Modelos Gráficos Probabilísticos
Parte III: Aprendizaje

Serafín Moral

smc@decsai.ugr.es
Departamento de Ciencias de la Computación e IA
Universidad de Granada
18071 - Granada

Contenido (I)

- Tema 1: Introducción a los modelos probabilísticos
  - Introducción al razonamiento probabilístico
  - Independencia condicional
  - Elicitación de probabilidades
- Tema 2: Redes bayesianas
  - Propiedades de las relaciones de independencia condicional
  - Representación de independencias mediante grafos
  - Parametrización de redes bayesianas
- Tema 3: Otros modelos gráficos probabilísticos
  - Grafos no dirigidos
  - Grafos cadena
- Tema 4: Construcción de redes bayesianas
  - Método general de construcción
  - Causalidad
  - Determinación y representación de las probabilidades condicionales
  - Puertas OR
  - Variables continuas
  - Modelos temporales

Contenido (II)

- Tema 5: Problemas de inferencia en modelos gráficos
  - Cálculo de probabilidades condicionales
  - Cálculo' de la explicación más probable
  - Cálculo de la cantidad de información
  - Análisis de sensibilidad
- Tema 6: Estimación de probabilidades
  - Máxima verosimilitud
  - Modelos bayesianos
  - El problema del tamaño muestral equivalente
  - Datos incompletos
- Tema 7: Aprendizaje estructural
  - Algoritmos basados en tests de independencia
  - Algoritmos basados en optimización de métricas
  - Combinando información de expertos con aprendizaje automático
  - El espacio de búsqueda
- Tema 8: Modelos gráficos en problemas de clasificación y clustering
  - Naive Bayes
  - Modelo TAN
  - Modelo BAN
  - AODE AODE
  - Multiclasificadores
  - El algoritmo EM para clustering

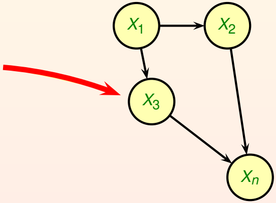
Referencias

- D. Heckerman (2008). A tutorial on learning with Bayesian networks. Innovations in Bayesian networks, 33-82.
- D. Koller, N. Friedman (2009). *Probabilistic Graphical Models: Principles and Techniques.* MIT Press.
- R.E. Neapolitan (2004) *Learning Bayesian Networks*. Prentice Hall, Upper Saddle River, NJ.

Aprendizaje de Redes Bayesianas

Aprendizaje de redes Bayesianas es el proceso de inducir un modelo a partir de una base de observaciones.

X_1	X_2	$\dots$	X_n
x_1^1	x_2^1	$\dots$	x_n^1
x_1^2	x_2^2	$\dots$	x_n^2
x_1^3	x_2^3	$\dots$	x_n^3
x_1^4	x_2^4	$\dots$	x_n^4

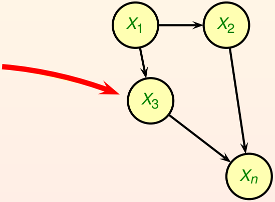


Aprendizaje = Inducir un grafo + Estimar los parámetros

Aprendizaje de Redes Bayesianas

Aprendizaje de redes Bayesianas es el proceso de inducir un modelo a partir de una base de observaciones.

X_1	X_2	$\dots$	X_n
x_1^1	x_2^1	$\dots$	x_n^1
x_1^2	x_2^2	$\dots$	x_n^2
x_1^3	x_2^3	$\dots$	x_n^3
x_1^4	x_2^4	$\dots$	x_n^4



Aprendizaje = Inducir un grafo + Estimar los parámetros

Motivación

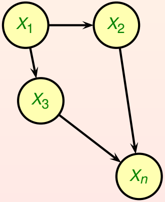
- La construcción a partir de expertos puede ser difícil
  - Dificultades de los expertos sobre el significado de grafos y números
  - Dominios muy grandes. Los expertos lo conocen de forma parcial
  - Es caro y lento
  - No hay expertos: queremos descubrir
- El aprendizaje con redes bayesianas ofrece un rango de posibilidades
  - Puede integrar el conocimiento expertos cuando esté disponible.
  - Pueden estar orientado a clasificación
  - Es posible descubrir relaciones causales
  - Es barato

Modalidades: Grafo Conocido

Grafo Conocido

X_1	X_2	$\dots$	X_n
x_1^1	x_2^1	$\dots$	x_n^1
x_1^2	x_2^2	$\dots$	x_n^2
x_1^3	x_2^3	$\dots$	x_n^3
x_1^4	x_2^4	$\dots$	x_n^4

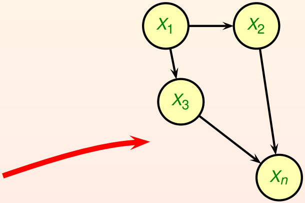
$$p(x_2|x_1), p(x_3|x_1), p(x_n|x_2, x_3)$$



Modalidades: Grafo desconocido

- Grafo desconocido

X_1	X_2	$\dots$	X_n
x_1^1	x_2^1	$\dots$	x_n^1
x_1^2	x_2^2	$\dots$	x_n^2
x_1^3	x_2^3	$\dots$	x_n^3
x_1^4	x_2^4	$\dots$	x_n^4

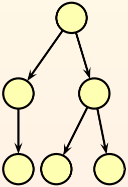


$$p(x_2|x_1), p(x_3|x_1), p(x_n|x_2, x_3)$$

Modalidades: Conocimiento Parcial

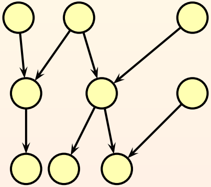
- Conocemos la presencia (ausencia) de enlaces
- Conocemos una relación de orden parcial: X_1 precede X_n ó X_1 no puede ser un descendiente de X_n.
- Conocemos que algunas relaciones de independencia condicional se dan (o que no se dan)
- Tenemos una distribución 'a priori' sobre el conjunto de los grafos posibles.

Conjunto Restringido de Estructuras



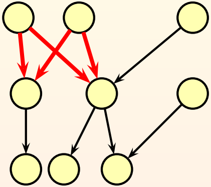
Árbol

Cada nodo, a lo más, un padre.



Hiperárbol

No ciclos no-dirigidos.



Grafo Simple

Cada ciclo no-dirigido tiene, al memos, dos nodos *cabeza-cabeza*.

Datos Incompletos

- Los datos incompletos no son un problema cuando se usa una red bayesiana (*Pensad en otros modelos como regresión o redes neuronales*).
- Los datos incompletos cuando se aprende una red hacen todo más difícil (*hay una hipótesis importante que no se satisface en ese caso*).

  Podemos tener dos modalidades: estimar los parámetros o estimar los parámetros y la estructura.

X_1	X_2	$\dots$	X_n
?	x_2^1	$\dots$	x_n^1
x_1^2	?	$\dots$	?
x_1^3	?	$\dots$	x_n^3
x_1^4	x_2^4	$\dots$	x_n^4

Datos Perdidos

En general, necesitamos alguna condición de cómo se pierden los datos:

- Missing completely at random (MCAR): *El proceso por el cual una variable se pierde es independiente del valor de la misma y del resto de las variables (sean o no observadas)*.
- Missing at random (MAR): *El proceso por el cual una variable no es observada es independiente de todo lo demás condicionado a las variables observadas*. Ejemplo: los pacientes sin enfermedad cardiovascular (variable observada) serán más propensos a no tener datos de hipertensión arterial.

En otro caso, se podría considerar el valor de 'no observado' como un valor más de la variable, pero en ese caso el proceso mediante el cual una variable no se observa debería de ser el mismo cuando se usa que cuando se aprende.

El Problema de la Clasificación Supervisada

Tenemos un conjunto de variables o atributos $\mathbf{X} = (X_1, \ldots, X_n)$.

Cada X_i toma valores en un conjunto finito Ω_{X_i}.

Tenemos un variable clase C, con valores en Ω_C.

Tenemos un conjunto de datos para estas variables:

X_1	X_2	$\ldots$	X_n	C
x_1^1	x_2^1	$\ldots$	x_n^1	c_1
x_1^2	x_2^2	$\ldots$	x_n^2	c_2
x_1^3	x_2^3	$\ldots$	x_n^3	c_3
x_1^4	x_2^4	$\ldots$	x_n^4	c_4

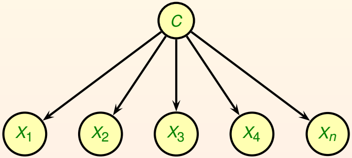
Queremos inducir un modelo M tal que si $\mathbf{x}$ es un valor de $\mathbf{X}$:

$$\mathbf{x} \longrightarrow \boxed{M} \longrightarrow c \in \Omega_C$$

obtenemos un valor estimado $c \in \Omega_C$

Clasificador Naive Bayes

Hipótesis:
Los atributos son condicionalmente independientes dada la clase.



La distribución de probabilidad factoriza de la siguiente forma:

$$P(C=c).\prod_{i=1}^{n} P(X_i=x_i|C=c)$$

Clustering: Variables ocultas

We have a learning problem in which a variable (the class) is never observed.
We know the number of possible values of this variable

X_1	X_2	$\dots$	X_n	C
x_1^1	x_2^1	$\dots$	x_n^1	?
x_1^2	x_2^2	$\dots$	x_n^2	?
x_1^3	x_2^3	$\dots$	x_n^3	?
x_1^4	x_2^4	$\dots$	x_n^4	?

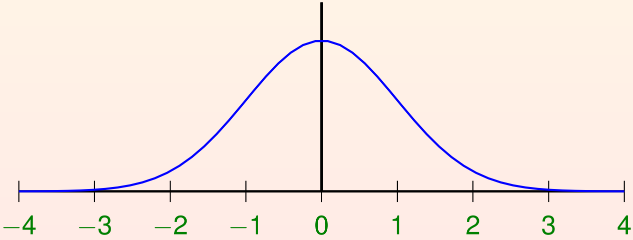
The value of the class variable is the group to which each case belongs.

Clustering con redes Bayesianas

- Construye una red bayesiana con una variable oculta con distintos valores que representan los clusters.
- En un nuevo caso, se puede calcular la probabilidad de pertenecer a cada uno de los clusters.
- El aprendizaje tratará de dividir los casos de tal forma que condicionado a cada grupo la red resultante se pueda aproximar por un tipo sencillo de red. De esta forma la estructura final es:
  - Naïve Bayes (Cheeseman, Stutz, 1995)
  - Semi Naïve Bayes, Árboles, otros modelos (Peña et al., 1999, 2000, 2002).

Variables Continuas

- Hay modelos para manejar variables discretas y continuas: el modelo condional Gaussiano. Tiene algunas restricciones importantes: *Por ejemplo, una variable discreta no puede ser hija de una variable continua*. X sigue una densidad gaussiana $N(\mu,\sigma)$.



- Sin embargo, el método más común es discretizar las variables.

Discretización: métodos

- Dividir el rango en k intervalos del mismo tamaño.
- Dividir el rango en k con la misma frecuencia.
- Enfoques cluster (k medias).
- Opiniones de expertos
- Discretización dinámica (Kovlov, Koller, UAI97)
- Discretización al aprender (Monti, Cooper, UAI98)

Aprendizaje: Casos Particulares

- Aprendizaje a partir de bases de datos estáticas o aprendizaje a partir de datos que llegan de forma continua (*data streams* ó *online*).
- Redes Bayesianas Dinámicas
- Diagramas de influencia o grafos de decisión: Es muy importante determinar el resultado de las acciones, ya que no es lo mismo observar $X = a$ que intervenir y hacer que $X = a$. Intervenir y hacer que $X = a$ se denota como $X = do(a)$

  Si las relaciones son causales, se puede determinar el efecto de las intervenciones, en otro caso no se puede.

Parameter Estimation

Serafín Moral
Dpt. Computer Science and Artificial Intelligence
University of Granada, Spain

Outline

- Learning a binomial probability
- Maximum likelihood
- The Bayesian approach
- The multinomial case
- Estimating parameters in a Bayesian network

Binomial Probability

We have a variable, X, with two possible values 1 and 0, and we want to estimate $\theta = P(X = 1)$.

For example, imagine a thumbtack of a given size and shape and imagine that we want to estimate the probability that if we throw the thumbtack up in the air, it will come to rest either in its point (heads, 1) or in its head (tails, 0).

We have a series of independent observations of this distribution, $(x[1], x[2], \ldots, x[m])$ and we want to estimate θ.

For example, imagine that we have observed $D = (1, 1, 0, 0, 1)$.

Sufficient Statistics

Let N_0 number of times in which $X = 0$ and N_1 the number of times in which $X = 1$.

The likelihood function is the probability of having observed the sequence given the parameter:

$$L(\theta : D) = \theta^{N_1}.(1-\theta)^{N_0}$$

In our case,

$$L(\theta : D) = \theta^{3}.(1-\theta)^{2}$$

N_0 and N_1 are called sufficient statistics for the parameter θ as the likelihood depends only of the data through these values.

Maximum Likelihood

The maximum likelihood estimation proposes as estimator for θ, the value $\hat{\theta}$ that maximizes the likelihood given the data $L(\theta : D)$.

- It is intuitively appealing
- It has good asymptotic properties

For binomial samples:

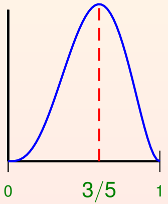
$$\hat{\theta} = \frac{N_1}{N_0 + N_1}$$

It coincides with the relative frequency of the event in the sample.

Valores	$X = 0$	$X = 1$
Frecuencias	2	3
Probabilidades	2/5	3/5

The Likelihood Function Shape

In the particular case of the sequence $D = (1,1,0,0,1)$, we have that the likelihood function is $\theta^3.(1-\theta)^2$.



And the maximum likelihood estimation is $\hat{\theta} = 3/5$, the relative frequency of heads in the sequence.

Problems

The main problem with maximum likelihood is when we have small samples.

Imagine that, in former example we have only observed the two first cases $D = (1,1)$, the estimated value would be $\hat{\theta} = 1$.

Valores	$X = 0$	$X = 1$
Frecuencias	0	2
Probabilidades	0.0	1.0

This looks very hard for future behavior as we consider that tail is completely impossible and this is based on a sample of size 2.

This is even worst for a sample of size 1, in which we always obtain 0-1 estimations.

Relevance of this problem

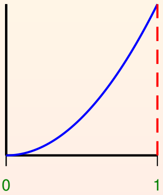
In Bayesian networks this problem is always relevant even if we have a very large database.

We have to estimate a conditional probability of a variable for each configuration of values of its parents.

For each configuration, we only consider the part of the database with values of variables in such configuration. In some cases, we can get a very small subset of the original database.

Shape of Likelihood

For the case $D = (1,1)$, the likelihood is as follows:



We give as maximum likelihood the supremum of all the possible values: 1.0.

It looks more reasonable a more intermediate value, of course closer to 1 than to 0.

The Bayesian Approach

The Bayesian approach assumes that the parameter θ is another random variable with a distribution $p(\theta)$.
If we observe data D, then the 'a posteriori' probability for the parameter is computed:

$$P(\theta|D) = \frac{P(D|\theta).p(\theta)}{P(D)} = \frac{P(D|\theta).p(\theta)}{\int P(D|\theta).p(\theta)d\theta}$$

As parameter estimation we compute

$$\theta^* = P(X[m+1]=1|D) = \int \theta p(\theta|D)d\theta = E_{P(\theta|D)}[\theta]$$

The Uniform Distribution

Even assuming the uniform distribution for the parameter:

$$f(\theta) = 1$$

the situation changes. Now, the estimation is

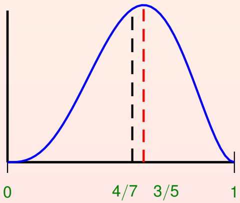
$$\theta^* = \frac{N_1 + 1}{N_0 + N_1 + 2}$$

This is also known as the Laplace correction.
The 'a posteriori' distribution for the parameter is proportional to the likelihood function. But now, the estimation is not the maximum, but the mean value.

The Uniform Distribution

In the case of $D = (1,1,0,0,1)$, we have $\theta^* = 4/7$:

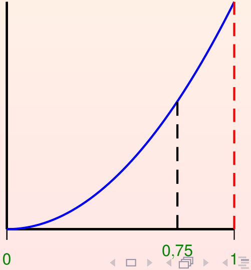
Valores	$X = 0$	$X = 1$
Frecuencias	2	3
Extra	1	1
Total	3	4
Probabilidades	3/7	4/7



The Uniform Distribution

In the case of $D = (1, 1)$, we have $\theta^* = 3/4 = 0{,}75$:

Valores	$X = 0$	$X = 1$
Frecuencias	0	2
Extra	1	1
Total	1	3
Probabilidades	1/4	3/4



The Beta Distribution

The uniform ’a priori’ distribution is a particular case of the Beta Distribution.
Its general form is:

$$f(\theta)=\frac{\Gamma(s)}{\Gamma(\alpha_1).\Gamma(\alpha_0)}\theta^{\alpha_1-1}.(1-\theta)^{\alpha_0-1}$$

where $s=\alpha_1+\alpha_0$ and $\Gamma(x)$ is the gamma function, a generalization of the factorial function. For integers $\Gamma(N)=(N-1)!$.
This is the Beta, $Beta(\alpha_1,\alpha_0)$.

The expected value of the parameter is: $\frac{\alpha_1}{\alpha_1+\alpha_0}$.

The uniform is the $Beta(1,1)$

Conjugate Distributions

There are important theoretical reasons for using the Beta 'a priori' distribution.

One of them has also important practical consequences: it is the conjugate distribution of binomial sampling. This means that if the 'a priori' is $\mathit{Beta}(\alpha_1, \alpha_0)$ and we have observed some data with N_1 and N_0 cases for the two possible values of the variable, then the 'a posteriori' is also Beta with parameters $\mathit{Beta}(\alpha_1 + N_1, \alpha_0 + N_0)$.

The expected value for the 'a posteriori' distribution is

$$\frac{\alpha_1 + N_1}{\alpha_1 + N_1 + \alpha_0 + N_0}$$

Intuitive Interpretation

The values

$$\left(\frac{\alpha_1}{\alpha_1+\alpha_0},\frac{\alpha_0}{\alpha_1+\alpha_0}\right)$$

represent the 'a priori' probabilities for the values of the variables based in our past experience.

The value $s=\alpha_0+\alpha_1$ is called the equivalent sample size measures the importance of our past experience. Larger values make that 'a priori' probabilities have more importance.

Uso Práctico

Valores	$X = 0$	$X = 1$
Frecuencias	N_0	N_1
Extra	α_0	α_1
Total	$N_0 + \alpha_0$	$N_1 + \alpha_1$
Probabilidades	$\frac{N_0+\alpha_0}{N_0+\alpha_0+N_1+\alpha_1}$	$\frac{N_1+\alpha_1}{N_0+\alpha_0+N_1+\alpha_1}$

Uso Práctico

Valores	$X = 0$	$X = 1$
Frecuencias	N_0	N_1
Extra	α_0	α_1
Total	$N_0 + \alpha_0$	$N_1 + \alpha_1$
Probabilidades	$\frac{N_0+\alpha_0}{N_0+\alpha_0+N_1+\alpha_1}$	$\frac{N_1+\alpha_1}{N_0+\alpha_0+N_1+\alpha_1}$

$N_0 = 5, N_1 = 2, \alpha_0 = \alpha_1 = 0{,}5$

Valores	$X = 0$	$X = 1$
Frecuencias	5	2
Extra	0.5	0.5
Total	5.5	2.5
Probabilidades	5.5/8	2.5/8

Uso Práctico

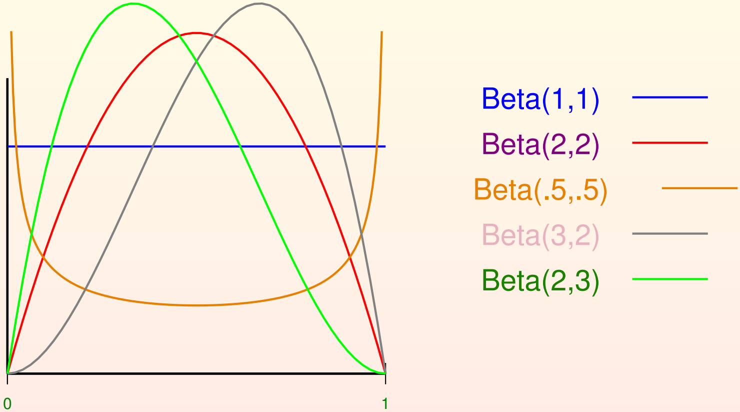
Valores	$X = 0$	$X = 1$
Frecuencias	N_0	N_1
Extra	α_0	α_1
Total	$N_0 + \alpha_0$	$N_1 + \alpha_1$
Probabilidades	$\frac{N_0+\alpha_0}{N_0+\alpha_0+N_1+\alpha_1}$	$\frac{N_1+\alpha_1}{N_0+\alpha_0+N_1+\alpha_1}$

$N_0 = 5, N_1 = 2, \alpha_0 = \alpha_1 = 2$

Valores	$X = 0$	$X = 1$
Frecuencias	5	2
Extra	2	2
Total	7	4
Probabilidades	7/11	4/11

Son más cercanos a la uniforme: 0.5 para todos los valores.

Examples



When $\alpha_0, \alpha_1 \to 0$, then we have maximum likelihood estimation.

The Multinomial Case

Now, assume that we have a variable X taking values on a finite set $\{a_1, \ldots, a_n\}$ and we have a series of independent observations of this distribution, $(x[1], x[2], \ldots, x[m])$ and we want to estimate the values $\theta_i = p(a_i), \quad i = 1, \ldots, n$.

Sufficient statistics are N_i, the number of cases in the sample in which we have obtained the value a_i ($i = 1, \ldots, n$).

The maximum likelihood estimation of θ_i is

$$\hat{\theta}_i = \frac{N_i}{m}$$

The problems with small samples are completely analogous.

Ejemplo: Máxima Verosimilitud

En un dado hemos tirado 10 veces y ha salido (1,3,2,5,5,6,1,2,1,6). Tenemos:

Valores	1	2	3	4	5	6
Frecuencias	3	2	1	0	2	2
Estimación	0.3	0.2	0.1	0.0	0.2	0.2

The Dirichlet 'a priori' distribution

We can also follow the Bayesian approach, but the 'a priori' distribution is the Dirichlet distribution, a generalization of the Beta distribution for more than 2 cases: $(\theta_1, \ldots, \theta_n)$.

The expression of $D(\alpha_1, \ldots, \alpha_n)$ is

$$f(\theta_1, \ldots, \theta_n) = \frac{\Gamma(s)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_n)} \prod_{i=1}^{n} \theta^{\alpha_i - 1}$$

where $s = \sum_{i=1}^{n} \alpha_i$ is the equivalent sample size.
The expected vector is

$$E(\theta_1, \ldots, \theta_n) = \left(\frac{\alpha_1}{s}, \ldots, \frac{\alpha_n}{s} \right)$$

Greater values of s makes this distribution more concentrated around the mean vector.

Dirichlet 'a posteriori'

If we have a set of data with counts $(N_1, \ldots, N_n)$, then the 'a posteriori' distribution is also Dirichlet with parameters

$$D(\alpha_1 + N_1, \ldots, \alpha_n + N_n)$$

The Bayesian estimation of probabilities are:

$$\left(\frac{\alpha_1 + N_1}{s+m}, \ldots, \frac{\alpha_n + N_n}{s+m}\right)$$

where $m = \sum_{i=1}^{n} N_i, \quad s = \sum_{i=1}^{n} \alpha_i$.

Example

Imagine that we have an urn with balls of different colors: red (R), blue (B), and green (G); but on an unknown quantity.

Assume that we picked up balls with replacement, with the following sequence: (B, B, R, R, B).

Máxima verosimilitud

Valores	R	B	G
Frecuencias	2	3	0
Estimación	0.4	0.6	0.0

Example

Imagine that we have an urn with balls of different colors: red (R), blue (B), and green (G); but on an unknown quantity.

Assume that we picked up balls with replacement, with the following sequence: (B, B, R, R, B).

Laplace or Dirichlet(1,1,1)

Valores	R	B	G
Frecuencias	2	3	0
Extra	1	1	1
Total	3	4	1
Estimación	3/8	0.5	1/8

Example

Imagine that we have an urn with balls of different colors: red (R), blue (B), and green (G); but on an unknown quantity.

Assume that we picked up balls with replacement, with the following sequence: (B, B, R, R, B).

Dirichlet(2,2,2)

Valores	R	B	G
Frecuencias	2	3	0
Extra	2	2	2
Total	4	5	2
Estimación	4/11	5/11	2/11

Ejemplo: Máxima Verosimilitud

En un dado hemos tirado 10 veces y ha salido (1,3,2,5,5,6,1,2,1,6). Tenemos:

Valores	1	2	3	4	5	6
Frecuencias	3	2	1	0	2	2
Estimación	0.3	0.2	0.1	0.0	0.2	0.2

Dirichlet(100,100,100,100,100,100)

Valores	1	2	3	4	5	6
Frecuencias	3	2	1	0	2	2
Extra	100	100	100	190	100	100
Total	103	102	101	100	102	2
Estimación	$\frac{103}{610}$	$\frac{102}{610}$	$\frac{101}{610}$	$\frac{100}{610}$	$\frac{102}{610}$	$\frac{102}{610}$

Cómo elegimos los α (valores extra)?

En un problema de aprendizaje donde no tenemos información previa de las variables, se suelen elegir unos valores con uno de los siguientes criterios:

- Se elige un valor constante para los $\alpha_i = t$ (estilo Laplace en el que $\alpha = 1$).
  El tamaño muestral equivalente es $S = tn$, donde n es el número de casos de la variable X
- Se elige S y se calcula $\alpha_i = S/n$ donde n es el número de casos de la variable X

A Bayesian Network

- The basic approach is to apply the Bayesian approach with Dirichlet ’a priori’ distributions for each conditional probability distribution.
- In each case, we should only consider the part of the database that is compatible with the values of the parents to which we are conditioning.
- We have to be conscious of the basic hypothesis allowing us to do it, and when it does not make sense to apply them.
- We can have problems when selecting the equivalent sample size of the Dirichlet distributions.

Parametrization

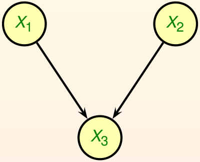
- For each variable X_i let $x_i^1, \ldots, x_i^{r_i}$ the set of possible values where r_i is the number of possible values.
- The number of configurations for the parents of X_i will be denoted by q_i. The configuration number j will be denoted by pa_j^i.
- The parameters necessary to specify a Bayesian network are

$$\theta_{ijk} = P(x_i^k | pa_j^i), \quad i = 1, \ldots, n, \quad j = 1, \ldots, q_i, \quad k = 1, \ldots, r_i$$

- $\boldsymbol{\theta_{ij}}$ will denote the vector of multinomial probabilities $(\theta_{ij1}, \ldots, \theta_{ijr_i})$.

Example

Assume the following network where all the variables are binary

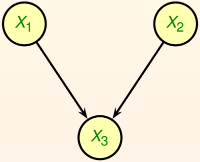


Example

Assume the following network where all the variables are binary

$\theta_{111} = P(X_1 = 0)$

$\theta_{112} = P(X_1 = 1)$



$\theta_{211} = P(X_2 = 0)$

$\theta_{212} = P(X_2 = 1)$

$$\theta_{311} = P(X_3 = 0|0,0), \quad \theta_{312} = P(X_3 = 1|0,0)$$
$$\theta_{321} = P(X_3 = 0|0,1), \quad \theta_{322} = P(X_3 = 1|0,1)$$
$$\theta_{331} = P(X_3 = 0|1,0), \quad \theta_{332} = P(X_3 = 1|1,0)$$
$$\theta_{341} = P(X_3 = 0|1,1), \quad \theta_{342} = P(X_3 = 1|1,1)$$

	1	2	3	4
	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$X_3 = 0$	θ_{311}	θ_{321}	θ_{331}	θ_{341}
$X_3 = 1$	θ_{312}	θ_{322}	θ_{332}	θ_{342}
	θ_{31}	θ_{32}	θ_{33}	θ_{34}

Parameter Independence

A basic hypothesis that is convenient and in some situations real is that the parameters distributions are independent.

$$p(\theta) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} p(\theta_{ij})$$

where θ denotes the vector of all the parameters.

Main Result

If we observe all the values of the variables in the sample, then the 'a posteriori' distributions of the parameters θ_{ij} are also independent.

$$P(\theta|D) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} P(\theta_{ij}|D)$$

The Consequence:

We can apply the Dirichlet model to each conditional probability on an independent way and update each of them in an independent way.

The Dirichlet Updating

- Under conditions of independence
- If the distribution of the parameters θ_{ij} is a Dirichlet $D(\alpha_{ij1},\ldots,\alpha_{ijr_i})$
- If in the database, there are N_{ijk} cases in which the variable X_i takes the value x_k^i, and the parents of this variable are in configuration pa_j^i
- Then the 'a posteriori' distributions for the parameters θ_{ij} are

$$D(\alpha_{ij1}+N_{ij1},\ldots,\alpha_{ijr_i}+N_{ijr_i})$$

En forma de tabla

	1	2	3	4
	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$X_3 = 0$	θ_{311}	θ_{321}	θ_{331}	θ_{341}
$X_3 = 1$	θ_{312}	θ_{322}	θ_{332}	θ_{342}
	$\theta_{\mathbf{31}}$	$\theta_{\mathbf{32}}$	$\theta_{\mathbf{33}}$	$\theta_{\mathbf{34}}$

Frecuencias:

	1	2	3	4
	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$X_3 = 0$	N_{311}	N_{321}	N_{331}	N_{341}
$X_3 = 1$	N_{312}	N_{322}	N_{332}	N_{342}
	N_{31}	N_{32}	N_{33}	N_{34}

$N_{ei} = N_{3i1} + N_{3i2}$ y Tamaño muestral total $N = N_{31} + N_{32} + N_{33} + N_{34}$

Estimación Máxima verosimilitud

Frecuencias:

	1	2	3	4
	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$X_3 = 0$	N_{311}	N_{321}	N_{331}	N_{341}
$X_3 = 1$	N_{312}	N_{322}	N_{332}	N_{342}
	N_{31}	N_{32}	N_{33}	N_{34}

Estimación Condicional

	1	2	3	4
	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$X_3 = 0$	$\frac{N_{311}}{N_{31}}$	$\frac{N_{321}}{N_{32}}$	$\frac{N_{331}}{N_{33}}$	$\frac{N_{341}}{N_{34}}$
$X_3 = 1$	$\frac{N_{311}}{N_{31}}$	$\frac{N_{321}}{N_{32}}$	$\frac{N_{332}}{N_{33}}$	$\frac{N_{341}}{N_{34}}$
	N_{31}	N_{32}	N_{33}	N_{34}

Estimación conjunta

No hay que confundir la estimación condicional:

	1	2	3	4
	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$X_3 = 0$	$\frac{N_{311}}{N_{31}}$	$\frac{N_{321}}{N_{32}}$	$\frac{N_{331}}{N_{33}}$	$\frac{N_{341}}{N_{34}}$
$X_3 = 1$	$\frac{N_{311}}{N_{31}}$	$\frac{N_{321}}{N_{32}}$	$\frac{N_{332}}{N_{33}}$	$\frac{N_{341}}{N_{34}}$
	N_{31}	N_{32}	N_{33}	N_{34}

Con la estimación de la probabilidad conjunta en la que hay que dividir por el total N

	1	2	3	4
	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$X_3 = 0$	$\frac{N_{311}}{N}$	$\frac{N_{321}}{N}$	$\frac{N_{331}}{N}$	$\frac{N_{341}}{N}$
$X_3 = 1$	$\frac{N_{311}}{N}$	$\frac{N_{321}}{N}$	$\frac{N_{332}}{N}$	$\frac{N_{341}}{N}$
	N_{31}	N_{32}	N_{33}	N_{34}

Estimación Bayesiana

Frecuencias y parámetros α_{ijk} (extra):

	1	2	3	4
	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$X_3 = 0$	$N_{311} + \alpha_{311}$	$N_{321} + \alpha_{321}$	$N_{331} + \alpha_{331}$	$N_{341} + \alpha_{341}$
$X_3 = 1$	$N_{312} + \alpha_{312}$	$N_{322} + \alpha_{322}$	$N_{332} + \alpha_{332}$	$N_{342} + \alpha_{342}$
	$N_{31} + \alpha_{31}$	$N_{32} + \alpha_{32}$	$N_{33} + \alpha_{33}$	$N_{34} + \alpha_{34}$

$\alpha_{3j} = \alpha_{3j1} + \alpha_{3j2}$. Estimación Condicional

	1	2	3	4
	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$X_3 = 0$	$\frac{N_{311}+\alpha_{311}}{N_{31}+\alpha_{31}}$	$\frac{N_{321}+\alpha_{321}}{N_{32}+\alpha_{32}}$	$\frac{N_{331}+\alpha_{331}}{N_{33}+\alpha_{33}}$	$\frac{N_{341}+\alpha_{341}}{N_{34}+\alpha_{34}}$
$X_3 = 1$	$\frac{N_{311}+\alpha_{312}}{N_{31}+\alpha_{31}}$	$\frac{N_{321}+\alpha_{322}}{N_{32}+\alpha_{32}}$	$\frac{N_{332}+\alpha_{332}}{N_{33}+\alpha_{33}}$	$\frac{N_{342}+\alpha_{342}}{N_{34}+\alpha_{34}}$
	$N_{31} + \alpha_{31}$	$N_{32} + \alpha_{32}$	$N_{33} + \alpha_{33}$	$N_{34} + \alpha_{34}$

Determinación de los α_{ijk}

hay dos enfoques:

- Seleccinamos el mismo valor para todos $\alpha_{ijk} = t$ (estilo Laplace en el que $t = 1$).
- Seleccionados un tamaño muestral equivalente S y calculamos $\alpha_{ijk} = (q_i r_i)$ donde q_i es el número de configuraciones de los padres y r_i el número de casos de la variable: $q_i.r_i$ es el tamño de la tabla.

Example

Assume two variables:
X_1: Smoking; values 1 (yes) 2 (no)
X_2: Lung Cancer; values 1 (yes) 2 (no)
Imagine the following sequence of observations and network:

	$c,1$	$c,2$	$c,3$	$c,4$	$c,5$	$c,6$	$c,7$	$c,8$
X_1	1	1	2	2	2	2	1	2
X_2	2	1	1	2	1	1	2	2

Assuming that all the parameters $\alpha_{ijk} = 1$ (Laplace) the estimations are:

$$\theta^*_{111} = 0{,}4, \quad \theta^*_{121} = 0{,}6,$$
$$\theta^*_{211} = 0{,}4, \quad \theta^*_{212} = 0{,}6, \quad \theta^*_{221} = 4/7, \quad \theta^*_{222} = 3/7$$

Equivalent Sample Sizes

If we compute the probabilities of X_2 before the data, we have $P(X_2 = 1) = 0{,}5$.

However, after the data

$$P(X_2 = 1) = P(X_2 = 1|X_1 = 1).P(X_1 = 1) + P(X_2 = 1|X_1 = 2).P(X_1 = 2) =$$

The question is: Is there anything in the data increasing the probability of Lunch cancer?

In the data exactly half of the cases have lunch cancer.

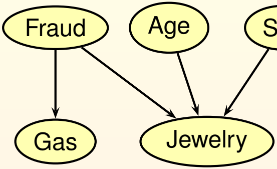
The problem is with the equivalent sample size

Método del tamaño muestral equivalente y consistencia

Si elegimos un tamaño muestral equivalente $S = 2$ y elegimos los alfas dividiendo este valor por el tamaño de la tabla, entonces:

- Diferentes probabilidades condicionadastendrán distintos valores de alfas: cuanto más grande es la tabla, menores son los alfas.
- El problema de la inconsistencia del ejemplo anterior desaparece
- Se considera que es la opción mejor justificada desde el punto de vista teórico, pero puede tener problemas desde el punto de vista práctico para una variable con muchos padres (alfas muy pequeños y estimación muy paracida a máxima verosimilitud).

Equivalent Sample Sizes



If the equivalent sample size (the sum of the parameters of the Dirichlet distribution) represents the strength of our past experience, Can we claim the same same experience when estimating the probabilities of sex that when estimating the probabilities of using credit car in a jewelry in a fraudulent way, for males above 50?

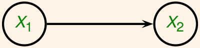
This is a difficult problem.

Exercise

Assume that we throw a thumbtack (X_1) with results: heads (1) and tails (2) and that depending of the result, we pick up a ball from one of two urns with balls of different colors.

Let X_2 the color of the ball, which can be Red (R), Blue (B) and Green (G).

Assume the network:



Determine appropriate 'a priori' Dirichlet distributions

Obtain Bayesian estimations under the following sequence of observations:

X_1	1	2	1	1	2	1	1	1	2	2
X_2	R	G	R	B	B	R	G	R	G	G

Modelos

- Estimar tablas de probabilidad puede tener el problema de presentar un elevado número de parámetros: es exponencial en el número de padres.
  $P(CompraJoya|Fraude = SI, Edad = 50, Hombre)$ no tiene nada que ver con
  $P(CompraJoya|Fraude = SI, Edad = 50, Mujer)$
- Un modelo básico es el de la puerta NOISY-OR. Si hay n padres, hay n parámetros $p_i, i = 1, \ldots, n$ que es la probabilidad de que ess padre sólo produzca el efecto y una probabilidad p_0 de que el efecto se produzca sin que ningún padre esté activo. Si tenemos que una combinación de padres está activo, entonces la probabilidad de que se produzca el efecto es
  $1 - (1 - p_0) \prod_{Padresactivos} \frac{1-p_i}{1-p_0}$
  Este modelo tiene muchos menos parámetros y siempre se puede optimizar por máxima verosimilitud.

Regresión logística

- Es un modelo que sirve para una probabilidad condicionada de una variable discreta a una continua. Pero también se puede usar para dependencias entre variables discretas. Si Y es una variable y $Z_1, \ldots, Z_k$ son sus padres, supongamos que todas las variables son binarias 0-1, entonces se supone que la probabilidad

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 Z_1 + \cdots + \beta_k Z_k)}}$$

- El número de parámetros es $k + 1$
- Se puede estimar por máxima verosimilitud
- Se puede extender a variables multinomiales (no binarias) de forma sencilla.
- Se pueden mezclar variables continuas y discretas al mismo tiempo

Missing Data

Serafín Moral
Dpt. Computer Science and Artificial Intelligence
University of Granada, Spain

Outline

- Missing Data
  - Simple procedures
  - EM Learning for estimating parameters
  - Score Approximation
- Variational Approaches
  - Equal Size
  - Equal Frequency
  - Meaningful Intervals

Missing Data

This occurs very often in databases.
It can be the case in which a variable is never observed. This is called a Hidden Variable.

A basic property is the MAR (Missing at Random) property: *the fact that one of several variables are missing is independent of the values of the missing variables given the observed ones*.

This is not always true *a person does not reply an inquiry because is fearful of saying her thinking*.

But this hypothesis is also usually done when using Bayesian networks and introducing observations.

Dependence on Missing

If we have dependence on missing values, statistical techniques are more sophisticated, but there are also another approaches we can take, especially if future cases follow the same pattern for non observed variables.

We could add a new case to each variable saying that it has not been observed and then run algorithms for complete data.

X_1	1	?	1	1	2	?	1
X_2	R	G	R	?	B	R	G
X_3	W	?	T	?	T	W	W

X_1	1	ND	1	1	2	ND	1
X_2	R	G	R	ND	B	R	G
X_3	W	ND	T	ND	T	W	W

Dependence

Another possibility is to add variables saying if the initial variables have been observed.

In these conditions the MAR condition is always verified (if we know whether a variable has been observed, then we do not get any information about the fact that this variable is observed). It is difficult to learn.

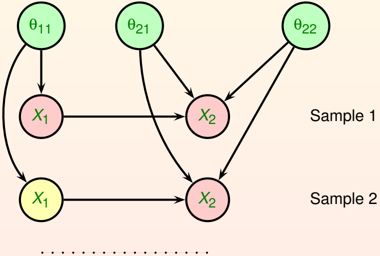
X_1	1	?	1	1	2	?	1
X_2	*R*	*G*	*R*	?	*B*	*R*	*G*
X_3	*W*	?	*T*	?	*T*	*W*	*W*

X_1	1	?	1	1	2	?	1
X_2	*R*	*G*	*R*	?	*B*	*R*	*G*
X_3	*W*	?	*T*	?	*T*	*W*	*W*
OX_1	*Y*	*N*	*Y*	*Y*	*Y*	*N*	1
OX_2	*Y*	*Y*	*Y*	*N*	*Y*	*Y*	*Y*
OX_3	*Y*	*N*	*Y*	*N*	*Y*	*Y*	*Y*

MAR Hypothesis

Under the MAR hypothesis the main problem is the 'a posteriori' distribution for the parameters are not independent (and not Dirichlet).

Remember the parameter model for two variables:



If X_1 is not observed parameters become dependent.

$P(x_2) = \sum_{x_1} P(x_2|x_1).P(x_1)$.

Simple Approaches

- Not to use the missing cases. If one case has a missing value we could not to use any part of it or to use the always the useful parts: if X_1 and X_2 are observed we would use its values to estimate $p(x_2|x_1)$ even if X_3 is not observed.
- Complete the missing cases in some way. For example, with the most probable case of this variable; or looking for a variable with a high degree of dependence and then estimating the missing value according to the observed value in that variable; or in a classification problem with the most probable value in the corresponding class.

The EM Algorithm

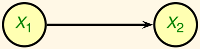
The EM algorithm complete the data in a more sophisticated way. It was proposed by Dempster et al. (1977) as a method for likelihood maximization (local maximum).

Intuition: It starts with some arbitrary estimation of the parameters (former procedures)

- Obtain a completion of the data with the current model (Expectation Step). Usually, the data are not complete with a single value, but a probability is assigned to each one of the possible values.
- Get a new estimation of the parameters given the complete data (Maximization Step)

It can be used for other cases as maximum 'a posteriori' or expected value of parameters.

Example



Assume that X_1 and X_2 are binary. The 'a priori' distribution for X_1 is $D(2,2)$ and the 'a priori' distributions for the conditional distributions are $D(1,1)$.

If we have the observations:

Case	1	2	3	4	5
X_1	1	1	1	1	2
X_2	1	?	1	2	?

Example

From the initial model, we complete the cases (expectation step):

Case	1	2	2	3	4	5	5
X_1	1	1	1	1	1	2	2
X_2	1	1	2	1	2	1	2
Numb.	1	1/2	1/2	1	1	1/2	1/2

Then, with these values we can estimate the probabilities of parameters (propagation algorithms):

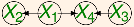
$$P^*(X_1 = 1) = 6/9, \quad P^*(x_1 = 2) = 3/9$$
$$P^*(X_2 = 1|X_1 = 1) = 3{,}5/6, \quad P^*(X_2 = 2|X_1 = 1) = 2{,}5/6$$
$$P^*(X_2 = 1|X_1 = 2) = 1{,}5/3, \quad P^*(X_2 = 2|X_1 = 2) = 1{,}5/3$$

We can complete the data again with these probabilities, and continue with this until there is no changes.

Algoritmo EM

Datos incompletos: Algoritmo EM

Partimos de una red completa (con parámetros) y de los datos

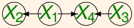


D				
Caso	X_1	X_2	X_3	X_4
1	1	0	0	?
2	0	1	0	1
3	?	1	1	?
4	0	?	0	0
5	1	0	0	1
6	1	1	0	1

Algoritmo EM

Datos incompletos: Algoritmo EM

Partimos de una red completa (con parámetros) y de los datos



D				
Caso	X_1	X_2	X_3	X_4
1	1	0	0	?
2	0	1	0	1
3	?	1	1	?
4	0	?	0	0
5	1	0	0	1
6	1	1	0	1

Mediante un algoritmo de inferencia calculamos

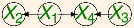
$P(X_4|X_1=1,X_2=0,X_3=0)$ (0,6,0,4),
$P(X_1,X_4|X_2=1,X_3=1)$ (0,08,0,02,0,72,0,18),

$P(X_2|X_1=0,X_3=0,X_4=0)$ (0,3,0,7)

Algoritmo EM

Datos incompletos: Algoritmo EM

Partimos de una red completa (con parámetros) y de los datos



D

Caso	X_1	X_2	X_3	X_4
1	1	0	0	?
2	0	1	0	1
3	?	1	1	?
4	0	?	0	0
5	1	0	0	1
6	1	1	0	1

D

Caso	X_1	X_2	X_3	X_4	peso
1	1	0	0	0	0.6
1	1	0	0	1	0.4
2	0	1	0	1	1
3	0	1	1	0	0.08
3	0	1	1	1	0.02
3	1	1	1	0	0.72
3	1	1	1	1	0.18
4	0	0	0	0	0.3
4	0	1	0	0	0.7
5	1	0	0	1	1
6	1	1	0	1	1

Mediante un algoritmo de inferencia calculamos

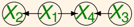
$P(X_4|X_1=1,X_2=0,X_3=0)$ (0,6,0,4),
$P(X_1,X_4|X_2=1,X_3=1)$ (0,08,0,02,0,72,0,18),

$P(X_2|X_1=0,X_3=0,X_4=0)$ (0,3,0,7)

Algoritmo EM

Datos incompletos: Algoritmo EM

Partimos de una red completa (con parámetros) y de los datos



D				
Caso	X_1	X_2	X_3	X_4
1	1	0	0	?
2	0	1	0	1
3	?	1	1	?
4	0	?	0	0
5	1	0	0	1
6	1	1	0	1

Mediante un algoritmo de inferencia calculamos

$P(X_4|X_1 = 1, X_2 = 0, X_3 = 0)$ $(0{,}6, 0{,}4)$,
$P(X_1, X_4|X_2 = 1, X_3 = 1)$ $(0{,}08, 0{,}02, 0{,}72, 0{,}18)$,

$P(X_2|X_1 = 0, X_3 = 0, X_4 = 0)$ $(0{,}3, 0{,}7)$

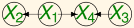
Los datos “completados” son:

D					
Caso	X_1	X_2	X_3	X_4	peso
1	1	0	0	0	0.6
1	1	0	0	1	0.4
2	0	1	0	1	1
3	0	1	1	0	0.08
3	0	1	1	1	0.02
3	1	1	1	0	0.72
3	1	1	1	1	0.18
4	0	0	0	0	0.3
4	0	1	0	0	0.7
5	1	0	0	1	1
6	1	1	0	1	1

Algoritmo EM

Datos incompletos: Algoritmo EM

Partimos de una red completa (con parámetros) y de los datos



		D		
Caso	X_1	X_2	X_3	X_4
1	1	0	0	?
2	0	1	0	1
3	?	1	1	?
4	0	?	0	0
5	1	0	0	1
6	1	1	0	1

Mediante un algoritmo de inferencia calculamos

$P(X_4|X_1=1,X_2=0,X_3=0)$ $(0{,}6,0{,}4)$,
$P(X_1,X_4|X_2=1,X_3=1)$ $(0{,}08,0{,}02,0{,}72,0{,}18)$,

$P(X_2|X_1=0,X_3=0,X_4=0)$ $(0{,}3,0{,}7)$

Los datos “completados” son:

			D		
Caso	X_1	X_2	X_3	X_4	peso
1	1	0	0	0	0.6
1	1	0	0	1	0.4
2	0	1	0	1	1
3	0	1	1	0	0.08
3	0	1	1	1	0.02
3	1	1	1	0	0.72
3	1	1	1	1	0.18
4	0	0	0	0	0.3
4	0	1	0	0	0.7
5	1	0	0	1	1
6	1	1	0	1	1

A partir de esos “datos” calculamos las “frecuencias”

X_1	X_2	frec.
0	0	0.3
0	1	1.8
1	0	2
1	1	1.9

X_1	frec.
0	2.1
1	3.9

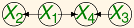
X_3	frec.
0	5
1	1

X_1	X_3	X_4	frec.
0	0	0	1
0	0	1	1
0	1	0	0.08
0	1	1	0.02
1	0	0	0.6
1	0	1	1.4
1	1	0	0.72
1	1	1	1.18

Algoritmo EM

Datos incompletos: Algoritmo EM

Partimos de una red completa (con parámetros) y de los datos



D				
Caso	X_1	X_2	X_3	X_4
1	1	0	0	?
2	0	1	0	1
3	?	1	1	?
4	0	?	0	0
5	1	0	0	1
6	1	1	0	1

Mediante un algoritmo de inferencia calculamos

$P(X_4|X_1=1,X_2=0,X_3=0)$ (0,6,0,4),

$P(X_1,X_4|X_2=1,X_3=1)$ (0,08,0,02,0,72,0,18),

$P(X_2|X_1=0,X_3=0,X_4=0)$ (0,3,0,7)

Los datos “completados” son:

D					
Caso	X_1	X_2	X_3	X_4	peso
1	1	0	0	0	0.6
1	1	0	0	1	0.4
2	0	1	0	1	1
3	0	1	1	0	0.08
3	0	1	1	1	0.02
3	1	1	1	0	0.72
3	1	1	1	1	0.18
4	0	0	0	0	0.3
4	0	1	0	0	0.7
5	1	0	0	1	1
6	1	1	0	1	1

A partir de esos “datos” calculamos las “frecuencias”

X_1	X_2	frec.
0	0	0.3
0	1	1.8
1	0	2
1	1	1.9

X_1	frec.
0	2.1
1	3.9

X_3	frec.
0	5
1	1

X_1	X_3	X_4	frec.
0	0	0	1
0	0	1	1
0	1	0	0.08
0	1	1	0.02
1	0	0	0.6
1	0	1	1.4
1	1	0	0.72
1	1	1	1.18

Reestimamos los parámetros

Iteramos

Métodos Variacionales

- Los métodos variacionales siguen un enfoque bayesiano.
- El problema es que ahora las probabilidades ’a posteriori’ de los parámetros no son fáciles de calcular y no son independientes.
- Supongamos que tenemos un conjunto de parámetros θ y que hemos observado en D las variables $\mathbf{X}$, mientras que las variables $\mathbf{Z}$ están ocultas.

  Entonces la verosimilitud:

$$L(\theta) = P(D|\theta) = \int_z P(z, D|\theta) dz$$

Métodos Variacionales

- En lugar de calcular de forma exacta la distribución a posteriori de los parámetros esta se aproxima con la mejor distribución de una famila $Q(z,\theta) \in \mathcal{Q}$
- Se usa la divergencia de Kuback-Leibler como criterio:

$$Q^*(z,\theta) = \arg \min_{Q\in\mathcal{Q}} KL(Q(z,\theta), P(z,\theta|D))$$

- Tenemos

$$KL(Q(z,\theta), P(z,\theta|D)) = E_Q[log(Q)] - E_Q[\log(P(z,\theta|D))] =$$

$$E_Q[\log(Q)] - E_Q[\log(P(z,\theta,D))] + \log(P(D))$$

Métodos Variacionales

- $KL(Q(z,\theta), P(z,\theta|D) = E_Q[\log(Q)] - E_Q[\log(P(z,\theta,D))] + \log(P(D))$
- Luego minimizar la divergencia es equivalente a maximizar:

$$-E_Q[\log(Q)] + E_Q[\log(P(z,\theta,D))]$$

- Como la divergencia es siempre mayor o igual a 0, este valor es siempre menor o igual a $\log(P(D))$ (el logaritmo de la probabilidad de los datos) y por eso se conoce como ELBO (*Evidence Lower Bound*).

Métodos De Campo Medio

- En este caso se usa una familia $\mathcal{Q}$ en la que todas las variables unidimensionales son independientes: por ejemplo si hay dos parámetros (θ_1, θ_2) se supone que
$$Q(\theta_1, \theta_2, z) = Q(\theta_1)Q(\theta_2)Q(z)$$

- En el caso de redes bayesianas discretas, si las distribuciones 'a priori' de los parámetros son Dirichlet, existe un método iterativo para minimizar el ELBO que es muy rápido.

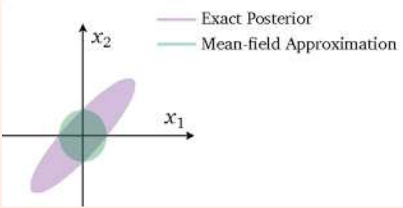
Como Maximización EM

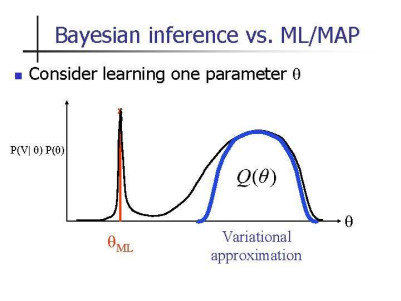
Como Esperanza en EM

$$\begin{aligned} Q(\theta_1) &\propto E_{Q(\theta_2),Q(z)}[\log P(\theta_1, \theta_2, z, D)] \\ Q(\theta_2) &\propto E_{Q(\theta_1),Q(z)}[\log P(\theta_1, \theta_2, z, D)] \qquad (1) \\ Q(z) &\propto E_{Q(\theta_1),Q(\theta_2)}[\log P(\theta_1, \theta_2, z, D)] \end{aligned}$$

Aproximaciones Variacionales

No aproximan correlaciones (por ser de campo medio) y se concentran en un máximo local (por el orden la divergencia).





Bayesian inference vs. ML/MAP

- Consider learning one parameter θ

Pero son muy rápidos y efectivos