

Elvira II - Reunión de Albacete

Serafín Moral

Elvira II

- Entorno Elvira
 - Extensión del formato - Interfaz gráfica
 - Algoritmos de Inferencia
 - Preprocesamiento de datos
 - Algoritmos de aprendizaje

Elvira II

- Entorno Elvira
 - Extensión del formato - Interfaz gráfica
 - Algoritmos de Inferencia
 - Preprocesamiento de datos
 - Algoritmos de aprendizaje
- Aplicaciones
 - Datos agrícolas
 - Aplicaciones Médicas
 - Datos de expresión genética
 - Filtrado cooperativo

Tarea A.1: Formato Elvira

- Optimización del lector: M1-M3
- Lectura de otros formatos: M4-M12
- Lector de redes en formato XML: M15-M18

Falta por hacer:

- Letra ñ
- Variables gaussianas
- Restricciones entre variables

Propuestas

- Leer fichero en memoria
- Utilizar estructuras intermedias

Propuestas

- Leer fichero en memoria
- Utilizar estructuras intermedias
- Variables gaussianas
 - representations
 - $\text{values} = G(\text{LinearFunction}, \text{Float});$
 - $\text{values} = \text{float1} * G(\text{lf1}, \text{float1}) + \dots + \text{floatk} * G(\text{lfk}, \text{floatk})$

Propuestas

- Leer fichero en memoria
- Utilizar estructuras intermedias
- Variables gaussianas
 - representations
 - $\text{values} = G(\text{LinearFunction}, \text{Float});$
 - $\text{values} = \text{float1} * G(\text{lf1}, \text{float1}) + \dots + \text{floatk} * G(\text{lfk}, \text{floatk})$
- Restricciones: Manolo

A.2 Diagramas de Influencia con Árboles

M1-M5 Estudio

M6-M12 Implementación

- Representar los potenciales con árboles
- Útiles para las restricciones
- Permiten aproximaciones
- La utilidad se aproxima de forma distinta a la probabilidad

A.3 Preprocesamiento

Responsable: UPV

M1-M12 Estudio

M13-M18 Implementación

- Implementación de algoritmos de discretización: misma frecuencia con intervalos dados.
- Haremos otros como los basados en el principio de mínimo tamaño de descripción.
- Estamos interesados en imputación de valores y selección de variables.

A.4 Aprendizaje

- Aprendizaje con conocimiento parcial:
preórdenes, estructura.
M1-M3 Estudio
M4-M12 Implementación
- Aprendizaje de Modelos Gráficos Dinámicos
M19-M28 Estudio
M28-M36 Implementación

Conocimiento Parcial

- Alguna forma de almacenar información parcial
- Restricciones entre variables (si una variable toma un valor, entonces otra es imposible).
- Jerarquías

A5 Interfaz

- Especificación y visualización de variables continuas en el interfaz gráfico
M4-M9 Implementación
- Preguntar por la forma (MixExpTree o Mixtura de Gaussianas)
- Para MixExpTree determinar la estructura de árbol. Cada nodo una ventana. Cada ventana permite crear las ventanas de sus hijos.

A5 Interfaz 2

- Preguntar por el número de sumandos.
- Para cada uno de ellos, pedir el coeficiente y la función lineal de los padres (un coeficiente para cada variable padre).
- En el caso de normal se pide la varianza.

B Aplicaciones

B.4: Aplicaciones de los Modelos Gráficos al Análisis de Datos de Microarrays

- M7-M12 Discretización
- M13-M18 Mixtura de Gaussianas
- M19-M24 Mixtura de Exponenciales Truncadas
- M25-M31 Modelos temporales, hibridación
- M32-M36 Experimentación

Trabajos Previos

Nir Friedman y otros (2000-2001) Using Bayesian Networks to Analyze Expression Data.

- Datos del ciclo de la levadura.
- Modelos multinomiales y Gaussianos.
- Para los multinomiales se divide en tres intervalos: un control x y valores $2^{-0,5} * x$ y $2^{0,5} * x$ como puntos de corte.
- Utilizan el score bayesiano con la condición de equivalencia

Friedman

- Restringen los posibles candidatos de un padre mediante correlación
- Es iterativo readaptando los posibles candidatos: los padres de una iteración son candidatos en la siguiente
- Usan *bootstrap* para estimar hechos a partir de los datos. Generan versiones ruidosas de los datos y aprenden a partir de ellas.
- Variable el estado del ciclo de la levadura
- Comprueba la frontera de Markov y relaciones de orden.

Yoo, Thorsson, Cooper, 2002

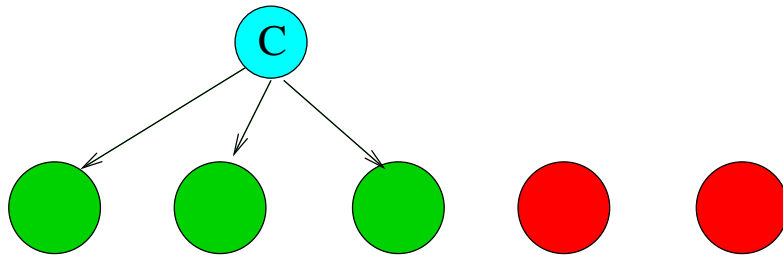
- Utilizan datos manipulados y observados para inferir relaciones causales
- Comprueban la existencia de relaciones causales entre cada par de variables de forma independiente
- Discriminan el tipo de relación y la posible existencia de una variable oculta, mediante una combinación de los scores con los dos tipos de datos

Imoto, Goto, Miyano, 2002

- Usan un modelo de variables continuas similar al gaussiano lineal, pero en el que existen unas funciones fijas (bases de Fourier, splines,...) que se usan en la combinación lineal.
- Estas funciones pueden variar de nodo a nodo
- Desarrollan una aproximación para la verosimilitud marginal
- Resultados (redes) similares a las de Friedman.

Brash, Friedman, 2001

- Un método de cluster
- Incluye variables que cuentan en número de promotores de los genes
- Consideran un modelo naive bayes con selección de variables



Independencias Asimétricas

	X=0	X=1	X=2	X=3
C=1	0.1	0.1	0.5	0.3
C=2	0.1	0.1	0.2	0.6
C=3	0.1	0.1	0.2	0.6

Independencias Asimétricas

	X=0	X=1	X=2	X=3
C=1	0.1	0.1	0.5	0.3
C=2	0.1	0.1	0.2	0.6
C=3	0.1	0.1	0.2	0.6

	X=0	X=1	X=2	X=3
C=1	0.1	0.1	0.5	0.3
C=*	0.1	0.1	0.2	0.6

Algoritmo de aprendizaje

- Aprenden una red con 3, ..., 7 clusters y se quedan con la mejor.
- Variables gaussianas y discretas
- Para cada caso se aplica un algoritmo EM estructural: comienza con un modelo; calcula la esperanza de los estadísticos suficientes para el Score a partir de los datos; aprende un modelo y unos datos (un EM paramétrico); vuelve a repetir.
- Tiene procedimientos para salir de máximos locales.

Plan para este año

- Procedimientos de discretización.
- Estudio de las variables continuas: test de normalidad, mixtura de dos gaussianas.
- Un modelo sencillo para imputar valores: naive bayes, árboles de clasificación
- Estudio si se pueden imputar valores (sólo un valor con probabilidad alta).
- Algunas redes discretas
- Selección de variables (pasos previos).