



Aprendizaje Discriminativo Clasificadores Bayesianos

Grupo de Sistemas Inteligentes
Departamento de Ciencias de la Computación e Inteligencia Artificial
Universidad del País Vasco

Desarrollo de la presentación

- Introducción
- Aprendizaje discriminativo de parámetros:
Algoritmo TM
- Aprendizaje discriminativo de estructuras:
Algoritmo TM estructural
- Algunos resultados
- Conclusiones

Introducción: Aprendizaje Generativo

- Clasificadores Bayesianos son típicamente generativos
- Se maximiza la verosimilitud conjunta
 $p(c, \mathbf{x})$
- Para clasificar se utiliza la prob. condicional
 $p(c|\mathbf{x})$
- Estimación de parámetros sencilla
- Se estiman más parámetros de los necesarios
- Si el modelo no es correcto \rightarrow mayor error

Introducción: Aprendizaje Discriminativo

- Ejemplo de clasificador discriminativo:
Regresión logística
- Se maximiza la verosimilitud condicional
 $p(c|\mathbf{x})$
- Aproximación más natural
- Menos parámetros a estimar
- Estimación de parámetros compleja
- Si el modelo no es correcto \rightarrow menor error
- Ignoramos información disponible sobre $p(\mathbf{x})$

Aprendizaje Discriminativo: Algoritmo TM

- Propuesto por Edwards y Lauritzen (2001)
- Algoritmo general para maximizar la verosimilitud condicional cuando la conjunta es más sencilla de maximizar
- Es un algoritmo iterativo
- Consta de dos pasos: T y M
- Lo hemos adaptado para el aprendizaje de clasificadores Bayesianos
- Disponible en Elvira para modelos hasta TAN

Aprendizaje Discriminativo: Algoritmo TM. Estructura general

$$l(\boldsymbol{\theta}) = \log f(\mathbf{x}, c|\boldsymbol{\theta}), \quad l_{\mathbf{x}}(\boldsymbol{\theta}) = \log f(\mathbf{x}|\boldsymbol{\theta}), \quad l^{\mathbf{x}}(\boldsymbol{\theta}) = \log f(c|\mathbf{x}, \boldsymbol{\theta})$$

- Tenemos en cuenta que $l^{\mathbf{x}}(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) - l_{\mathbf{x}}(\boldsymbol{\theta})$
- Obtenemos una aproximación a la verosimilitud condicional usando sólo la verosimilitud conjunta:

$$l^{\mathbf{x}}(\boldsymbol{\theta}) \approx q(\boldsymbol{\theta}|\boldsymbol{\theta}_r) = l(\boldsymbol{\theta}) - \boldsymbol{\theta}^T \dot{l}_{\mathbf{x}}(\boldsymbol{\theta}_r) = l(\boldsymbol{\theta}) - \boldsymbol{\theta}^T E_{\boldsymbol{\theta}}\{\dot{l}(\boldsymbol{\theta})|\mathbf{x}\}$$

- Pasos del algoritmo TM:

Paso T: dado $\boldsymbol{\theta}_r$, construir la función $q(\boldsymbol{\theta}|\boldsymbol{\theta}_r)$

Paso M: obtener $\boldsymbol{\theta}_{r+1} = \arg \max_{\boldsymbol{\theta}} q(\boldsymbol{\theta}|\boldsymbol{\theta}_r)$

Aprendizaje Discriminativo: Algoritmo TM. Familia Exponencial

- La verosimilitud conjunta y condicional se pueden escribir como:

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

$$l^x(\boldsymbol{\theta}) = \boldsymbol{\alpha}^T u(c, \mathbf{x}) - \psi^x(\boldsymbol{\alpha})$$

donde

$$\psi(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \log \int e^{\boldsymbol{\alpha}^T u(c, \mathbf{x}) + \boldsymbol{\beta}^T v(\mathbf{x})} \mu(dc|\mathbf{x}) \mu(d\mathbf{x})$$

$$\psi^x(\boldsymbol{\alpha}) = \log \int e^{\boldsymbol{\alpha}^T u(c, \mathbf{x})} \mu(dc|\mathbf{x})$$

y con

$\boldsymbol{\alpha}$ parámetros modelo condicional

$\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ parámetros modelo conjunto

$\boldsymbol{u} = u(C, \mathbf{X})$ estad. suf. modelo condicional

$(\boldsymbol{u}, \boldsymbol{v}) = (u(C, \mathbf{X}), v(\mathbf{X}))$ estad. suf. modelo conjunto

Aprendizaje Discriminativo: Algoritmo TM. Familia Exponencial

- El algoritmo queda de la siguiente forma:

$$\mathbf{u}_{r+1} = \mathbf{u}_r + \mathbf{u}_0 - E_{\boldsymbol{\theta}_r} \{\mathcal{U} | \mathbf{x}\} \quad \text{con } u_0 = u(c, \mathbf{x})$$

$$\boldsymbol{\theta}_{r+1} = \hat{\boldsymbol{\theta}}(\mathbf{u}_{r+1}, \mathbf{v})$$

- Las estimaciones máximo verosímiles de $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, se obtienen resolviendo las siguientes ecuaciones:

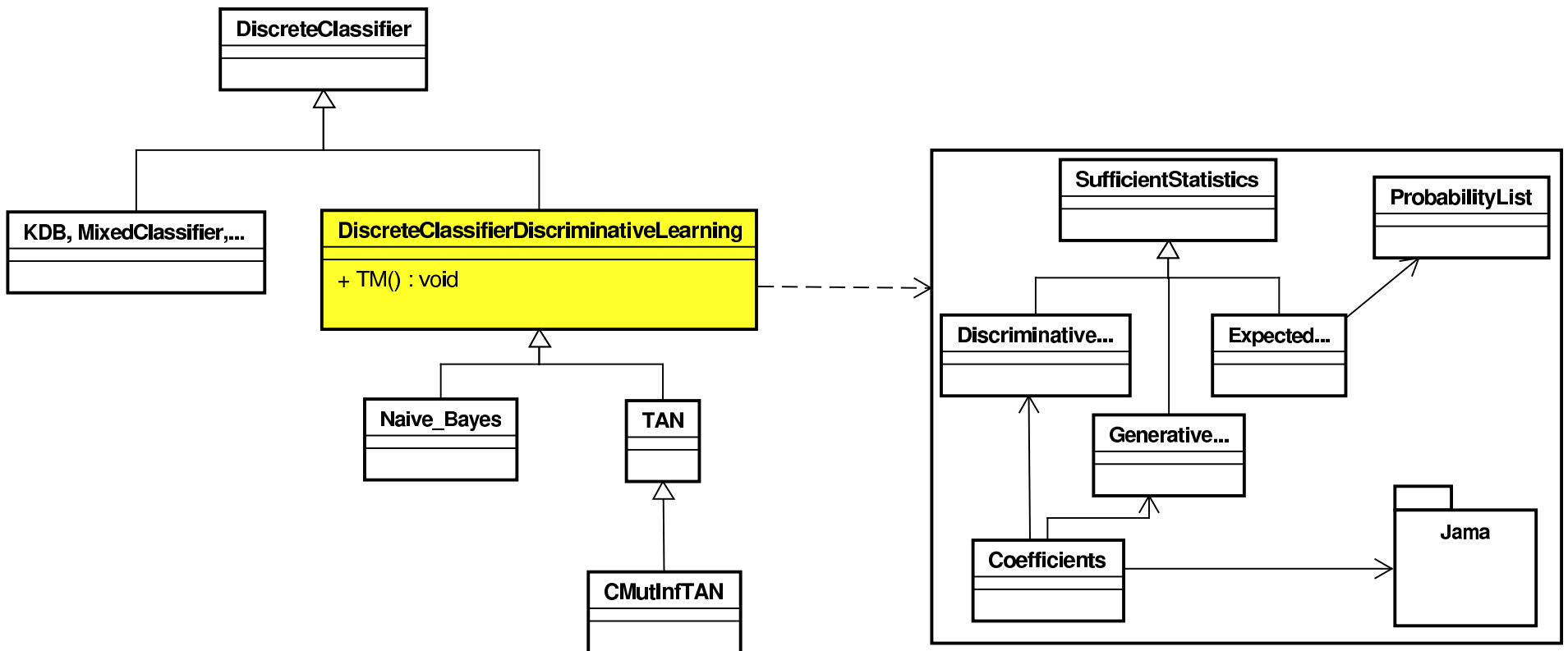
$$u(c, \mathbf{x}) = E_{\boldsymbol{\theta}} \{\mathcal{U}\}$$

$$v(\mathbf{x}) = E_{\boldsymbol{\theta}} \{\mathcal{V}\}$$

- Como primer conjunto de parámetros, $\boldsymbol{\theta}_0$, se toman los estimadores máximo verosímiles dado el conjunto de datos
- A veces en alguna iteración es necesaria una búsqueda lineal

$$\mathbf{u}_{r+1} = \mathbf{u}_r + \lambda(\mathbf{u}_0 - E_{\boldsymbol{\theta}_r} \{\mathcal{U} | \mathbf{x}\})$$

Aprendizaje Discriminativo: Algoritmo TM en Elvira



Aprendizaje Discriminativo: Algoritmo TM Estructural

- Aprendizaje discriminativo de la estructura del clasificador
- Utiliza el aprendizaje discriminativo de parámetros (TM) para maximizar la verosimilitud condicionada
- Método voraz de búsqueda
- Se parte de una red vacía
- En cada paso se añade el arco que maximice la medida de calidad BICd

Aprendizaje Discriminativo: Algoritmo TM Estructural

- La métrica BICd es una verosimilitud condicional penalizada, variante de la BIC:

$$BICd(D|\mathcal{M}) = \sum_{i=1}^N \log p(c^i|\mathbf{x}^i, \mathcal{M}) - \frac{\log N}{2}d$$

- El número de parámetros del modelo, d , son los parámetros del modelo condicional
- Se busca un equilibrio entre máxima verosimilitud condicional y complejidad

Resultados TM. Aprendizaje paramétrico

	<i>NB</i>	<i>NB-TM</i>	<i>NB</i> vs. <i>NB-TM</i>	<i>TAN</i>	<i>TAN-TM</i>	<i>TAN</i> vs. <i>TAN-TM</i>
Breast	97,37 ± 1,64	98,98 ± 0,74	• 0,036	97,37 ± 1,64	95,46 ± 1,41	• 0,016
Cleve	83,14 ± 4,89	87,53 ± 4,72	○ 0,072	82,77 ± 1,61	87,85 ± 3,24	• 0,043
Iris	94,67 ± 3,40	95,33 ± 3,40	0,746	93,33 ± 2,11	96,00 ± 2,49	0,142
German	75,40 ± 3,50	78,90 ± 4,00	○ 0,059	72,80 ± 2,22	84,00 ± 0,89	• 0,009
Corral	86,77 ± 9,27	90,61 ± 6,27	0,197	100,00 ± 0,00	99,20 ± 1,60	0,317
Crx	86,68 ± 4,70	88,52 ± 1,59	0,600	86,06 ± 1,33	89,59 ± 1,56	○ 0,075
Lymphography	83,77 ± 4,97	91,22 ± 3,49	0,141	79,08 ± 2,28	98,98 ± 1,65	• 0,008
Hepatitis	85,00 ± 10,15	93,75 ± 5,56	0,316	87,50 ± 6,84	100,00 ± 0,00	• 0,004
Vote	89,88 ± 2,45	98,39 ± 1,17	• 0,008	93,56 ± 1,55	99,08 ± 0,86	• 0,008
Heart	83,33 ± 6,73	86,67 ± 4,44	0,390	72,90 ± 2,74	81,75 ± 3,87	• 0,036

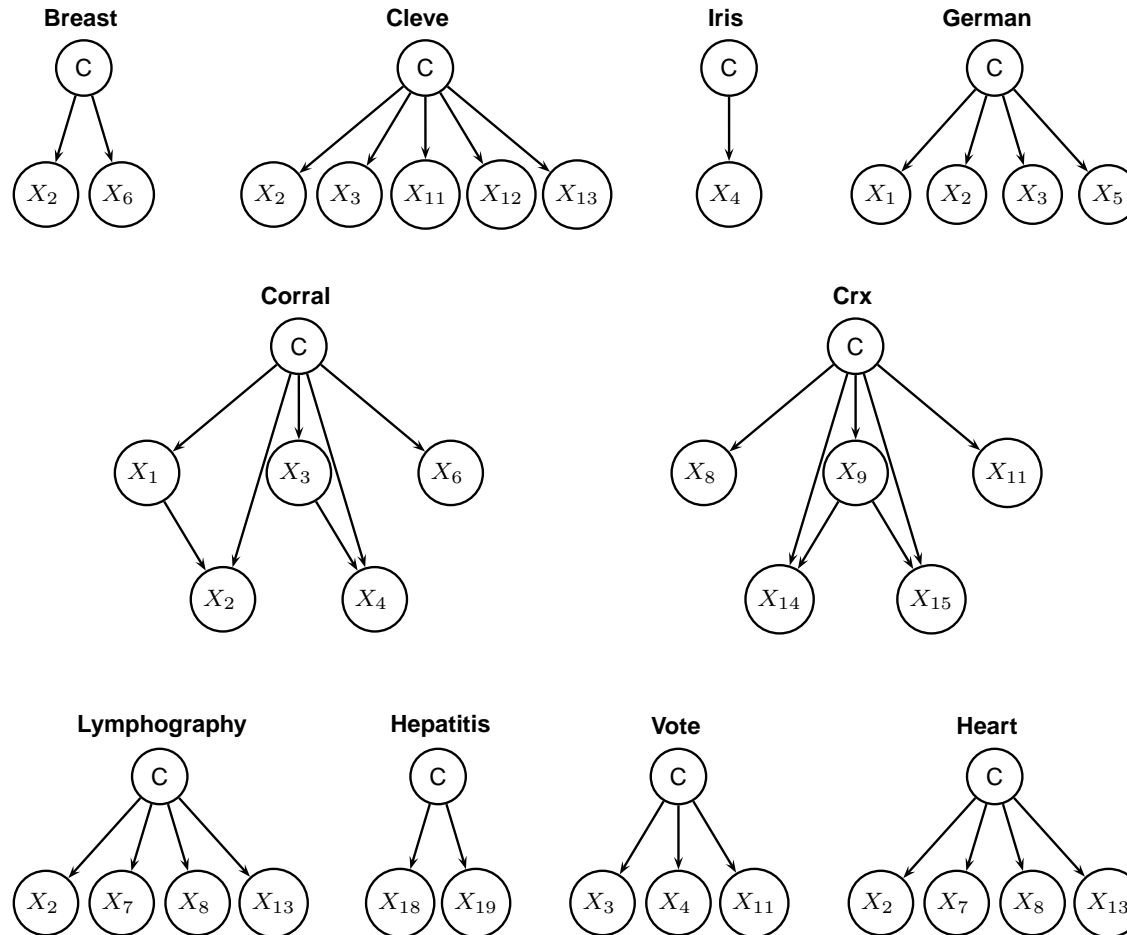
Tabla 1: Precisión estimada obtenidas con modelos naive Bayes y TAN

Resultados Estructural TM. Aprendizaje estructural

	<i>TM</i> <i>Estructural</i>	<i>NB-TM</i>	<i>TAN-TM</i>
Breast	95,02±8,55	98,98± 0,74	95,46± 1,41
Cleve	82,42± 4,89	87,53± 4,72	87,85± 3,24
Iris	95,33± 3,40	95,33± 3,40	96,00± 2,49
German	73,30± 3,75	78,90± 4,00	84,00± 0,89
Corral	98,46± 3,08	90,61± 6,27	99,20± 1,60
Crx	86,06± 3,32	88,52± 1,59	89,59± 1,56
Lymphography	71,61± 10,92	91,22± 3,49	98,98± 1,65
Hepatitis	87,50± 7,90	93,75± 5,56	100,00± 0,00
Vote	95,63± 2,45	98,39± 1,17	99,08± 0,86
Heart	85,56± 2,46	86,67± 4,44	81,75± 3,87

Tabla 2: Precisión estimada obtenida en la búsqueda estructural

Resultados TM Estructural. Estructuras aprendidas



Resultados Estructural TM. Reducción en n° de variables

	<i>Variables Base Datos</i>	<i>Variables Modelo Discriminativo</i>
Breast	9 + Clase	2 + Clase
Cleve	13 + Clase	5 + Clase
Iris	4 + Clase	1 + Clase
German	20 + Clase	4 + Clase
Corral	6 + Clase	5 + Clase
Crx	15 + Clase	5 + Clase
Lymphography	18 + Clase	4 + Clase
Hepatitis	19 + Clase	2 + Clase
Vote	16 + Clase	3 + Clase
Heart	13 + Clase	4 + Clase

Tabla 3: Número de Variables en cada una de las bases de datos utilizadas y número de variables seleccionadas por el algoritmo TM Estructural

Conclusiones

- Se proponen nuevos métodos de aprendizaje discriminativo tanto de parámetros como de estructura
- El aprendizaje discriminativo de parámetros es ligeramente superior al generativo en los experimentos realizados
- Esto no siempre es verdad: depende del modelo y de los datos
- El aprendizaje discriminativo estructural obtiene peor precisión pero reduce la complejidad del modelo

Trabajo futuro

- Extender el TM y TM estructural para tratar con valores perdidos
- Extender el TM y TM estructural a problemas de clustering
- Paralelización del código del TM Estructural
- Evaluación más exhaustiva de los algoritmos sobre más bases de datos
- Aplicación de los algoritmos propuestos a problemas reales