Algortimo TM aplicado a modelos gráficos probabilísticos

Grupo de Sistemas Inteligentes

Departamento de Ciencias de la Computación e Inteligencia Artificial

Universidad del País Vasco

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

Reunión ELVIRA. Mayo 2003

Indice

- Introducción
- Introducción al algoritmo TM
- El algoritmo TM en familias exponenciales
- Aplicación del algoritmo TM en modelos gráficos probabilísticos: un ejemplo

Introducción

Medidas para aprender un clasificador basado en MGPs:

- Basadas en verosimilitud conjunta $p(c, \mathbf{x})$
- Basadas en verosimilitud condicionada $p(c|\mathbf{x})$
- Basadas en la teoría de la información
- Basadas en estimaciones del % de bien clasificados

Distinción: generativos versus discriminativos

Introducción al algoritmo TM

- Introducido por Edwards y Lauritzen (2001) (Biometrika)
- Objetivo: optimizar una verosimilitud condicional cuando la conjunta es más fácil de optimizar
- Resultados sobre convergencia: Sundberg (2002) (Biometrika)
- Similitud con el algoritmo EM, se aumentan los parámetros en lugar de los datos
- Consta de dos pasos: el paso T y el paso M

Introducción al algoritmo TM

$$l(\theta) = \log p(x, c|\theta) \qquad l_x(\theta) = \log p(x|\theta) \qquad l^x(\theta) = \log p(c|x, \theta)$$

- En el paso T se construye la función $q(\theta)$:

$$\begin{aligned} q(\theta) = q(\theta|\theta_r) &= l(\theta) - \theta^T \dot{l}_x(\theta_r) \\ &= l(\theta) - \theta^T E_{\theta_r}\left[\dot{l}(\theta_r)|x\right] \end{aligned}$$

- La función anterior es una aproximación a la verosimilitud condicional ($l^x = l - l_x$)

Introducción al algoritmo TM

- El paso M consiste en la maximización en θ de la función $q(\theta|\theta_r)$:

$$E_{\theta_r}\left[\dot{l}(\theta_r)|x\right] = \dot{l}(\theta)$$

- El algoritmo se puede escribir como:

$$\theta_0 = arg\max_{\theta} l(\theta) \qquad \theta_{r+1} = arg\max_{\theta} q(\theta|\theta_r)$$

TM en la familia exponencial

- La verosimilitud conjunta se puede escribir como ($\theta = (\alpha, \beta)$):

$$l(\theta) = \alpha^T u(c, x) + \beta^T v(x) - \psi(\alpha, \beta)$$

con

$$\psi(\alpha, \beta) = \log \int exp\left[\alpha^T u(c, x) + \beta^T v(x)\right] \mu(dc|x)\mu dx$$

TM en la familia exponencial

- Definimos dos nuevas funciones de variables aleatorias: $\mathcal{U} = u(C, X)$ y $\mathcal{V} = v(X)$
- Las estimaciones máximo verosímiles $\hat{\theta}$ de los parámetros θ se obtienen al resolver las ecuaciones:

$$E_\theta[\mathcal{U}] = u(c, x) \qquad E_\theta[\mathcal{V}] = v(x)$$

TM en la familia exponencial

- El algoritmo queda como sigue:

$$\begin{aligned} u_{r+1} &= u_r + u - E_{\theta_r}[\mathcal{U}|x] \text{ con } u_0 = u(c,x) \\ \theta_{r+1} &= \hat{\theta}(u_{r+1}, v) \end{aligned}$$

- Como primer conjunto de parámetros θ_0 se toman los estimadores máximo verosímiles

Algoritmo TM en MGPs

- Veamos un ejemplo con el Naive-Bayes con variables binarias
- Modelo condicional:

$$
\begin{aligned}
p(c|x) &= \frac{p(c,x)}{p(x)} = \frac{p(c)\prod_{i=1}^{n} p(x_i|c)}{p(x)} \\
&= \frac{p(c)\prod_{i=1}^{n} p(x_i|c)}{p(c)\prod_{i=1}^{n} p(x_i|c) + p(\bar{c})\prod_{i=1}^{n} p(x_i|\bar{c})}
\end{aligned}
$$

Algoritmo TM en MGPs

- Introducimos el modelo condicional en un modelo conjunto:

$$p(c, x) = \frac{1}{(p(c))^{n-1}} \prod_{i=1}^{n} p(x_i, c)$$

Algoritmo TM en MGPs

- Finalmente puede expresarse en función de los parámetros conjuntos:

$$
\begin{aligned}
p(c,x) &= \left(p(C=0)^{(1-c)}(1-p(C=0))^{c}\right)^{-(n-1)} \\
&\quad \prod_{i=1}^{n} \Big[p(C=0,X_i=0)^{(1-c)(1-x_i)}p(C=0,X_i=1)^{(1-c
\\
&\quad p(C=1,X_i=0)^{c(1-x_i)}p(C=1,X_i=1)^{cx_i}\Big] \\
&= \exp\Big\{\log\Big(\left(p(C=0)^{(1-c)}(1-p(C=0))^{c}\right)^{-(n-1)} \\
&\quad \prod_{i=1}^{n} \Big[p(C=0,X_i=0)^{(1-c)(1-x_i)}p(C=0,X_i=1)^{(1-c
\\
&\quad p(C=1,X_i=0)^{c(1-x_i)}p(C=1,X_i=1)^{cx_i}\Big]\Big)\Big\}.
\end{aligned}
$$

Algoritmo TM en MGPs

- Expresamos la log-verosimilitud conjunta como en la familia exponencial: $l(\theta)$

$$
\begin{aligned}
= \quad & \sum_{k=1}^{N} c^{(k)} \left((n-1)(\log(p(C=0)) - \log(1-p(C=0)))\right) \\
+ \quad & \sum_{k=1}^{N} \sum_{i=1}^{n} c^{(k)} x_i^{(k)} \left(\log(p(C=0, X_i=0)) \log(p(C=1, X_i=1))\right. \\
- \quad & \left.\log(p(C=0, X_i=1)) - \log(p(C=1, X_i=0))\right) \\
+ \quad & \sum_{k=1}^{N} \sum_{i=1}^{n} x_i^{(k)} \left(\log(p(C=0, X_i=1)) - \log(p(C=0, X_i=0))\right) \\
+ \quad & N(\sum_{i=1}^{n} \log(p(C=0, X_i=0)) - (n-1)\log(p(C=0)))
\end{aligned}
$$

Algoritmo TM en MGPs

- Identificando términos obtengo los siguientes valores:

$$\mathcal{U} = u(C, X) = \left(\sum_{k=1}^{N} C^{(k)}, \sum_{k=1}^{N} C^{(k)} X_1^{(k)}, \ldots, \sum_{k=1}^{N} C^{(k)} X_n^{(k)} \right)$$

$$\mathcal{V} = v(X) = \left(\sum_{k=1}^{N} X_1^{(k)}, \ldots, \sum_{k=1}^{N} X_n^{(k)} \right)$$

Algoritmo TM en MGPs

- Los estadísticos suficientes a considerar son:

$$\begin{aligned}\mathcal{U} &= (N_0, N_{01}, \ldots, N_{0i}, \ldots, N_{0n}) \\ \mathcal{V} &= (N_1, \ldots, N_i, \ldots, N_n)\end{aligned}$$

Algoritmo TM en MGPs

- Para completar el algoritmo debemos calcular $E_{\theta_r}[\mathcal{U}|x]$:

$$\begin{aligned} E_{\theta_r}[N_0|x] &= \sum_{k=1}^{N} E_{\theta_r}[C^{(k)}|x] = \sum_{k=1}^{N} p_{\theta_r}(C=1|X=x^{(k)}) \\ E_{\theta_r}[N_{0i}|x] &= \sum_{k=1}^{N} E_{\theta_r}[C^{(k)}X_i^{(k)}|x] \\ &= \sum_{k=1}^{N} p_{\theta_r}(C=1|X=x^{(k)})x_i^{(k)} \quad i=1,\ldots,n \end{aligned}$$

Algoritmo TM en MGPs

- En el cáculo anterior se utilizaron probabilidades condicionadas estas se calculan como:

$$
\begin{aligned}
p_r(C=1) &= \frac{N_{0,(r)}}{N} \\
p_r(X_i=1|C=1) &= \frac{N_{0i,(r)}}{N_{0,(r)}} \\
p_r(X_i=1|C=0) &= \frac{N_i - N_{0i,(r)}}{N - N_{0,(r)}}
\end{aligned}
$$

Algoritmo TM en MGPs

i) A partir de $N_{0,(r)}$y $N_{0i,(r)}$ calcular las probabilidades: $p_{(r)}(C=1)$ y $p_{(r)}(X_i=1|C=c)$

ii) Utilizando las probabilidades anteriores calcular las esperanzas de los estadísticos suficientes condicionadas a la muestra: $E_{\theta_r}[N_0|x]$ y $E_{\theta_r}[N_{0i}|x]$

iii) Calcular los nuevos estadísticos suficientes mediante la fórmula: $u_{r+1} = u_r + u - E_{\theta_r}[\mathcal{U}|x]$ siendo $u_r = \left(N_{0,(r)}, N_{01,(r)}, \ldots, N_{0n,(r)}\right)$