0.880024 0.119976

Conditional probabilities:

malhist

Y	no	yes
0	0.94819359	0.05180641
1	0.76000000	0.24000000

auton

Y	no	yes
0	0.8936605	0.1063395
1	0.8350000	0.1650000

soltero

Y	no	yes
0	0.603272	0.396728
1	0.495000	0.505000

gasto	rat	rat.desem	malhist	auton	soltero	denegado
0.265	0.922	3.20	no	no	yes	0

```
mod2 <- naiveBayes(denegado ~gasto+rat+rat.desem, data = dataTrain,laplace=0)
```

```
A-priori probabilities:
Y
       0        1
0.880024 0.119976

Conditional probabilities:
   gasto
Y        [,1]       [,2]
  0 0.3223320 0.08225523
  1 0.3918405 0.22898055

   rat
Y           [,1]         [,2]
  0 5.227914e+11 8.259187e+12
  1 4.087431e+12 2.142508e+13

   rat.desem
Y       [,1]     [,2]
  0 2.988534 1.211648
  1 3.000800 1.395574
```