Curso de Aprendizaje Automático
para el INE

Conceptos generales de aprendizaje bayesiano y simulación Montecarlo

David Rios Insua
AXA-ICMAT Chair y Real Academia de Ciencias
18 y 22 de Marzo, INE

ICMAT INSTITUTO DE CIENCIAS MATEMÁTICAS

Agenda

- **Conceptos generales de aprendizaje bayesiano 18 y 22 de Marzo**
- Conceptos generales de aprendizaje supervisado
- Conceptos generales sobre aprendizaje supervisado

Agenda

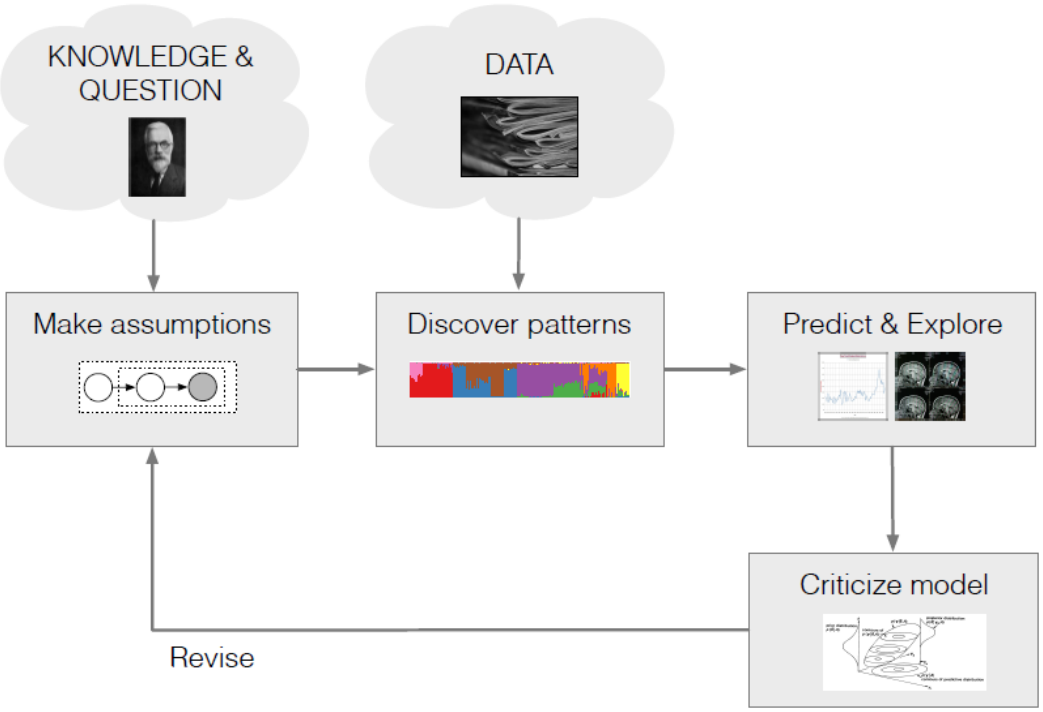
- Aproximaciones a Inferencia en Aprendizaje Automático
- Conceptos básicos de aprendizaje bayesiano (incluyendo asignación de probabilidades subjetivas)
- Montecarlo (incluyendo métodos de cadena de Markov)
- Ejemplos de aprendizaje bayesiano
- Introducción a aprendizaje bayesiano en problemas de grandes dimensiones
- Referencias a estadística oficial
- Ejemplos computacionales

Inferencia en AA

Incertidumbre casi ubicua en AA:

- Dada cierta transacción con determinadas características, es fraudulenta o no?
- Dada la traza de monitorización de cierto dispositivo en Internet, estamos asistiendo a un ataque?
- Se corresponde esta imagen médica a la de una persona con cierta enfermedad?
- Se corresponde este manuscrito a alguna de estas personas?
- Si mi sistema realiza esta acción? Cómo responderá un usuario? Y el entorno? En consecuencia, qué acción debe realizar?

Inferencia en AA



Hipótesis Computación Aplicación

Inferencia

Qué dice este modelo sobre estos datos?

General, Escalable

Inferencia en AA

- Modelo probabilístico de variables observadas **x** y variables latentes (incluye parámetros) **z** $p(\mathbf{z}, \mathbf{x})$

- Estimación máxima verosimilitud $\mathbf{z}^\star = \arg\max_{\mathbf{z}} p(\mathbf{x}|\mathbf{z})$

- Estimación MAP $\mathbf{z}^\star = \arg\max_{\mathbf{z}} p(\mathbf{z}|\mathbf{x}) = \arg\max_{\mathbf{z}} p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$

- Estimación bayesiana $p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{\int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}}$

  Incorpora información a priori

  Estima toda la distribución

  El denominador (la evidencia) a menudo es intratable ---→ MCMC, Inferencia aproximada

Inferencia clásica

- Una vez fijado el modelo sobre el que queremos aprender (sus parámetros)
- Suponemos fijo el parámetro
- Dados los datos, formulamos la verosimilitud
- Maximizamos la verosimilitud para encontrar el estimador de máxima verosimilitud (EML, MLE) tal vez regularizado
  - Mínimos cuadrados + regularizador
- Resolvemos el problema

Inferencia bayesiana

- Una vez fijado el modelo sobre el que queremos aprender (sus parámetros)
- Suponemos incierto el parámetro, recogemos la información a priori
- Dados los datos, formulamos la verosimilitud
- Combinamos la verosimilitud y la información a priori para calcular la distribución a posteriori
- Resolvemos el problema

Inferencia bayesiana

- La distribución a posteriori resume toda la información disponible para resolver los problemas estándar de inferencia (ciencia):
  - **Estimación puntual**
  - **Estimación por intervalos**
  - **Contraste de hipótesis**
- También es clave para resolver dos problemas principales en ingeniería, negocios y administración:
  - **Predicción**. Distribución predictiva (predicción puntual, intervalo predictivo, contraste de hipótesis predictiva)
  - **Apoyo a la toma de decisiones**. Maximizando utilidad esperada a posteriori (o predictiva).

Un ejemplo típico

Consideramos un protocolo de recuperación de un servicio informático ante un cibertataque. Introducimos un procedimiento y deseamos evaluar su eficacia en cierta forma (para compararlo, p.ej., con otro protocolo).

Hemos probado el protocolo ante 12 ataques y en 9 de ellos ha resultado efectivo (p.ej. la duración de la caída es menor de 1 hora)

Comenzamos formulando un modelo

Un ejemplo típico

- Número X de éxitos en n ensayos (idénticos, independientes)
- Probabilidad de éxito en un ensayo θ_1
- Distribución del número de éxitos en n ensayos con probabilidad de éxito

$$X|\theta_1 \sim Bin(12, \theta_1)$$

- Para X=9,

$$Pr(X = 9|\theta_1) \propto \theta_1^9(1-\theta_1)^3,\ \theta_1 \in [0,1]$$

Un ejemplo típico

- Número X de éxitos en n ensayos (idénticos, independientes)
- Probabilidad de éxito en un ensayo θ_1
- Distribución del número de éxitos en n ensayos con probabilidad de éxito $X|\theta_1 \sim Bin(12, \theta_1)$
- Para X=9,

Verosimilitud $$Pr(X = 9|\theta_1) \propto \theta_1^9(1-\theta_1)^3,\ \theta_1 \in [0,1]$$

Una primera aproximación intenta encontrar el valor del parámetro que maximiza la probabilidad anterior. ***Estimador de máxima verosimilitud***

En nuestro caso, el EMV es 9/12

Pero el EMV tiene varios defectos ….

Un ejemplo típico

- Podemos utilizar otra fuente de información relativa al parámetro de interés. La información a priori. Pongamos que en este caso

$$p(\theta_1) = 1, \theta_1 \in [0,1].$$

- Se actualiza mediante la fórmula de Bayes, para obtener la distribución a posteriori

$$p(\theta_1|x=9) \propto p(\theta_1) \times Pr(X=9|\theta_1) \propto \theta_1^9(1-\theta_1)^3, \ \theta_1 \in [0,1]$$

que resume toda la información disponible sobre el parámetro que se corresponde a una distribución

$$\text{Beta } (10,4)$$

http://en.wikipedia.org/wiki/Beta_distribution

Un ejemplo típico

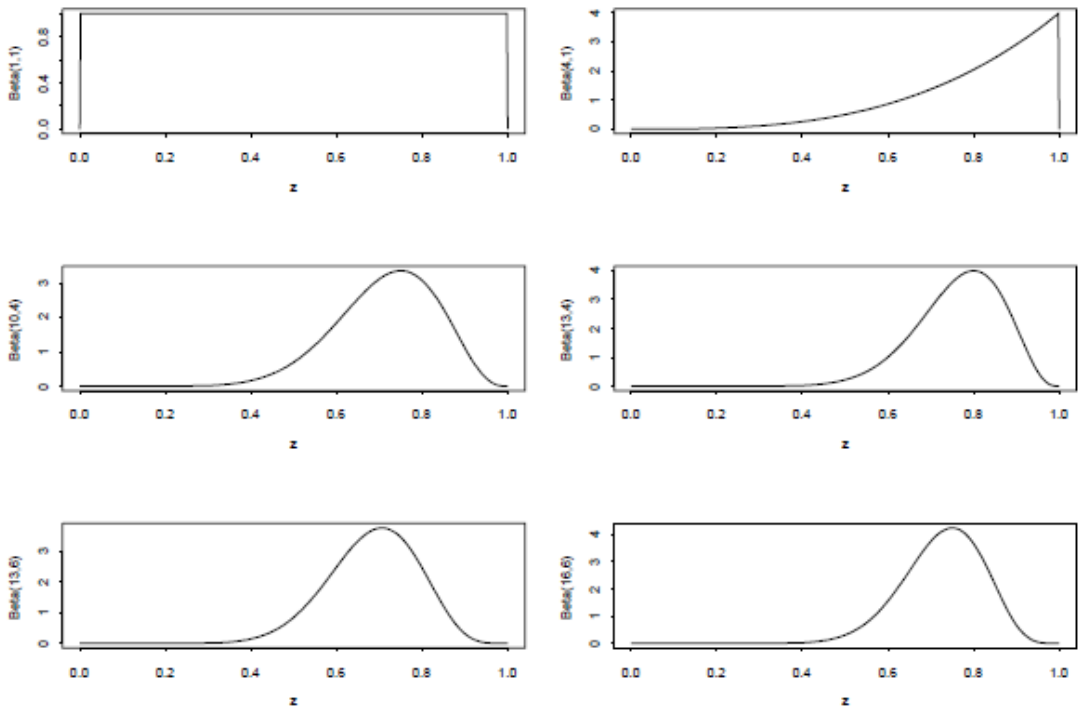
La distribución a posteriori sirve como distribución a priori para estudios siguientes. Por ejemplo, si en las siguientes 5 aplicaciones hay 3 éxitos la nueva distribución a posteriori es

$$p(\theta_1|x=3) \propto [\theta_1^0(1-\theta_1)^3] \times [\theta_1^3(1-\theta_1)^2] = \theta_1^{12}(1-\theta_1)^5,\ \theta_1 \in [0,1] \qquad \text{Beta } (13,6)$$

Si a priori sabemos que está en torno al 80% y mayores valores son más verosímiles, tendríamos el aprendizaje

$$\text{Beta } (4,1) \longrightarrow \text{Beta}(13,4) \longrightarrow \text{Beta}(16,6),$$

Un ejemplo típico



Convergencia en el aprendizaje, Consenso, Consistencia, Comportamiento asintótico

Un ejemplo típico

- Caso Beta (10,4)

Estimador puntual. Resumir en un valor, e.g. la media a posteriori

$$\frac{10}{10+4} = 0.72$$

Estimador por intervalos. Resumir en un intervalo, e.g. un intervalo de probabilidad 0.90

$$[0.505, .887]$$

Predicciones. P.ej., probabilidad de que haya más de 4 éxitos en 7 ensayos

$$\begin{aligned} Pr(X = k|x = 9) &= \int Pr(X = k|\theta_1)p(\theta_1|x = 9)d\theta_1 = \\ &= \int \binom{7}{k} \theta_1^k (1-\theta_1)^{7-k} \binom{13}{3} \theta_1^9 (1-\theta_1)^3 d\theta_1 = \\ &= \frac{\binom{7}{k}\binom{13}{3}}{\binom{20}{9+k}}. \end{aligned}$$

$$Pr(X \geq 5|x = 9) = \sum_{k=5}^{7} Pr(X = k|x = 9) = 0.6641.$$

Un ejemplo típico

Consideremos ahora un segundo protocolo de seguridad que se ha aplicado en 10 ocasiones, habiendo sido exitoso en 6 de ellas. Si es θ_2 la probabilidad de éxito con el segundo protocolo, tenemos

$$X|\theta_1 \sim Bin(12, \theta_1)$$

$$Y|\theta_2 \sim Bin(10, \theta_2)$$

$$\theta_1, \theta_2 \sim Unif[0, 1]$$ independientes

Queremos calcular

$$r = Pr(\theta_1 \geq \theta_2 | x = 9, y = 6)$$

Un ejemplo típico

$$\theta_1 \sim Beta(10, 4),\ \theta_2 \sim Beta(7, 5)$$

- Distribución de $\theta_1 - \theta_2$?
- Por simulación. P.ej. simulamos 1000 observaciones de las a posteriori, calculamos las diferencias, contamos cuántas son mayores que 0, dividimos por 1000.

$$r \approx 0.772.$$

- Cuál de los dos protocolos es mejor?

Un ejemplo típico

- Estructura de utilidades

	succeeds	does not succeed
Plan A	0.8	0
Plan B	1	0.2

- Utilidades esperadas dadas las probabilidades

$$0.8\theta_1 + 0(1-\theta_1) = 0.8\theta_1 \qquad \theta_2 + 0.2(1-\theta_2) = 0.2 + 0.8\theta_2$$

- Utilidades esperadas

$$0.8E(\theta_1|x=9) = 0.8 \times \tfrac{10}{14} = \tfrac{4}{7} \qquad 0.2 + 0.8E(\theta_2|y=6) = 0.2 + 0.8 \times \tfrac{7}{12} = \tfrac{2}{3}.$$

Aprendizaje bayesiano

1. Incertidumbre sobre parámetros, medida probabilísticamente
2. Probabilidad subjetiva (Estadística oficial…)
3. Actualización de la información mediante la fórmula de Bayes
4. Problemas de inferencia como problemas de decisión

Incertidumbre

- ***La incertidumbre es la falta de conocimiento completo de lo que es o lo que puede ocurrir***.
- Es casi inherente en nuestras vidas. Considerad los ejemplos siguientes
  - Fumar produce cáncer
  - Madrid será sede olímpica en 2036
  - México alcanzó su independencia en 1826
  - Peso más de 90 kilos
  - Aníbal cruzó los Alpes por San Gotardo

Incertidumbre

Debido a la incertidumbre, no conocerás qué consecuencias tendrás después de tomar tus decisiones. Considerad los siguientes ejemplos relevantes en distintos problemas de toma de decisiones

- El precio del petróleo Brent al final del año superará los 80 dólares/barril
- La economía española entrará en recesión al final del 2020
- La demanda de agua en Madrid en el año 2035 será superior a….
- No se levantará temporal en la costa de Ferrol en los próximos dos días

Incertidumbre

Usaremos probabilidades para cuantificar la incertidumbre. Hay tres conceptos básicos de probabilidad

- *Concepto clásico*. En una situación con n casos igualmente verosímiles, por razones físicas o lógicas, la probabilidad de un suceso se define mediante

  Casos favorables/Casos posibles

- *Concepto frecuentista*. En una situación en la que es posible repetir un experimento indefinidamente en idénticas condiciones, la probabilidad de un suceso se define mediante

  Límite de frecuencia relativa de aparición del suceso

  Operacionalmente, frecuencia relativa en un número grande pruebas

Incertidumbre

- El tercer concepto básico de probabilidad es
  - ***Concepto subjetivo***. La probabilidad de un suceso es

una medida del grado de creencia en la verdad de una proposición.

Incertidumbre

En general, al ir a modelizar incertidumbre, las cuestiones a responder serán

- Cuáles son las incertidumbres clave?
- Cuáles son los posibles resultados de esas incertidumbres?
- Cuáles son las probabilidades de ocurrencia de los posibles resultados?
- Cuáles son las consecuencias asociadas a cada resultado?

Estructura de problemas: Diagrama de influencia

- Grafo acíclico dirigido que modeliza un problema de toma de decisiones con

tres tipos de Nodos

  - Decisión. Cuadrado
  - Azar. Círculo
  - Valor. Hexágono (a veces aparecen como diamantes)

tres tipos de Arcos

  - A un nodo de decisión. La decisión se toma conociendo valor del antecesor
  - A un nodo de azar. La incertidumbre depende del antecesor
  - A un nodo de valor. La utilidad depende del antecesor

Diagramas de Influencia. Trozos

Diagramas de Influencia. Trozos

Tomamos la decisión 2 conociendo cómo se ha resuelto la incertidumbre 1

Tomamos la decisión 6, sabiendo qué decisión hemos tomado en la decisión 5

Las probabilidades sobre los resultados de la incertidumbre 4 dependen de los resultados de la incertidumbre 3

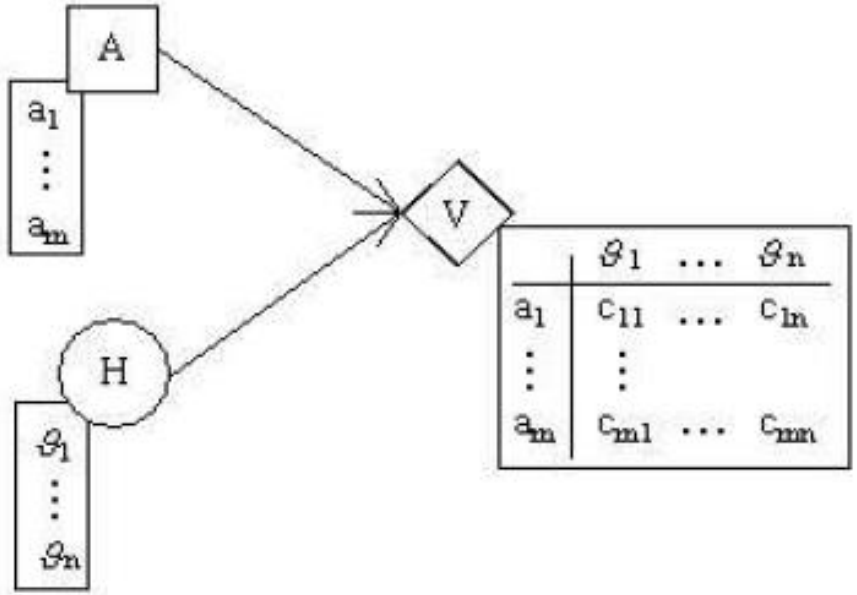
La decisión 7 que tomamos influye sobre el resultado de la incertidumbre que, eventualmente, se resolverá en el futuro

Diagramas de influencia

- Los DIs han de estar bien definidos. Para ello
  - Deben ser acíclicos
  - Deben tener 1 nodo de valor
  - Deben tener Memoria (ie un camino dirigido que pase por todos los nodos de decisión y llegue al nodo de valor)
- Además, los DIs llevan una información oculta: Por debajo de cada nodo hay tablas con información relativa a
  - Alternativas disponibles si el nodo es de decisión
  - Estados si el nodo es de azar
  - Consecuencias si es de valor

Os ilustramos ahora los DIs de una tabla de decisión (las estructuras de datos que aparecen debajo de cada nodo)

DI de una tabla



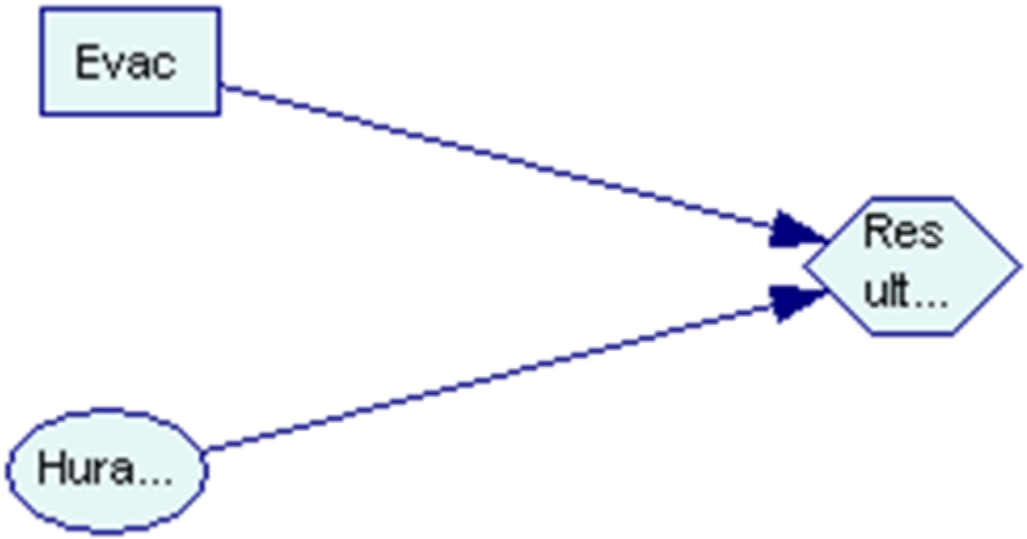
Interprétalo!!!

Decisiones en riesgo

Frente a la amenaza de un tornado, una ciudad debe decidir si desaloja o no a los habitantes. Si no lo hace y se produce el tornado, puede que dé tiempo a realizar un desalojo de emergencia (a mayor coste que el desalojo normal), y, si no da tiempo, puede que se produzcan víctimas mortales, lo que supondría enormes pérdidas económicas y de imagen para la ciudad. Representa el problema de decisión.

Decisiones en riesgo

Nosotros lo representamos así



Tomamos la decisión de evacuar o no, sin saber si va a llegar el huracán. Las consecuencias dependen de la decisión de evacuar o no y de si el huracán toca o no.

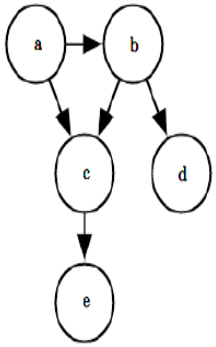
Diagramas probabilísticos

- Como herramienta básica para modelizar, cualitativamente, la incertidumbre usamos los diagramas de influencias probabilísticos, también llamados redes causales, redes bayesianas, redes de creencias, modelos gráficos probabilísticos,…

Se asimilan a diagramas de influencia en los que sólo hay nodos de azar.
Cualitativamente definen un modelo probabilístico mediante

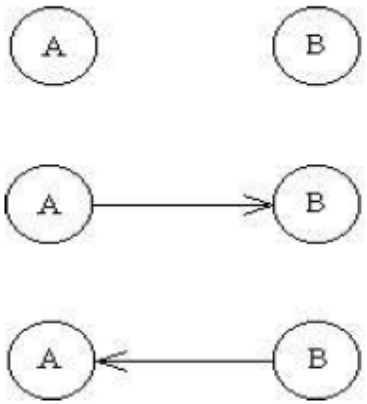
$$P(A1, A2, \ldots, An) = P(A1 \mid ant(A1)) \ldots P(An \mid ant(An))$$

donde ant (Ai) son los antecesores del nodo Ai.



$$p(a, b, c, d, e) = p(a)p(b \mid a)p(c \mid a, b)p(d \mid b)p(e \mid c)$$

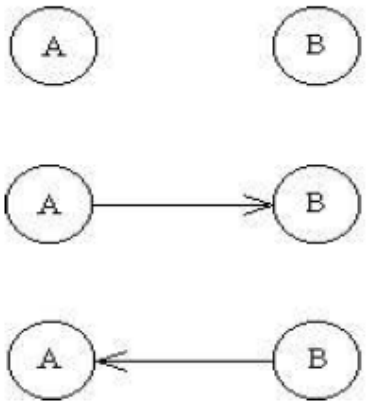
DIPs con dos nodos



Antes de pasar a la siguiente transparencia, escribe los modelos probabilísticos asociados

DIPs con dos nodos

Modelo para P(A,B)



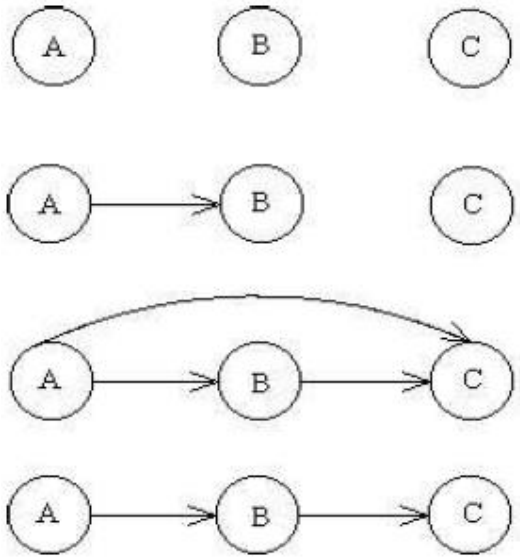
P(A)P(B)

P(A) P(B|A)

P(B) P(A|B)

El primer caso corresponde a A y B independientes. Del segundo se pasa al tercero, o viceversa, por inversión mediante la fórmula de Bayes.

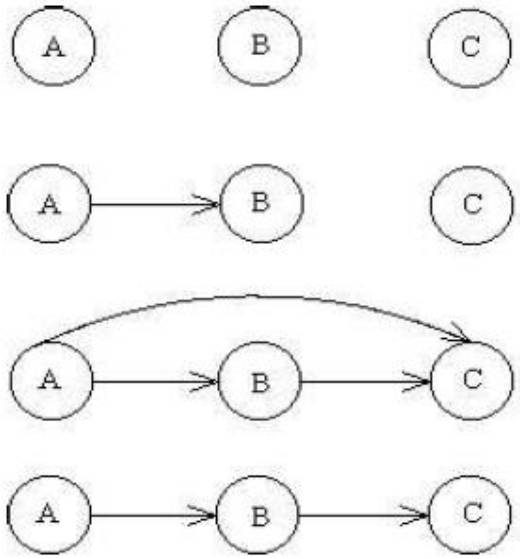
DIPs con tres nodos



Antes de pasar construye los correspondientes modelos probabilísticos

DIPs con tres nodos

Modelo P(A ,B ,C)



P(A)P(B)P(C)

P(A) P(B|A) P(C)

P(A)P(B|A)P(C|A,B)

P(A)P(B|A)P(C|B)

El primer caso se corresponde al de independencia. El tercer caso se corresponde a independencia condicional de A y C dado B. Para este importante concepto leed http://en.wikipedia.org/wiki/Conditional_independence

Independencia condicional

- Probabilidad condicionada

$$P(\mathrm{y} = y \mid \mathrm{x} = x) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}$$

- Independencia $\mathrm{x} \perp \mathrm{y}$

$$\forall x \in \mathrm{x}, y \in \mathrm{y},\ p(\mathrm{x} = x, \mathrm{y} = y) = p(\mathrm{x} = x)p(\mathrm{y} = y)$$

- Independencia condicional

$$\mathrm{x} \perp \mathrm{y} \mid \mathrm{z}$$

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z},\ p(\mathrm{x} = x, \mathrm{y} = y \mid \mathrm{z} = z) = p(\mathrm{x} = x \mid \mathrm{z} = z)p(\mathrm{y} = y \mid \mathrm{z} = z)$$

Fórmula de Bayes

- Distribución marginal

$$P(\mathrm{x}, \mathrm{y})$$

$$\forall x \in \mathrm{x}, P(\mathrm{x} = x) = \sum_{y} P(\mathrm{x} = x, \mathrm{y} = y)$$

$$p(x) = \int p(x, y) dy$$

- Regla de Bayes

$$P(\mathrm{x} \mid \mathrm{y}) = \frac{P(\mathrm{x}) P(\mathrm{y} \mid \mathrm{x})}{P(\mathrm{y})}$$

$$P(\mathrm{y}) = \sum_{x} P(\mathrm{y} \mid x) P(x)$$

DIP Asia

Damos un ejemplo complejo que recoge información médica relativa a enfermedades pulmonares:

Un problema respiratoria (dispnea) puede deberse a tuberculosis, cáncer de pulmón o bronquitis, a ninguna de ellas o a varias de ellas. Una reciente visita a Asia aumenta las posibilidades de tuberculosis, mientras que fumar es un factor de riesgo para el cáncer de pulmón y la bronquitis. Los resultados de una radiografía no discriminan entre el cáncer y la tuberculosis, como tampoco la presencia o la ausencia de dispnea.

DIP Asia

Damos un ejemplo complejo que recoge información médica relativa a enfermedades pulmonares:

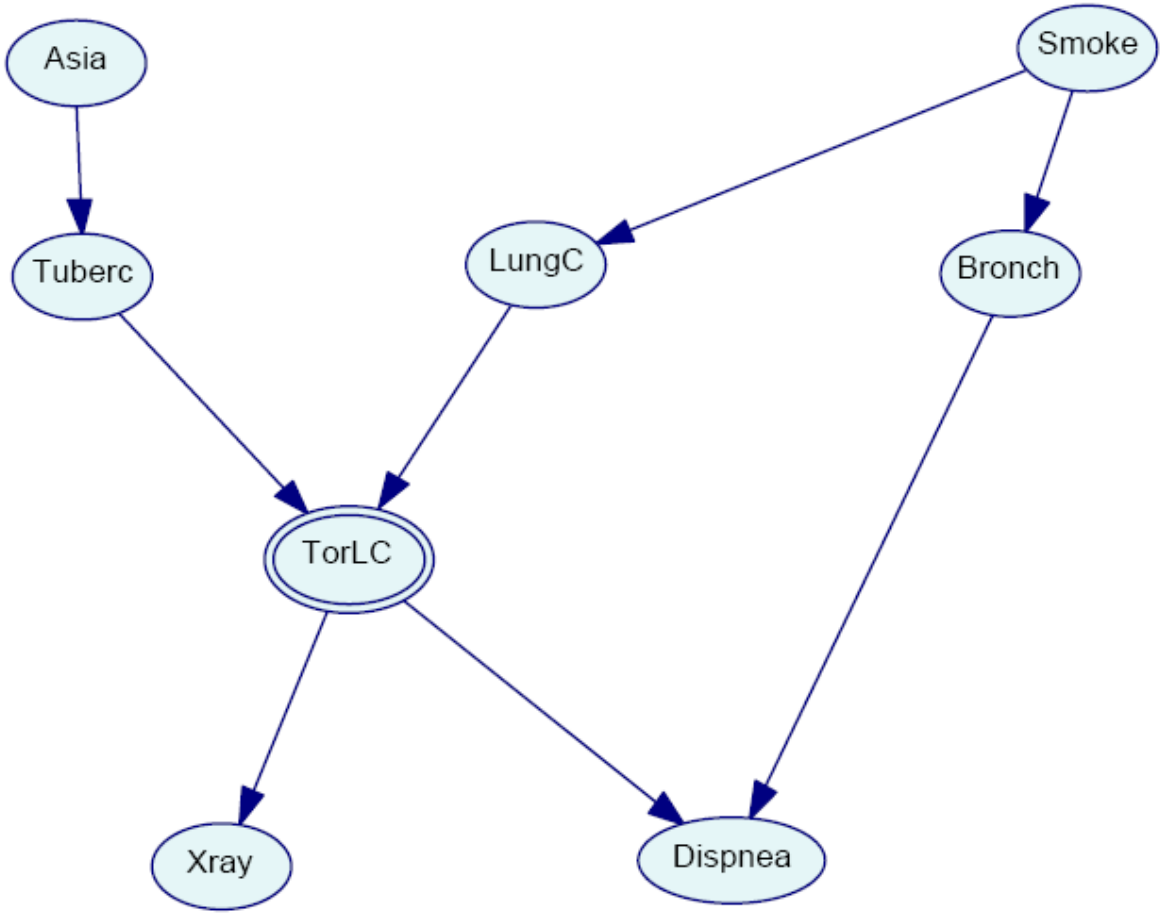
Un problema respiratorio (dispnea) puede deberse a tuberculosis, cáncer de pulmón o bronquitis, a ninguna de ellas o a varias de ellas. Una reciente visita a Asia aumenta las posibilidades de tuberculosis, mientras que fumar es un factor de riesgo para el cáncer de pulmón y la bronquitis. Los resultados de una radiografía no discriminan entre el cáncer y la tuberculosis, como tampoco la presencia o la ausencia de dispnea.

DIP Asia

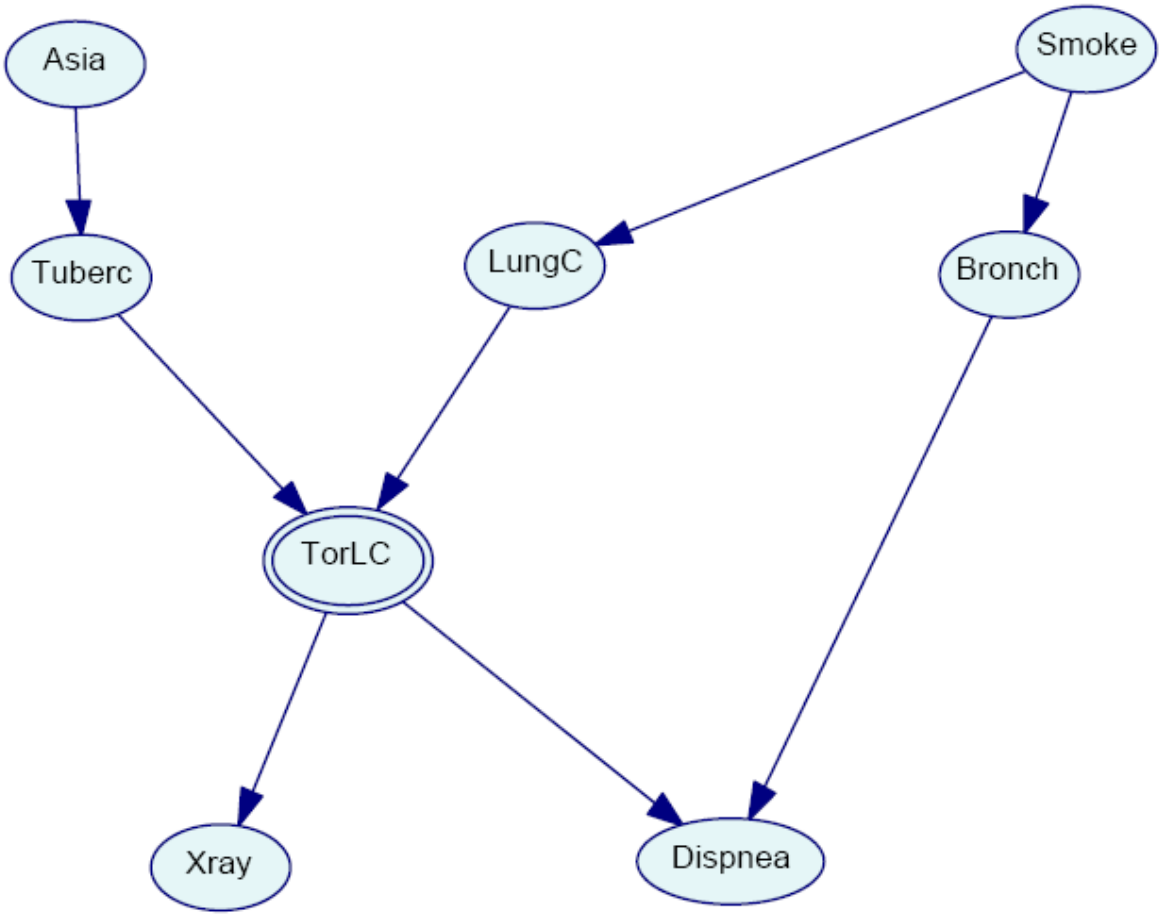
Damos un ejemplo complejo que recoge información médica relativa a enfermedades pulmonares:

Un problema respiratorio (dispnea) puede deberse a tuberculosis, cáncer de pulmón o bronquitis, a ninguna de ellas o a varias de ellas. Una reciente visita a Asia aumenta las posibilidades de tuberculosis, mientras que fumar es un factor de riesgo para el cáncer de pulmón y la bronquitis. Los resultados de una radiografía no discriminan entre el cáncer y la tuberculosis, como tampoco la presencia o la ausencia de dispnea.

DIP Asia

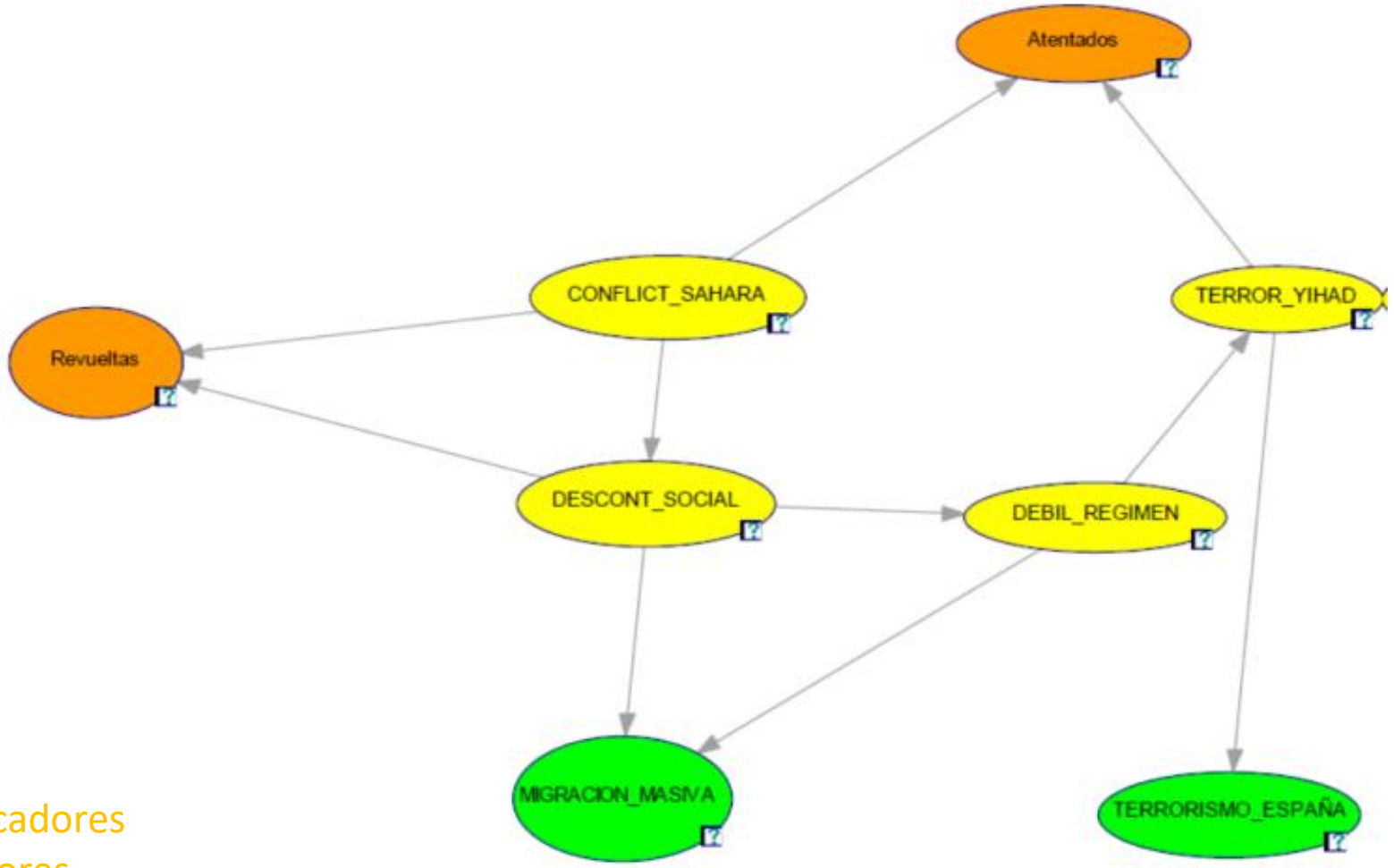


DIP Asia



$$P(A,T,S,L,B,O,X,D) = P(A)P(T|A)P(S)P(L|S)P(B|S)P(0|T,L)P(X|O)P(D|O,B)$$

DIP. Seguridad Nacional



Indicadores
Factores
Impactos

Usos DIP

1. **Inicialización.** Predicciones sobre impactos en seguridad nacional
2. **Absorción de evidencia y propagación**. Observamos indicadores y actualizamos predicciones sobre impactos en seguridad nacional.
3. **Formulación de hipótesis**, para determinar eventuales impactos sobre seguridad nacional.
4. **Planificación de la investigación**, para determinar los factores que más ayudan a clarificar el panorama de seguridad nacional.
5. **Hechos influyentes**, en una investigación forense para aprender de cara a futuros análisis.

Intercambiabilidad

En Inferencia hablamos de observaciones de fenómenos aleatorios. En muchas ocasiones tales observaciones se suponen independientes dado cierto parámetro (AKA condicionalmente independientes) lo que va asociado al concepto importante de intercambiabilidad.

http://en.wikipedia.org/wiki/Exchangeable_random_variables

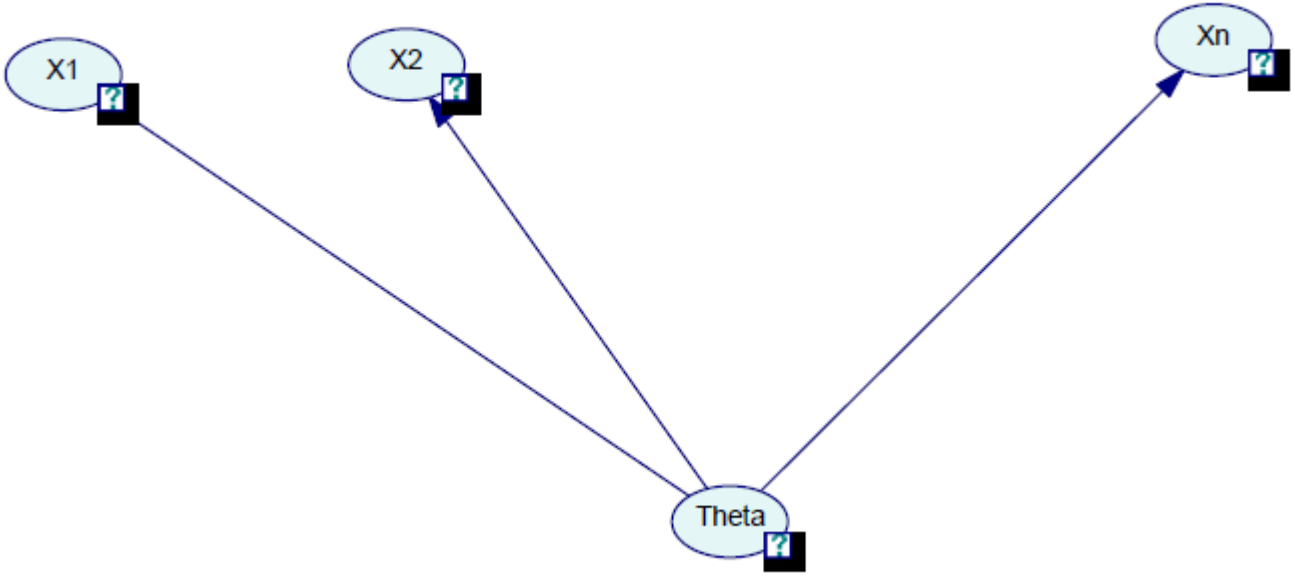
Intercambiabilidad

- Un conjunto finito de variables aleatorias es **intercambiable** si cualquier permutación suya, tiene la misma distribución conjunta que cualquier otra permutación.
- Un conjunto infinito de variables aleatorias es intercambiable, si cualquier subconjunto finito es intercambiable.
- El Teorema de De Finetti muestra que un conjunto de variables aleatorias es intercambiable si y sólo si son condicionalmente i.i.d dada cierta distribución…

Theorem 10 *If the* DM *holds* $X_1, X_2, ...$ *to be an infinitely exchangeable sequence of real random variables, then her beliefs about any finite sequence* $X_1, X_2, ..., X_n$ *has the form,*

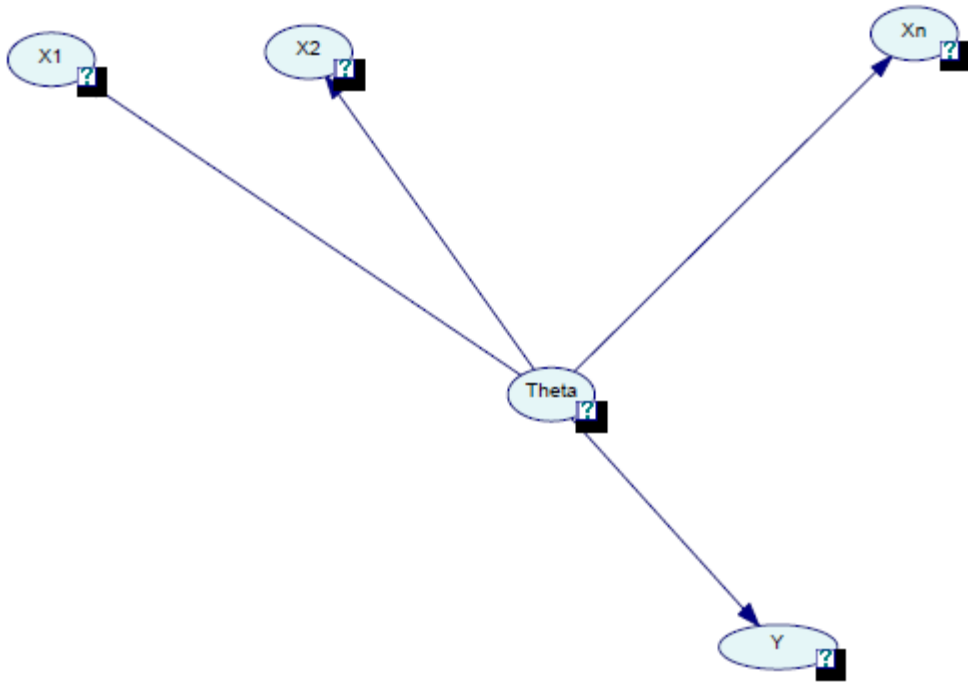
$$P_{X_1,X_2,...,X_n}(X_1 \leq x_1, X_2 \leq x_2, ..., X_n \leq x_n) = \int \prod_{i=1}^{n} F(x_i)\, d\Pi(F) \quad (3.14)$$

Modelos paramétricos. Inferencia



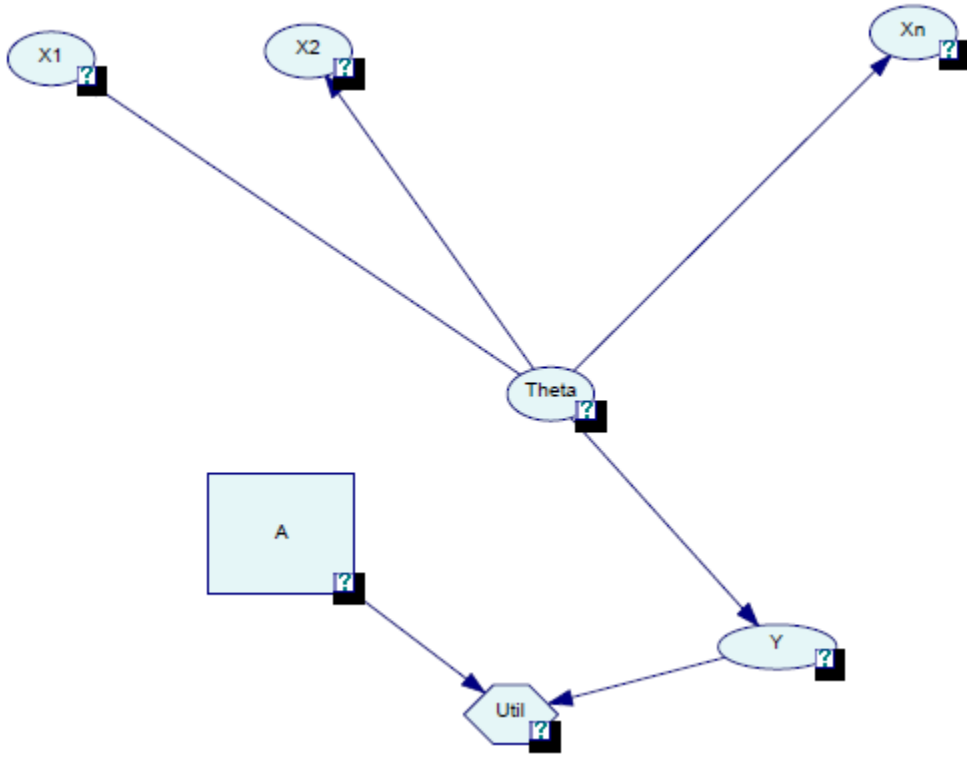
Formula el modelo

Modelos paramétricos. Predicción

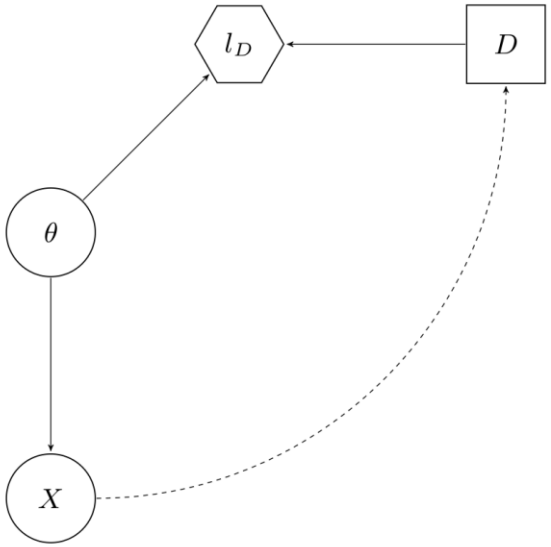


Formula el modelo

Modelos paramétricos. Predicción



Inferencia como Teoría de la Decisión



$$d^*(x) = \arg\min_d \int l_D(d,\theta)\, p_D(\theta \mid x)\, \mathrm{d}\theta.$$

$$d^*(x) = \arg\min_d \int l_D(d,\theta)\, p_D(x \mid \theta)\, p_D(\theta)\, \mathrm{d}\theta.$$

Estimación puntual. TD con pérdida cuadrática

$$l_D(d,\theta) = (\theta - d)^2$$

$$d^*(x) = \frac{1}{p_D(x)} \int \theta\, p_D(x \mid \theta)\, p_D(\theta)\, \mathrm{d}\theta = \int \theta\, p_D(\theta \mid x)\, \mathrm{d}\theta = E\left[\theta \mid x\right]$$

Funciones de pérdida estándar

En Inferencia se han introducido una series de funciones de pérdida estándar, algunas de las cuales veremos a continuación:

Función de pérdida	Alternativa Bayes
Pérdida cuadrática	Media a posteriori (si existe)
Pérdida absoluta	Mediana a posteriori (si existe)
Pérdida indicador	Moda a posteriori (si existe)

Funciones de pérdida estándar

$$\mathcal{A} = \{[\theta_1, \theta_2] : \theta_1 \leq \theta_2\}$$

$$l([\theta_1, \theta_2], \theta) = \theta_2 - \theta_1 + \begin{cases} \frac{2}{\alpha}(\theta_1 - \theta) & \text{if } \theta < \theta_1 \\ 0 & \text{if } \theta_1 \leq \theta < \theta_2 \\ \frac{2}{\alpha}(\theta - \theta_2) & \text{if } \theta_2 < \theta \end{cases}$$

$$l(A, \theta) = \nu(A) + c \times (1 - I_A(\theta))$$

Funciones de pérdida estándar

- Contraste de hipótesis $H_0 : \theta \in \Theta_0 \text{ and } H_1 : \theta \in \Theta_1$

	Es H0	Es H1
Dices H0	0	L01
Dices H1	L10	0

Pérdidas esperadas $P(H_1|\mathbf{x})l_{01}$ $P(H_0|\mathbf{x})l_{10}$

Escoger hipótesis de menor pérdida esperada

Funciones de pérdida estándar

- Factor Bayes

$$B_1^0 = \frac{P(H_1)P(H_0|\mathbf{x})}{P(H_0)P(H_1|\mathbf{x})}.$$

B_1^0	$2\log_{10} B_1^0$	Evidence against H_1
1 to 3	0 to 2	Hardly worth commenting
3 to 20	2 to 6	Positive
20 to 150	6 to 10	Strong
> 150	> 10	Very strong

Funciones de pérdida estándar

- Predicción

$$\mathcal{A} = \{q_Y(. \mid x) : q_Y(. \mid \ x) \geq 0, \int q_Y(y \mid x)dy = 1\}$$

$$l_2(q_Y(\cdot \mid x), y) = -2q_Y(y \mid x) + \int q_Y^2(y| \mid x)dy$$

$$f(\mathbf{y}|\mathbf{x}) = \int f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta}$$

Asignación de probabilidades

Una vez construido el modelo gráfico, debemos asignar las probabilidades correspondientes. A veces, tenemos acceso a buenas bases de datos y aproximamos las probabilidades mediante frecuencias relativas.

Hay problemas con datos insuficientes:

- No hay datos
- No se hacen públicos
- Existen, pero son (muy) incompletos

En tal caso, podemos utilizar juicios de expertos. Para ello …

https://www.expertsinuncertainty.net/

Asignación de probabilidades subjetivas. Experimento de referencia

- Necesito una 'regla' para medir creencias. Tales reglas se denominan experimentos de referencia o experimentos de calibración
- Un experimento es de referencia para alguien si esta persona encuentra todos los resultados de este experimento igualmente verosímiles. Algunos ejemplos, para mí, son:
  - Una bolsa con seis bolas idénticas numeradas 1,2,…,6. Este experimento me permitirá medir probabilidades con valores entre 0, 1/6, 2/6, 3/6,4/6,5/6,6/6=1
  - Lanzar cuatro monedas equilibradas, para medir probabilidades 0, 1/16, 2/16,…., 15/16, 16/16=1
  - Una rueda de la fortuna con 14 sectores iguales, para medir probabilidades 0, 1/14, 2/14,…., 14/14=1.
- En los ejemplos anteriores podemos cambiar 6, 4 y 14 para obtener 'reglas' para medir otras probabilidades

Asignación de probabilidades subjetivas. Protocolo

- Una vez identificado un experimento de calibración, lo empleamos para calibrar la probabilidad del suceso de interés.
- La idea es ir comparando el suceso de interés con sucesos de referencia hasta encontrar uno ‘igual’ de verosímil.
- Como esto no es fácil de hacer con un usuario novel, podemos recurrir a distintos protocolos.

Asignación de probabilidades subjetivas Protocolo

Dime un intervalo de probabilidad 90% (definido por su cota inferior y su cota superior) para la fecha en que Newton publicó sus leyes de Gravitación Universal.

Tu respuesta debe ser un intervalo, por ejemplo, [1620, 1680]. Para calibrarlo adecuadamente podemos proceder como sigue. Os ofrezco la posibilidad de ganar $1000 de alguna de estas dos maneras:

- Ganas $1000 si el verdadero año en que Newton publicó sus leyes cae en el intervalo que indicas. Si no, ganas $0.
- Dividimos una ruleta en dos zonas como la abajo indicada. Una es tal que la zona marcada de gris cubre el 10% y la zona de blanco cubre el 90%. Si el puntero marca al parar la zona blanca ganas $1000; si no, ganas $0.

Sesgos

- Multitud de experimentos han identificado sesgos en nuestra mente a la hora de procesar información y tomar decisiones.
- Sesgo identificado $\rightarrow$ Remedio (para propósitos prácticos)
- Sistema 1, Sistema 2. Pensando rápido y lento (Kahneman).

Sesgos

Considera esta situación:

- Un sistema de detección de cáncer de mama puede detectar el 80% de las mujeres con cáncer no diagnosticado, cometiendo errores con sólo el 5% de las mujeres que no tienen cáncer. Se estima que la tasa de cáncer de mama es de 30 casos por 10000. Intuitivamente, cuál crees que es la probabilidad de que una mujer que dé positivo tenga, de hecho, cáncer.

Sesgos

- Mucha gente cree que esta probabilidad está entre el 70 y el 75%, cuando la verdadera probabilidad está alrededor del 5%, como se sigue de una simple aplicación de la fórmula de Bayes. Tendrás de hecho que hacerlo como actividad de este capítulo.

- Este sesgo se denomina de *ignorancia de las tasas básicas*. Aquí se está ignorando la información básica de presencia de cáncer de mama, que es de 30 casos por cada 10000.

REMEDIO: Emplear explícitamente reglas del cálculo de probabilidades

Sesgos

- Tanto los expertos como los no expertos pueden tener dificultades en calibrar lo que conocen y lo que no conocen. Un experimento típico es hacer preguntas de almanaque del estilo

- Dame un rango de valores de manera que estés seguro al 90% de que la población actual de Móstoles esté entre los valores inferior y superior que has indicado

Sesgos

- Según el INE (2017) la población era de 206.589 habitantes.
- En una serie de cuestiones de ese estilo, podemos estimar la proporción p de veces que un individuo acierta. Si p=0.9 está bien calibrado; si p <0.9, está sobreconfiado (intervalos demasiado estrechos); si p>0.9, es demasiado cauteloso (intervalos demasiado anchos)
- Los estudios de calibración indican que la gente tiende a ser sobreconfiada, ie. tiende a estar demasiado segura de lo que cree saber.

Asignación de probabilidades subjetivas
Sesgos

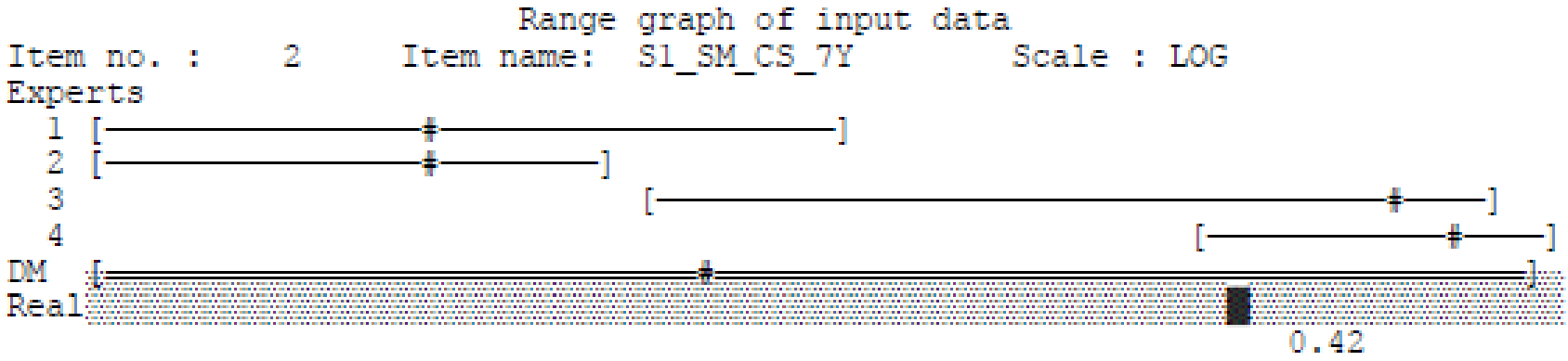
- Durante una fiesta te presento a Román, un chico tímido. ¿Será comercial o bibliotecario?

Asignación de probabilidades subjetivas
Sesgos

- Los comerciales tienden a ser personas extrovertidas, pongamos que 9 de cada 10 lo son.
- Por el contrario, hay muchos bibliotecarios tímidos, digamos que 5 de cada 10 lo son.
- Pero por cada 10 comerciales sólo hay un bibliotecario.
- Todos aquellos que piensan en un primer momento que Román será bibliotecario están ignorando este hecho.

Heurísticos y sesgos

Queremos ser calibrados e informativos

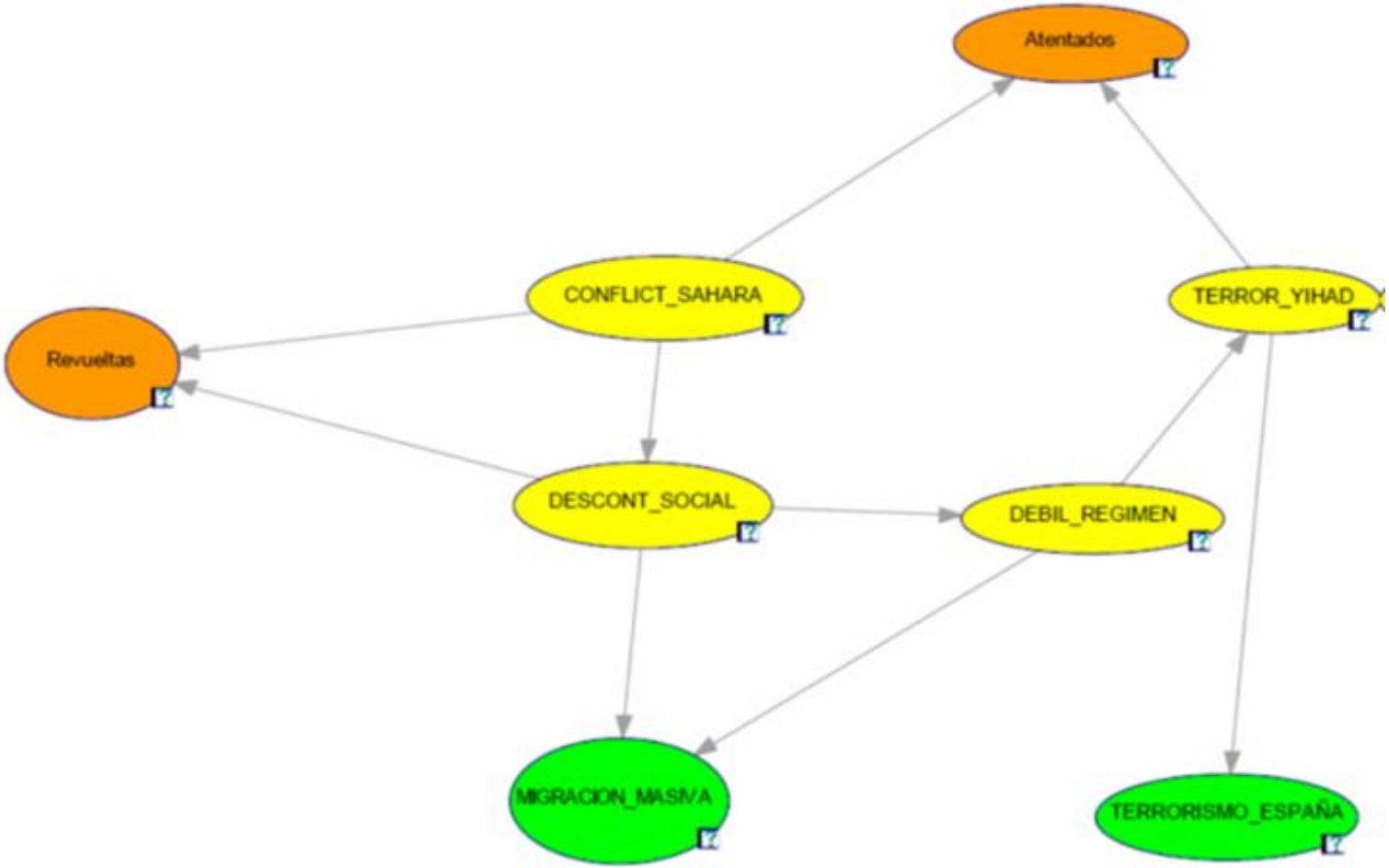


Agregación de expertos: p.ej., Método clásico de Cooke. EFSA Guidelines

Ejemplo

#	Pregunta	Cota inferior (probabilidad 95% valor es mayor)	Cota superior (probabilidad 95% valor es menor)	Valor verdadero
1	¿Cuántos turistas recibió España en 2017?			
2	¿Cuántos países no reconocen a Israel?			
3	¿Qué porcentaje de la población española vive en los primeros 5 kilómetros de costa?			
4	¿Cuántos años tiene Abdelaziz Buteflika?			
5	¿Qué porcentaje de la energía consumida en España es importada? (2016)			
6	¿Cuántos muertos por terrorismo ha habido en la UE desde el año 2000?			
7	¿Cuál fue el valor máximo de la prima de riesgo española durante la crisis?			
8	¿Cuál es la edad mediana en Marruecos?			
9	¿Cuántos muertos por accidentes de tráfico hubo en España en 2017?			
10	¿Cuántas cabezas nucleares tiene China?			

DIP SN


Atentados
CONFLICT_SAHARA
TERROR_YIHAD
Revueltas
DESCONT_SOCIAL
DEBIL_REGIMEN
MIGRACION_MASIVA
TERRORISMO_ESPAÑA

DIP SN

Conflicto Sahara	Probabilidad
Bajo	0.3
Medio	0.5
Alto	0.2

DebRg|DscScl	Nulo	Medio	Alto
Debil	0.05	0.3	0.6
Fuerte	0.95	0.7	0.4

Fijar la definición de los estados
Horizonte temporal
No tienen por qué ser discretos (eg., tasa de desempleo)

Asignación de probabilidades

- Entrenamiento
- Asignación individual
  - Motivación
  - Sesgos
  - Pruebas
  - Asignación
  - Consistencia
- Agregación
  - Discusión, Asignación individual, Agregación
    - Mismo peso a todos
    - Pesos dependientes de resultados
    - …
  - Discusión y agregación en grupo
  - Varias fases

Procesos de asignación con expertos

- Identificación de expertos. Proceso de identificación con diversidad de opiniones
- Separación de roles: expertos, decisores, analistas.
- Neutralidad: Todos los expertos igualmente tratados a priori.
- Mitigar ambiguedad en cómo se formulan las preguntas
- Mantener la diversidad plausible, incluyendo reconocer entendimiento científico incompleto
- Entrenar para que entiendan lo que se les pide (e.g. rango de probabilidad)
- Neutralidad: Promover que digan su verdadera opinión, e.g. no groupthink, no agendas ocultas.
- Control empírico
- Transparencia: Razones de las asignaciones, cálculos reproducibles, hipótesis registradas
- Responsabilidad: Fuentes de opinión de expertos identificables.

https://www.expertsinuncertainty.net/

Actualización de probabilidades

Muchas veces, disponemos, además, de evidencia (datos) que aportan información adicional sobre el suceso de interés. Nuestras creencias se actualizan mediante la fórmula de Bayes que adopta la siguiente forma

P(Suceso|Evidencia)=P(Evidencia|Suceso) P(Suceso)/ P(Evidencia)

$$p_{\theta,X}(\theta, x) = p_X(x \mid \theta) \times p_\theta(\theta).$$

$$p_\theta(\theta \mid x) = \frac{p_X(x \mid \theta) \times p_\theta(\theta)}{\int_\Theta p_X(x \mid \theta') \times p_\theta(\theta')\, d\theta'}$$

$$\underset{\theta}{\propto} p_X(x \mid \theta) \times p_\theta(\theta).$$

Un ejemplo

El modelo beta-binomial surge de manera natural cuando se desea aprender y proporcionar información sobre la proporción de éxitos p que se observan al repetir cierto experimento, con dos posibles resultados, éxito y fracaso.
Suponemos que tenemos acceso a información adicional a través de un experimento que consiste en observar n experimentos de Bernoulli independientes, registrándose el número de éxitos x (por ejemplo, tenemos un registro de las veces que ha fallado un sensor en las últimas 50 operaciones en las que ha sido activado). La verosimilitud (o el modelo) es, en este caso, binomial,

$$Pr(X = x|p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots n.$$

Un ejemplo

Se dispone de creencias iniciales sobre p, sabemos que toma valores entre 0 y 1. Dichas creencias se modelizan con la **distribución a priori**, que, en este caso, es la distribución beta:

$$f(p|\alpha,\beta) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, & \text{para } 0 \leq p \leq 1 \\ 0, & \text{en otro caso.} \end{cases}$$

cuya media y varianza son $E(p) = \frac{\alpha}{\alpha+\beta}$ y $V(p) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Una vez realizado el experimento aleatorio, y tomados los datos, se puede actualizar la información acerca de la probabilidad $p \rightarrow$ **distribución a posteriori**.

Un ejemplo

Un aspecto importante es la **asignación de distribuciones a priori**. Escoger una $\mathscr{B}e(\alpha,\beta)$ a priori implica asignar los valores α y β, es decir, conocer dos juicios de un experto:

- Su creencia acerca de la probabilidad r_1 de obtener éxito en el primer ensayo. Sabemos que $r_1 \equiv \frac{\alpha}{\alpha+\beta}$.
- Suponiendo que el primer ensayo fue un éxito, su creencia acerca de la probabilidad r_2 de éxito en el segundo ensayo ($r_2 \geq r_1$, sino supondría que las observaciones no son independientes). Sabemos que $r_2 \equiv \frac{\alpha+1}{\alpha+\beta+1}$.

Resolviendo el sistema obtenemos

$$\alpha = \frac{r_1(1-r_2)}{r_2-r_1}, \quad \beta = \frac{(1-r_1)(1-r_2)}{r_2-r_1}.$$

Un ejemplo

Todos los cálculos que hagamos quedarán afectados por r_1 y r_2, a través de α y β, especialmente si el experimento proporciona pocos datos. Debemos, por tanto, evaluar la *consistencia* de tales asignaciones.

Una vez calculadas (α, β) a partir de (r_1, r_2), pedimos al experto que asigne la probabilidad r_3 de éxito en el segundo ensayo, si el primero fue un fracaso. Sabemos que $r_3 \equiv \frac{\alpha}{\alpha+\beta+1}$.

Si ocurre que $r_3 \approx \frac{\alpha}{\alpha+\beta+1}$, la asignación es consistente; si no, deben reasignarse r_1, r_2 y r_3.

En casos con mucha incertidumbre y poca información a priori sobre qué valores de p pueden ser más probables, asignar valores pequeños de α y β, pues facilitan la absorción de evidencia. Como caso límite, tenemos la $\mathscr{B}e(1,1)$, que coincide con la $\mathscr{U}(0,1)$.

Un ejemplo

Estamos interesados en aprender acerca de la proporción p de aterrizajes que finalizan dentro de los límites de la pista cuando se ha producido una aproximación desestabilizada. Deseamos asignar una distribución a priori a tal proporción.

Preguntamos a un experto qué valores asigna a:

- r_1 (probabilidad de "éxito" en el primer aterrizaje), y nos contesta $r_1 = \frac{1}{5}$,
- r_2 (probabilidad de "éxito" en el segundo aterrizaje, sabiendo que el primero fue un "éxito"), y nos contesta $r_2 = \frac{1}{3}$.

Despejando los valores de α y β, obtenemos $\alpha = 1$ y $\beta = 4$.

Si preguntamos al experto por el valor de r_3 (probabilidad de "éxito" en el segundo aterrizaje, sabiendo que el primero fue un "fracaso") y nos contesta $r_3 = 0.15$ ¿es consistente este juicio con su afirmaciones previas sobre r_1 y r_2?

Un ejemplo

Actualizamos nuestras creencias sobre p aplicando la **fórmula de Bayes**, que se extiende al caso continuo de la forma

$$f(p|x) = \frac{f(p)Pr(X=x|p)}{\int f(p)Pr(X=x|p)dp} \Rightarrow f(p|x) \propto f(p)Pr(X=x|p).$$

Se trata de un caso mixto, discreto para x y continuo para p, y se tiene que:

$$\begin{aligned} f(p|x) &\propto f(p)\,Pr(x|p) \\ &\propto p^{\alpha-1}(1-p)^{\beta-1}p^{x}(1-p)^{n-x} \\ &\propto p^{x+\alpha-1}(1-p)^{n-x+\beta-1}, \end{aligned}$$

siendo $p \in (0,1)$ y $f(p|x) = 0$ para $p \notin (0,1)$.

Se observa, pues, que $p|x \sim \mathscr{B}e(x+\alpha, n-x+\beta)$.

Distribuciones conjugadas

El caso beta-binomial es paradigmático del concepto importante de distribución conjugada.

Una distribución a priori es conjugada para un modelo si la distribución a posteriori es de la misma familia.

Tenéis una tabla de distribuciones conjugadas en

http://en.wikipedia.org/wiki/Conjugate_prior

Recordad que no siempre utilizaremos distribuciones conjugadas!!!!

Distribuciones no informativas

Otro concepto importante de distribución a priori es el de distribución no informativa. Estadística Oficial

Una distribución a priori es no informativa (objetiva, de referencia, plana) si no aporta información sobre el parámetro de interés, 'dejando hablar a los datos por sí mismos'

Tenéis abundante información sobre distribuciones no informativas en

http://www.stats.org.uk/priors/noninformative/

Recordad que no siempre utilizaremos distribuciones no informativas!!!!

Modelo normal-normal

Observaciones ciid dados θ, σ^2 son $\mathrm{N}(\theta, \sigma^2)$

A priori impropia $f(\theta) \propto 1$

$$f(\theta \mid \mathbf{x}) \propto \exp\left(-\frac{n}{2}\left(\frac{\theta^2}{\sigma^2} - 2\frac{\theta\bar{x}}{\sigma^2}\right)\right) \qquad \theta \mid \mathbf{x} \sim \mathrm{N}\left(\bar{x}, \frac{\sigma^2}{n}\right)$$

A priori propia $\theta \sim \mathrm{N}(\mu_0, \sigma_0^2)$

$$f(\theta \mid \mathbf{x}) \propto \exp\left(-\frac{1}{2}\theta^2\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right) - 2\theta\left(\frac{\sum x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right)\right)$$

$$\theta \mid \mathbf{x} \sim \mathrm{N}\left(\frac{n\bar{x}/\sigma^2 + \mu_0/\sigma_0^2}{n/\sigma^2 + 1/\sigma_0^2}, \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}\right)$$

Modelo normal-normal

- Estimación puntual

$$\bar{x} \qquad \frac{n\bar{x}/\sigma^2 + \mu_0/\sigma_0^2}{n/\sigma^2 + 1/\sigma_0^2}$$

- Estimación por intervalo. Intervalo creíble

$$[\mu_1 - 1.96\sigma_1, \mu_1 + 1.96\sigma_1]$$

(de probabilidad 0.95, región de máxima densidad a posteriori)

Modelo normal-normal

Con a priori no informativa, la predictiva

$$X_{n+1} \mid x \sim \mathrm{N}\left(\bar{x}, \tfrac{n+1}{n}\sigma^2\right)$$

Intervalo predictivo

$$\left[\bar{x} - z_{\alpha/2}\sigma\sqrt{(n+1)/n}, \bar{x} + z_{\alpha/2}\sigma\sqrt{(n+1)/n}\right]$$

Introducción a Métodos de Montecarlo

Ejemplo simplificado de ordenador, (E/S) y (CPU). El ordenador falla cuando lo hace alguno de los dos componentes.

El tiempo hasta fallo del sistema $T = \text{mín}(X_1, X_2)$, con X_1, X_2 v.a. que siguen ciertas distribuciones

El tiempo esperado $E(T)$, suponiendo $X_i \sim \mathcal{E}(\mu_i),\ i = 1, 2$ es

$$E(T) = \int_0^\infty \int_0^\infty \text{mín}(x_1, x_2)\mu_1 e^{-\mu_1 x_1}\mu_2 e^{-\mu_2 x_2} dx_1 dx_2$$

Se puede calcular $E(T)$ de tres formas:

- *Aproximación analítica*. Bajo ciertas suposiciones (exponencialidad e independencia de X_1, X_2)
- *Aproximación numérica*. Regla del trapecio, Simpson, etc... Más robusta, pues se puede prescindir de alguna de las hipótesis. Problema: dimensión.
- *Aproximación basada en simulación*. Consiste en construir un programa que describa el comportamiento del sistema y realizar experimentos con él.

Introducción a Métodos de Montecarlo

En el ejemplo anterior:

Algoritmo para calcular $T = \text{mín}(X_1, X_2)$

Entrada: Hacer $\text{tiempo\_fallo} = 0$

Desde $i = 1$ *hasta* n

  Desde $j = 1$ *hasta* 2

    Generar $U_j \sim \mathcal{U}(0,1)$

    Hacer $X_j = -\log(1 - U_j)/\mu_j$

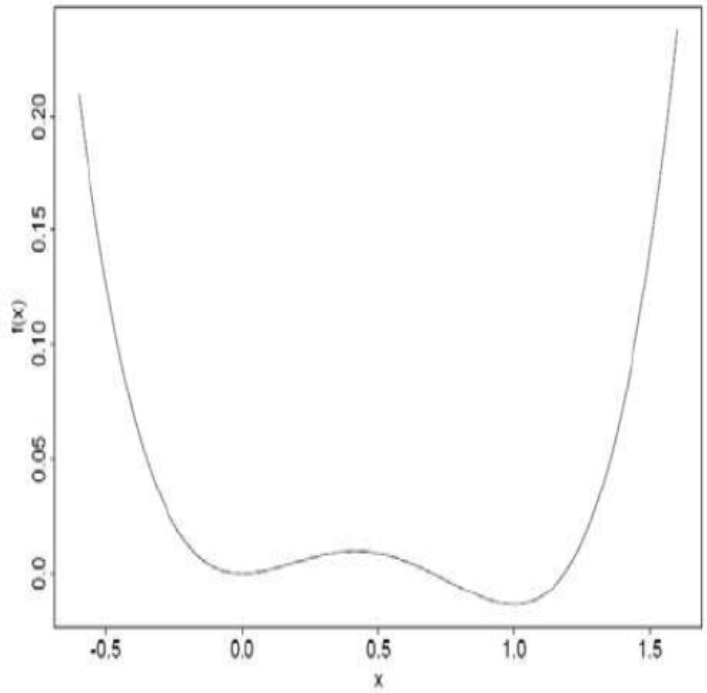
  Hacer $\text{tiempo\_fallo} = \text{tiempo\_fallo} + \text{mín}(X_1, X_2)$

Hacer $E[\text{tiempo\_fallo}] = \text{tiempo\_fallo}/n$

El modelo propuesto no depende de las condiciones de independencia y exponencialidad.

Introducción a Métodos de Montecarlo

Consideramos la función $f(x) = x^4/4 - 17x^3/36 + 5x^2/24$, cuya gráfica es



f puede ser la función de coste de cierto sistema, que depende de la variable de decisión x.

En este ejemplo tenemos solución analítica y métodos numéricos (que suelen conducir a óptimos locales).

Introducción a Métodos de Montecarlo

Como alternativa (que además permite tratar problemas con funciones no diferenciables) es utilizar simulación Montecarlo.

El ejemplo más sencillo es el de búsqueda aleatoria pura, que se adapta al siguiente esquema

16	82	39	86	86	73	07	32	72	35	12	82	87	21	30	30	60	53	89	92
99	77	85	43	72	34	52	99	30	86	81	40	18	61	20	16	92	39	34	44
01	48	69	32	37	05	99	27	23	55	88	47	38	48	53	79	41	08	73	95
08	34	04	83	42	92	64	49	51	23	44	62	75	43	09	22	55	51	38	18
09	22	76	47	23	99	48	66	26	73	95	53	37	47	00	43	66	80	44	18
64	18	83	13	90	95	64	21	97	09	86	59	99	70	20	73	35	11	81	57
77	67	94	40	00	12	26	45	77	54	21	51	91	91	28	74	47	00	53	95
73	01	20	47	86	40	71	03	13	36	98	50	48	45	30	23	40	85	76	63
72	67	37	77	52	79	93	67	57	78	77	07	58	19	48	32	72	94	66	11
58	96	79	92	08	88	46	62	58	96	75	18	57	89	21	17	26	92	26	63
41	69	24	18	81	29	14	06	67	15	23	70	27	89	40	77	31	98	71	15
16	45	84	78	49	17	84	92	51	12	08	78	30	35	63	84	34	68	97	10
92	09	48	47	40	81	30	44	03	98	19	38	33	07	00	55	70	65	24	19
26	92	58	75	64	61	49	58	68	45	09	32	76	03	29	08	73	11	33	79
17	33	86	83	91	26	51	12	57	73	21	12	09	58	24	64	91	53	24	92
91	41	97	06	57	45	39	16	64	92	66	18	78	71	55	99	29	18	02	56
52	00	40	81	14	06	94	03	71	39	09	33	74	08	42	15	85	08	35	30
10	07	82	40	99	00	91	31	44	73	51	42	08	28	31	35	20	07	85	96
70	70	14	24	43	71	60	86	17	85	61	05	65	10	68	73	03	31	45	23
96	50	32	72	89	62	28	76	60	00	52	94	67	16	32	08	88	27	01	94
20	89	41	95	46	28	45	92	21	97	51	15	02	82	11	95	65	27	50	08
30	99	84	95	47	72	38	22	55	44	50	61	71	58	86	49	25	60	69	17
94	53	29	42	38	74	90	06	18	71	99	27	84	88	03	43	07	53	96	02
50	61	25	57	55	50	92	14	39	77	29	17	73	75	83	38	40	02	06	47
38	63	63	30	36	25	66	30	53	98	49	78	40	92	80	97	67	46	38	34

Búsqueda aleatoria pura Montecarlo

Entrada: Hacer $f^* = \infty$

Desde $i = 1$ *hasta* N,

  Generar $x_i \sim \mathcal{U}[-0.5, 1.5]$

  Si $f(x_i) < f^*$ **entonces**

    hacer $f^* = f(x_i)$, $x^* = x_i$

Utilizando los números de la tabla anterior y transformando cada u mediante $x = \frac{2}{99}u - 0.5$, obtenemos (con los veinte primeros números) como estimación del mínimo 1.015152.

Introducción a Métodos de Montecarlo

- Muchos problemas en AA son tan complejos que resulta:
  - Imposible una solución analítica o exacta
  - Ineficiente una solución numérica aproximada

  P.ej., con Big Data resulta ineficiente aplicar descenso del gradiente
  en Inferencia Bayesiana calcular la constante de normalización es complejo

- Alternativa métodos Montecarlo
  - Introducir elementos aleatorios que facilitan la solución del problema (p.ej sustituir Integral por suma Montecarlo)
  - Muestrear de las distribuciones de interés
  - Resolver el problema

  P.ej., con BigData aplicamos descenso de gradiente estocástico: sustituimos el gradiente sobre toda la muestra por el gradiente sobre un minilote (aleatorio)
  en Inferencia Bayesiana sustituimos integrales por sumas Montecarlo

Métodos Montecarlo

- Nuestro interés: Integración y Optimización
- Deseamos resolver

$$I_S = \int_{[0,1]^s} f(u)du$$

- Es ineficiente hacerlo numéricamente

$$I_S \simeq \sum_{n_1=0}^{m} \cdots \sum_{n_s=0}^{m} w_{n_1} \cdots w_{n_s} \, f\left(\frac{n_1}{m}, \ldots, \frac{n_s}{m}\right)$$

- Visualizamos el problema como

$$I_S = E(f)$$

- Resolvemos como

Generar $u_1, \ldots, u_N \sim \mathcal{U}[0,1]^s$
Hacer $\widehat{I}_S = \frac{1}{N} \sum_{i=1}^{N} f(u_i)$

Fundamento Método Montecarlo

- Ley Fuerte de los Grandes Números

$$\widehat{I}_S = \frac{1}{N}\sum_{i=1}^{N} f(u_i) \xrightarrow{c.s.} E(f) = I_S$$

- Cota de error

$$\int_{[0,1]^s} (\widehat{I}_S - I_S)^2 du = \frac{\sigma^2(f)}{N} = Var(\widehat{I}_s) \qquad con\ \sigma^2(f) = \int_{[0,1]^s} (f - E(f))^2 du \qquad O(N^{-\frac{1}{2}})$$

- Estimamos con

$$EE(\widehat{I}_S) = \frac{1}{\sqrt{N}} \sqrt{\frac{\sum_{i=1}^{N} (f(u_i) - \widehat{I}_S)^2}{N-1}}$$

Generalizaciones

- Para $I_B = \int_B f(u)du$

$$I_B = vol(B) \int_B f(u) \frac{du}{vol(B)} = vol(B)E(f)$$

- Para $I_g = \int f(u)g(u)du = E_g(f)$

Generar $u_1, \ldots, u_N \sim g$

Hacer $\widehat{I}_g = \frac{1}{N} \sum_{i=1}^{N} f(u_i)$

Ejemplo

Deseamos calcular $I = \int_{-\infty}^{\infty}(x + x^2)g(x)dx$, donde g es la función de densidad de una distribución normal de media 1 y desviación típica 2. Entonces, $I = E(X^2 + X) = E(X^2) + E(X)$. Como $Var(X) = E(X^2) - E(X)$, obtenemos que $E(X^2) = 5$, luego $I = 6$.

Aplicamos ahora integración Montecarlo. Generamos una muestra de tamaño 30 de la distribución $\mathcal{N}(1, 4)$:

1.467	2.261	-0.657	-0.351	-0.562	3.113
1.536	2.477	-0.246	1.701	0.972	-0.609
5.352	-2.509	5.690	0.005	3.622	-1.458
-0.967	-0.286	-2.824	-2.751	-1.136	-1.483
3.211	1.671	-0.004	0.152	0.509	3.849

y la aproximación queda

$$\widehat{I} = \frac{1}{N}\sum_{i=1}^{N} f(u_i) = \frac{1}{30}\left((1.467^2 + 1.467) + ... + (3.849^2 + 3.849)\right) = 6.1048,$$

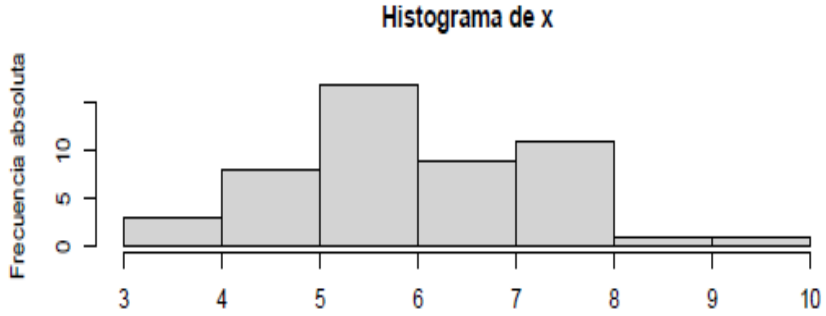
mientras que la estimación del error de la aproximación es

$$\frac{1}{\sqrt{30}}\sqrt{\frac{((1.467^2 + 1.467) - 6.1048)^2 + ... + ((3.849^2 + 3.849) - 6.1048)^2}{29}} = 0.323$$

Ejemplo

- 50 réplicas de tamaño 50

6.769	5.603	5.614	9.229	7.189	3.277	4.312	7.070	5.195	4.496
5.775	4.646	5.670	7.134	4.931	4.403	6.783	7.152	5.834	4.958
7.159	7.270	8.379	5.037	5.143	5.757	7.399	5.236	4.749	5.729
7.015	6.156	3.985	5.643	5.720	6.878	6.367	7.520	7.093	6.605
6.356	6.567	7.784	5.256	6.302	5.460	4.808	5.880	3.846	5.962



- Media

Para la implementación

- Algoritmos de generación de números aleatorios
- Algoritmos de generación de variables aleatorias
- Métodos de análisis de resultados
- Métodos de reducción de varianza
- Diseño de experimentos de simulación

Ejemplo: SGD y minilotes

- MLE. Optimizar

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\hat{p}_{\mathrm{data}}} L(\boldsymbol{x}, y, \boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} L(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta})$$

$$L(\boldsymbol{x}, y, \boldsymbol{\theta}) = -\log p(y \mid \boldsymbol{x}; \boldsymbol{\theta})$$

- Gradiente

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta})$$

Si m billon…pero gradiente es una esperanza y puede estimarse por mini-lotes (SGD)

$$\boldsymbol{g} = \frac{1}{m'}\nabla_{\boldsymbol{\theta}}\sum_{i=1}^{m'} L(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta})$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon \boldsymbol{g}$$

Require: Learning rate ϵ_k.
Require: Initial parameter $\boldsymbol{\theta}$
while stopping criterion not met **do**
  Sample a minibatch of m examples from the training set $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ with corresponding targets $\boldsymbol{y}^{(i)}$.
  Compute gradient estimate: $\hat{\boldsymbol{g}} \leftarrow +\frac{1}{m}\nabla_{\boldsymbol{\theta}}\sum_i L(f(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}), \boldsymbol{y}^{(i)})$
  Apply update: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon\hat{\boldsymbol{g}}$
end while

$$\sum_{k=1}^{\infty} \epsilon_k = \infty$$

$$\sum_{k=1}^{\infty} \epsilon_k^2 < \infty$$

Computación bayesiana

En general, debemos calcular alternativas de máxima utilidad esperada

$$\max_a \int u(c(a,\theta))p(\theta|x)d\theta$$

A veces, resulta ser conveniente resolver

$$\max_a \int u(a,\theta)p(x|\theta)p(\theta)d\theta$$

Una posibilidad, aproximar utilidades esperadas por MC y optimizar. Muestrear de la a posteriori

1. Select a sample $\theta^1, ..., \theta^m \sim p(\theta|x)$.
2. Solve the optimisation problem

$$\max_{a \in A} \frac{1}{m} \sum_{i=1}^{m} u(a, \theta^i)$$

yielding $a_m(\theta)$.

Métodos Montecarlo basados en cadenas de Markov

En AA parecen distribuciones de alta dimensión no estándar. Alternativa: MCMC

Construimos un proceso de Markov $p(\cdot,\cdot)$ en tiempo discreto cuyo espacio de estados es $\mathcal{X}$, del que es sencillo muestrear, distribución de equilibrio es $\pi(x)$.

Partiendo de un valor arbitrario, si dejamos correr el proceso un periodo suficientemente largo, muestrearemos aproximadamente de π.

Algoritmo general de los métodos basados en Cadenas de Markov

Entrada: Escoger X_0 arbitrariamente
Hacer: $i = 0$
Repetir
  Generar: $X_{i+1} \sim p(X_i, \cdot)$
  Hacer: $i = i + 1$
hasta *que se juzgue convergencia*
Desde *j=1* **hasta** *N*
  Generar: $X_{i+j} \sim p(X_{i+j-1}, \cdot)$
  Salir: X_{i+j}
  Hacer: j=j+1

Muestreador de Gibbs. Motivación

Suponemos (X,Y)

son variables de Bernoulli con distribución conjunta

X	Y	$P(X,Y)$
0	0	p_1
1	0	p_2
0	1	p_3
1	1	p_4

con $p_i > 0$, $\sum_{i=1}^{4} p_i = 1$.

La marginal de X es una distribución de Bernoulli con probabilidad de éxito $p_2 + p_4$, es decir, $Pr(X = 1) = p_2 + p_4$.

Las distribuciones de $X|Y = y$, $Y|X = x$ son también de fácil cálculo. Por ejemplo, la distribución de $X|Y = 1$ es de Bernoulli con probabilidad de éxito $p_4/(p_3 + p_4)$, es decir, $Pr(X = 1|Y = 1) = p_4/(p_3 + p_4)$.

De hecho, todas las distribuciones condicionadas pueden expresarse mediante dos matrices

$$A_{yx} = \begin{pmatrix} \frac{p_1}{p_1+p_3} & \frac{p_3}{p_1+p_3} \\ \frac{p_2}{p_2+p_4} & \frac{p_4}{p_2+p_4} \end{pmatrix}; \qquad A_{xy} = \begin{pmatrix} \frac{p_1}{p_1+p_2} & \frac{p_2}{p_1+p_2} \\ \frac{p_3}{p_3+p_4} & \frac{p_4}{p_3+p_4} \end{pmatrix},$$

para las distribuciones $Y|X$ y $X|Y$, respectivamente.

Muestreador de Gibbs. Motivación

Generando de las distribuciones condicionadas

Hacer: $Y_0 = y_0,\ i = 1$

Repetir

  Generar: $X_i \sim X|Y = y_{i-1}$

  Generar: $Y_i \sim Y|X = X_i$

  Hacer: $i = i + 1$

hasta *que se juzgue convergencia*

Se comprueba fácilmente que la sucesión $\{X_n\}$ define una cadena de Markov con matriz de transición $A = A_{yx}A_{xy}$
Como las probabilidades p_i son positivas, esta cadena es ergódica y tiene, por tanto, distribución límite. Se comprueba que ésta es la marginal de X, esto es, $X_n \xrightarrow{d} X$, puesto que se verifica

$$(p_1 + p_3 \quad p_2 + p_4) = (p_1 + p_3 \quad p_2 + p_4)A$$

Análogamente, se comprueba $Y_n \xrightarrow{d} Y$ y $(X_n, Y_n) \xrightarrow{d} (X, Y)$.

Muestreador de Gibbs. Esquema general

Esquema general del muestreador de Gibbs

Escoger $X_1^0, X_2^0, ..., X_p^0$

Hacer: i=1

Repetir

- **Generar**: $X_1^i \sim X_1 | X_2^{i-1}, ..., X_p^{i-1}$
- **Generar**: $X_2^i \sim X_2 | X_1^i, X_3^{i-1}, ..., X_p^{i-1}$
- ...
- **Generar**: $X_p^i \sim X_p | X_1^i, X_2^i, ..., X_{p-1}^i$
- **Hacer**: $i = i + 1$

hasta *que se juzgue convergencia*

Muestreador de Gibbs. Ejemplo

Ejemplo. Supongamos que deseamos muestrear de la densidad

$$\pi(x_1, x_2) = \frac{1}{\pi} e^{-x_1(1+x_2^2)},$$

para $(x_1, x_2) \in (0, \infty) \times (-\infty, \infty)$.

Muestreador de Gibbs. Ejemplo

Ejemplo. Supongamos que deseamos muestrear de la densidad

$$\pi(x_1, x_2) = \frac{1}{\pi} e^{-x_1(1+x_2^2)},$$

para $(x_1, x_2) \in (0, \infty) \times (-\infty, \infty)$.

$$\pi(x_1|x_2) = \frac{\pi(x_1, x_2)}{\pi(x_2)} \propto \pi(x_1, x_2) \propto e^{-x_1(1+x_2^2)}$$

con lo que

$$X_1|X_2 = x_2 \sim \mathcal{E}xp(1 + x_2^2)$$

$$\pi(x_2|x_1) \propto \pi(x_1, x_2) \propto e^{-x_1 x_2^2},$$

con lo que

$$X_2|X_1 = x_1 \sim \mathcal{N}\left(0, \sigma^2 = \frac{1}{2x_1}\right)$$

Muestreador de Gibbs. Ejemplo

Gibbs sampler para el ejemplo

Escoger el valor inicial X_2^0

Hacer: i=1

Repetir

- **Generar**: $X_1^i \sim \mathcal{E}xp(1 + (X_2^{i-1})^2)$
- **Generar**: $X_2^i \sim \mathcal{N}\left(0, \frac{1}{2X_1^i}\right)$
- **Hacer**: $i = i + 1$

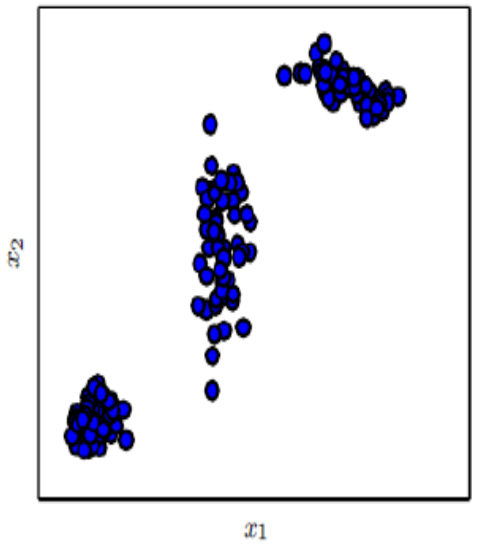
hasta *que se juzgue convergencia*

Modelos de mixturas

Modelos de mixturas

$$P(\mathrm{x}) = \sum_i P(\mathrm{c} = i) P(\mathrm{x} \mid \mathrm{c} = i)$$

Variables latentes

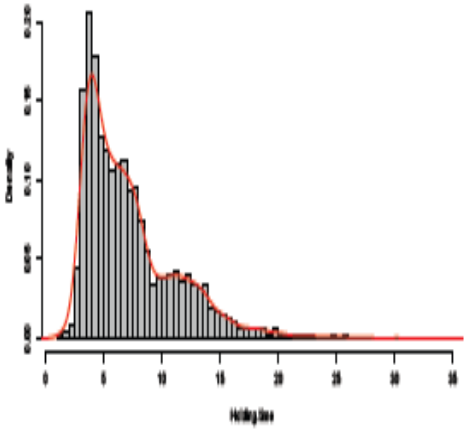


Computación en modelos de mixturas

Los modelos de mixturas proporcionan una forma aproximación flexible al tratamiento de la incertidumbre

- *Teoría*. Cualquier distribución positiva se puede aproximar mediante una distribución de gammas; cualquier distribución puede aproximarse mediante una mixtura de normales ➔ Una aproximación a estimación de densidades
- *Computación*. Muestreadores MCMC (incluyendo número incierto de componentes en la mixtura)
- *Aplicaciones*. Describe heterogeneidas (conglomerados),

  Incertidumbre en modelos,…



Computación en modelos de mixturas

Consideramos mixtura de exponenciales (con a priori Dirichlet-gamma)

$$f(t|\boldsymbol{\theta}) = q_1\mu_1 \exp(-\mu_1 t) + \ldots + q_k\mu_k \exp(-\mu_k t)$$

Introducimos etiquetas que describen pertenencia a componente

$$\mathbf{z}_j|\mathbf{q},\boldsymbol{\mu} \sim \mathcal{M}_k(1; q_1, ..., q_k) \qquad t_j|\mathbf{z}_j,\mathbf{q},\boldsymbol{\mu} \sim \mathcal{E}(\prod_{i=1}^{k} \mu_i^{z_{ij}}).$$

Deducimos las a posteriori condiconadas

$$\mathbf{z}_j|t_j,\mathbf{q},\boldsymbol{\mu} \sim \mathcal{M}_k\left(1; \frac{q_1\mu_1 \exp(-\mu_1 t_j)}{\sum_{i=1}^{k} q_i\mu_i \exp(-\mu_i t_j)}; ...; \frac{q_k\mu_k \exp(-\mu_k t_j)}{\sum_{i=1}^{k} q_i\mu_i \exp(-\mu_i t_j)}\right), j = 1, ..., n_s$$

$$\mu_j|\mathbf{t},\mathbf{z} \sim \mathcal{G}\left(a_j + \sum_{i=1}^{n} z_{ji}t_i, p_j + \sum_{i=1}^{n} z_{ji}\right), j = 1, ..., k$$

$$\mathbf{q}|\mathbf{t},\mathbf{z} \sim \mathcal{D}\left(\alpha_1 + \sum_{i=1}^{n} z_{1i}, ..., \alpha_k + \sum_{i=1}^{n} z_{ki}\right)$$

Computación con mixturas

1. Start with arbitrary values $(\mathbf{q}^0, \mu^0, \mathbf{z}^0)$, $i = 0$.
2. Until convergence, iterate through
   - Generate $\mathbf{z}_j^{i+1} \sim \mathbf{z}_j | t_j, \mathbf{q}^i, \mu^i$, $j = 1, ..., n_s$.
   - Generate $\mathbf{q}^{i+1} \sim \mathbf{q} | \mathbf{t}, \mathbf{z}^{i+1}$.
   - Generate $\mu_j^{i+1} \sim \mu_j | \mathbf{t}, \mathbf{z}^{i+1}$, $j = 1, ..., k$.
   - Set $i = i + 1$.

Se extiende a número desconocido de componentes

No escala a n muy grande….

Computación Bayesiana: Metropolis

A veces no podemos muestrear de las condicionadas.
Conocemos, salvo constante, la distribución meta. Escogemos distribución generadora de candidatos q(.|.), bajo condiciones adecuadas, este esquema se diseña para converger a la distribución meta

1. Choose initial values θ^0. $i = 0$

2. Until convergence is detected, iterate through
   - Generate a candidate $\theta^* \sim q(\theta|\theta^i)$.
   - If $p(\theta^i)q(\theta^i \mid \theta^*) > 0$, $\alpha(\theta^i, \theta^*) = \min\left(\frac{p(\theta^*)q(\theta^*|\theta^i)}{p(\theta^i)q(\theta^i|\theta^*)}, 1\right)$;
   - else, $\alpha(\theta^i, \theta^*) = 1$.
   - Do
   $$\theta^{i+1} = \begin{cases} \theta^* & \text{with prob } \alpha(\theta^i, \theta^*), \\ \theta^i & \text{with prob } 1 - \alpha(\theta^i, \theta^*) \end{cases}$$
   - $i = i + 1$.

Computación bayesiana: Simulación de probabilidades aumentadas

Frecuentemente, la a posteriori depende de la decisión que se tome. La siguiente observación resulta útil en ese contexto. Definimos una distribución artificial (supuesto u no negativa)

$$h(a,\theta) \propto u(a,\theta) \times p_\theta(\theta \mid x, a).$$

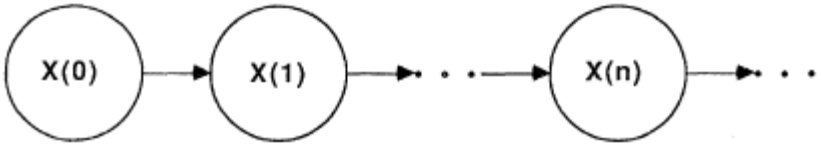
La marginal de la distribución artificial es proporcional a la UE

$$h(a) = \int h(a,\theta)\, d\theta_{a,\theta} \propto \Psi(a).$$

lo que sugiere

1. Generate a sample $((\theta^1, a^1), ..., (\theta^m, a^m))$ from density $h(a,\theta)$.
2. Convert it to a sample $(a^1, ..., a^m)$ from the marginal $h(a)$.
3. Find the sample mode.

Cadenas de Markov en tiempo discreto



Transición

$$P_{ij}^{(m,n)} = P\left(X_n = j \mid X_m = i\right)$$

$$P_{ij}^{(m,m+1)} = P\left(X_{m+1} = j \mid X_m = i\right)$$

Caso homogéneo

$$P_{ij} = P\left(X_{m+1} = j \mid X_m = i\right)$$

$$P_{ij}^{n} = P\left(X_{n+m} = j \mid X_m = i\right)$$

$$P_{ij}^{n+m} = \sum_{k=0}^{\infty} P_{ik}^{n} P_{kj}^{m}$$

Distribución estacionaria, si existe

$$\pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij}, j \geq 0, \text{ with } \sum_{i=0}^{\infty} \pi_i = 1$$

Inferencia y predicción con CMs

- Finita
- Homogénea

$$p_{ij} = P(X_n = j | X_{n-1} = i)$$

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathrm{P},\ \pi_i \geq 0, \sum p_i = 1$$

- Construir modelo gráfico para inferencia y predicción con CMs

Inferencia y predicción con CMs

- Suponemos estado inicial conocido. Observamos primeras m transiciones

$$X_1 = x_1, \ldots, X_m = x_m$$

- Verosimilitud

$$l(\mathbf{P}|\mathbf{x}) = \prod_{i=1}^{K} \prod_{j=1}^{K} p_{ij}^{n_{ij}}$$

- MLE

$$\hat{p}_{ij} = \frac{n_{ij}}{n_{i.}}$$

Inferencia y predicción con CMs

- A priori

$$\mathbf{p}_i \sim \mathrm{Dir}(\boldsymbol{\alpha}_i)$$

- A posteriori

$$\mathbf{p}_i|\mathbf{x} \sim \mathrm{Dir}(\boldsymbol{\alpha}_i') \quad \text{where } \alpha_{ij}' = \alpha_{ij} + n_{ij}$$

- A priori de Jeffreys

$$\boldsymbol{\alpha}_{ij} = 1/2$$

Ejemplo

- Lluvia en una estación (2, lluvia; 1, no lluvia) (Feb 1- Marzo 20)

2	2	2	2	2	2	2	2	2	2
1	1	2	1	1	1	1	1	1	1
2	2	1	1	1	1	2	2	2	1
2	1	1	1	1	1	2	1	1	1
1	1	1	1	1	1	1	1	1	

- Modelo

$$\mathrm{P} = \begin{pmatrix} p_{11} & 1-p_{11} \\ 1-p_{22} & p_{22} \end{pmatrix}.$$

- Prior (Jeffreys)

$$p_{ii} \sim \mathrm{Be}(1/2, 1/2).$$

Ejemplo

- Condicionando a que el primer llueve, las a posteriori son

$$p_{11}|\mathbf{x} \sim \text{Be}(25.5, 5.5) \quad p_{22}|\mathbf{x} \sim \text{Be}(12.5, 6.5).$$

- La matriz de transición esperada

$$E[\mathbf{P}|\mathbf{x}] = \begin{pmatrix} 0.823 & 0.177 \\ 0.342 & 0.658 \end{pmatrix}.$$

Inferencia y predicción con CMs

- Predicción a un paso

$$
\begin{aligned}
P(X_{n+1} = j|\mathbf{x}) &= \int P(X_{n+1} = j|\mathbf{x}, \mathbf{P}) f(\mathbf{P}|\mathbf{x})\, d\mathbf{P} \\
&= \int p_{x_n j} f(\mathbf{P}|\mathbf{x})\, d\mathbf{P} \\
&= \frac{\alpha_{x_n j} + n_{x_n j}}{\alpha_{x_n \cdot} + n_{x_n \cdot}}
\end{aligned}
$$

$\alpha_{i\cdot} = \sum_{j=1}^{K} \alpha_{ij}$.

- Predicción a T pasos

$$P(X_{n+t} = j|\mathbf{x}) = \int \left(\mathbf{P}^t\right)_{x_n j} f(\mathbf{P}|\mathbf{x})\, d\mathbf{P}$$

For $s = 1, \ldots, S$:

1. Generate $\mathbf{P}^{(s)}$ from $f(\mathbf{P}|\mathbf{x})$.
2. Generate $x_{n+1}^{(s)}, \ldots, x_{n+t}^{(s)}$ from the Markov chain with $\mathbf{P}^{(s)}$ and initial state x_n.

$$P(X_{n+t} = j|\mathbf{x}) \approx \frac{1}{S} \sum_{s=1}^{S} I\left(x_{n+t}^{(s)} = j\right)$$

$$E[X_{n+t}|\mathbf{x}] \approx \frac{1}{S} \sum_{s=1}^{S} x_{n+t}^{(s)}$$

Inferencia y predicción con CMs

- Predicción largo plazo

$$\mathbf{P} = \begin{pmatrix} p_{11} & 1-p_{11} \\ 1-p_{22} & p_{22} \end{pmatrix}.$$

$$\pi_1 = \frac{1-p_{22}}{2-p_{11}-p_{22}}$$

$$E[\pi_1|\mathbf{x}] = \int_0^1 \int_0^1 \frac{1-p_{22}}{2-p_{11}-p_{22}} f(p_{11}, p_{22}|\mathbf{x})\, dx$$

- Alternativamente, simulación

Ejemplo

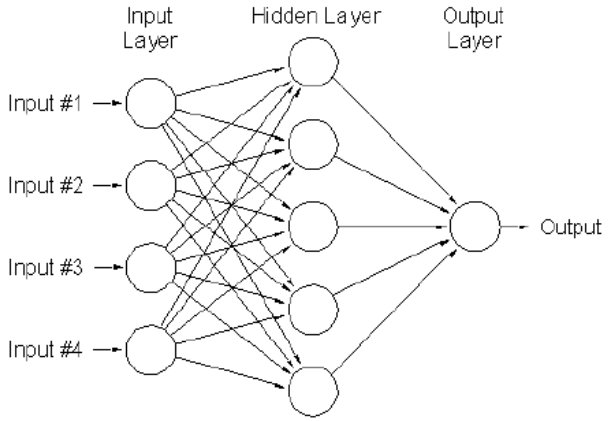
- No llueve 20 Marzo

$$
\begin{aligned}
P(\text{no rain on 21st March}|\mathbf{x}) &= E[p_{11}|\mathbf{x}] = 0.823, \\
P(\text{no rain on 22nd March}|\mathbf{x}) &= E\left[p_{11}^2 + p_{12}p_{21}|\mathbf{x}\right] = 0.742, \\
P(\text{no rain on both}) &= E[p_{11}^2|\mathbf{x}] = 0.681.
\end{aligned}
$$

$$
E[\pi_1|\mathbf{x}] = E\left[\frac{1-p_{22}}{2-p_{11}-p_{22}}\middle|\mathbf{x}\right] = 0.655
$$

Computación bayesiana para redes neuronales

$$\hat{y}(x) = \sum_{j=1}^{M} \beta_j \psi(x'\gamma_j + \delta_j)$$



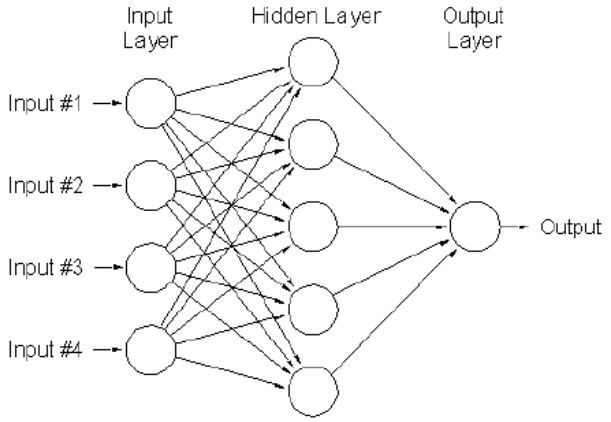
$$y_i = \sum_{j=1}^{M} \beta_j \psi(x_i'\gamma_j) + \epsilon_i, \quad i = 1, \ldots, N,$$

$$\epsilon_i \sim N(0, \sigma^2), \quad \psi(\eta) = \exp(\eta)/(1 + \exp(\eta))$$

$$\beta_j \sim N(\mu_\beta, \sigma_\beta^2), \quad \gamma_j \sim N(\mu_\gamma, S_\gamma), \quad j = 1, \ldots, M$$

Computación bayesiana para redes neuronales

$$\hat{y}(x) = \sum_{j=1}^{M} \beta_j \psi(x'\gamma_j + \delta_j)$$



$$y_i = \sum_{j=1}^{M} \beta_j \psi(x_i'\gamma_j) + \epsilon_i, \quad i = 1, \ldots, N,$$

$$\epsilon_i \sim N(0, \sigma^2), \quad \psi(\eta) = \exp(\eta)/(1 + \exp(\eta))$$

$$\beta_j \sim N(\mu_\beta, \sigma_\beta^2), \quad \gamma_j \sim N(\mu_\gamma, S_\gamma), \quad j = 1, \ldots, M.$$

$$\mu_\beta \sim N(a_\beta, A_\beta), \mu_\gamma \sim N(a_\gamma, A_\gamma), \sigma_\beta^{-2} \sim Gamma(c_b/2, c_b C_b/2).$$

$$S_\gamma^{-1} \sim Wish(c_\gamma, (c_\gamma C_\gamma)^{-1}), \text{ and } \sigma^{-2} \sim Gamma(s/2, sS/2).$$

Computación bayesiana para redes neuronales

1. Start with θ equal to some initial guess (for example, the prior means).

Until convergence is achieved, iterate through steps 2 through 4:

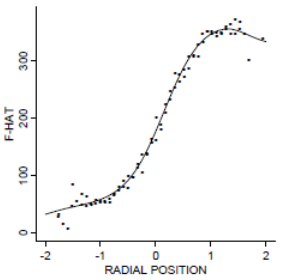
2. Given current values of ν only, (marginalizing over β) replace γ by Metropolis steps: For each γ_j, $j = 1, \ldots, M$, generate a proposal $\tilde{\gamma}_j \sim g_j(\gamma_j)$, with $g_j(\gamma_j)$ described below. Compute

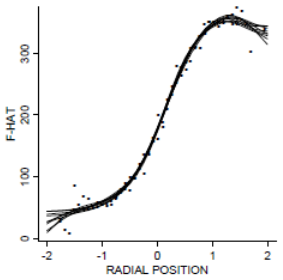
$$a(\gamma_j, \tilde{\gamma}_j) = \min\left[1, \frac{p(D|\tilde{\gamma}, \nu)p(\tilde{\gamma}|\nu)}{p(D|\gamma, \nu)p(\gamma|\nu)}\right], \tag{2.4}$$

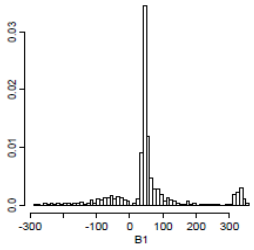
where $\tilde{\gamma} = (\gamma_1, \ldots, \gamma_{j-1}, \tilde{\gamma}_j, \gamma_{j+1}, \ldots, \gamma_M)$. With probability $a(\gamma_j, \tilde{\gamma}_j)$ replace γ_j by the new candidate $\tilde{\gamma}_j$. Otherwise leave γ_j unchanged. Use Lemma 2.1 to evaluate $p(D|\gamma, \nu)$.

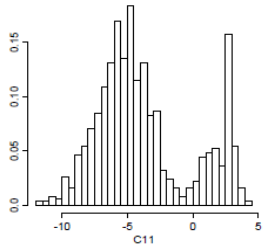
3. Given current values of (γ, ν), generate new values for β by a draw from the complete conditional $p(\beta|\gamma, \nu, D)$. This is a multivariate normal distribution with moments described in Lemma 2.1.

4. Given current values of (β, γ), replace the hyperparameters by a draw from the respective complete conditional posterior distributions: $p(\mu_\beta|\beta, \sigma_\beta)$ is a normal distribution, $p(\mu_\gamma|\gamma, S_\gamma)$ is multivariate normal, $p(\sigma_\beta^{-2}|\beta, \mu_\beta)$ is a Gamma distribution, $p(S_\gamma^{-1}|\gamma, \mu_\gamma)$ is Wishart, and $p(\sigma^{-2}|\beta, \gamma, y)$ is Gamma, as corresponds to a normal linear model. (See Bernardo & Smith, 1994).









$$\hat{f}(x) = \hat{E}(y_{n+1}|x_{n+1}, D) = \frac{1}{k}\sum_{t=1}^{k} E(y_{N+1}|x_{n+1}, \theta = \theta_t)$$

Computación bayesiana para redes neuronales

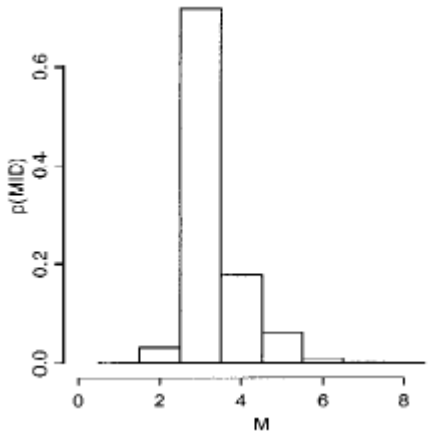
$$y_i = x_i'\lambda + \sum_{j=1}^{M^*} d_j \beta_j \psi(x_i'\gamma_j) + \epsilon_i,\ i = 1, \ldots, N,$$

$$\epsilon_i \sim N(0, \sigma^2), \quad \psi(\eta) = \exp(\eta)/(1 + \exp(\eta))$$

$$\gamma_{1p} \leq \gamma_{2p} \cdots \leq \gamma_{Mp}$$

$$Pr(d_j = d | d_{j-1} = 1) = \begin{cases} 1 - \alpha, & \text{for } d = 0 \\ \alpha & \text{for } d = 1 \end{cases} \quad j = 1, \ldots, M^*,$$

$$\beta_j \sim N(\mu_\beta, \sigma_\beta^2), \quad \lambda \sim N(\mu_\beta, \sigma_\beta^2),$$
$$\gamma_j \sim N(\mu_\gamma, S_\gamma), \quad \alpha \sim Beta(a_\alpha, b_\alpha).$$



Hyperpriors

Ventajas métodos bayesianos (French, DRI, 2000)

- Se tiene en cuenta toda la información t
- Base axiomática, marco coherente
- Incertidumbre repartida y reconocida
- Transparente al usuario
- Robusta, robusta frente a ataques
- Factible

Algunos debates recientes

- ASA statement on statistical significance and p-values (2016)

P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

4. Proper inference requires full reporting and transparency

5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.

Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning.

Ventajas métodos bayesianos (French, DRI, 2000)

longer major driving forces in model building. It is interesting to chart the history of applied Bayesian methods through the proceedings of the Valencia conferences from their beginnings in 1979 to their most recent in 1998. The balance has shifted from conceptual and analytical issues in theoretical models to computational aspects of applied studies. Today Bayesian methods are most certainly practicable. Indeed, for complicated models, Bayesian analysis has arguably now become the simplest (and often the only possible) method of analysis.

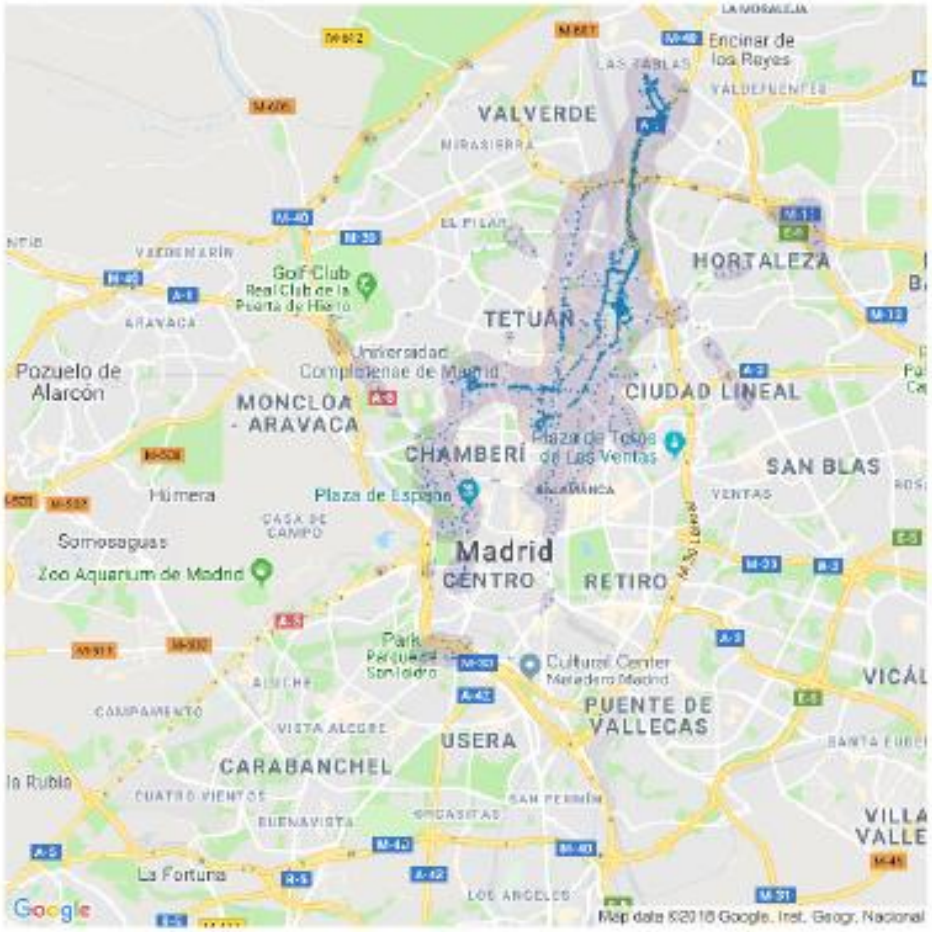
Ventajas métodos bayesianos (HOY)

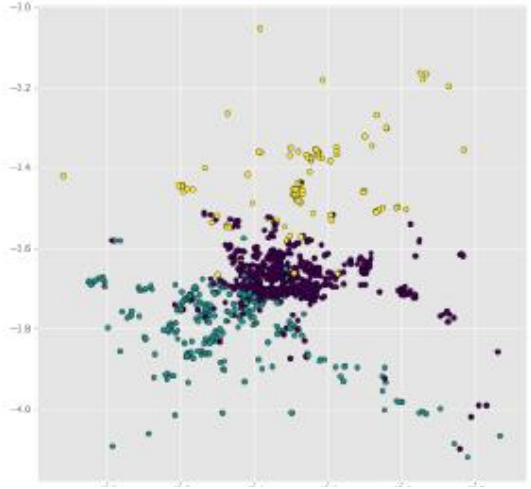
- Se tiene en cuenta toda la información t
- Base axiomática, marco coherente
- Incertidumbre repartida y reconocida
- Transparente al usuario
- Robusta, robusta frente a ataques
- Factible…

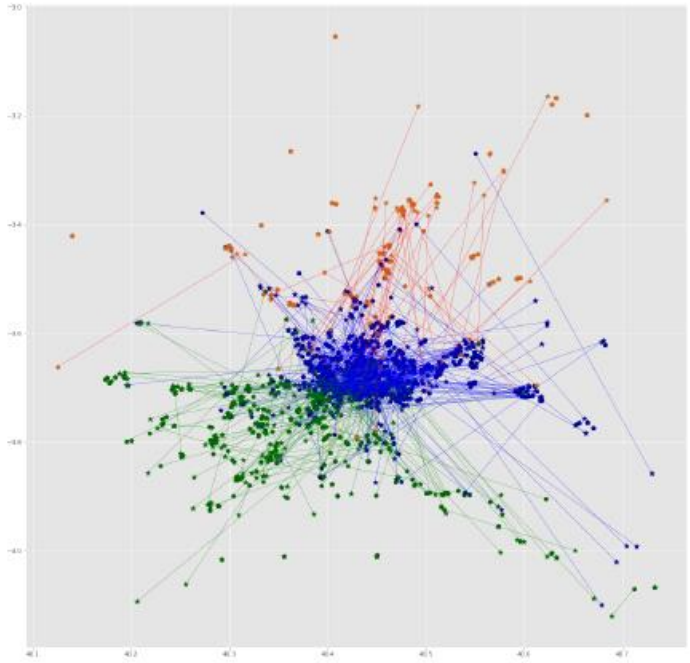
AA y BD

- Volumen
- Variedad. Texto, imágenes, sonido, video,….
- Velocidad. Alta frecuencia, series temporales, modelos dinámicos
- Valor

AA y BD


VALVERDE
Encinar de los Reyes
HORTALEZA
TETUÁN
CIUDAD LINEAL
Pozuelo de Alarcón
MONCLOA - ARAVACA
CHAMBERÍ
SAN BLAS
Húmera
Plaza de España
Somosaguas
Zoo Aquarium de Madrid
Madrid
CENTRO
RETIRO
PUENTE DE VALLECAS
USERA
CARABANCHEL
VILLA VALLE
La Fortuna
Google
Map data ©2018 Google, Inst. Geogr. Nacional





Bayes y BD

1. Start with arbitrary values $(\mathbf{q}^0, \boldsymbol{\mu}^0, \mathbf{z}^0)$, $i = 0$.
2. Until convergence, iterate through
   - Generate $\mathbf{z}_j^{i+1} \sim \mathbf{z}_j | t_j, \mathbf{q}^i, \boldsymbol{\mu}^i$, $j = 1, ..., n_s$.
   - Generate $\mathbf{q}^{i+1} \sim \mathbf{q} | \mathbf{t}, \mathbf{z}^{i+1}$.
   - Generate $\mu_j^{i+1} \sim \mu_j | \mathbf{t}, \mathbf{z}^{i+1}$, $j = 1, ..., k$.
   - Set $i = i + 1$.

Bayes y BD

1. Start with θ equal to some initial guess (for example, the prior means).

Until convergence is achieved, iterate through steps 2 through 4:

2. Given current values of ν only, (marginalizing over β) replace γ by Metropolis steps: For each γ_j, $j = 1, \ldots, M$, generate a proposal $\tilde{\gamma}_j \sim g_j(\gamma_j)$, with $g_j(\gamma_j)$ described below. Compute

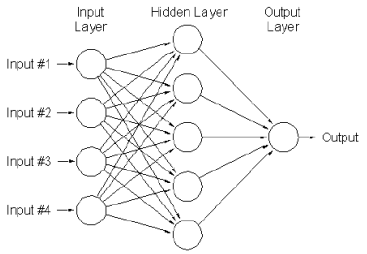
$$a(\gamma_j, \tilde{\gamma}_j) = \min\left[1, \frac{p(D|\tilde{\gamma}, \nu)p(\tilde{\gamma}|\nu)}{p(D|\gamma, \nu)p(\gamma|\nu)}\right], \tag{2.4}$$

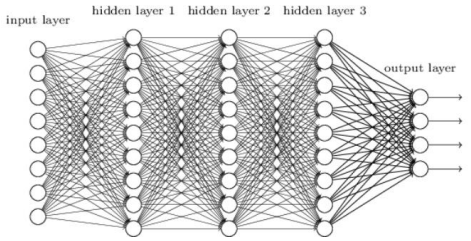
where $\tilde{\gamma} = (\gamma_1, \ldots, \gamma_{j-1}, \tilde{\gamma}_j, \gamma_{j+1}, \ldots, \gamma_M)$. With probability $a(\gamma_j, \tilde{\gamma}_j)$ replace γ_j by the new candidate $\tilde{\gamma}_j$. Otherwise leave γ_j unchanged. Use Lemma 2.1 to evaluate $p(D|\gamma, \nu)$.

3. Given current values of (γ, ν), generate new values for β by a draw from the complete conditional $p(\beta|\gamma, \nu, D)$. This is a multivariate normal distribution with moments described in Lemma 2.1.

4. Given current values of (β, γ), replace the hyperparameters by a draw from the respective complete conditional posterior distributions: $p(\mu_\beta|\beta, \sigma_\beta)$ is a normal distribution, $p(\mu_\gamma|\gamma, S_\gamma)$ is multivariate normal, $p(\sigma_\beta^{-2}|\beta, \mu_\beta)$ is a Gamma distribution, $p(S_\gamma^{-1}|\gamma, \mu_\gamma)$ is Wishart, and $p(\sigma^{-2}|\beta, \gamma, y)$ is Gamma, as corresponds to a normal linear model. (See Bernardo & Smith, 1994).

$$\hat{f}(x) = \hat{E}(y_{n+1}|x_{n+1}, D) = \frac{1}{k}\sum_{t=1}^{k} E(y_{N+1}|x_{n+1}, \theta = \theta_t)$$





Bayes y BD

- Cuellos de botella computacionales por iteración
  - Evaluar la verosimilitud
  - Visitar todos los parámetros
  - Visitar todos los datos
- Tasa lenta de mezcla

Bayes y BD

- MLE . Optimizar

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\hat{p}_{\mathrm{data}}} L(\boldsymbol{x}, y, \boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} L(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta})$$

$$L(\boldsymbol{x}, y, \boldsymbol{\theta}) = -\log p(y \mid \boldsymbol{x}; \boldsymbol{\theta})$$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta})$$

Bayes y BD

- MLE +regularizador . Optimiza

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\hat{p}_{\mathrm{data}}} L(\boldsymbol{x}, y, \boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} L(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta}) \quad +h(\theta)$$

$$L(\boldsymbol{x}, y, \boldsymbol{\theta}) = -\log p(y \mid \boldsymbol{x}; \boldsymbol{\theta})$$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta}) \quad +\nabla_{\theta} h(\theta)$$

MAP. Incertidumbre….

Inferencia aproximada

- Convierte el problema de inferencia en uno de optimización
- Más rápido que MCMC (tal y como se conoce hoy)
- Escala mejor para conjuntos grandes de datos (tal y como se conoce hoy)
- Puede subestimar incertidumbre

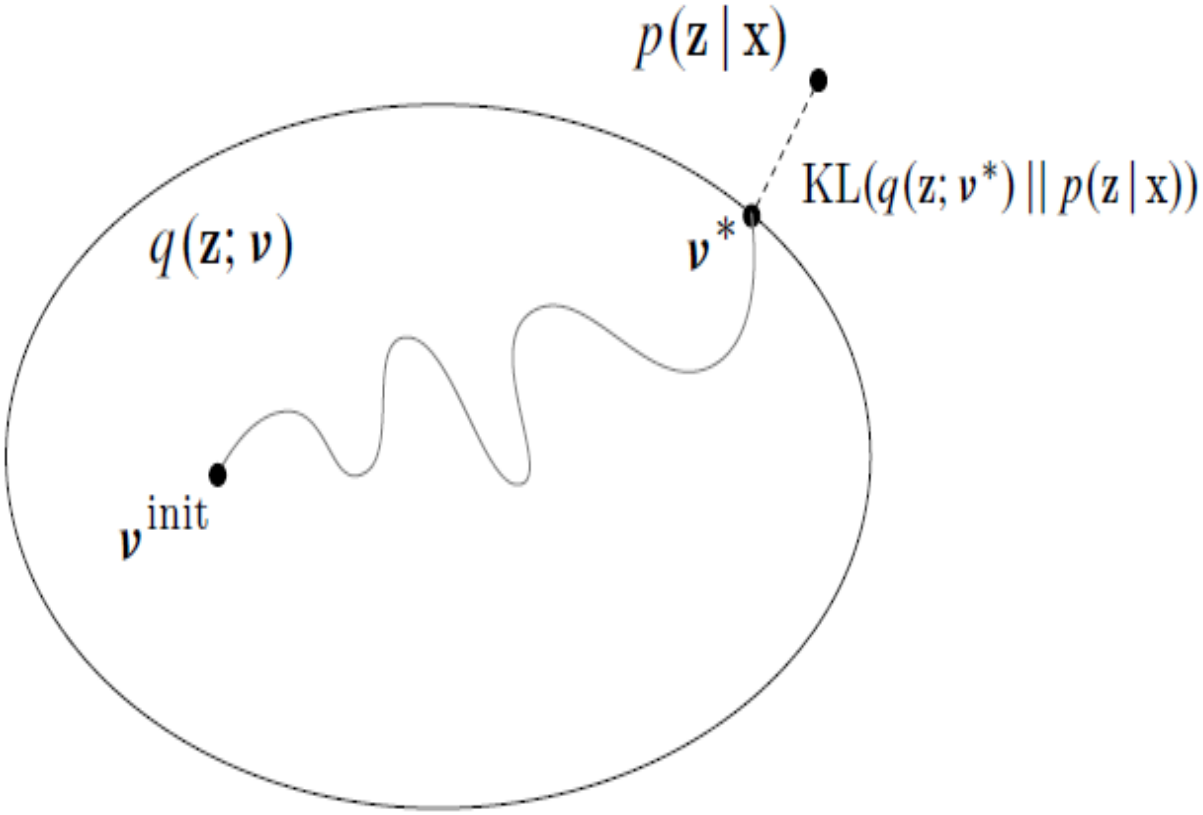
Concepto de inferencia variacional

- Queremos estimar $p(\mathbf{z}|\mathbf{x})$
- Aproximamos mediante una distribución que sea de cálculo sencillo $q(\mathbf{z})$
- Empleamos optimización para que se parezcan
- Usamos la divergencia de Kullback-Leibler

Medida de disimilaridad entre dos distribuciones
No negativa
0 si y sólo si coinciden
Lo que vamos a querer hacer es computable!!!

$$KL(q||p) = \int q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})} d\mathbf{z}$$

Concepto de inferencia variacional. Inferencia como optimización



Familia variacional $q(\mathbf{z}; \nu)$

Encontrar los parámetros variacionales ν para aproximarse lo más posible en KL a la distribución a posteriori

Concepto de inferencia variacional. ELBO

La optimización directa de KL no es posible pues aparece la evidencia $KL(q||p) \equiv \mathbb{E}_q[\log q(\mathbf{z})] - \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{z})] + \log p(\mathbf{x})$

$$ELBO(q) = \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{z})] - \mathbb{E}_q[\log q(\mathbf{z})]$$

Podemos quitar el último término. Evidence lower bound (Cota inferior de la evidencia)

$$\log p(\mathbf{x}) \geq ELBO(q)$$

$$q^\star(\mathbf{z}) = \arg\min_{q \in \mathcal{Q}} KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}))$$

$$q^\star(\mathbf{z}) = \arg\max_{q \in \mathcal{Q}} ELBO(q)$$

Algoritmos de inferencia variacional

1. Formular el modelo

2. Formular la familia variacional
3. Formular el problema de optimización
4. Resolver el problema de optimización

$$q^{\star}(\mathbf{z}) = \arg\max_{q \in \mathcal{Q}} ELBO(q)$$

Receta para inferencia variacional

Formulamos modelo

$$p(\mathbf{z}, \mathbf{x})$$

Escogemos aproximación variacional

$$q(\mathbf{z}; \boldsymbol{\nu})$$

Escribimos y calculamos ELBO

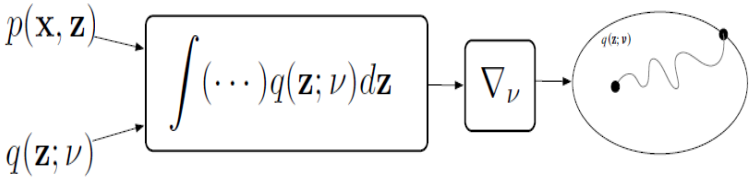
$$\mathcal{L}(\boldsymbol{\nu}) = \mathbb{E}_{q(\mathbf{z};\boldsymbol{\nu})}[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z}; \boldsymbol{\nu})]$$

Calculamos gradiente

$$\nabla_{\nu} \mathcal{L}(\boldsymbol{\nu})$$

Optimizamos

$$\boldsymbol{\nu}_{t+1} = \boldsymbol{\nu}_t + \rho_t \nabla_{\nu} \mathcal{L}$$



Paralelizar MCMC

- Procesamiento paralelo: divide tareas en subtareas ejecutadas en paralelo
- MC se paraleliza de forma trivial

$$\widehat{h} = \frac{1}{N} \sum_{i=1}^{N} h(X_i)$$

- Divide la suma en P>2 componentes y asigna un procesador para evaluar cada componente
- MCMC más difícil de paralelizar. Los elementos de la sucesión no son independientes. Necesitamos X_i para calcular X_{i+1}
- Ejecutar varias cadenas en varios procesadores y mezclar (Scott et al 2013)
- Partir datos en conjuntos no solapantes y mezclar (Wang y Dunson, 2016)

Acelerar MCMC

Stochastic gradient Langevin dynamics (SGLD) (Teh et al, 2016)

Hamiltonian Monte Carlo (Chen et al, 2014)

SGLD+Repulsion (Gallego et al, 2018)

Algorithm 1 Bayesian Inference via SGLD+R

Input: A target distribution with density function $\pi(\boldsymbol{z}) \propto \exp(-H(\boldsymbol{z}))$.
Output: A set of particles $\{\boldsymbol{z}_i\}_{i=1}^{MK}$ that approximates the target distribution.
Sample initial set of particles from prior: $\boldsymbol{z}_1^0, \boldsymbol{z}_2^0, \dots \boldsymbol{z}_K^0 \sim \pi(\boldsymbol{z})$.
for each iteration t **do**

$$\boldsymbol{z}_i^{t+1} \leftarrow \boldsymbol{z}_i^t - \epsilon_t \frac{1}{K} \sum_{j=1}^{K} \left[k(\boldsymbol{z}_j^t, \boldsymbol{z}_i^t) \nabla_{\boldsymbol{z}_j^t} H(\boldsymbol{z}_j^t) + \nabla_{\boldsymbol{z}_j^t} k(\boldsymbol{z}_j^t, \boldsymbol{z}_i^t) \right] + \boldsymbol{\eta}_i^t \tag{6}$$

where $\boldsymbol{\eta}_i^t$ is the noise for particle i defined as in Eq (5).
After a burn-in period, start collecting particles: $\{\boldsymbol{z}_i\}_{i=1}^{NK} \leftarrow \{\boldsymbol{z}_i\}_{i=1}^{(N-1)K} \cup \{\boldsymbol{z}_1^{t+1}, \dots, \boldsymbol{z}_K^{t+1}\}$
end for

Explotar información del gradiente para generar muestras alejadas del valor actual y con alta densidad

Inferencia bayesiana y estadística oficial

- Inferencia bayesiana, subjetiva
- Estadística oficial, objetiva
- "Nonetheless, the principal challenge to Bayesian methods that remains is the need to constantly rebut the notion that frequentist methods are 'objective' and thus are more appropriate for use in the public domain."

(Fienberg Statist Sci, 2011, 26, 212-226.)

- A prioris no informativas
- Bayesiano calibrado (Little) para EO. A priori poco informativa+ Análisis de sensibilidad
- A priori informativa, justificable y abierta a críticas

Inferencia bayesiana y estadística oficial

- Estimación áreas pequeñas
- Datos administrativos que no provienen de muestreo aleatorio
- Demografía
- Diseño muestral
- Predicción
- Imputación
- Monitorización de producción de datos

Infromación adicional

- Modelos gráficos probabilísticos, 10
- DLMs, 16
- Algos bayesianos para clasificación y regresión, 17

- Redes neuronales
- Naive Bayes (Algos clasificación…)

Información adicional

Jaynes. Probability theory. The logic of science
French y Rios Insua. Statistical Decision Theory
Geman, Carlin, Stern, Dunson, Vehtari, Rubin Bayesian Data Analysis
Bishop Pattern Recognition and Machine Learning
Rios Insua, Ruggeri, Wiper Bayesian Analysis of Stochastic Processes
DRI, Ríos, Martín, Jiménez Simulación: Métodos y aplicaciones

Blei, Kucucelbir, MacAulliffe Variational inference a review for statisticians

Jordan et al. An introduction to variational methods for graphical models

Fox, Roberts. A tutorial on variational Bayesian inference

Blei, Kucucelbir, MacAulliffe Variational inference a review for statisticians

Jordan et al. An introduction to variational methods for graphical models

Fox, Roberts. A tutorial on variational Bayesian inference

http://bayesiandeeplearning.org/ Neurips 2018
https://arxiv.org/pdf/1601.00670.pdf Variational inference
https://www.youtube.com/watch?v=KxV5kckOVeA
https://www.youtube.com/watch?v=L1Q7w3ch3
https://www.youtube.com/watch?v=OWjWYyG4Oys

Gracias!!

david.rios@icmat.es

SPOR DataLab https://www.icmat.es/spor/
It's a risky life @YouTube
Aisoy Robotics https://www.aisoy.com
CYBECO https://www.cybeco.eu/