

-
-
-

Clasificación no supervisada de datos de expresión genética con modelos gráficos probabilísticos

Intelligent System Group

Departamento de Ciencias de la Computación e Inteligencia Artificial

Universidad del País Vasco

<http://www.sc.ehu.es/isg/>

Albacete, 17-19 de Junio de 2002

Presentación

- Datos de expresión genética
- Aprendizaje no supervisado de:
 - Modelos discretos: naive-Bayes extendidos, naive-Bayes aumentados a árboles, multiredes Bayesianas recursivas
 - Modelos continuos: redes condicionales Gaussianas

Introducción I

- Clasificación no supervisada: Búsqueda de la mejor descripción de los grupos en d
- Suposición básica:
 - Desconocimiento del clúster al que cada instancia en d realmente pertenece.
- $d = \{x_1, \dots, x_N\}$ con $x_l = (x_{l1}, \dots, x_{ln+1}) = (c_l, y_l)$, donde:
 - c_l denota el clúster (desconocido) al que la instancia realmente pertenece.
 - $y_l = (y_{l1}, \dots, y_{ln})$ denota el vector, discreto o continuo, de atributos predictivos.

Introducción II

- Existencia de una variable aleatoria $\mathbf{X} = (X_1, \dots, X_{n+1}) = (C, \mathbf{Y})$ donde:
 - C denota la variable aleatoria clúster.
 - $\mathbf{Y} = (Y_1, \dots, Y_n)$ denota la variable aleatoria, discreta o continua, de atributos predictivos

Aprendizaje no supervisado de MGP

- Fases del aprendizaje: estructural y paramétrico
- Criterio o score Bayesiano: verosimilitud marginal $L(\mathbf{d} | \mathbf{s}^h)$
- Dificultades:
 - Inexistencia de fórmulas cerradas y separables para $L(\mathbf{d} | \mathbf{s}^h)$
 - Inexistencia de fórmulas cerradas para el aprendizaje paramétrico
- Soluciones:
 - Aproximaciones para $L(\mathbf{d} | \mathbf{s}^h)$
 - Algoritmo EM (aprendizaje estructural)
 - Algoritmo EM (aprendizaje paramétrico)

Algoritmo EM en el aprendizaje paramétrico

- Asignar un valor inicial a los parámetros del modelo $\hat{\theta}_0$
- Estimar la pertenencia de cada dato a cada clase, mediante inferencia probabilística $p(C = c_g | \mathbf{y}_l, \hat{\theta}_u, \mathbf{s}^h)$ (Expectation)
- Calcular el valor del nuevo conjunto de parámetros $\hat{\theta}_{u+1}$ a partir de los valores anteriores por máxima verosimilitud (Maximization)

Algoritmo EM estructural Bayesiano BS-EM

- Seleccionar una estructura y unos parámetros iniciales
- Aplicar el algoritmo EM de aprendizaje paramétrico
- Realizar una búsqueda estructural evaluando cada estructura como una esperanza sobre todas las posibles complexiones de la base de datos
- Aplicar el algoritmo EM paramétrico a la estructura obtenida

Modelos aprendidos con BS-EM

- modelos Naive-Bayes
- modelos Naive-Bayes extendidos a árboles
- multiredes Bayesianas recursivas
- redes condicionales Gaussianas

Aprendizaje de RG mediante exclusión de arcos

- Test de la razón de verosimilitud (Smith and Whittaker, 1998):

$$T_l = -N(1 - r_{ij|rest}^2)$$

- $r_{ij|rest}$ es la correlación parcial entre las variables X_i y X_j ajustada por el resto de las variables
- $r_{ij|rest} = -\hat{w}_{ij}(\hat{w}_{ii}\hat{w}_{jj})^{-\frac{1}{2}}$

Redes condicionales Gaussianas por medio de tests

- Seleccionar una estructura y unos parámetros iniciales
- Aplicar el algoritmo EM de aprendizaje paramétrico y completar la base de datos con los valores MAP
- Aprender una red Gaussiana para cada clase mediante test de exclusión de arcos
- Combinar las estructuras de las redes Gaussianas en una única estructura

Aprendizaje de RCG. Reducción de la dimensionalidad

- Aproximación filter
- Medida de relevancia:

$$\sum_{\substack{j=1 \\ j \neq i}} \frac{-N \log(1 - r_{ij|rest}^2)}{n - 1}$$

- Umbral de relevancia: la frontera de la region de rechazo de un test de eliminación de un arco en una red Gaussiana