



Clasificadores Probabilísticos en Aprendizaje Automático

Día 2: Modelos Probabilísticos

Daniel Ramos Castro

daniel.ramos@uam.es

Audias – Audio, Data Intelligence and Speech
Universidad Autónoma de Madrid

<http://audias.ii.uam.es>

audias

Audio, Data Intelligence and Speech

UAM

Sumario del Día

- Modelos probabilísticos
 - Generativos
 - Discriminativos
- Modelos probabilísticos gráficos
- Ajuste de funciones de densidad de probabilidad
 - Paramétrico
 - No paramétrico
- Inferencia bayesiana

Modelos Probabilísticos

Clasificador Probabilístico

- Problema fundamental a resolver
 - Saber los valores de las variables latentes a partir de las variables observadas

Clasificador Probabilístico

- Problema fundamental a resolver
 - Saber los valores de las variables latentes a partir de las variables observadas
- No podemos resolver este problema sin incertidumbre
 - Problema esencialmente probabilístico
 - $p(\mathbf{Z}|\mathbf{X})$

Clasificador Probabilístico

- Problema fundamental a resolver
 - Saber los valores de las variables latentes a partir de las variables observadas
- No podemos resolver este problema sin incertidumbre
 - Problema esencialmente probabilístico
 - $p(Z|X)$
- Ejemplos:
 - Conocer el idioma hablado (Z) a partir de la voz (X)
 - Conocer si los vidrios comparados (X) pertenecen o no a la misma fuente (Z)

Clasificador Probabilístico: Modelos

- Modelos que obtienen directamente una decisión
 - Se intenta obtener la decisión directamente, a través de una función directa del dato observado a la clase
 - $z = f(x)$: función discriminante
 - Minúscula: valor concreto de la variable aleatoria en mayúscula

Clasificador Probabilístico: Modelos

- Modelos que obtienen directamente una decisión
 - Se intenta obtener la decisión directamente, a través de una función directa del dato observado a la clase
 - $z = f(x)$: función discriminante
 - Minúscula: valor concreto de la variable aleatoria en mayúscula
- Problema: perdemos la información probabilística
 - No podemos tomar decisiones en un marco riguroso
 - No podemos independizarnos de nuestra aplicación
 - No podemos combinar modelos fácilmente
 - No podemos tomar decisiones con opción “no decisión”
 - ...

Clasificador Probabilístico: Modelos

- Modelos probabilísticos **discriminativos**
 - Intentan obtener la probabilidad de interés directamente
 - $p(\mathbf{Z}|\mathbf{X})$

Clasificador Probabilístico: Modelos

- Modelos probabilísticos **discriminativos**
 - Intentan obtener la probabilidad de interés directamente
 - $p(\mathbf{Z}|\mathbf{X})$
- Ventajas:
 - Obtengo la información que necesito para mi decisión
 - Sin preocuparme de nada más
 - Ej: DNNs

Clasificador Probabilístico: Modelos

- Modelos probabilísticos **discriminativos**
 - Intentan obtener la probabilidad de interés directamente
 - $p(\mathbf{Z}|\mathbf{X})$
- Ventajas:
 - Obtengo la información que necesito para mi decisión
 - Sin preocuparme de nada más
 - Ej: DNNs
- Inconvenientes
 - Se pierde la información del modelo generador $p(\mathbf{X}|\mathbf{Z})$
 - Tiene ventajas conocerlo (ver discusión Sutton-Welling)
 - No permiten generar datos nuevos (*data augmentation*)

Clasificador Probabilístico: Modelos

- Modelos probabilísticos **generativos**
 - Se intenta obtener la representación probabilística completa
 - O bien buscando la probabilidad generadora $p(X|Z)$
 - O bien modelando el problema completo $p(X, Z)$
 - Y obteniendo más adelante $p(X|Z)$, $p(Z|X)$, $p(X)$

Clasificador Probabilístico: Modelos

- Modelos probabilísticos **generativos**
 - Se intenta obtener la representación probabilística completa
 - O bien buscando la probabilidad generadora $p(X|Z)$
 - O bien modelando el problema completo $p(X, Z)$
 - Y obteniendo más adelante $p(X|Z)$, $p(Z|X)$, $p(X)$
- Ventaja: problema probabilístico definido completamente
 - $p(X)$: *marginal likelihood*: ajuste del modelo a los datos
 - Selección de modelos
 - Detección de *outliers*
 - ...

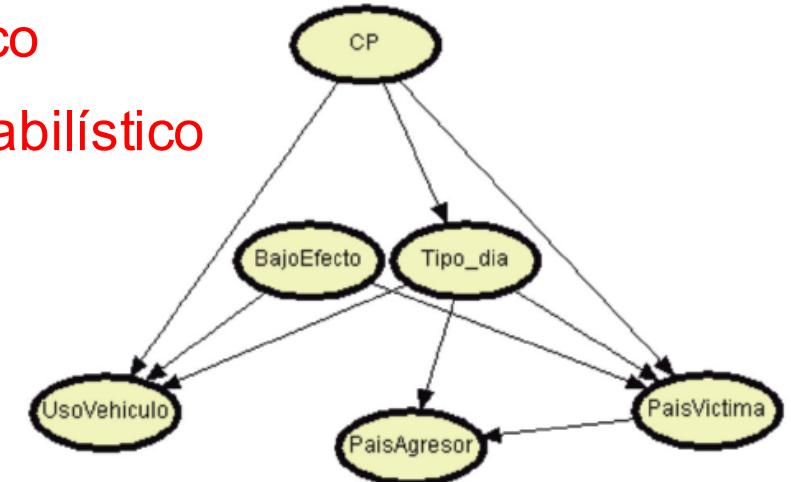
Clasificador Probabilístico: Modelos

- Modelos probabilísticos **generativos**
 - Se intenta obtener la representación probabilística completa
 - O bien buscando la probabilidad generadora $p(X|Z)$
 - O bien modelando el problema completo $p(X, Z)$
 - Y obteniendo más adelante $p(X|Z)$, $p(Z|X)$, $p(X)$
- Ventaja: problema probabilístico definido completamente
 - $p(X)$: *marginal likelihood*: ajuste del modelo a los datos
 - Selección de modelos
 - Detección de *outliers*
 - ...
- Inconvenientes
 - Es el problema más complejo de todos

Modelos Probabilísticos Gráficos: Redes Bayesianas

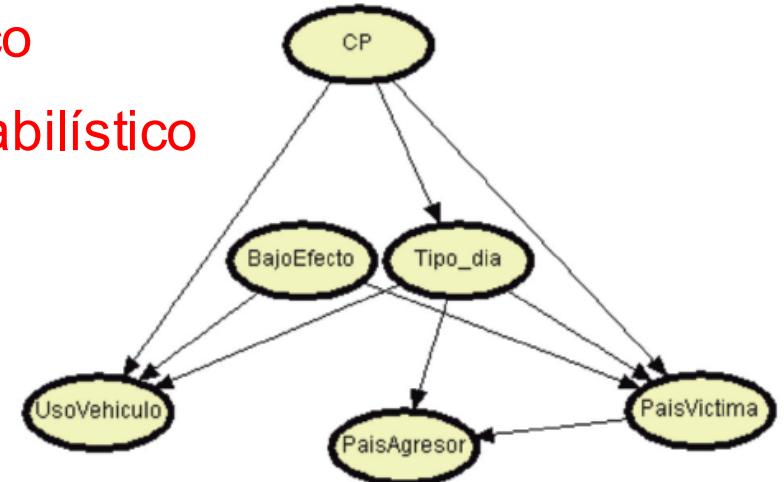
Redes Bayesianas

- Red bayesiana
 - Un tipo de modelo gráfico probabilístico
- Representación visual de un problema probabilístico



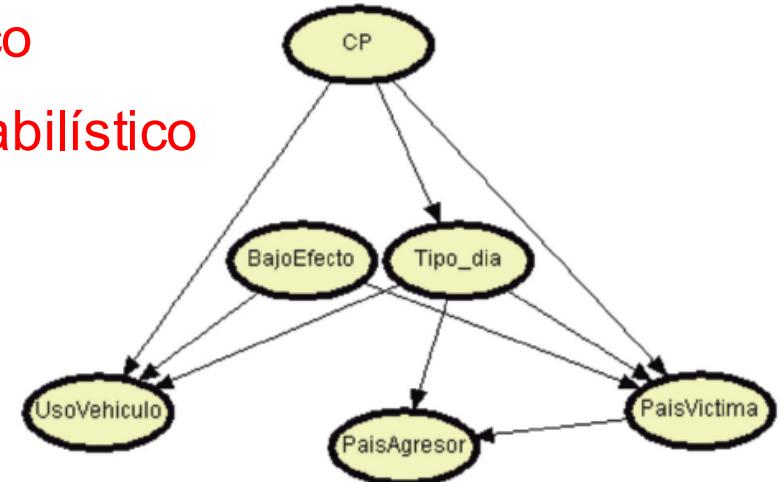
Redes Bayesianas

- Red bayesiana
 - Un tipo de modelo gráfico probabilístico
- Representación visual de un problema probabilístico
- Ventajas de las redes bayesianas
 - Visuales
 - Facilitan la interpretación



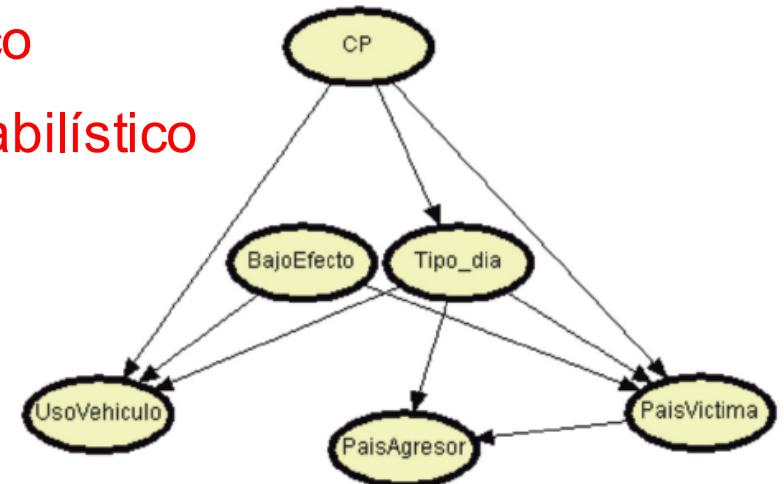
Redes Bayesianas

- Red bayesiana
 - Un tipo de modelo gráfico probabilístico
- Representación visual de un problema probabilístico
- Ventajas de las redes bayesianas
 - Visuales
 - Facilitan la interpretación
 - Holísticas
 - Relacionan todas las variables entre ellas



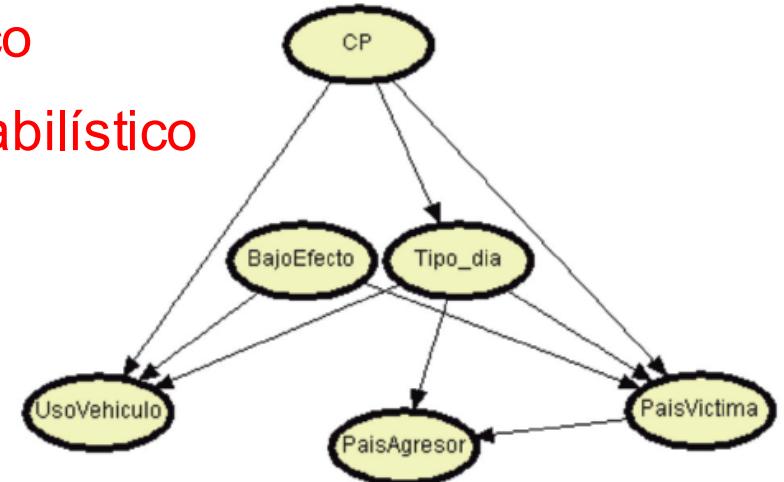
Redes Bayesianas

- Red bayesiana
 - Un tipo de modelo gráfico probabilístico
- Representación visual de un problema probabilístico
- Ventajas de las redes bayesianas
 - Visuales
 - Facilitan la interpretación
 - Holísticas
 - Relacionan todas las variables entre ellas
 - Entrenables
 - Pueden aprender de datos



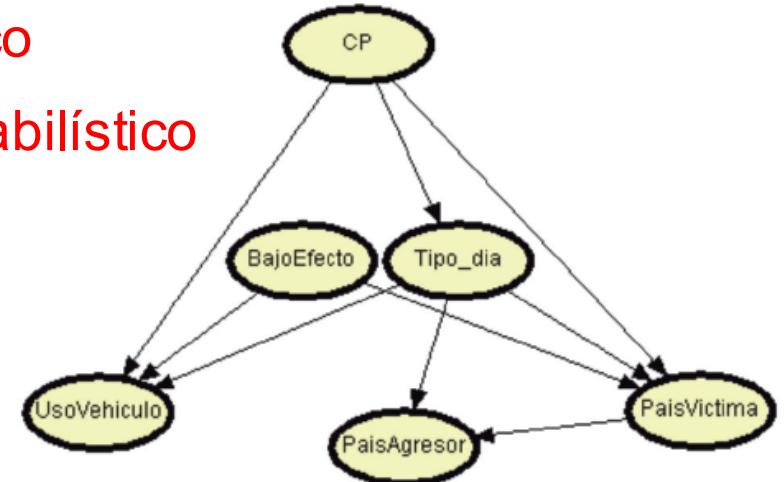
Redes Bayesianas

- Red bayesiana
 - Un tipo de modelo gráfico probabilístico
- Representación visual de un problema probabilístico
- Ventajas de las redes bayesianas
 - Visuales
 - Facilitan la interpretación
 - Holísticas
 - Relacionan todas las variables entre ellas
 - Entrenables
 - Pueden aprender de datos
 - Funcionales
 - Predicen cual(es)quier(a) variable(s) a partir del resto



Redes Bayesianas

- Red bayesiana
 - Un tipo de modelo gráfico probabilístico
- Representación visual de un problema probabilístico
- Ventajas de las redes bayesianas
 - Visuales
 - Facilitan la interpretación
 - Holísticas
 - Relacionan todas las variables entre ellas
 - Entrenables
 - Pueden aprender de datos
 - Funcionales
 - Predicen cual(es)quier(a) variable(s) a partir del resto
 - Flexibles
 - Permiten múltiples configuraciones, incluir conocimiento experto...

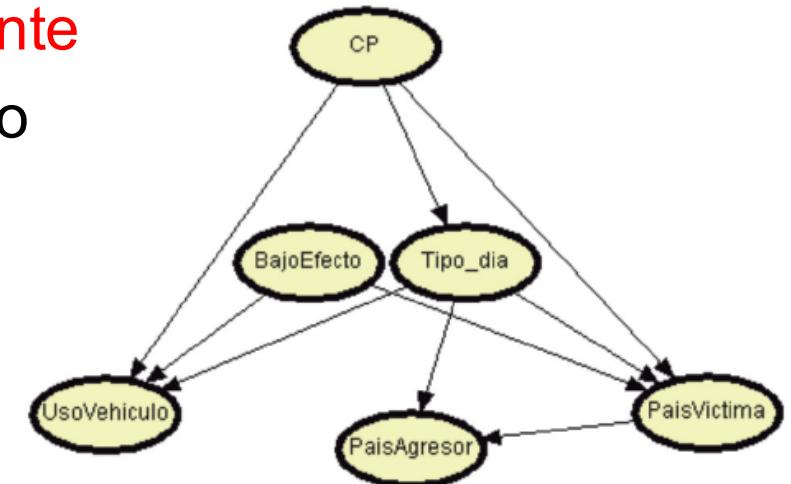


Redes Bayesianas

■ Inconvenientes

□ Complejas de modelar algorítmicamente

- Sobre todo si las bases de datos no son muy grandes
 - Por otra parte, es lo que ocurre en otros muchos modelos...
 - Existe literatura para solucionarlo



Redes Bayesianas

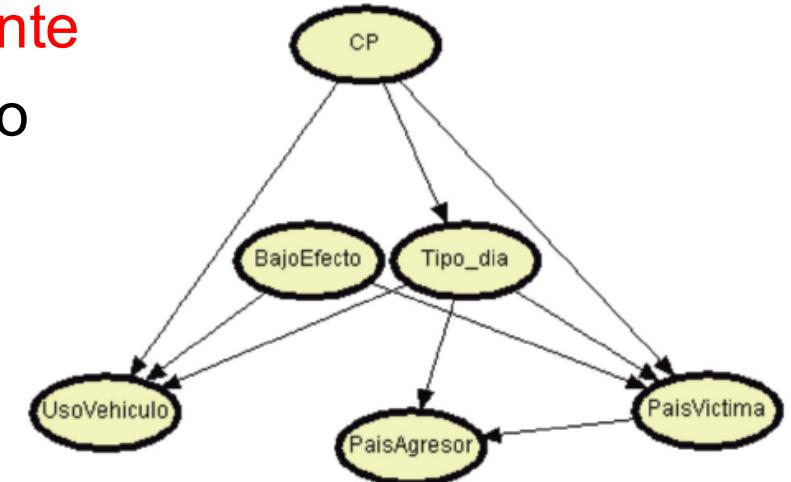
■ Inconvenientes

□ Complejas de modelar algorítmicamente

- Sobre todo si las bases de datos no son muy grandes
 - Por otra parte, es lo que ocurre en otros muchos modelos...
 - Existe literatura para solucionarlo

□ No modelan causalidad

- Solo modelan dependencia estadística
 - La causalidad requiere de suposiciones



Redes Bayesianas

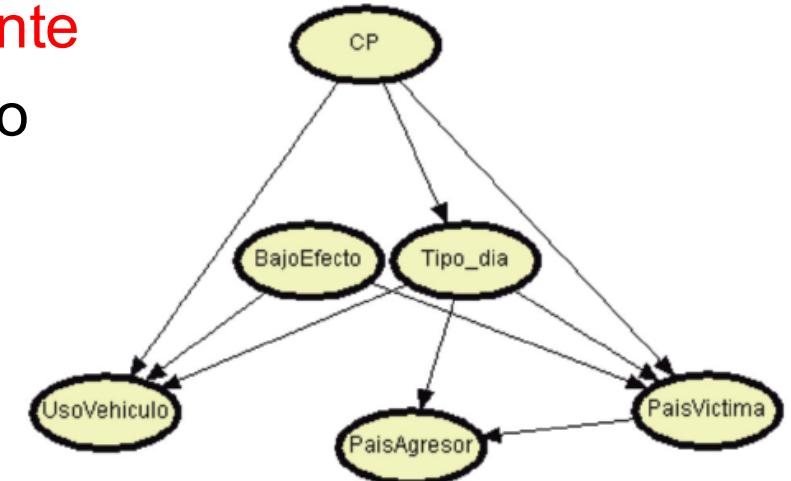
■ Inconvenientes

□ Complejas de modelar algorítmicamente

- Sobre todo si las bases de datos no son muy grandes
 - Por otra parte, es lo que ocurre en otros muchos modelos...
 - Existe literatura para solucionarlo

□ No modelan causalidad

- Solo modelan dependencia estadística
 - La causalidad requiere de suposiciones



D. Koehler, N. Friedman. "Probabilistic Graphical Models". MIT Press, 2009.

J. Pearl. "The Book of Why: The New Science of Cause and Effect". Basic Books, 2018.

Notación

- V.A. en mayúscula (X), valor concreto en minúscula (x)
- Probabilidad conjunta (de que sucedan a la vez X e Y)
 - $p(X, Y)$
- Regla de la suma (marginalización discreta, continua)
 - $p(X) = \sum_Y p(X, Y); \quad p(X) = \int p(X, y) dy$
- Regla del producto (probabilidad condicional)
 - $p(X, Y) = p(X|Y)p(Y) = p(Y|X)p(X)$
- Regla de Bayes
 - $p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)}$

Redes Bayesianas: Modelo Estadístico

- Se tiene un problema con N variables continuas: $\{X_1, \dots, X_N\}$
- Cualquier problema de predicción se puede plantear de forma óptima si se conoce la función densidad de probabilidad **conjunta**:

$$p(X_1 = x_1, \dots, X_N = x_N) \equiv p(x_1, \dots, x_N)$$

Redes Bayesianas: Modelo Estadístico

- Se tiene un problema con N variables continuas: $\{X_1, \dots, X_N\}$
- Cualquier problema de predicción se puede plantear de forma óptima si se conoce la función densidad de probabilidad **conjunta**:

$$p(X_1 = x_1, \dots, X_N = x_N) \equiv p(x_1, \dots, x_N)$$

- Porque con ella se tienen el resto:

- Marginales $p(X_i) = \int \dots \int_{\setminus i} p(X_1, \dots, X_N) dX_{\setminus i}$; con $\setminus i \equiv "Not\ i"$

Redes Bayesianas: Modelo Estadístico

- Se tiene un problema con N variables continuas: $\{X_1, \dots, X_N\}$
- Cualquier problema de predicción se puede plantear de forma óptima si se conoce la función densidad de probabilidad **conjunta**:

$$p(X_1 = x_1, \dots, X_N = x_N) \equiv p(x_1, \dots, x_N)$$

- Porque con ella se tienen el resto:

- Marginales $p(X_i) = \int \dots \int_{\setminus i} p(X_1, \dots, X_N) dX_{\setminus i}$; con $\setminus i \equiv "Not\ i"$

- Condicionadas

$$p(X_i | X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_N) = \frac{p(X_1, \dots, X_N)}{p(X_1, \dots, X_{i-1}, X_{i+1}, X_N)}$$

$$p(X_i | X_j) = \frac{p(X_j | X_i)p(X_i)}{p(X_j)}$$

Redes Bayesianas: Modelo Estadístico

- Problema
 - La distribución conjunta es muy compleja (problema “np-hard”)
 - Requiere de $O(2^N)$ parámetros
 - Computacionalmente desafiante incluso con los modelos más sencillos
 - Necesidad muy grande de datos

Redes Bayesianas: Modelo Estadístico

- Problema
 - La distribución conjunta es muy compleja (problema “np-hard”)
 - Requiere de $O(2^N)$ parámetros
 - Computacionalmente desafiante incluso con los modelos más sencillos
 - Necesidad muy grande de datos
- Posible solución
 - Limitar el número de dependencias

Redes Bayesianas: Modelo Estadístico

- Problema
 - La distribución conjunta es muy compleja (problema “np-hard”)
 - Requiere de $O(2^N)$ parámetros
 - Computacionalmente desafiante incluso con los modelos más sencillos
 - Necesidad muy grande de datos
- Posible solución
 - Limitar el número de dependencias
- Término clave
 - Independencia condicional
 - Dos conjuntos de variables son independientes una vez observado un tercero

Independencia Condicional

- Ejemplo:
 - Tres variables
 - X_1 : “ADN de una persona”
 - X_2 : “AND de su padre”
 - X_3 : “AND de su abuelo”

Independencia Condicional

- Ejemplo:
 - Tres variables
 - X_1 : “ADN de una persona”
 - X_2 : “ADN de su padre”
 - X_3 : “ADN de su abuelo”
 - Si X_2 es conocida (observada), X_1 no depende de X_3
 - Si conozco el ADN del padre, el ADN del abuelo no da información sobre el ADN de la persona

Independencia Condicional

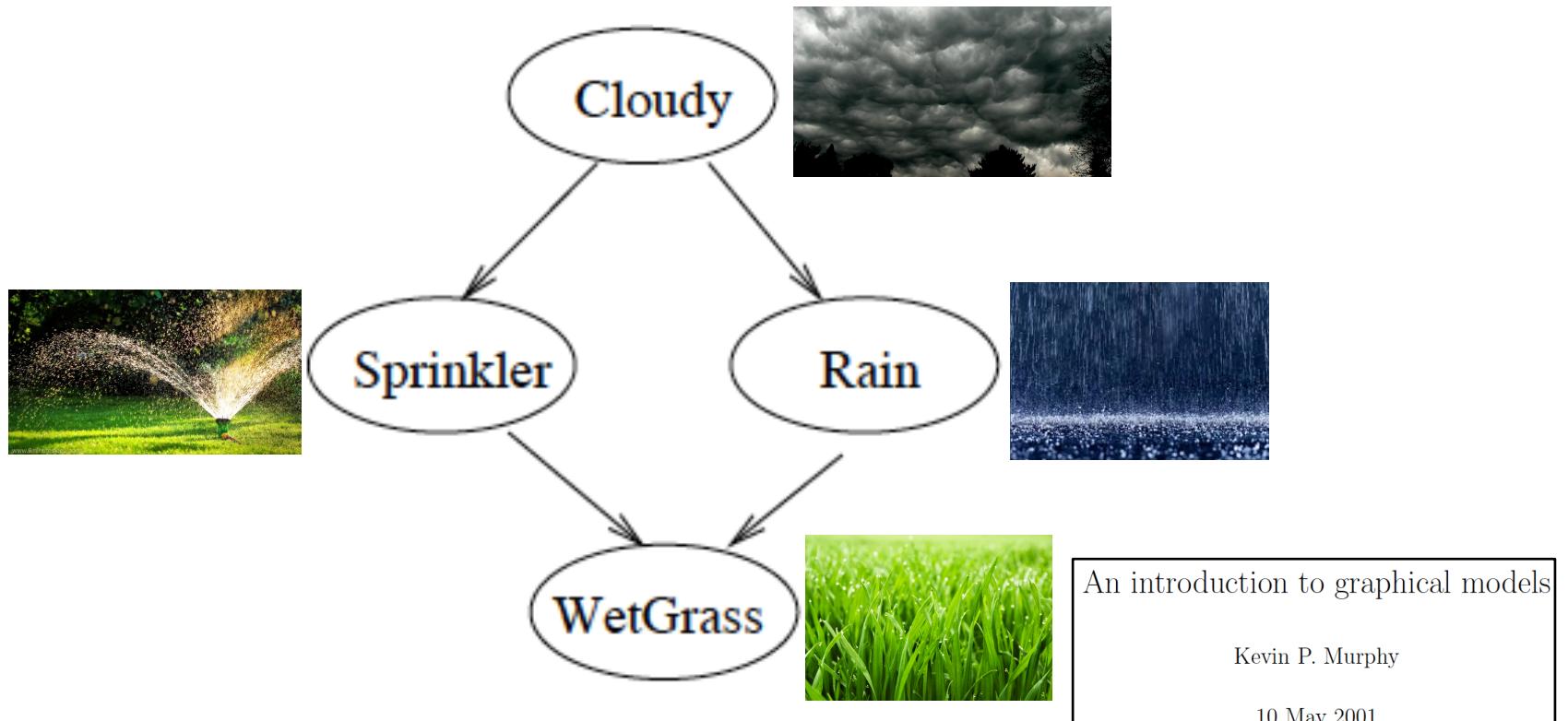
- Ejemplo:
 - Tres variables
 - X_1 : “ADN de una persona”
 - X_2 : “ADN de su padre”
 - X_3 : “ADN de su abuelo”
 - Si X_2 es conocida (observada), X_1 no depende de X_3
 - Si conozco el ADN del padre, el ADN del abuelo no da información sobre el ADN de la persona
 - Si X_2 es desconocida (latente), X_1 sí depende de X_3
 - Si desconozco el ADN del padre, el ADN del abuelo sí da información sobre el ADN de la persona

Independencia Condicional

- Ejemplo:
 - Tres variables
 - X_1 : “ADN de una persona”
 - X_2 : “ADN de su padre”
 - X_3 : “ADN de su abuelo”
 - Si X_2 es conocida (observada), X_1 no depende de X_3
 - Si conozco el ADN del padre, el ADN del abuelo no da información sobre el ADN de la persona
 - Si X_2 es desconocida (latente), X_1 sí depende de X_3
 - Si desconozco el ADN del padre, el ADN del abuelo sí da información sobre el ADN de la persona
 - X_1 es **condicionalmente independiente** de X_3
 - La condición es si observo o no X_2

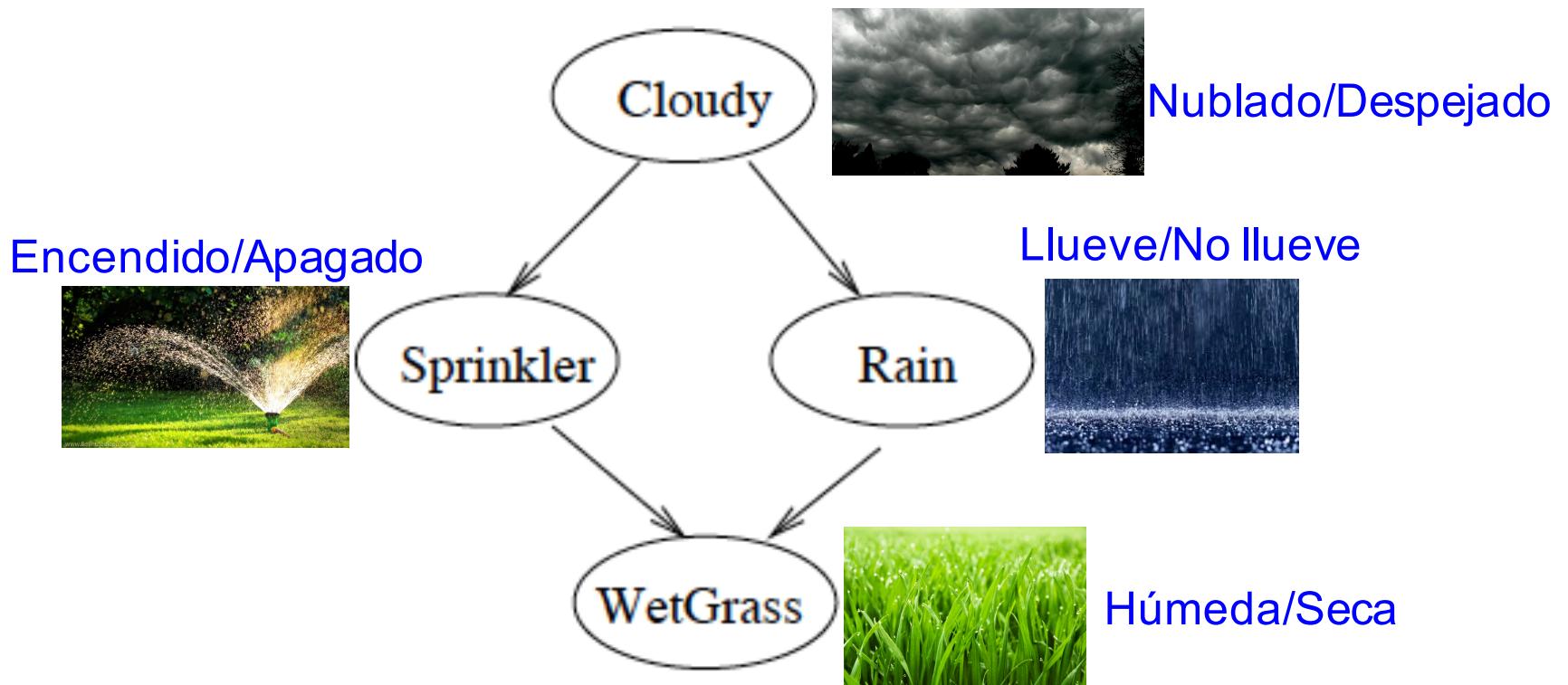
Redes Bayesianas: Ejemplo Sencillo (y Clásico)

- Las direcciones de las flechas se pueden ver como relaciones “causales”
 - Pero en realidad indican “dependencia”
 - No necesariamente causalidad

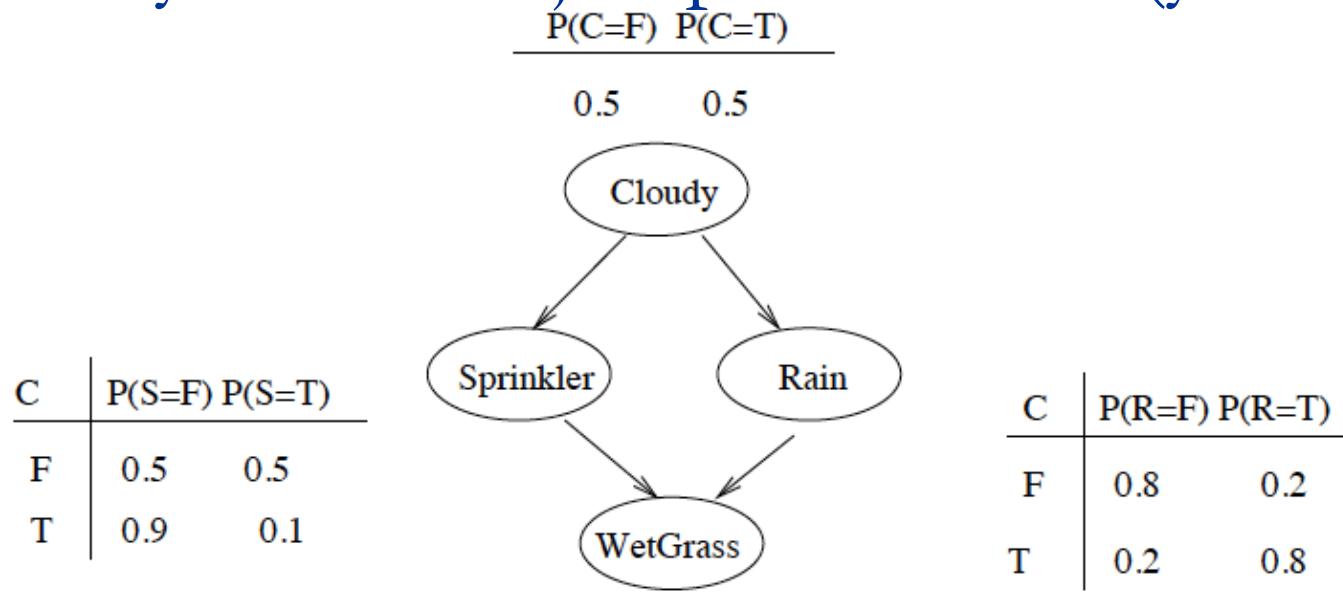


Redes Bayesianas: Ejemplo Sencillo (y Clásico)

- Cada variable tiene varios valores posibles
 - Conocer el valor de una variable...
 - Influye en el resto de variables desconocidas
 - Dependerá de las relaciones (flechas) en el gráfico...

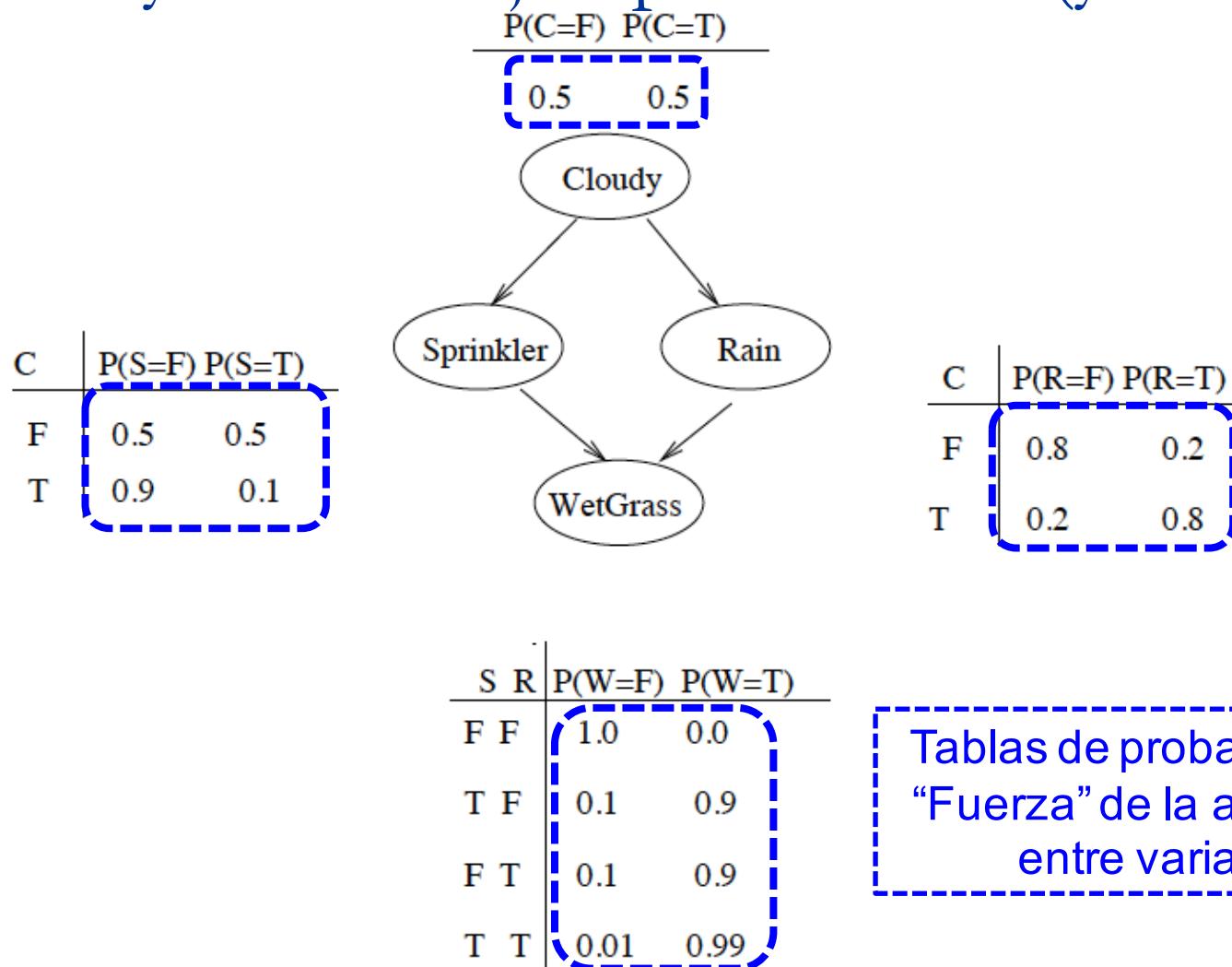


Redes Bayesianas: Ejemplo Sencillo (y Clásico)

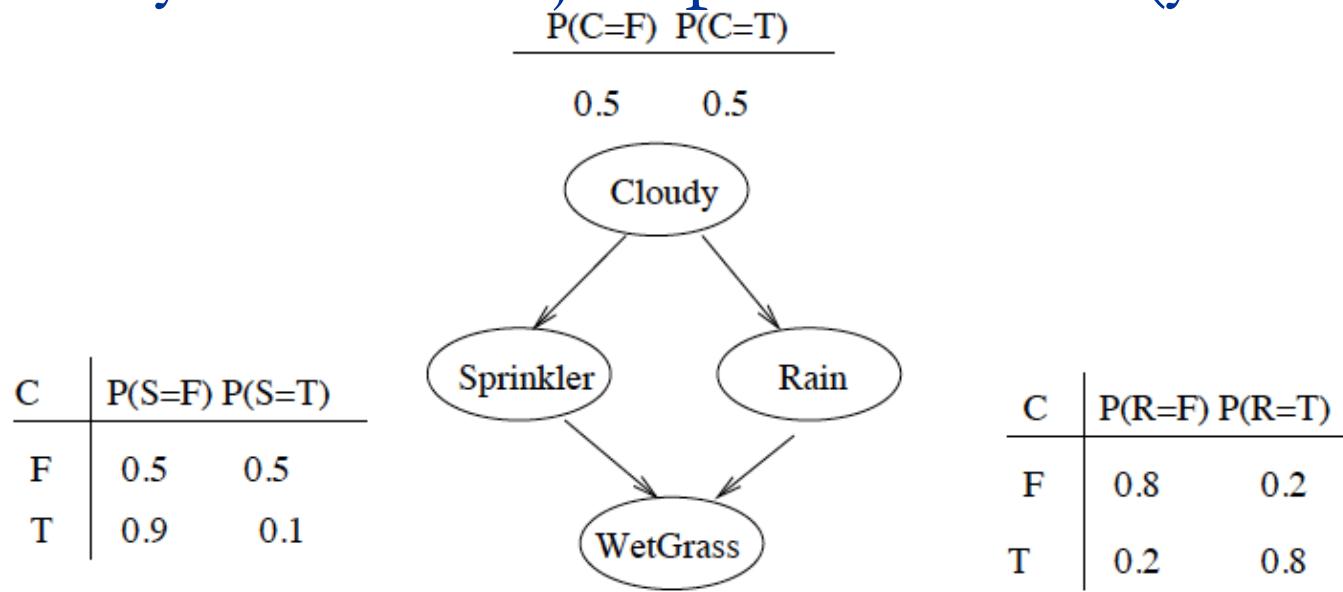


S	R	$P(W=F)$	$P(W=T)$
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

Redes Bayesanas: Ejemplo Sencillo (y Clásico)

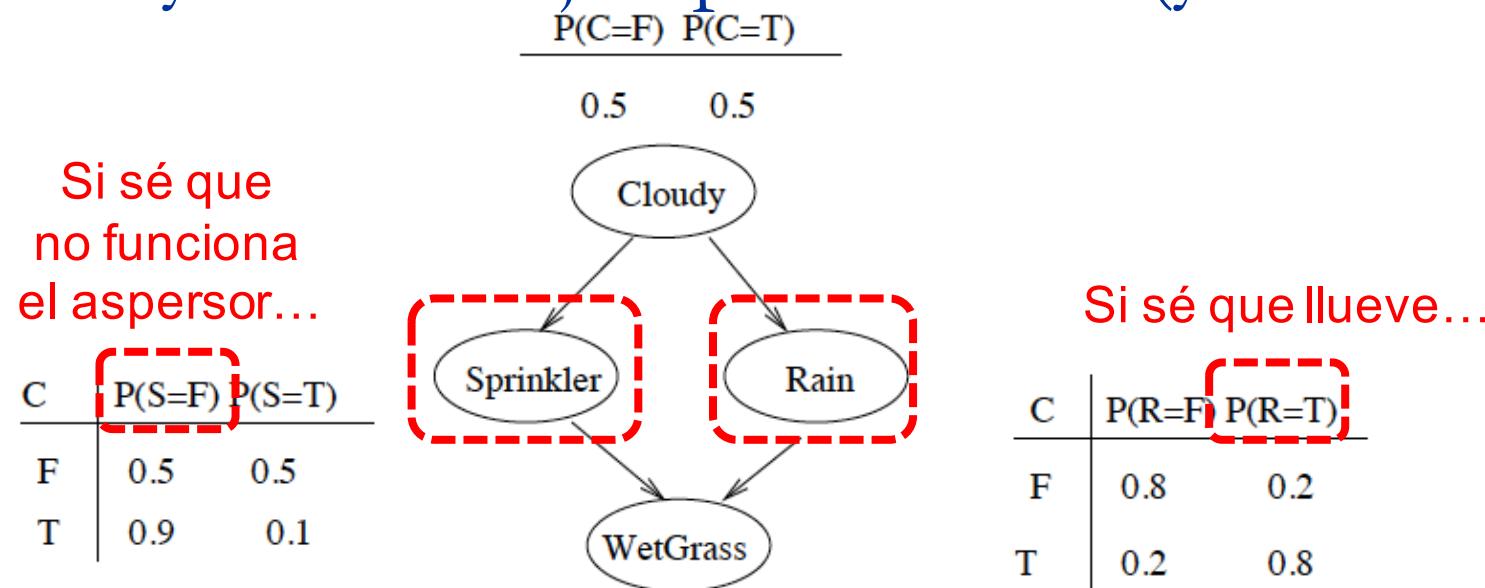


Redes Bayesianas: Ejemplo Sencillo (y Clásico)



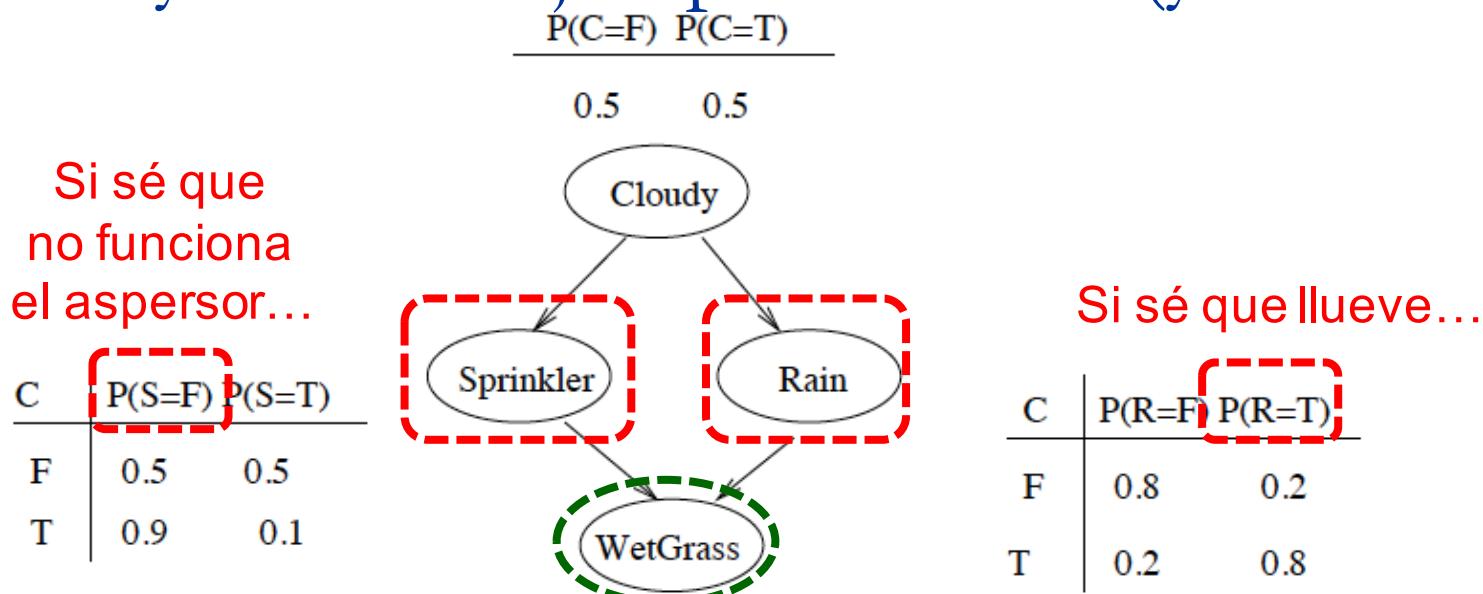
S	R	$P(W=F)$	$P(W=T)$
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

Redes Bayesianas: Ejemplo Sencillo (y Clásico)



S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

Redes Bayesianas: Ejemplo Sencillo (y Clásico)

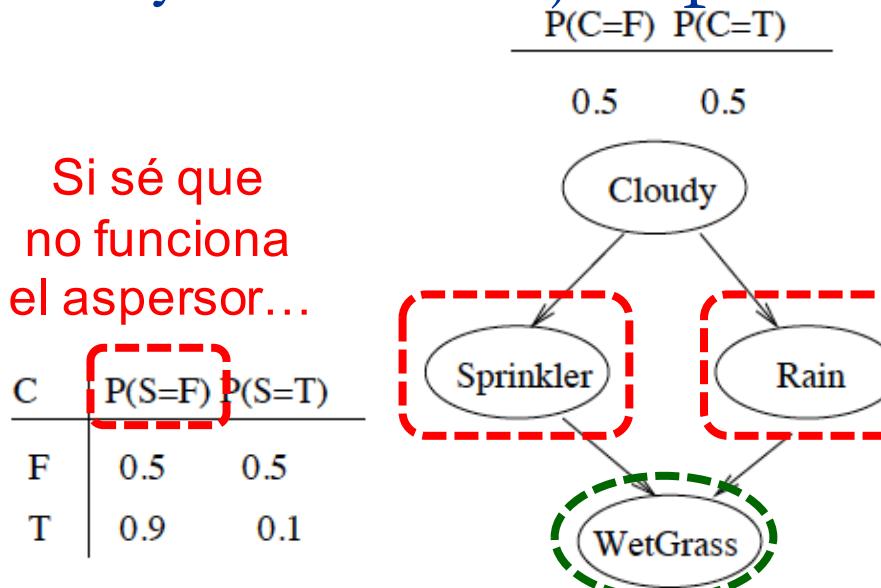


Entonces...

Hay un 90% de probabilidad de que la hierba esté húmeda

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
	T	0.01	0.99

Redes Bayesinas: Ejemplo Sencillo (y Clásico)



Si sé que llueve...

C	$P(R=F)$	$P(R=T)$
F	0.8	0.2
T	0.2	0.8

Independencia Condicional
(concepto clave):

S	R	$P(W=F)$	$P(W=T)$
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
	T	0.01	0.99

Entonces...

Hay un 90% de probabilidad de que la hierba esté húmeda

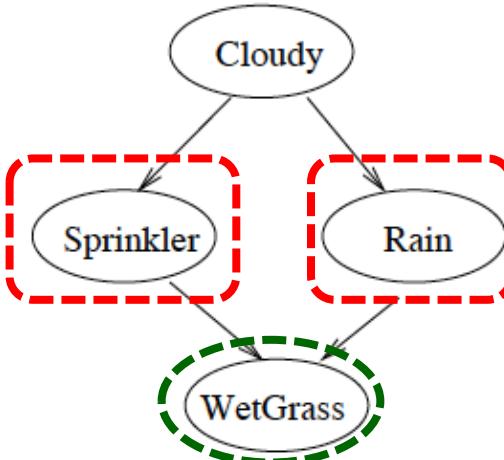
Redes Bayesinas: Ejemplo Sencillo (y Clásico)

$$\begin{array}{c} P(C=F) \quad P(C=T) \\ \hline \end{array}$$

0.5 0.5

Si sé que
no funciona
el aspersor...

C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1



Si sé que llueve...

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

Independencia Condicional
(concepto clave):

Conozco si llueve o no

Conozco el estado del aspersor

Entonces...

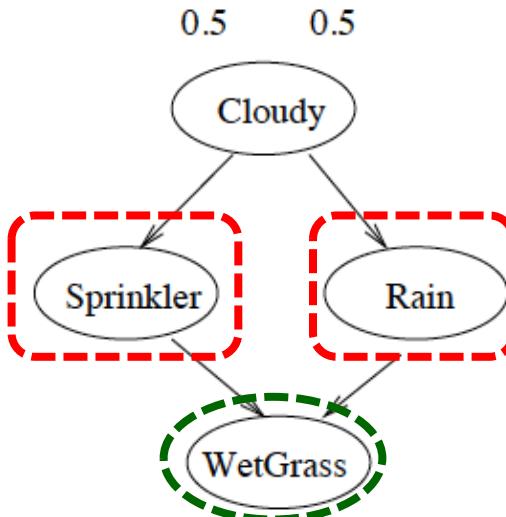
Hay un 90% de probabilidad
de que la hierba esté húmeda

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
	T	0.01	0.99

Redes Bayesianas: Ejemplo Sencillo (y Clásico)

Si sé que no funciona el aspersor...

C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1



Si sé que llueve...

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

Independencia Condicional
(concepto clave):

Conozco si llueve o no

Conozco el estado del aspersor

Entonces el estado de la hierba es independiente de si hay nubes o no...

Entonces...

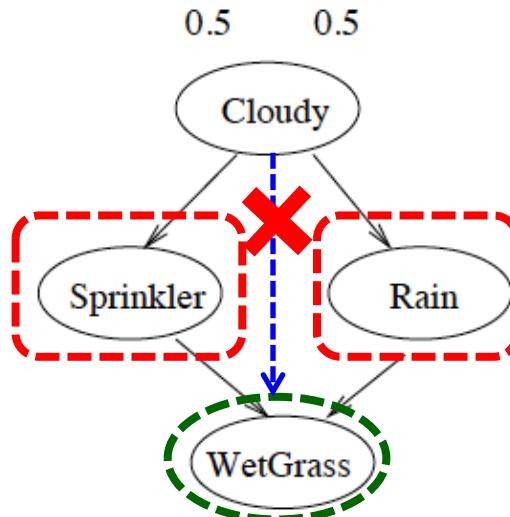
Hay un 90% de probabilidad de que la hierba esté húmeda

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

Redes Bayesianas: Ejemplo Sencillo (y Clásico)

Si sé que no funciona el aspersor...

C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1



Si sé que llueve...

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

Independencia Condicional
(concepto clave):

Conozco si llueve o no

Conozco el estado del aspersor

Entonces el estado de la hierba es independiente de si hay nubes o no...

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

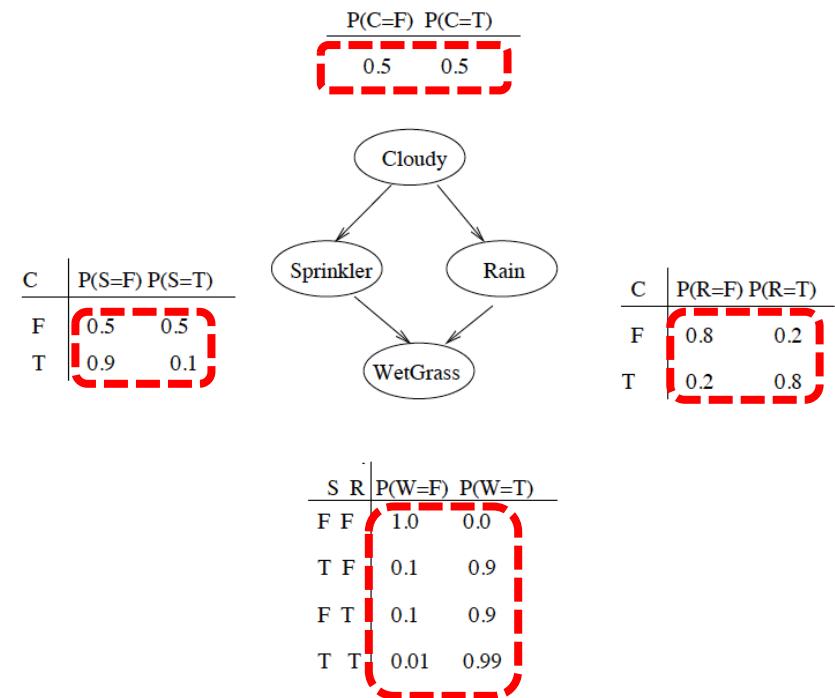
Entonces...

Hay un 90% de probabilidad de que la hierba esté húmeda

Redes Bayesianas: Ejemplo Sencillo (y Clásico)

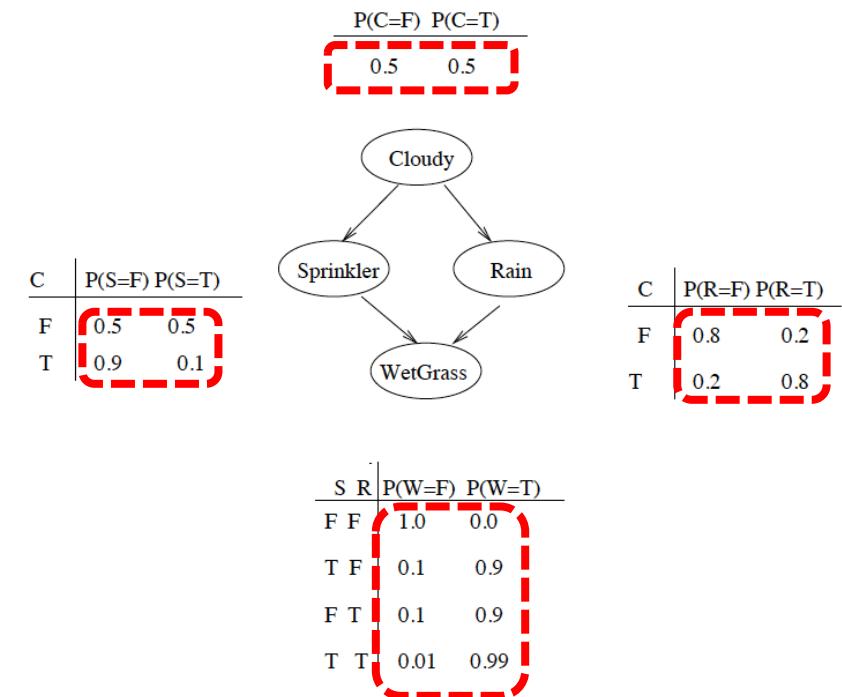
■ En el ejemplo

- Hay que aprender 18 números



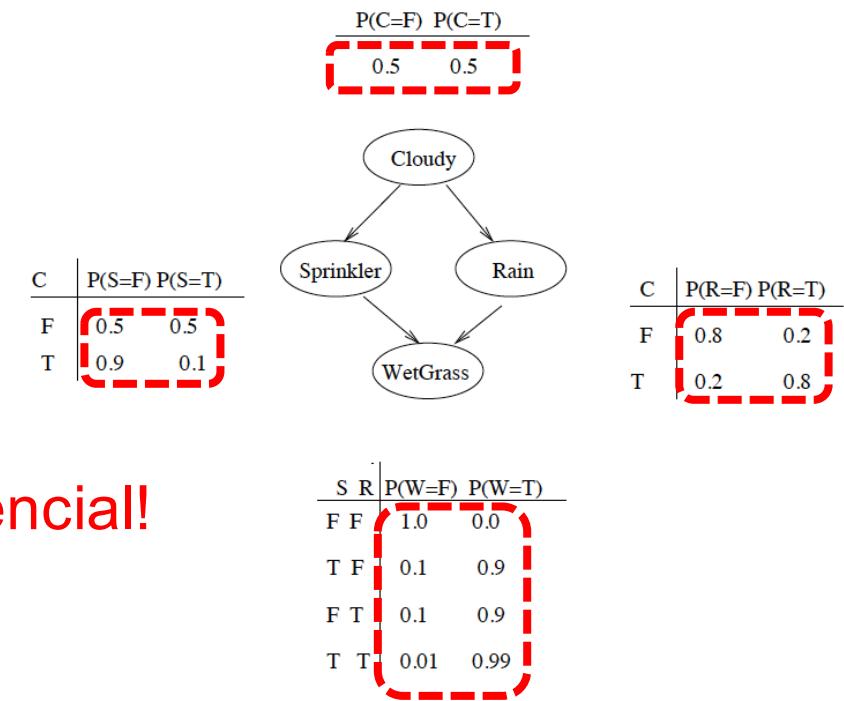
Redes Bayesanas: Ejemplo Sencillo (y Clásico)

- En el ejemplo
 - Hay que aprender 18 números
- Si estuvieran todas las relaciones (“flechas”) posibles...
 - Habría que calcular (aprender) 26 números



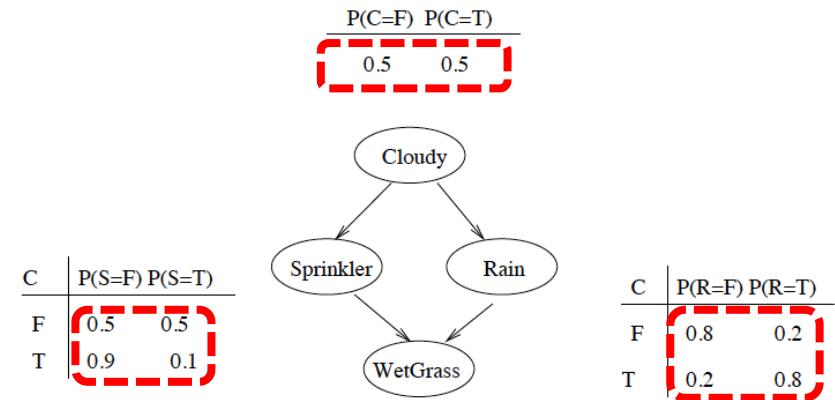
Redes Bayesinas: Ejemplo Sencillo (y Clásico)

- En el ejemplo
 - Hay que aprender 18 números
- Si estuvieran todas las relaciones (“flechas”) posibles...
 - Habría que calcular (aprender) 26 números
- A más variables, ¡crecimiento exponencial!
 - Más tiempo de aprendizaje
 - Necesarios más datos



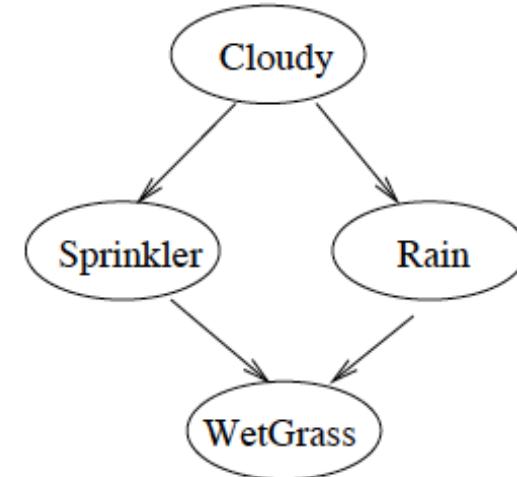
Redes Bayesinas: Ejemplo Sencillo (y Clásico)

- En el ejemplo
 - Hay que aprender 18 números
- Si estuvieran todas las relaciones (“flechas”) posibles...
 - Habría que calcular (aprender) 26 números
- A más variables, ¡crecimiento exponencial!
 - Más tiempo de aprendizaje
 - Necesarios más datos
- Independencia condicional
 - Elimina relaciones (“flechas”): hay que aprender menos números
 - Aprendizaje más rápido
 - Necesarios menos datos



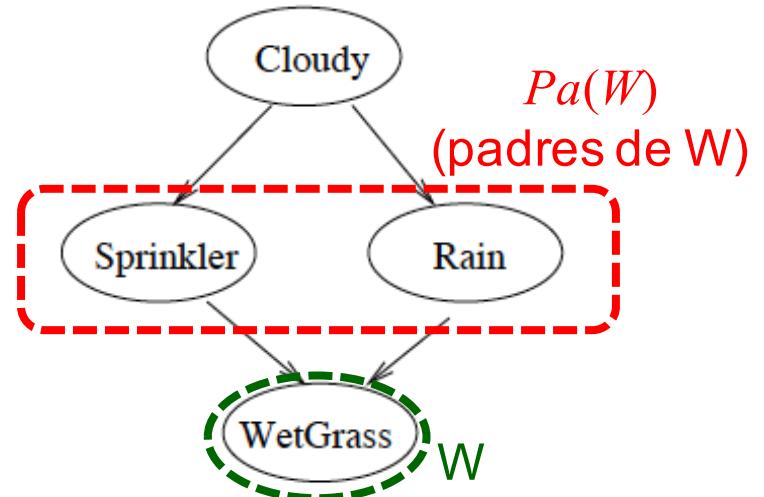
Redes Bayesianas: Representación

- Las direcciones de las flechas se pueden ver como relaciones “causales”
 - Aunque en realidad no lo sean



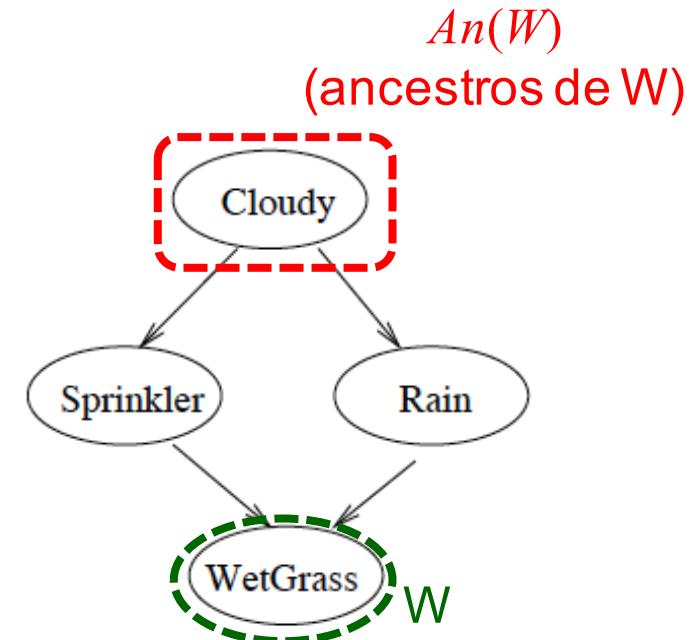
Redes Bayesanas: Representación

- Las direcciones de las flechas se pueden ver como relaciones “causales”
 - Aunque en realidad no lo sean
- Un nodo tiene “padres”
 - Nodos con flechas hacia él



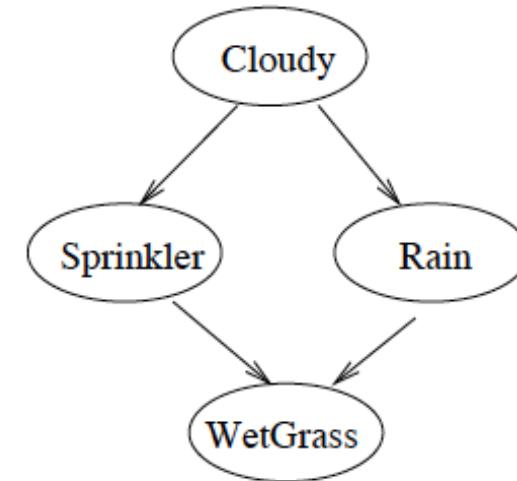
Redes Bayesinas: Representación

- Las direcciones de las flechas se pueden ver como relaciones “causales”
 - Aunque en realidad no lo sean
- Un nodo tiene “padres”
 - Nodos con flechas hacia ellos
- Y ancestros
 - Padres de los padres
 - Ancestros de los padres



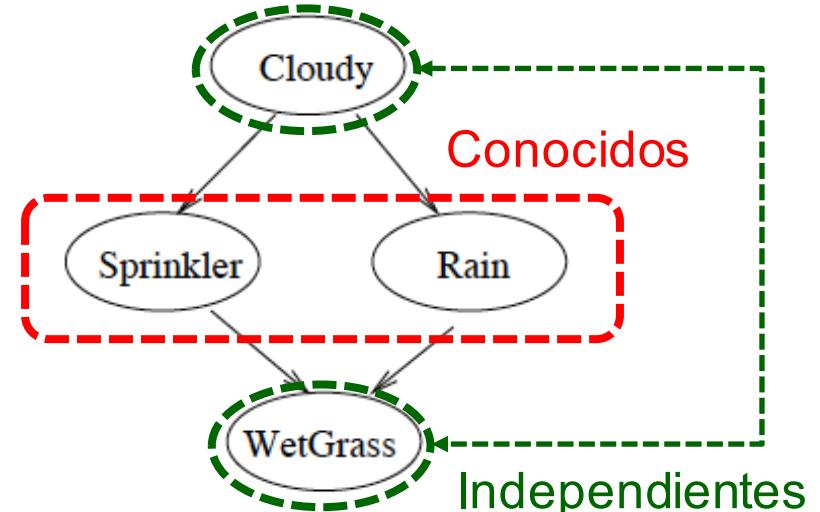
Redes Bayesianas: Representación

- Las direcciones de las flechas se pueden ver como relaciones “causales”
 - Aunque en realidad no lo sean
- Un nodo tiene “padres”
 - Nodos con flechas hacia ellos
- Y ancestros
 - Padres de los padres
 - Ancestros de los padres
- Todo depende del grafo en cuestión
 - Y de la orientación de las flechas



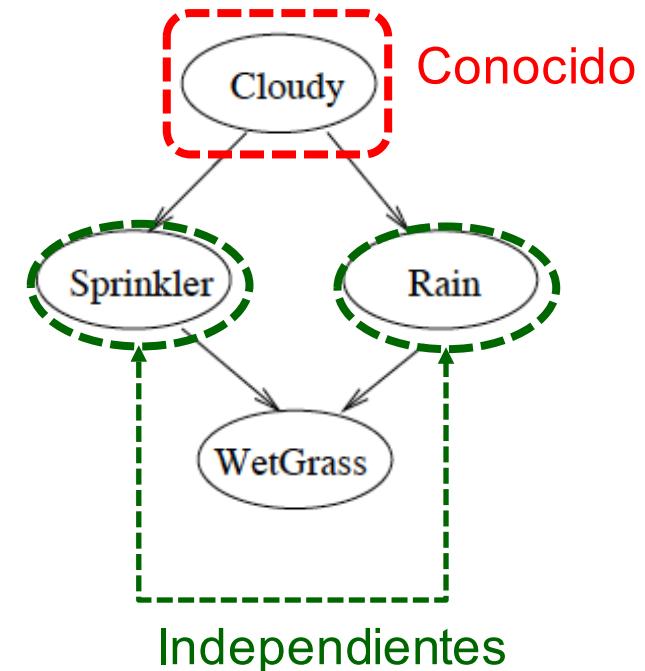
Redes Bayesinas: Representación

- Las direcciones de las flechas se pueden ver como relaciones “causales”
 - Aunque en realidad no lo sean
- Un nodo tiene “padres”
 - Nodos con flechas hacia ellos
- Y ancestros
 - Padres de los padres
 - Ancestros de los padres
- Todo depende del grafo en cuestión
 - Y de la orientación de las flechas
- Propiedad de Markov (fundamental)
 - Un nodo es independiente de sus ancestros dados sus padres



Redes Bayesinas: Representación

- Las direcciones de las flechas se pueden ver como relaciones “causales”
 - Aunque en realidad no lo sean
- Un nodo tiene “padres”
 - Nodos con flechas hacia ellos
- Y ancestros
 - Padres de los padres
 - Ancestros de los padres
- Todo depende del grafo en cuestión
 - Y de la orientación de las flechas
- Propiedad de Markov (fundamental)
 - Un nodo es independiente de sus ancestros dados sus padres
 - ...y de otros nodos no conectados



Redes Bayesianas: Independencia Condicional

- Supone una gran simplificación de la probabilidad conjunta
- Sin considerar independencia condicional
 - Sin considerar el modelo gráfico

$$P(C, S, R, W) = P(C) \times P(S|C) \times P(R|C, S) \times P(W|R, C, S)$$

Redes Bayesianas: Independencia Condicional

- Supone una gran simplificación de la probabilidad conjunta
- Sin considerar independencia condicional
 - Sin considerar el modelo gráfico

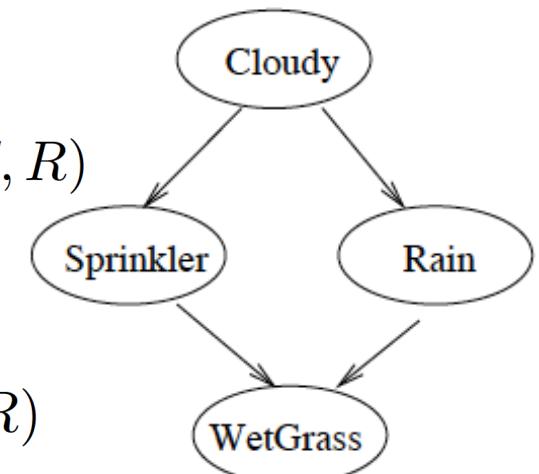
$$P(C, S, R, W) = P(C) \times P(S|C) \times P(R|C, S) \times P(W|R, S, C)$$

- Considerando independencia condicional
 - Considerando “las flechas del grafo”
 - Reducción considerable de la complejidad

$$P(C, S, R, W) = P(C) \times P(S|C) \times P(R|C, \cancel{S}) \times P(W|\cancel{C}, \cancel{S}, R)$$



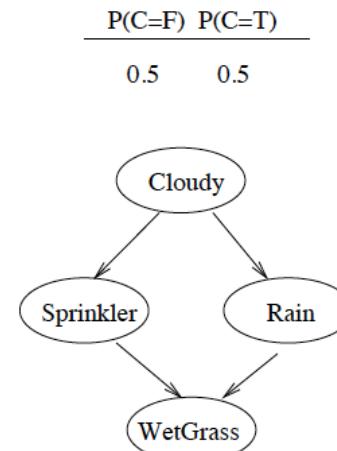
$$P(C, S, R, W) = P(C) \times P(S|C) \times P(R|C) \times P(W|S, R)$$



Redes Bayesianas: Inferencia

- A partir de la red, con todos sus parámetros...
- Calcular probabilidad de variables latentes
 - Puede que no todas...
- A partir de variables observadas
- Ejemplos:

C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

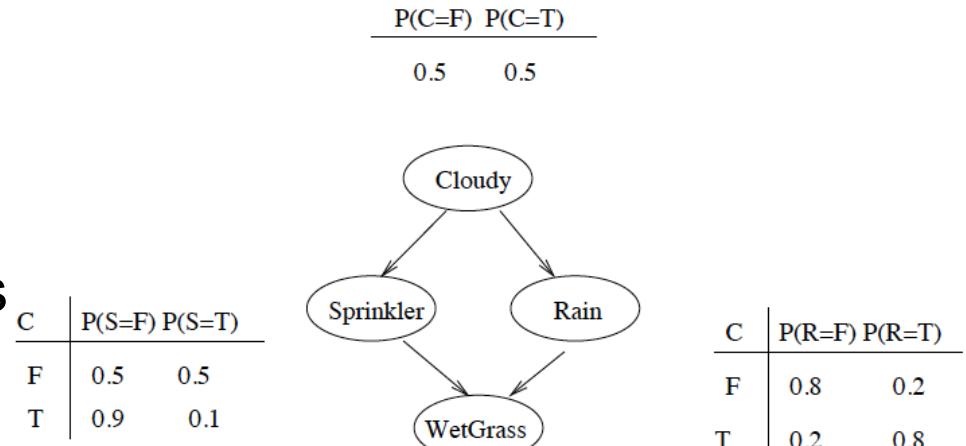


C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

Redes Bayesianas: Inferencia

- A partir de la red, con todos sus parámetros...
- Calcular probabilidad de variables latentes
 - Puede que no todas...
- A partir de variables observadas
- Ejemplos:

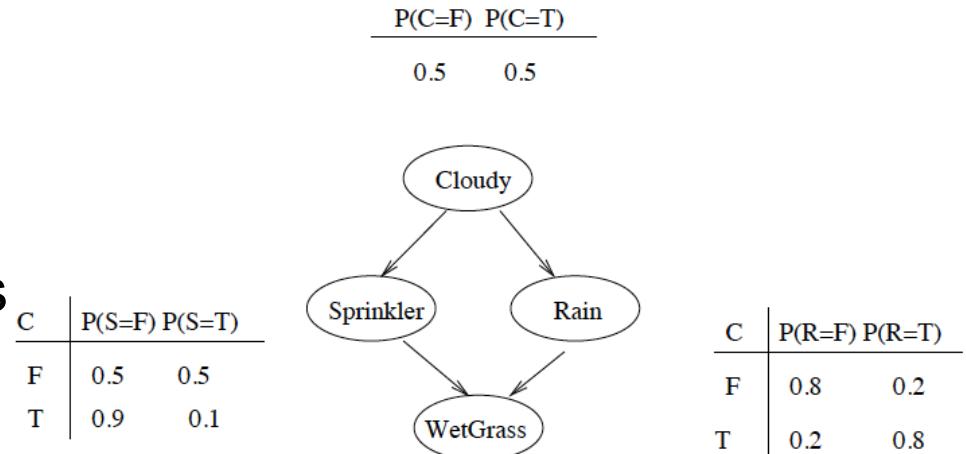


S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

$$P(S=1|W=1) = \frac{P(S=1, W=1)}{P(W=1)} = \frac{\sum_{c,r} P(C=c, S=1, R=r, W=1)}{P(W=1)} = \frac{0.2781}{0.6471} = 0.430$$

Redes Bayesianas: Inferencia

- A partir de la red, con todos sus parámetros...
- Calcular probabilidad de variables latentes
 - Puede que no todas...
- A partir de variables observadas
- Ejemplos:



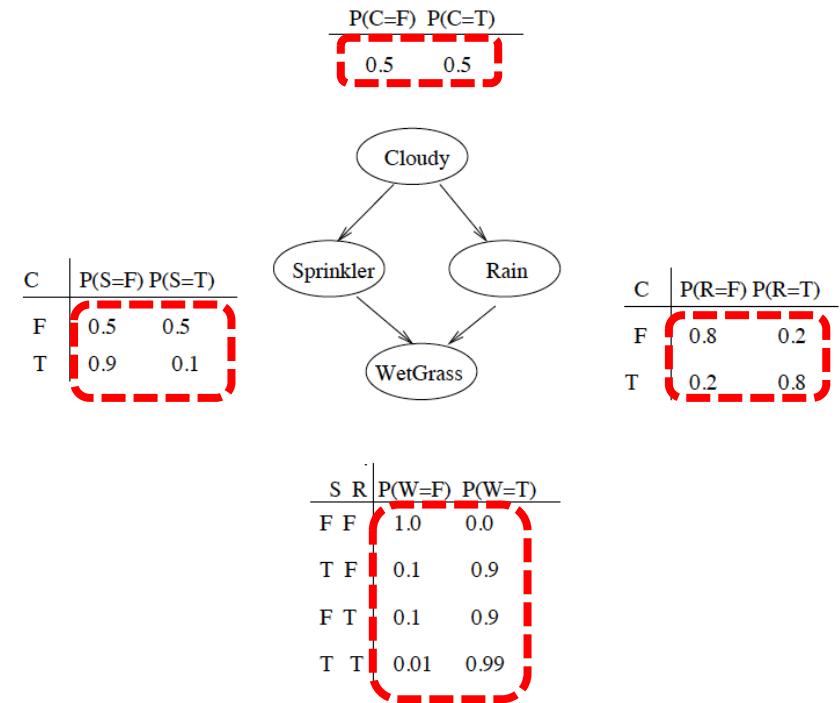
S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

$$P(S=1|W=1) = \frac{P(S=1, W=1)}{P(W=1)} = \frac{\sum_{c,r} P(C=c, S=1, R=r, W=1)}{P(W=1)} = \frac{0.2781}{0.6471} = 0.430$$

$$P(R=1|W=1) = \frac{P(R=1, W=1)}{P(W=1)} = \frac{\sum_{c,s} P(C=c, S=s, R=1, W=1)}{P(W=1)} = \frac{0.4581}{0.6471} = 0.708$$

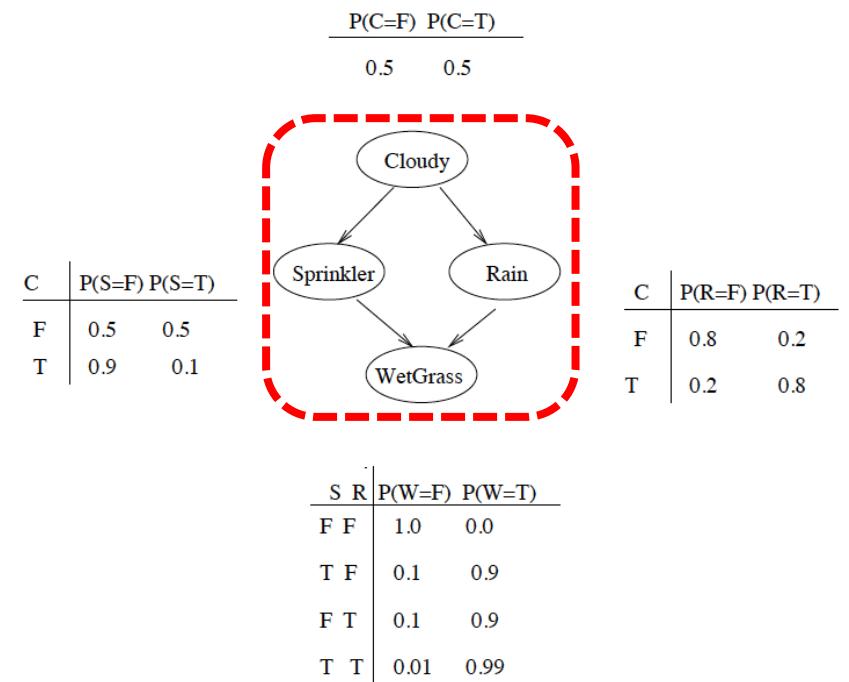
Redes Bayesianas: Aprendizaje

- A partir de datos obtener la red bayesiana
 - Aprendizaje de parámetros
 - La estructura ya es conocida
 - Tenemos que aprender los parámetros de las distribuciones probabilísticas



Redes Bayesinas: Aprendizaje

- Aprendizaje estructural
 - Aprender la estructura de dependencia de las variables



Ajuste Puntual Paramétrico *(Point-Estimates)*

Ajuste Puntual Paramétrico



- Supongamos que medimos la altura de 10 individuos de la población (media 170, desviación típica 20)
 - Muestra:
 $\{161.3, 175.7, 147.0, 193.8, 193.7\}$
- Objetivo: intentar ajustar una función densidad de probabilidad a esos datos
 - De acuerdo con un **criterio** a optimizar
- Suponemos que es una población con distribución gaussiana
 - **Sólo tenemos que hallar los parámetros μ y σ**
 - Ajuste paramétrico

Ajuste de media y varianza

- Ajuste de la media: media muestral

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N x_i = \frac{x_1 + x_2 + \dots + x_N}{N}$$

- Ajuste insesgado de la varianza: **cuasivarianza**

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \hat{\mu})^2 = \frac{(x_1 - \hat{\mu})^2 + (x_2 - \hat{\mu})^2 + \dots + (x_N - \hat{\mu})^2}{N-1}$$

- Cuasi-desviación típica

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$$



Ejemplo: Altura

- Tenemos los valores de antes: $x_i: \{161.3, 175.7, 147.0, 193.8, 193.7\}$
- Media muestral: 174,3
- Ajuste de la **cuasivarianza**:

$$\hat{\sigma}^2 = \frac{(161,3 - 174,3)^2 + (175,7 - 174,3)^2 + (147,0 - 174,3)^2 + (193,8 - 174,3)^2 + (193,7 - 174,3)^2}{4 - 1}$$
$$= 420,3$$

- La **cuasidesviación típica** es la raíz de la **cuasivarianza**:

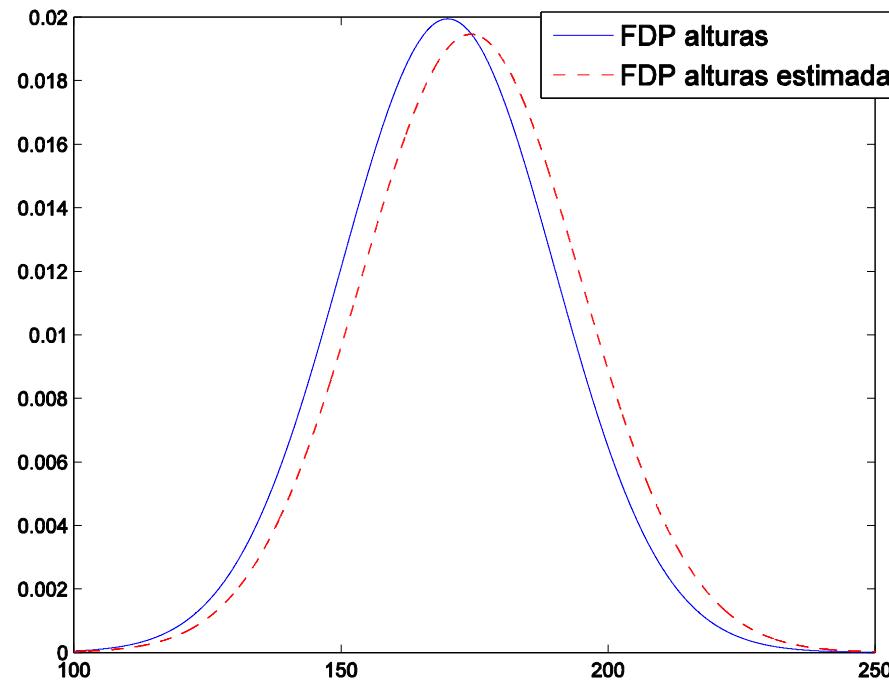
$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{407,3} = 20,5$$

- La desviación típica de la gaussiana generadora de alturas era 20...

Estimación



- Tomamos esos valores como la media y la varianza de la gaussiana:
 - Podemos hacer cálculos con ella



Problema



- Tenemos una muestra de longitud de fémures en hombres de España.
 - Esa muestra es (en mm.) 433.8, 383.6, 233.4, 436.3, 414.5, 323.0, 306.7, 276.7, 366.2 y 444.7 mm.
- Suponiendo una distribución gaussiana:
 - Ajustar la media.
 - Ajustar la cuasivarianza.
- Con esos valores de ajuste:
 - Calcular la probabilidad de que un hombre tenga el fémur más corto que 435 mm.

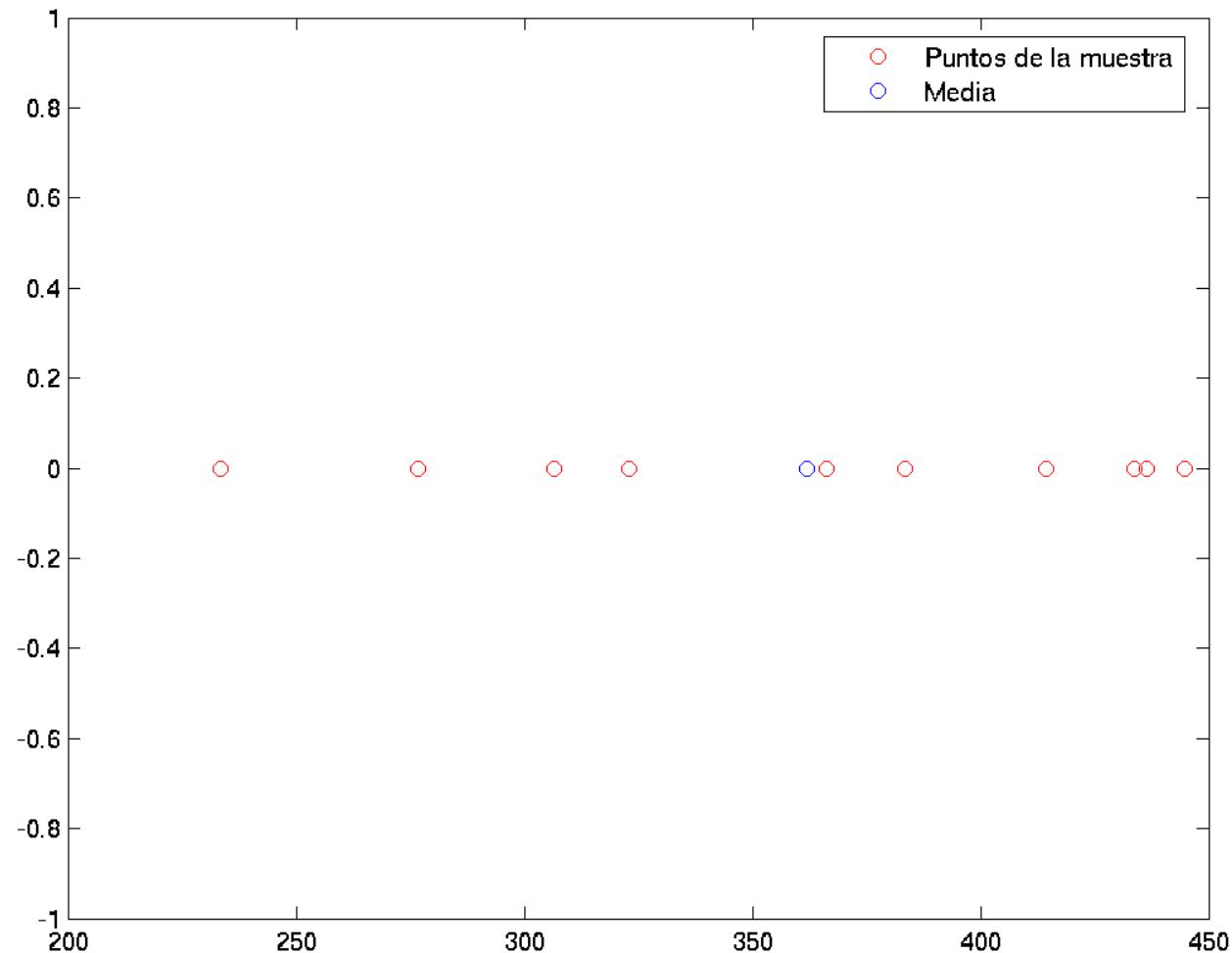
Ajuste de la media



- Muestra: 433.8, 383.6, 233.4, 436.3, 414.5, 323.0, 306.7, 276.7, 366.2 y 444.7 mm.

$$\hat{\mu} = \frac{(433.8 + 383.6 + 233.4 + \dots)}{10} = 361.9$$

Ajuste de la media



Ajuste de la cuasidesviación típica y la varianza

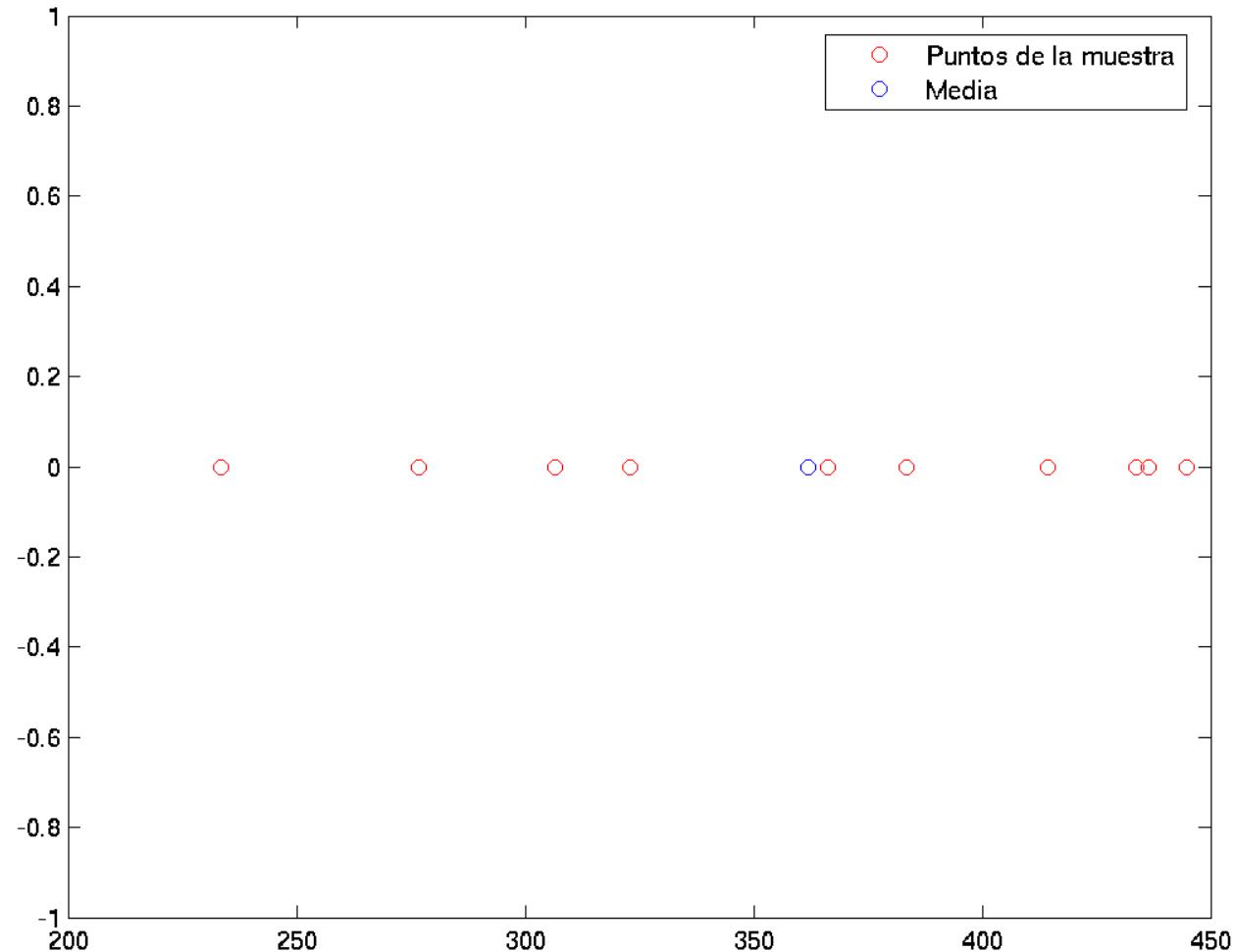


- Muestra: 433.8, 383.6, 233.4, 436.3, 414.5, 323.0, 306.7, 276.7, 366.2 y 444.7 mm.

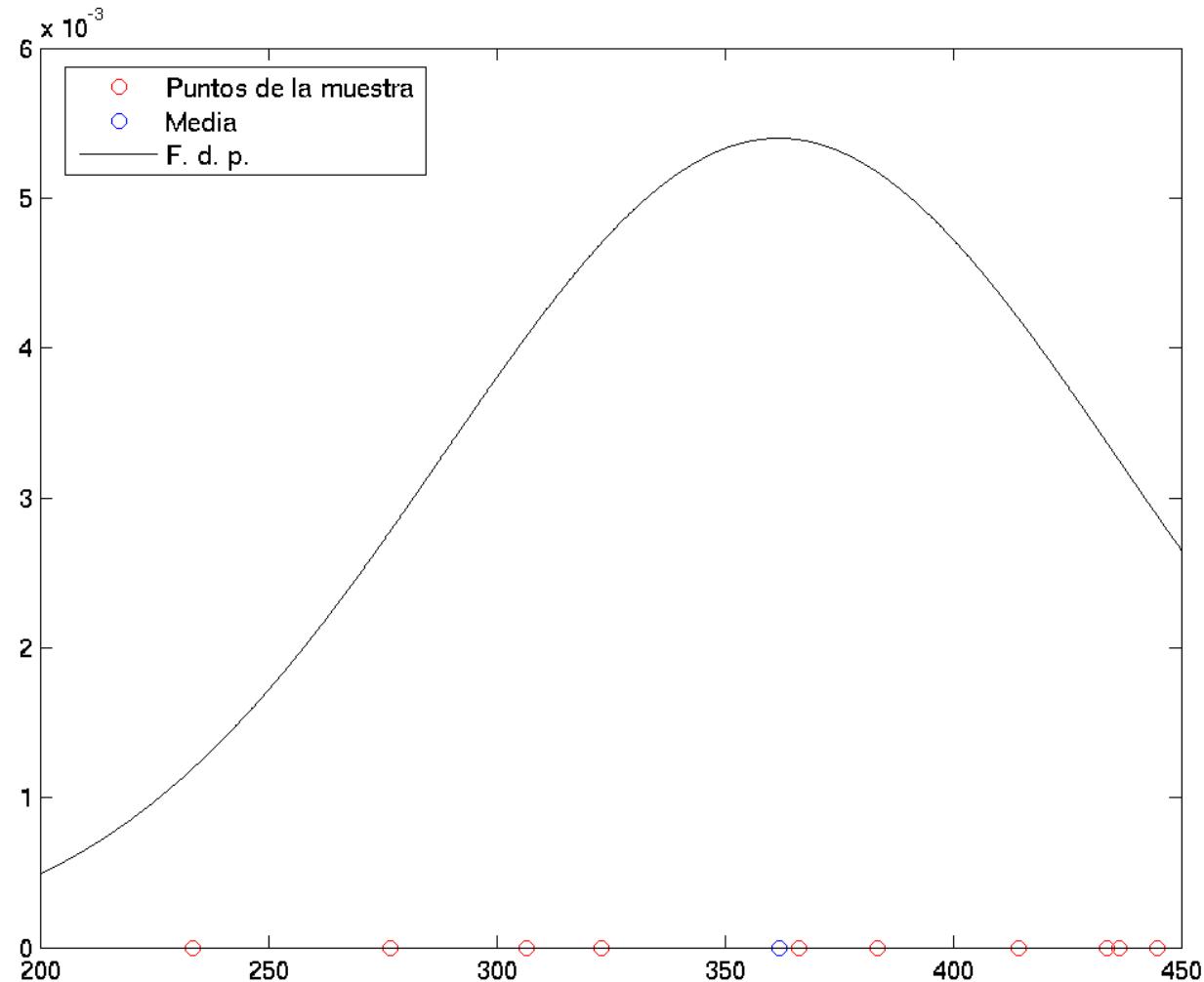
$$\hat{\sigma}^2 = \frac{(433.8 - 361.9)^2 + (383.6 - 361.9)^2 + (233.4 - 361.9)^2 + \dots}{9} = 5461$$

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = 73.9$$

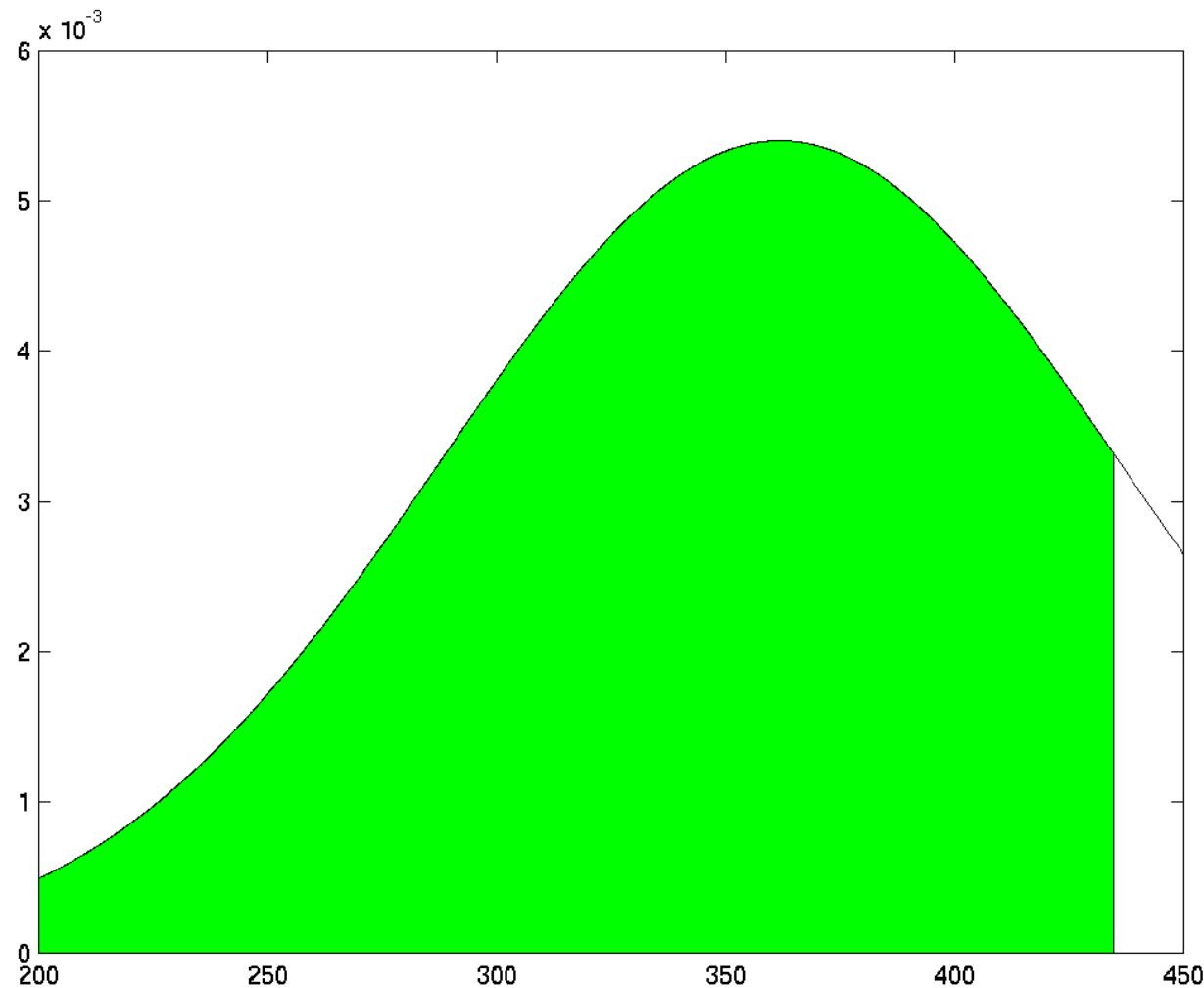
Ajuste de la desviación típica



Ajuste de la desviación típica



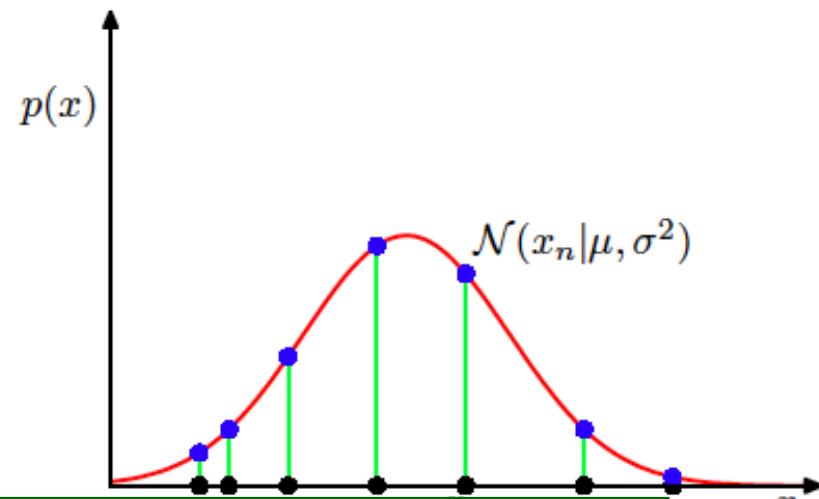
Cálculo de probabilidad



Máxima Verosimilitud

- Un criterio a maximizar para obtener estos ajustes
- Maximizar la *verosimilitud* de los datos
 - Producto de densidades de probabilidad de los datos observados
 - De ese modo (casi) se obtienen los ajustes antes mencionados

Figure 1.14 Illustration of the likelihood function for a Gaussian distribution, shown by the red curve. Here the black points denote a data set of values $\{x_n\}$, and the likelihood function given by (1.53) corresponds to the product of the blue values. Maximizing the likelihood involves adjusting the mean and variance of the Gaussian so as to maximize this product.



C. Bishop. "Pattern Recognition and Machine Learning". Springer, 2006.

Máxima Verosimilitud

- Ejemplo: ajuste de la media con varianza conocida

R. Duda, P. Hart, D. Stork.
“Pattern Classification, 2nd Ed.”. Wiley, 2001.

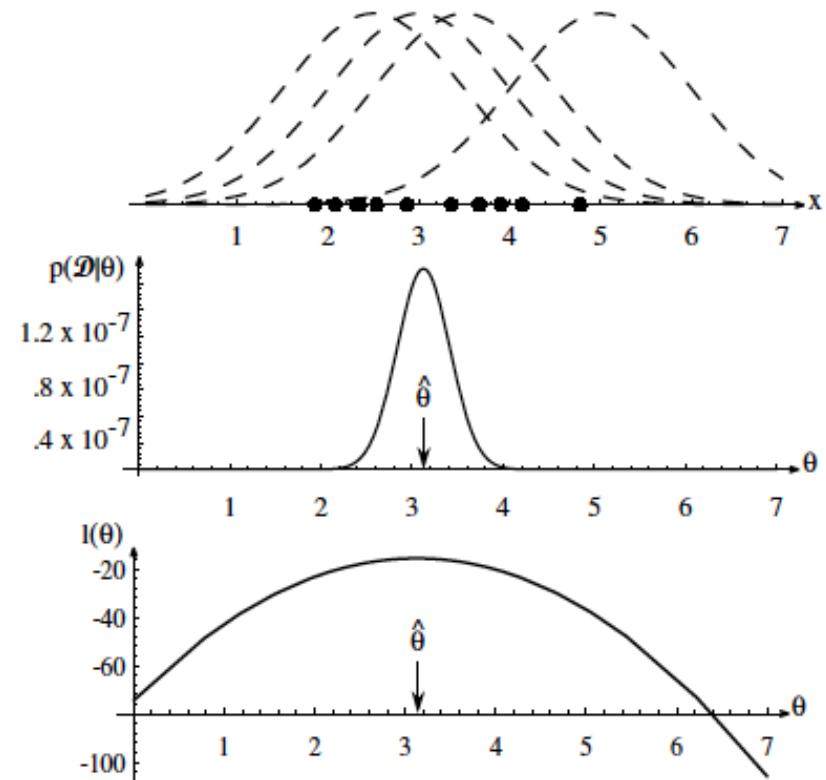


Figure 3.1: The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figures shows the likelihood $p(\mathcal{D}|\theta)$ as a function of the mean. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked $\hat{\theta}$; it also maximizes the logarithm of the likelihood — i.e., the log-likelihood $l(\theta)$, shown at the bottom. Note especially that the likelihood lies in a different space from $p(x|\hat{\theta})$, and the two can have different functional forms.

Máxima Verosimilitud

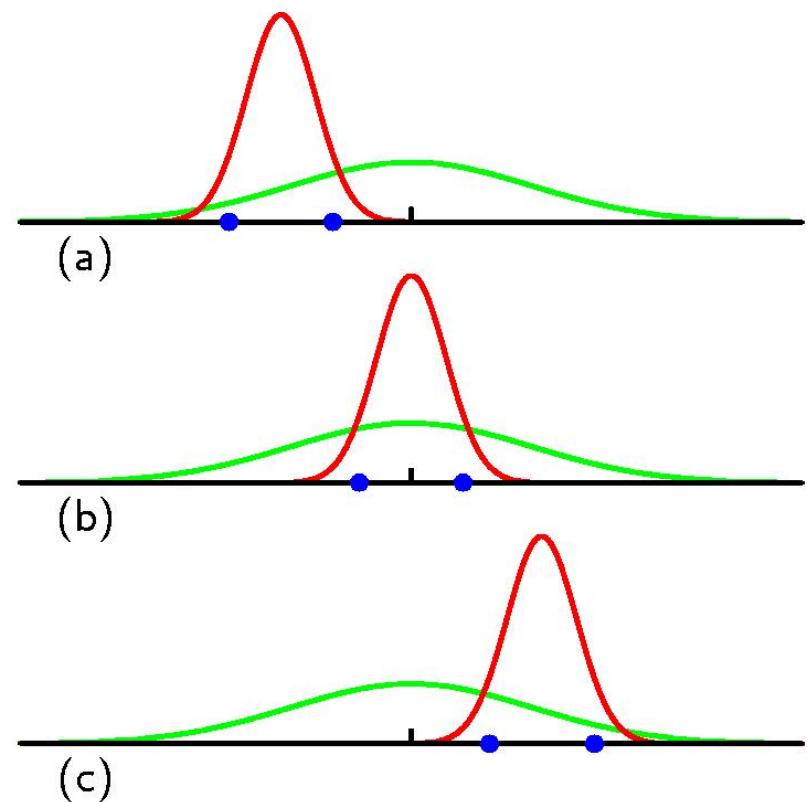
- En algunos casos, se puede obtener una solución analítica del ajuste por máxima verosimilitud
 - Ejemplo: f.d.p. Gaussiana (notación Bishop)
 - Verosimilitud: $p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^N \mathcal{N}(x_n|\mu, \sigma^2)$
 - Log-likelihood $\ln p(\mathbf{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi)$
 - Ajuste ML $\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n$ $\sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2$
 $\tilde{\sigma}^2 = \frac{N}{N-1} \sigma_{\text{ML}}^2$
 - Cuasivarianza (insesgado) $= \frac{1}{N-1} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2$

Máxima Verosimilitud: Problemas

- ¡Máxima verosimilitud (ML) solo tiene en cuenta los datos!
- Si los datos son escasos...
 - Existe incertidumbre
 - Que ML no tiene en cuenta

Figure 1.15 Illustration of how bias arises in using maximum likelihood to determine the variance of a Gaussian. The green curve shows the true Gaussian distribution from which data is generated, and the three red curves show the Gaussian distributions obtained by fitting to three data sets, each consisting of two data points shown in blue, using the maximum likelihood results (1.55) and (1.56). Averaged across the three data sets, the mean is correct, but the variance is systematically under-estimated because it is measured relative to the sample mean and not relative to the true mean.

C. Bishop. "Pattern Recognition and Machine Learning". Springer, 2006.



Máximo A Posteriori (MAP)

- Podemos incluir un poco más de contexto sobre el parámetro
 - Probabilidad a priori del parámetro (*prior* del parámetro)
 - Nos dice dónde están los valores esperados del parámetro

Máximo A Posteriori (MAP)

- Podemos incluir un poco más de contexto sobre el parámetro
 - Probabilidad a priori del parámetro (*prior* del parámetro)
 - Nos dice dónde están los valores esperados del parámetro
- Se optimiza likelihood multiplicado por el prior $p(\mu, \sigma)$

$$p(x|\mu, \sigma^2)p(\mu, \sigma^2)$$

- Equivalente a optimizar $p(\mu, \sigma^2|x)$
 - Probabilidad a posteriori: variables latentes dadas las variables observadas

Máximo A Posteriori (MAP)

- Podemos incluir un poco más de contexto sobre el parámetro
 - Probabilidad a priori del parámetro (*prior* del parámetro)
 - Nos dice dónde están los valores esperados del parámetro
- Se optimiza likelihood multiplicado por el prior $p(\mu, \sigma)$

$$p(x|\mu, \sigma^2)p(\mu, \sigma^2)$$

- Equivalente a optimizar $p(\mu, \sigma^2|x)$
 - Probabilidad a posteriori: variables latentes dadas las variables observadas
- MAP sigue siendo ajuste puntual
 - Problemas parecidos a ML
 - Reducidos, por efecto del prior (regularización)

Ajuste Puntual vs. Estadística Bayesiana

- En estadística bayesiana, a estas técnicas de estimación paramétrica se les suele llamar “estimación puntual”
 - Máxima verosimilitud (*Maximum Likelihood, ML*)
 - Máximo a posteriori (MAP)

Ajuste Puntual vs. Estadística Bayesiana

- En estadística bayesiana, a estas técnicas de estimación paramétrica se les suele llamar “estimación puntual”
 - Máxima verosimilitud (*Maximum Likelihood*, ML)
 - Máximo a posteriori (MAP)
- Son casos particulares de la estadística bayesiana
 - Con los múltiples problemas asociados comentados

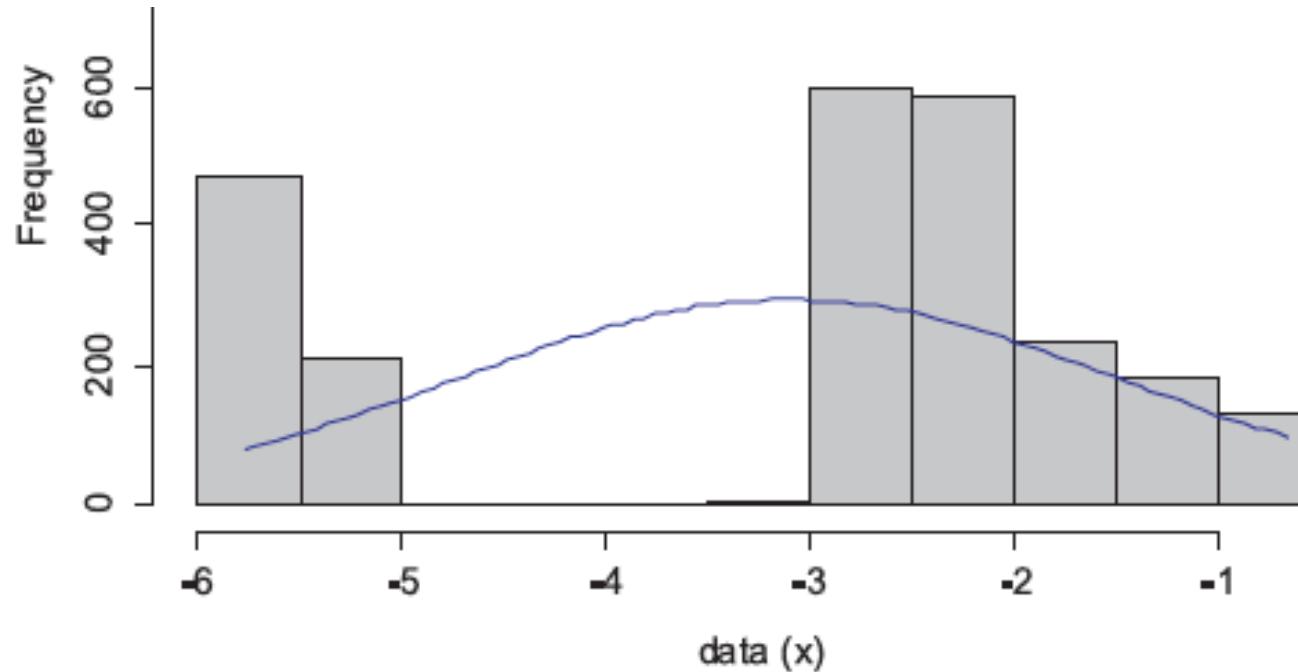
Ajuste Puntual vs. Estadística Bayesiana

- En estadística bayesiana, a estas técnicas de estimación paramétrica se les suele llamar “estimación puntual”
 - Máxima verosimilitud (*Maximum Likelihood*, ML)
 - Máximo a posteriori (MAP)
- Son casos particulares de la estadística bayesiana
 - Con los múltiples problemas asociados comentados
- Sin embargo, son métodos típicos en entrenamiento de DNNs
 - Criterio: ML, MAP
 - Obtención de parámetros de la red (pesos): ajuste
 - Optimización por gradiente (SGD, *backpropagation*)

Ajuste No Paramétrico

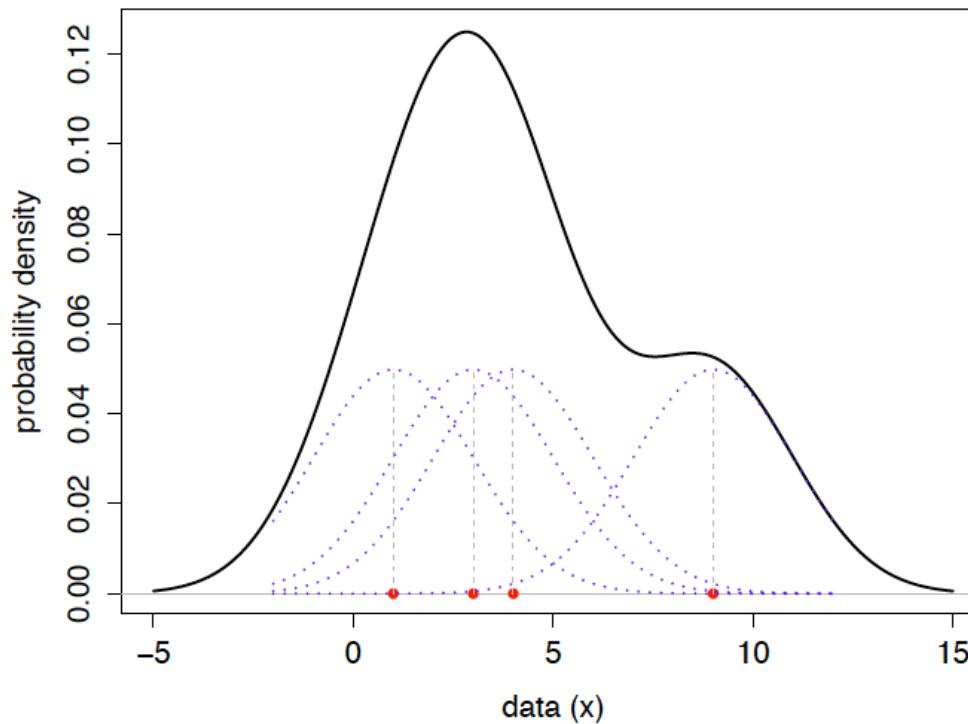
Problemas del Ajuste Paramétrico

- Es necesario elegir una función densidad de probabilidad adecuada a nuestros datos
 - “Seleccionar un modelo” (gaussiano, uniforme, etc.)
- En ocasiones, ningún modelo parece cuadrar con los datos...
 - Ejemplo ([\[Zadora14\]](#))



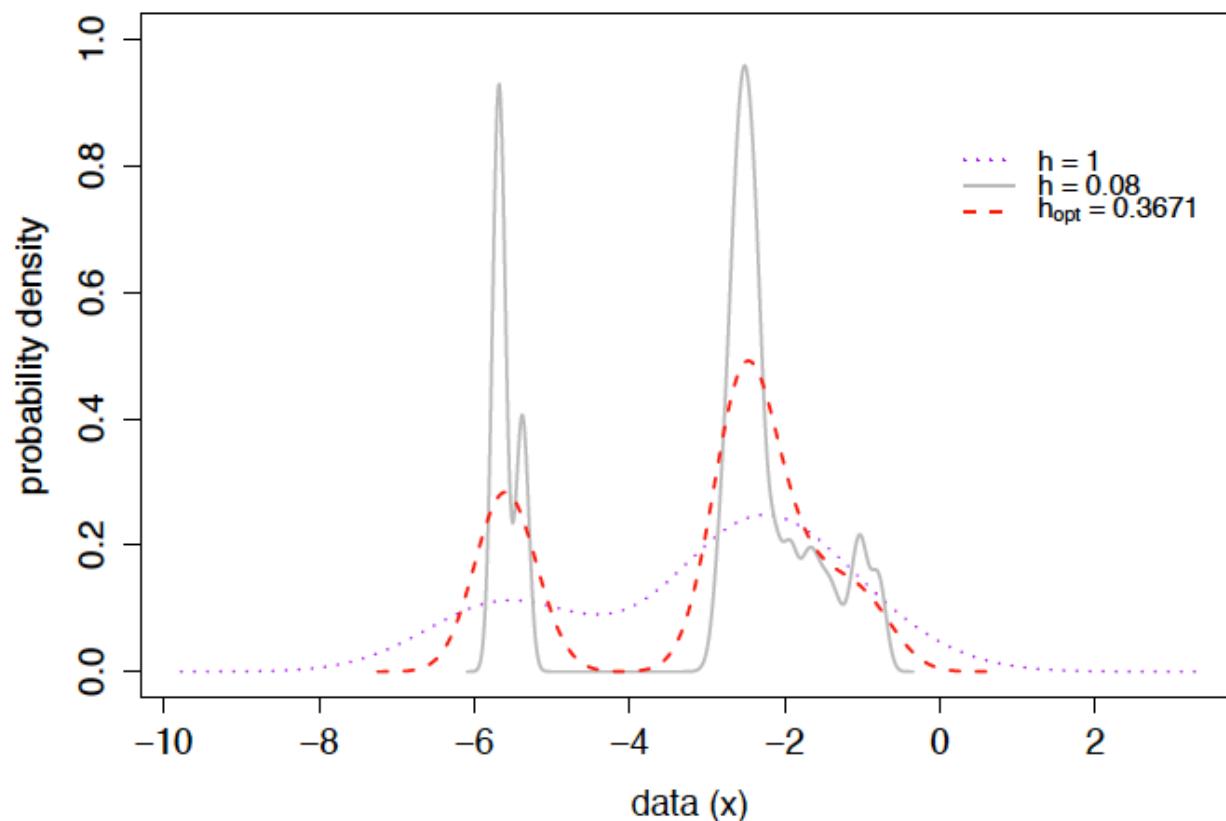
Ajuste No Paramétrico

- Mejor ajuste si hay muchos datos
 - Pero con pocos datos, suele funcionar mucho peor
- Ejemplo típico: funciones kernel (Kernel Density Functions)
 - **Ejemplo ([Zadora14])**



Ajuste No Paramétrico

- Importante: selección de la *anchura* del kernel
 - Hay opciones óptimas a partir de criterios estadísticos
 - **Ejemplo ([Zadora14])**



Inferencia Bayesiana Paramétrica

Inferencia Bayesiana

- Enfoque mucho más natural que el ajuste puntual
 - Considérese la relación entre variables
 - Modelo gráfico probabilístico
 - Úsense las leyes de la probabilidad

Inferencia Bayesiana

- Enfoque mucho más natural que el ajuste puntual
 - Considérese la relación entre variables
 - Modelo gráfico probabilístico
 - Úsense las leyes de la probabilidad
- Ejemplo típico:
 - Se tiene un conjunto de datos observados $x = \{x_1, \dots, x_N\}$
 - Se quiere saber cuál es la probabilidad de un nuevo dato dados los datos observados
 - $p(X = x|x) \equiv p(x|x)$
 - Probabilidad **predictiva**
 - Por ejemplo, quiero saber cuál será probablemente el índice de refracción de los vidrios que mida en el futuro

Inferencia Bayesiana Paramétrica

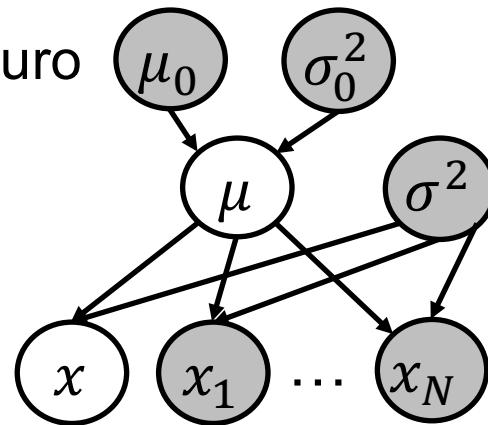
- En inferencia bayesiana, 2 suposiciones básicas
 - Si conociéramos el modelo, ¿cómo se distribuirían los datos?
 - **Verosimilitud (*likelihood*)**
 - $p(x|\theta)$, siendo θ los parámetros que dominan el modelo
 - En nuestro ejemplo, $p(x|\theta) = \mathcal{N}(x|\mu, \sigma^2)$
 - “Si conozco media y varianza, el índice de refracción es gaussiano”

Inferencia Bayesiana Paramétrica

- En inferencia bayesiana, 2 suposiciones básicas
 - Si conociéramos el modelo, ¿cómo se distribuirían los datos?
 - **Verosimilitud (*likelihood*)**
 - $p(x|\theta)$, siendo θ los parámetros que dominan el modelo
 - En nuestro ejemplo, $p(x|\theta) = \mathcal{N}(x|\mu, \sigma^2)$
 - “Si conozco media y varianza, el índice de refracción es gaussiano”
 - Si no he observado ningún dato, ¿cuál sería mi conocimiento sobre el modelo (sobre los parámetros)?
 - **Probabilidad a priori (*prior probability, prior*)**
 - $p(\theta)$. En nuestro ejemplo, $p(\mu) = N(\mu|\mu_0, \sigma_0^2)$, σ^2 conocida
 - “Conozco la varianza del índice de refracción, y la media debe estar en la región $N(\mu|\mu_0, \sigma_0^2)$ ”
 - μ_0, σ_0^2 : hiperparámetros (parámetros de los parámetros)

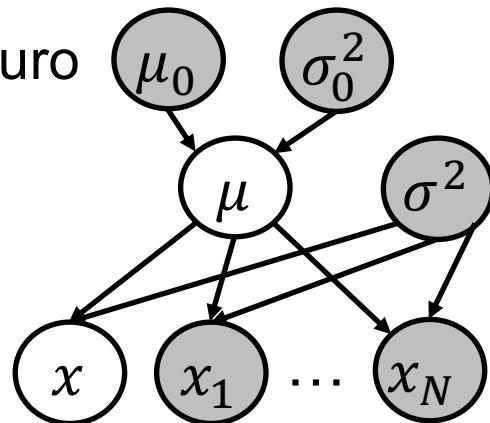
Inferencia Bayesiana Paramétrica

- Enfoque probabilístico puro
- Modelo gráfico:



Inferencia Bayesiana Paramétrica

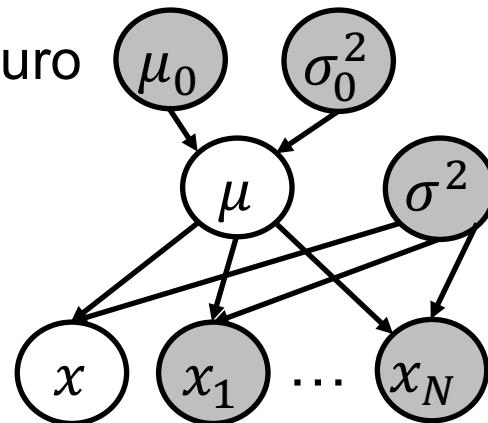
- Enfoque probabilístico puro
- Modelo gráfico:



- Consecuencias:
 - Factorización del modelo:
 - $p(x, x, \mu, \sigma^2, \sigma_0^2, \mu_0) = p(x|\mu, \sigma^2)p(x|\mu, \sigma^2)p(\mu|\mu_0, \sigma_0^2)$
 - Si conociera μ , los datos son independientes
 - $p(x|\mu, \sigma^2) = \prod_i p(x_i|\mu, \sigma^2)$

Inferencia Bayesiana Paramétrica

- Enfoque probabilístico puro
- Modelo gráfico:



- Consecuencias:
 - Factorización del modelo:
 - $p(x, \mathbf{x}, \mu, \sigma^2, \sigma_0^2, \mu_0) = p(\mathbf{x}|\mu, \sigma^2)p(\mathbf{x}|\mu, \sigma^2)p(\mu|\mu_0, \sigma_0^2)$
 - Si conociera μ , los datos son independientes
 - $p(\mathbf{x}|\mu, \sigma^2) = \prod_i p(x_i|\mu, \sigma^2)$
- Quiero hallar la probabilidad predictiva
 - $p(x|\mathbf{x})$

Inferencia Bayesiana Paramétrica

- La predictiva se descompone

$$p(x|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{x})}{p(\mathbf{x})} = \frac{\int p(\mathbf{x}, \mathbf{x}, \mu, \sigma^2, \sigma_0^2, \mu_0) d\mu}{\int p(\mathbf{x}, \mu, \sigma^2, \sigma_0^2, \mu_0) d\mu}$$

Inferencia Bayesiana Paramétrica

- La predictiva se descompone

$$p(x|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{x})}{p(\boldsymbol{x})} = \frac{\int p(\boldsymbol{x}, \boldsymbol{x}, \mu, \sigma^2, \sigma_0^2, \mu_0) d\mu}{\int p(\boldsymbol{x}, \mu, \sigma^2, \sigma_0^2, \mu_0) d\mu}$$

- Operando, se llega a:

$$p(x|\boldsymbol{x}) = \int p(\boldsymbol{x}|\mu, \sigma^2) \frac{\prod_i p(x_i|\mu, \sigma^2) p(\mu|\sigma_0^2, \mu_0)}{\int p(\boldsymbol{x}, \mu|\sigma^2, \mu_0, \sigma_0^2) d\mu} d\mu$$

Inferencia Bayesiana Paramétrica

- La predictiva se descompone

$$p(\mathbf{x}|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{x})}{p(\mathbf{x})} = \frac{\int p(\mathbf{x}, \mathbf{x}, \mu, \sigma^2, \sigma_0^2, \mu_0) d\mu}{\int p(\mathbf{x}, \mu, \sigma^2, \sigma_0^2, \mu_0) d\mu}$$

- Operando, se llega a:

$$p(\mathbf{x}|\mathbf{x}) = \int p(\mathbf{x}|\mu, \sigma^2) \frac{\prod_i p(x_i|\mu, \sigma^2) p(\mu|\sigma_0^2, \mu_0)}{\int p(\mathbf{x}, \mu|\sigma^2, \mu_0, \sigma_0^2) d\mu} d\mu$$

- Donde se define la probabilidad *a posteriori* del parámetro como:

$$p(\mu|\mathbf{x}, \sigma^2, \sigma_0^2, \mu_0) = \frac{\prod_i p(x_i|\mu, \sigma^2) p(\mu|\sigma_0^2, \mu_0)}{\int p(\mathbf{x}, \mu|\sigma^2, \mu_0, \sigma_0^2) d\mu} = \frac{\prod_i p(x_i|\mu, \sigma^2) p(\mu|\sigma_0^2, \mu_0)}{p(\mathbf{x}|\sigma^2, \mu_0, \sigma_0^2)}$$

Inferencia Bayesiana Paramétrica

- Simplificando la notación (eliminando términos constantes $\sigma^2, \sigma_0^2, \mu_0$)

- Probabilidad a priori: $p(\mu)$

- Probabilidad a posteriori: $p(\mu|x) = \frac{\prod_i p(x_i|\mu)p(\mu)}{p(x)} = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$

$$\mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \bar{x} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 \quad \sigma_N^2 = \frac{\sigma^2 \sigma_0^2}{N\sigma_0^2 + \sigma^2}$$

- Lo que sabemos de la media después de conocer los datos

Inferencia Bayesiana Paramétrica

- Simplificando la notación (eliminando términos constantes $\sigma^2, \sigma_0^2, \mu_0$)

- Probabilidad a priori: $p(\mu)$

- Probabilidad a posteriori: $p(\mu|x) = \frac{\prod_i p(x_i|\mu)p(\mu)}{p(x)} = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$

$$\mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \bar{x} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 \quad \sigma_N^2 = \frac{\sigma^2 \sigma_0^2}{N\sigma_0^2 + \sigma^2}$$

- Lo que sabemos de la media después de conocer los datos

R. Duda, P. Hart, D. Stork.
“Pattern Classification, 2nd Ed.”. Wiley, 2001.

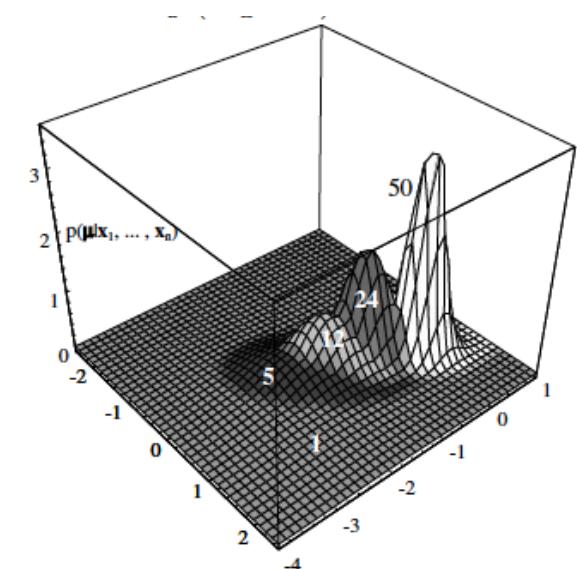
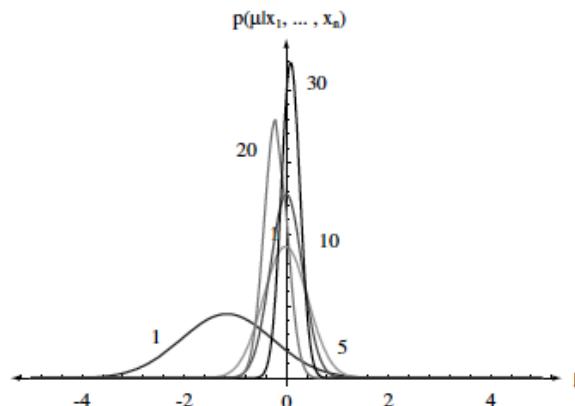


Figure 3.2: Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labelled by the number of training samples used in the estimation.

Inferencia Bayesiana Paramétrica

- Simplificando la notación (eliminando términos constantes $\sigma^2, \sigma_0^2, \mu_0$)

- Probabilidad a priori: $p(\mu)$

- Probabilidad a posteriori: $p(\mu|x) = \frac{\prod_i p(x_i|\mu)p(\mu)}{p(x)} = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$

$$\mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \bar{x} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 \quad \sigma_N^2 = \frac{\sigma^2 \sigma_0^2}{N\sigma_0^2 + \sigma^2}$$

- Lo que sabemos de la media después de conocer los datos

R. Duda, P. Hart, D. Stork.
“Pattern Classification, 2nd Ed.”. Wiley, 2001.

Infinitos datos:
Máximo (“Delta”)
en parámetro MAP

MAP: caso particular
de inferencia
Bayesiana

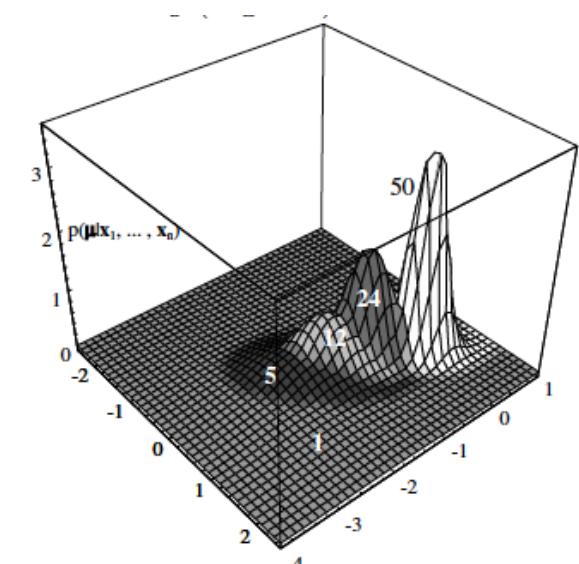
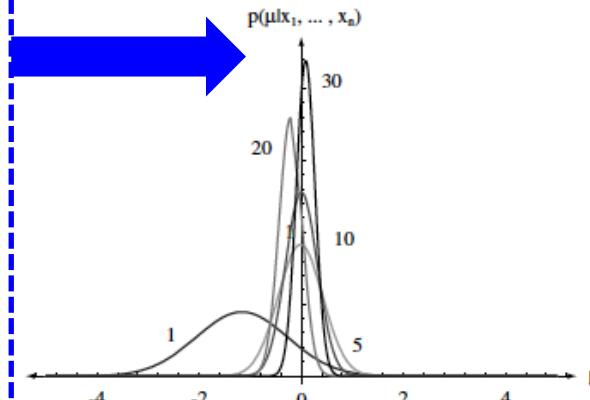


Figure 3.2: Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labelled by the number of training samples used in the estimation.

Densidades (distribuciones) conjugadas

- Se puede demostrar que si $p(\mu)$ es gaussiano, $p(\mu|x)$ es gaussiano
 - La gaussiana es la conjugada de la gaussiana para la media
 - Con *likelihood* gaussiana y varianza constante

Densidades (distribuciones) conjugadas

- Se puede demostrar que si $p(\mu)$ es gaussiano, $p(\mu|x)$ es gaussiano
 - La gaussiana es la conjugada de la gaussiana para la media
 - Con *likelihood* gaussiana y varianza constante
 - Probabilidades conjugadas permiten obtener fórmulas analíticas para el *posterior* a partir del *prior*
 - Simplifica muchísimo los cálculos
 - Sobre todo en el caso multivariado
 - Y en análisis *Fully-Bayesian*

Densidades (distribuciones) conjugadas

- Se puede demostrar que si $p(\mu)$ es gaussiano, $p(\mu|x)$ es gaussiano
 - La gaussiana es la conjugada de la gaussiana para la media
 - Con *likelihood* gaussiana y varianza constante
 - Probabilidades conjugadas permiten obtener fórmulas analíticas para el *posterior* a partir del *prior*
 - Simplifica muchísimo los cálculos
 - Sobre todo en el caso multivariado
 - Y en análisis *Fully-Bayesian*

Conjugate Bayesian analysis of the Gaussian distribution

Kevin P. Murphy*
murphyk@cs.ubc.ca

Last updated October 3, 2007

Densidad Predictiva

- Finalmente:

$$p(x|\mu) = \int p(x|\mu) p(\mu|x) d\mu = \mathcal{N}(\mu|\mu_N, \sigma^2 + \sigma_N^2)$$

Densidad Predictiva

- Finalmente:

$$p(x|\boldsymbol{x}) = \int p(x|\mu) p(\mu|\boldsymbol{x}) d\mu = \mathcal{N}(\mu|\mu_N, \sigma^2 + \sigma_N^2)$$

- La predictiva está centrada en la media del parámetro a posteriori
 - Pero con una varianza igual a la de la verosimilitud (σ^2)
 - Más una varianza adicional σ_N^2
 - Que se reduce cuantos más datos hay...
$$\sigma_N^2 = \frac{\sigma^2 \sigma_0^2}{N \sigma_0^2 + \sigma^2}$$

Densidad Predictiva

- Finalmente:

$$p(x|x) = \int p(x|\mu)p(\mu|x)d\mu = \mathcal{N}(\mu|\mu_N, \sigma^2 + \sigma_N^2)$$

- La predictiva está centrada en la media del parámetro a posteriori
 - Pero con una varianza igual a la de la verosimilitud (σ^2)
 - Más una varianza adicional σ_N^2
 - Que se reduce cuantos más datos hay...
$$\sigma_N^2 = \frac{\sigma^2 \sigma_0^2}{N \sigma_0^2 + \sigma^2}$$
- σ^2 es la incertidumbre inevitable (ej: por los errores de medida de vidrios)
 - Incertidumbre aleatoria
- σ_N^2 es la incertidumbre que tenemos porque no tenemos datos suficientes
 - Incertidumbre epistémica

Modelo *Fully-Bayesian* de la Gaussiana

Inferring a Gaussian distribution

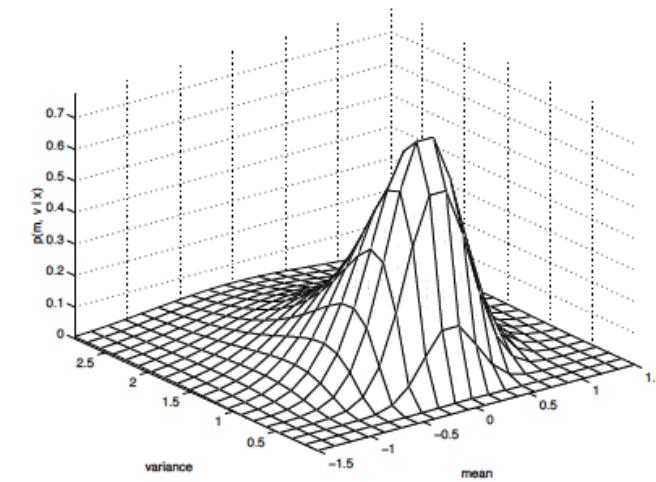
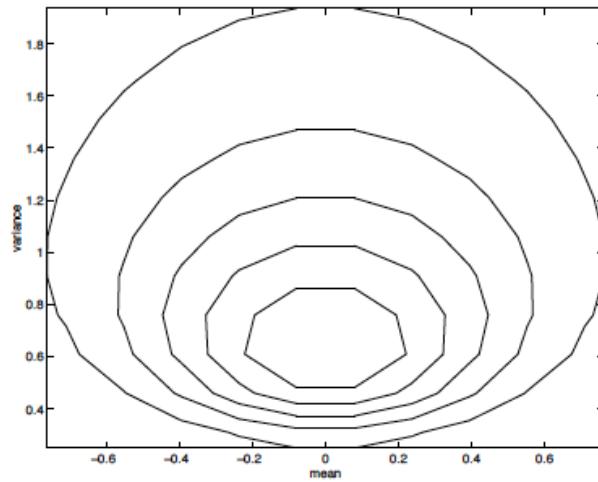
Thomas P. Minka

1998 (revised 2001)

- Verosimilitud:
 - gaussiana ([univariada](#) o [multivariada](#))
- Prior/posterior:
 - gaussiana-gamma ([gaussiana-Wishart inversa](#))
 - Gaussiana para la media
 - Gamma ([univariada](#)) o Wishart inversa ([multivariada](#)) para la varianza (matriz de covarianzas)
- Predictiva:
 - T de Student ([univariada](#) o [multivariada](#))

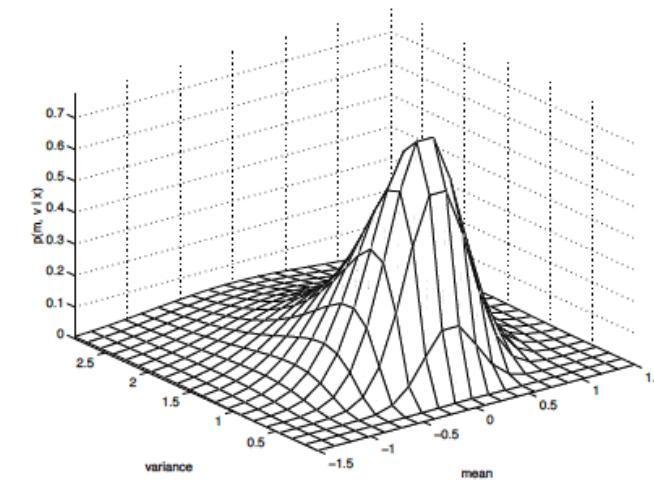
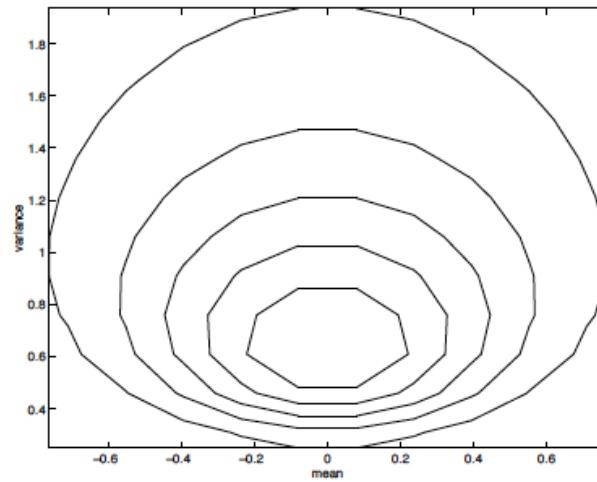
Modelo *Fully-Bayesian* de la Gaussiana

- Posterior Gaussiano-gamma (caso univariado):

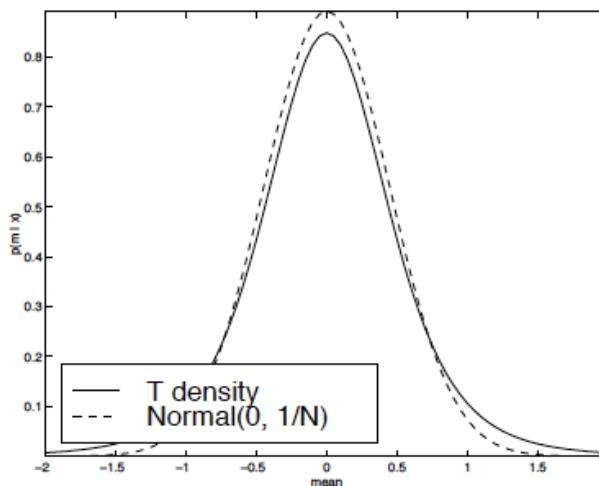


Modelo *Fully-Bayesian* de la Gaussiana

- Posterior Gaussiano-gamma (caso univariado):



- Predictiva: t de Student (univariada)



Verosimilitud Marginal (*Marginal Likelihood*)

- Recordemos:

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mu) d\mu = \int p(\mathbf{x}|\mu)p(\mu)d\mu$$

Verosimilitud Marginal (*Marginal Likelihood*)

- Recordemos:

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mu) d\mu = \int p(\mathbf{x}|\mu)p(\mu)d\mu$$

- Define el ajuste del modelo a los datos
 - Modelos que mejores se adapten a los datos funcionarán mejor

Verosimilitud Marginal (*Marginal Likelihood*)

- Recordemos:

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mu) d\mu = \int p(\mathbf{x}|\mu)p(\mu)d\mu$$

- Define el ajuste del modelo a los datos
 - Modelos que mejores se adapten a los datos funcionarán mejor
- Considera la incertidumbre
 - Suele contener un término que penaliza modelos muy complejos

Verosimilitud Marginal (*Marginal Likelihood*)

- Recordemos:

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mu) d\mu = \int p(\mathbf{x}|\mu)p(\mu)d\mu$$

- Define el ajuste del modelo a los datos
 - Modelos que mejores se adapten a los datos funcionarán mejor
- Considera la incertidumbre
 - Suele contener un término que penaliza modelos muy complejos
- Utilidad:
 - Selección de modelos (se elige el modelo con mayor $p(\mathbf{x})$)
 - Detección de outliers (se descartan valores con baja $p(\mathbf{x})$)

Verosimilitud Marginal (*Marginal Likelihood*)

- Recordemos:

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mu) d\mu = \int p(\mathbf{x}|\mu)p(\mu)d\mu$$

- Define el ajuste del modelo a los datos
 - Modelos que mejores se adapten a los datos funcionarán mejor
- Considera la incertidumbre
 - Suele contener un término que penaliza modelos muy complejos
- Utilidad:
 - Selección de modelos (se elige el modelo con mayor $p(\mathbf{x})$)
 - Detección de outliers (se descartan valores con baja $p(\mathbf{x})$)
- Además, es necesaria para calcular $p(\mu|\mathbf{x})$

$$p(\mu|\mathbf{x}) = \frac{\prod_i p(x_i|\mu)p(\mu)}{p(\mathbf{x})}$$

Verosimilitud Marginal (*Marginal Likelihood*)

- Problema:
 - Esta integral es muy difícil de calcular en algunos problemas

$$p(x) = \int p(x, z) dz$$

Verosimilitud Marginal (*Marginal Likelihood*)

- Problema:
 - Esta integral es muy difícil de calcular en algunos problemas
$$p(x) = \int p(x, z) dz$$
- Ejemplo:
 - En autoencoders variacionales, esa integral no se puede calcular directamente
 - ¡Computacionalmente intratable!

Verosimilitud Marginal (*Marginal Likelihood*)

- Problema:
 - Esta integral es muy difícil de calcular en algunos problemas
$$p(x) = \int p(x, z) dz$$
- Ejemplo:
 - En autoencoders variacionales, esa integral no se puede calcular directamente
 - ¡Computacionalmente intratable!
 - La verosimilitud marginal es normalmente la responsable de la no tratabilidad de modelos bayesianos complejos
 - Motiva el uso de aproximaciones al posterior
 - Inferencia Variacional
 - Métodos Monte Carlo

Inferencia Bayesiana Paramétrica

- Algunas conclusiones
 - Caso gaussiano *Fully-Bayesian*
 - Muy sencillo conceptualmente
 - Aun así, bastante complicado como modelo matemático

Inferencia Bayesiana Paramétrica

- Algunas conclusiones
 - Caso gaussiano *Fully-Bayesian*
 - Muy sencillo conceptualmente
 - Aun así, bastante complicado como modelo matemático
 - Conjugados permiten tratabilidad
 - Soluciones analíticas (cerradas) del *posterior*

Inferencia Bayesiana Paramétrica

- Algunas conclusiones
 - Caso gaussiano *Fully-Bayesian*
 - Muy sencillo conceptualmente
 - Aun así, bastante complicado como modelo matemático
 - Conjugados permiten tratabilidad
 - Soluciones analíticas (cerradas) del *posterior*
 - Si no tenemos conjugados o tenemos distribuciones más complejas
 - Los problemas rápidamente se vuelven intratables
 - Computacionalmente o analíticamente
 - Principal responsable computacional: la verosimilitud marginal

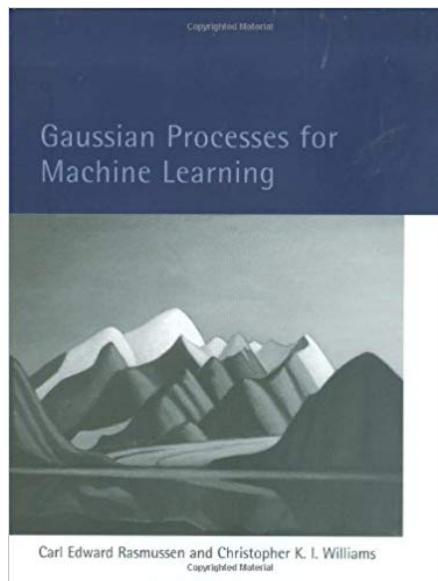
$$p(\mathbf{x}) = \int p(\mathbf{x}, \boldsymbol{\mu}) d\boldsymbol{\mu}$$



Inferencia Bayesiana No Paramétrica

Inferencia Bayesiana No Paramétrica

- “No paramétrico” suele querer decir que el número de parámetros no está limitado
 - Podría ser infinito
- Método bayesiano no paramétrico más usado actualmente
 - Procesos Gaussianos (no los veremos)



THE ROYAL SOCIETY PUBLISHING | PHILOSOPHICAL TRANSACTIONS A

ABOUT ▶ BROWSE BY SUBJECT ▶ ALERTS ▶ FREE TRIAL ▶

[Philos Trans A Math Phys Eng Sci.](#) 2013 Feb 13; 371(1984): 20110553. PMCID: PMC3538441
[doi: 10.1098/rsta.2011.0553](#) PMID: 23277609

Bayesian non-parametrics and the probabilistic approach to modelling

[Zoubin Ghahramani](#)



Clasificadores Probabilísticos en Aprendizaje Automático

Día 2: Modelos Probabilísticos

Daniel Ramos Castro

daniel.ramos@uam.es

Audias – Audio, Data Intelligence and Speech
Universidad Autónoma de Madrid

<http://audias.ii.uam.es>

audias

Audio, Data Intelligence and Speech

UAM