Modelos Gráficos Probabilísticos

Máster: Ciencia de Datos e Ingeniería de Computadores

Profesora: **Acid S.**

email: acid@decsai.ugr.es

Dpto. Ciencias de la Computación e Inteligencia Artificial

E.T.S. de Ingeniería Informática. Universidad de Granada

Contenido

- El problema de la clasificación
- Aplicaciones de la clasificación
- Las RRBB como clasificadores
  - El modelo de partida
  - Especialización de los algoritmos de aprendizaje
  - Algoritmos
  - Nuevos formalismos. Las multiredes
- Nuevos problemas de clasificación
- Recursos bibliográficos

El problema de la clasificación

¿Qué es la clasificación ?

- Objetivo
  Construir un modelo, el clasificador
  - Sea $(\mathbf{X}, C)$, donde C es una variable aleatoria unidimensional y $\mathbf{X} = X_1, X_2, \ldots, X_n$, n-dimensional, predictoras.
  - Asignación o predicción del valor para C, a partir de $X_1, \ldots X_i, X_n$ relevantes para discriminar entre las diferentes valores de C.

El problema de la clasificación

¿Qué es la clasificación ?

- Objetivo
  Construir un modelo, el clasificador
  - Sea $(\mathbf{X}, C)$, donde C es una variable aleatoria unidimensional y $\mathbf{X} = X_1, X_2, \ldots, X_n$, n-dimensional, predictoras.
  - Asignación o predicción del valor para C, a partir de $X_1, \ldots X_i, X_n$ relevantes para discriminar entre las diferentes valores de C.

- Paradigmas

El problema de la clasificación

¿Qué es la clasificación ?

- Objetivo
  Construir un modelo, el clasificador
  - Sea $(\mathbf{X}, C)$, donde C es una variable aleatoria unidimensional y $\mathbf{X} = X_1, X_2, \ldots, X_n$, n-dimensional, predictoras.
  - Asignación o predicción del valor para C, a partir de $X_1, \ldots X_i, X_n$ relevantes para discriminar entre las diferentes valores de C.

- Paradigmas
  - Clasificación supervisada. C representa la *Clase*

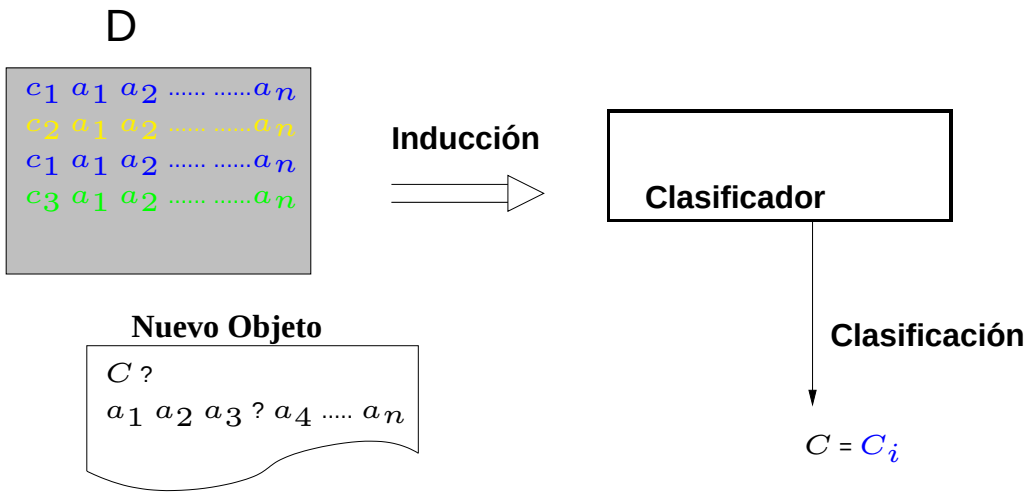
El problema de la clasificación

¿Qué es la clasificación ?

- Objetivo
  Construir un modelo, el clasificador
  - Sea $(\mathbf{X}, C)$, donde C es una variable aleatoria unidimensional y $\mathbf{X} = X_1, X_2, \ldots, X_n$, n-dimensional, predictoras.
  - Asignación o predicción del valor para C, a partir de $X_1, \ldots X_i, X_n$ relevantes para discriminar entre las diferentes valores de C.

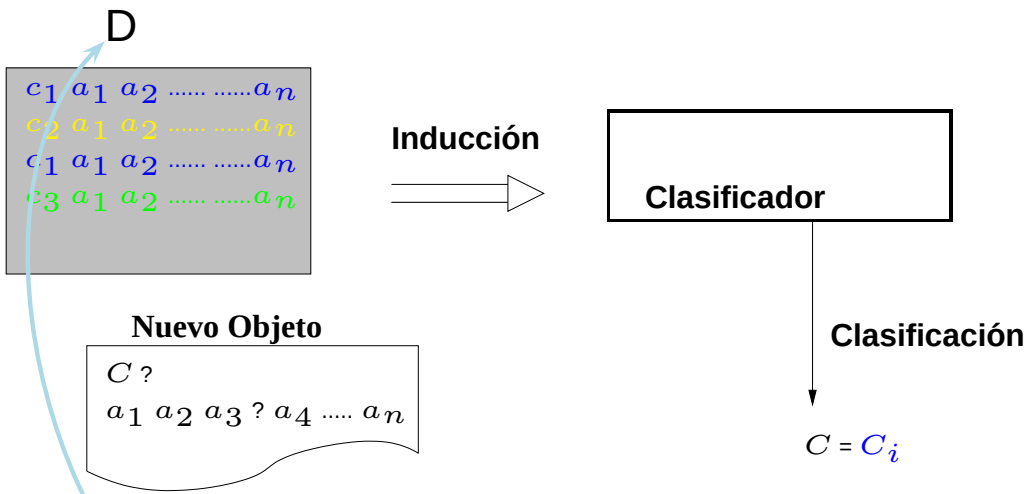
- Paradigmas
  - Clasificación supervisada. C representa la *Clase*
  - Clasificación no supervisada. C representa el *Clúster*

Clasificación supervisada.



- Entradas

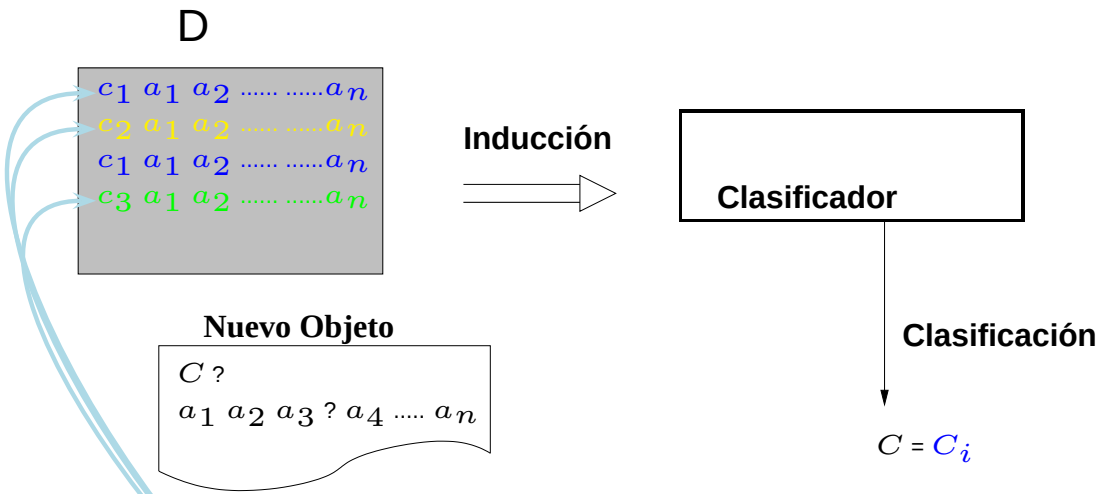
Clasificación supervisada.



- Entradas

  conjuntos de datos, los individuos correctamente clasificados.

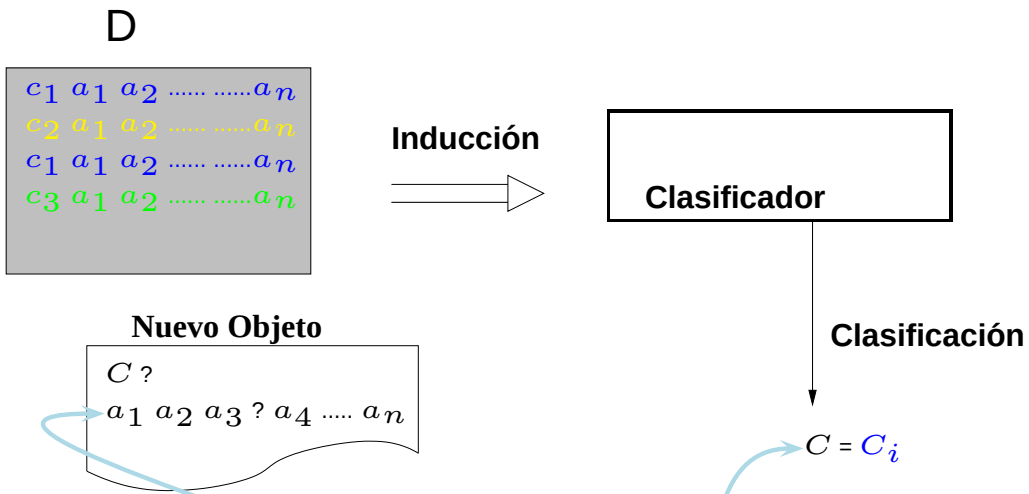
Clasificación supervisada.



- Entradas

  conjuntos de datos, los individuos correctamente clasificados.
  - Cada clase puede representar una población diferente.

Clasificación supervisada.



- Entradas

  conjuntos de datos, los individuos correctamente clasificados.
  - Cada clase puede representar una población diferente.
- Objetivo

  predecir la clase C a un nuevo individuo, Ev conociendo los valores de

  algunos de sus atributos $X_1 = a_1, X_2 = a_2, X_3?...X_n = a_n$

Clasificación no supervisada

Particionamiento de D en grupos o **clusters**.
Los elementos de la misma partición tienen características **comunes**.
Obtener la estructura subyascente del grupo apartir de los datos.

- Objetivo
  Construir un modelo, apartir de $X = \{X_1, X_2, \ldots, X_n\}$ y la v.a. C, *cluster*, **oculta** o desconocida.

$$D = \{(c^1?, x^1)(c^2?, x^2) \ldots (c^N?, x^N)\}$$

Clasificación no supervisada

Particionamiento de D en grupos o **clusters**.
Los elementos de la misma partición tienen características **comunes**.
Obtener la estructura subyascente del grupo apartir de los datos.

- Objetivo
  Construir un modelo, apartir de $X = \{X_1, X_2, \ldots, X_n\}$ y la v.a. C, *cluster*, **oculta** o desconocida.

$$D = \{(c^1?, x^1)(c^2?, x^2) \ldots (c^N?, x^N)\}$$

- Clustering, puede verse como:
  Aprendizaje con datos incompletos o datos perdidos.
  intrínsecamente no hace referencia a variables ocultas **.**

Clasificación no supervisada

Particionamiento de D en grupos o **clusters**.
Los elementos de la misma partición tienen características **comunes**.
Obtener la estructura subyascente del grupo apartir de los datos.

- Objetivo
  Construir un modelo, apartir de $X = \{X_1, X_2, \ldots, X_n\}$ y la v.a. C, *cluster*, **oculta** o desconocida.

$$D = \{(c^1?, x^1)(c^2?, x^2) \ldots (c^N?, x^N)\}$$

- Clustering, puede verse como:
  Aprendizaje con datos incompletos o datos perdidos.
  intrínsecamente no hace referencia a variables ocultas **.**
- El proceso de clustering se le conoce como:
  - aprendizaje apartir de datos sin etiquetar o
  - **Aprendizaje no supervisado**.

El problema de la clasificación. Los Datos.

Variables Discretas y Contínuas

En la distribución de datos $(\mathbf{X}, C)$, la v.a. n-dimensional $\mathbf{X}$ puede ser

- *discreta*
  $\forall X_i$ es unidimensional discreta, $X_i \in \mathbf{X}$
- *contínua*
  $\forall X_i$ es unidimensional contínua, $X_i \in \mathbf{X}$
- *mixta*
  $\mathbf{Y}, \mathbf{Z}$, $\forall Y_i \in \mathbf{Y}$ es r-dimensional discreta, $\forall Z_j \in \mathbf{Z}$ es s-dimensional contínua, con $\mathbf{Y}, \mathbf{Z} \in \mathbf{X}$ y $r + s = n$

C una variable aleatoria unidimensional discreta.

El problema de la clasificación. Los Datos.

Variables Discretas y Continuas

- Una variable es discreta si el conjunto de valores posibles es finito (Presencia o una enfermedad, Número de hijos, Sexo, Estudios realizados)
- Una variable es contínua si toma valores en un intervalo de los números reale (Altura, Peso, Luminosidad).

El problema de la clasificación. Los Datos.

Variables Discretas y Continuas

- Una variable es discreta si el conjunto de valores posibles es finito (Presencia o una enfermedad, Número de hijos, Sexo, Estudios realizados)
- Una variable es contínua si toma valores en un intervalo de los números reale (Altura, Peso, Luminosidad).
- Vamos a considerar sólo variables discretas en $\mathbf{X}$

El problema de la clasificación. Los Datos.

Variables Discretas y Continuas

- Una variable es discreta si el conjunto de valores posibles es finito (Presencia o una enfermedad, Número de hijos, Sexo, Estudios realizados)
- Una variable es contínua si toma valores en un intervalo de los números reale (Altura, Peso, Luminosidad).
- Vamos a considerar sólo variables discretas en $\mathbf{X}$
- Si hay contínuas $\Longrightarrow$ Preprocesamiento de datos
  - las discretizamos dividiéndolas en un conjunto finito de intervalos

Aplicaciones de Clasificación Automática

Tiene innumerables aplicaciones en multitud de dominios diferentes

- Médico
  - Diagnóstico
- Financiero
  - Prospecciones mineras. Estudio de viabilidad
  - Bancos Concesión de préstamos
  - Comerciales Clasificar clientes
  - Industriales Diagnóstico de fallos
- Instrumental. Recuperación de Información
  - Organización de documentos
  - Filtrado
  - Categorización jerarquizada
- Bioquímica. Microarray

Aplicaciones de Clasificación Automática. Ejemp

Diagnóstico médico.

Sistemas de ayuda al diagnóstico

Atributos predictores:

presión sanguínea, nivel de glucosa, edad, embarazo, fumador, etc.

Clase: padece o no padece (2) | alto, medio o bajo riesgo (3) |

lista de enfermedades excluyentes (15).

Aplicaciones de Clasificación Automática. Ejemp

Diagnóstico médico.

Sistemas de ayuda al diagnóstico

Atributos predictores:

presión sanguínea, nivel de glucosa, edad, embarazo, fumador, etc.

Clase: padece o no padece (2) | alto, medio o bajo riesgo (3) |

lista de enfermedades excluyentes (15).

Reconocimientos de terrenos de explotación minera

Atributos predictores:

longitud de perímetro, valores de perfiles, concentración de las muestras etc.

Clase:

apto o no apto |lista de tipos de terrenos (excluyentes)

Aplicaciones de Clasificación Automática. Ejemp

Financieros

Concesión de créditos o la detección de morosos de compañías

Atributos predictores:

edad, EstadoCivil, nivelEstudios, ingresos, gastosMedios etc.

Clase:

crédito aceptado, rechazado | moroso, no moroso

Aplicaciones de Clasificación Automática. Ejemp

Financieros

Concesión de créditos o la detección de morosos de compañías

Atributos predictores:

edad, EstadoCivil, nivelEstudios, ingresos, gastosMedios etc.

Clase:

crédito aceptado, rechazado | moroso, no moroso

Clasificación o categorización documental

Atributos predictores: términos de la colección

Clase: relevante, no relevante | deportes, ocio, noticias etc..

Filtrado de correo Spam

SpamAssassin `http://spamassassin.apache.org/`

SpamBayes `http://spambayes.sourceforge.net/`

Aplicaciones de Clasificación Automática. Ejemp

Financieros

Concesión de créditos o la detección de morosos de compañías

Atributos predictores:

edad, EstadoCivil, nivelEstudios, ingresos, gastosMedios etc.

Clase:

crédito aceptado, rechazado | moroso, no moroso

Clasificación o categorización documental

Atributos predictores: términos de la colección

Clase: relevante, no relevante | deportes, ocio, noticias etc..

Filtrado de correo Spam

SpamAssassin `http://spamassassin.apache.org/`

SpamBayes `http://spambayes.sourceforge.net/`

Microarray

Atributos predictores: lista de genes activos, clones

Clase: relevante o no dtdo proceso biológico| grupos que activan, inhiben, desarrollan.

Bases de Datos. UCI Machine Learning Repository

Data Set	Instances	Attributes	Classes
australian	690	14	2
breast	682	10	2
car	1728	6	4
chess	3196	36	2
cleve	296	13	2
corral	128	6	2
diabetes	768	8	2
DNA-nominal	3186	60	3
flare	1066	10	2
german	1000	20	2
heart	270	13	2
hepatitis	80	19	2
iris	150	4	3
letter	20000	16	26
mushroom	8124	22	2
nursery	12960	8	5
pima	768	8	2
soybean-large	562	35	19
vehicle	846	18	4
vote	435	16	2

Otros Datos

Data Set	Documentos	Términos	Classes
Reuters V1	806.791	47.219	1362
WebKB	8282	30403	7
Cora	4330	15753	7
Yahoo Science	14000	76000	264 (jerár.)

Modelos de clasificación

Cualquier función $f : \mathbf{X} \rightarrow C$ es un clasificador.

- medidas estadísticas clásicas

  obtenidas mediante *Análisis discriminante, Regresión logística,K-NN*

¿Cómo se representa el conocimiento en un clasificador ?
classification modeling

- reglas
- árboles de clasificación
- redes neuronales
- Modelos Gráficos Probabilísticos: nuestra elección

Las RRBB como clasificadores

Los MGP adecuados para clasificación, tratan de modelizar el mecanismo por el que una muestra fue generada.

Las RRBB como clasificadores

Los MGP adecuados para clasificación, tratan de modelizar el mecanismo por el que una muestra fue generada.

Supuesta $\mathbf{Y} = Y_1, ..Y_{n+1}$ discreta, partida de la forma $\mathbf{X}, C$, donde $\mathbf{X} = X_1, ..X_n$ y C la clase, $c \in \{1..r_C\}$ el modelo: las RRBB.

Las RRBB como clasificadores

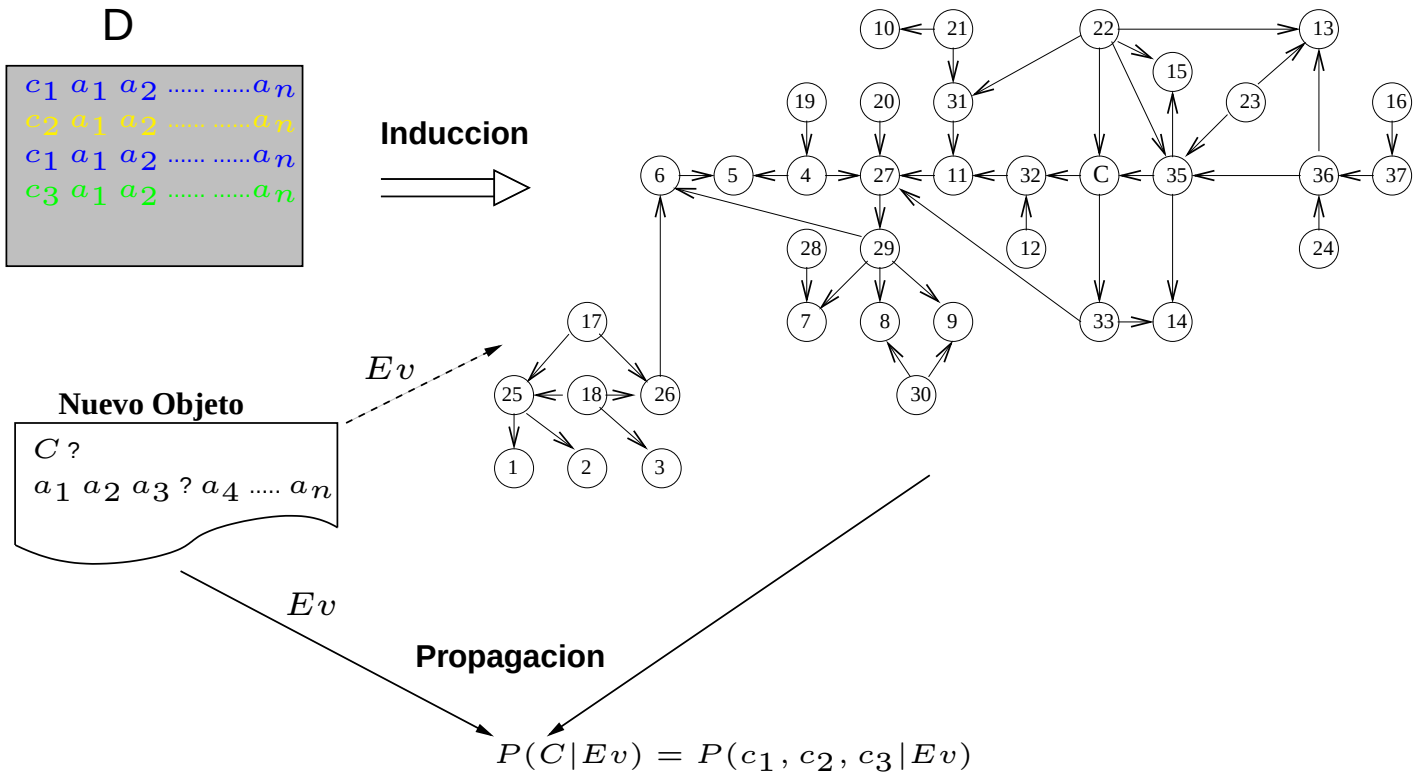
Los MGP adecuados para clasificación, tratan de modelizar el mecanismo por el que una muestra fue generada.

Supuesta $\mathbf{Y} = Y_1, ..Y_{n+1}$ discreta, partida de la forma $\mathbf{X}, C$, donde $\mathbf{X} = X_1, ..X_n$ y C la clase, $c \in \{1..r_C\}$ el modelo: las RRBB.

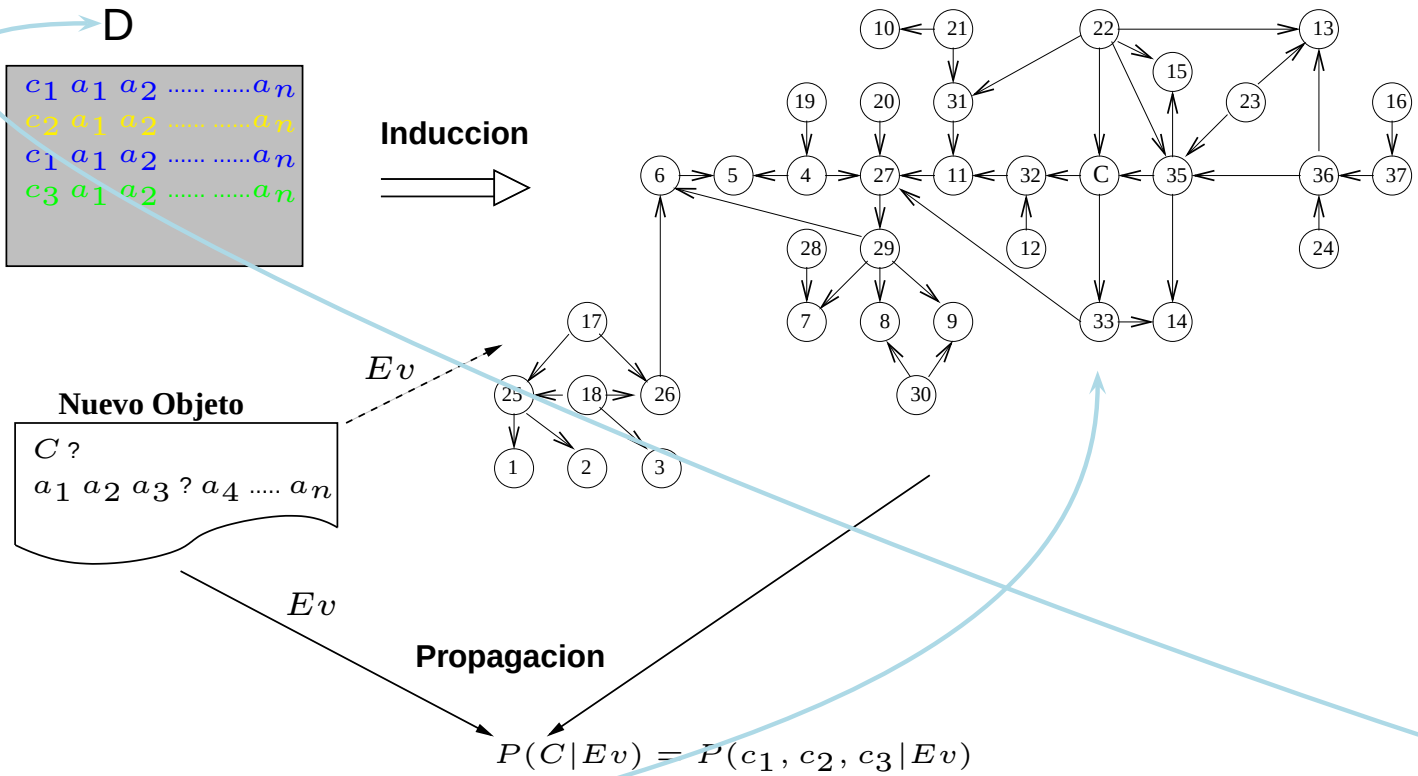
Se trata por tanto de

1. Aprender una RB apartir de D *(inducción)* $D = \{(c^1, \mathbf{x}^1)\}, \ldots (c^N, \mathbf{x}^N)\}$
   - Componente estructural un DAG g
     que representa las (in)dependencias entre las n+1 vars unidimensionales
   - Componente paramétrico
     un conjunto de distribuciones locales de probabilidad para g.
2. dada una nueva muestra, con C? predecir el valor para C *(inferencia)*

Las RRBB como clasificadores

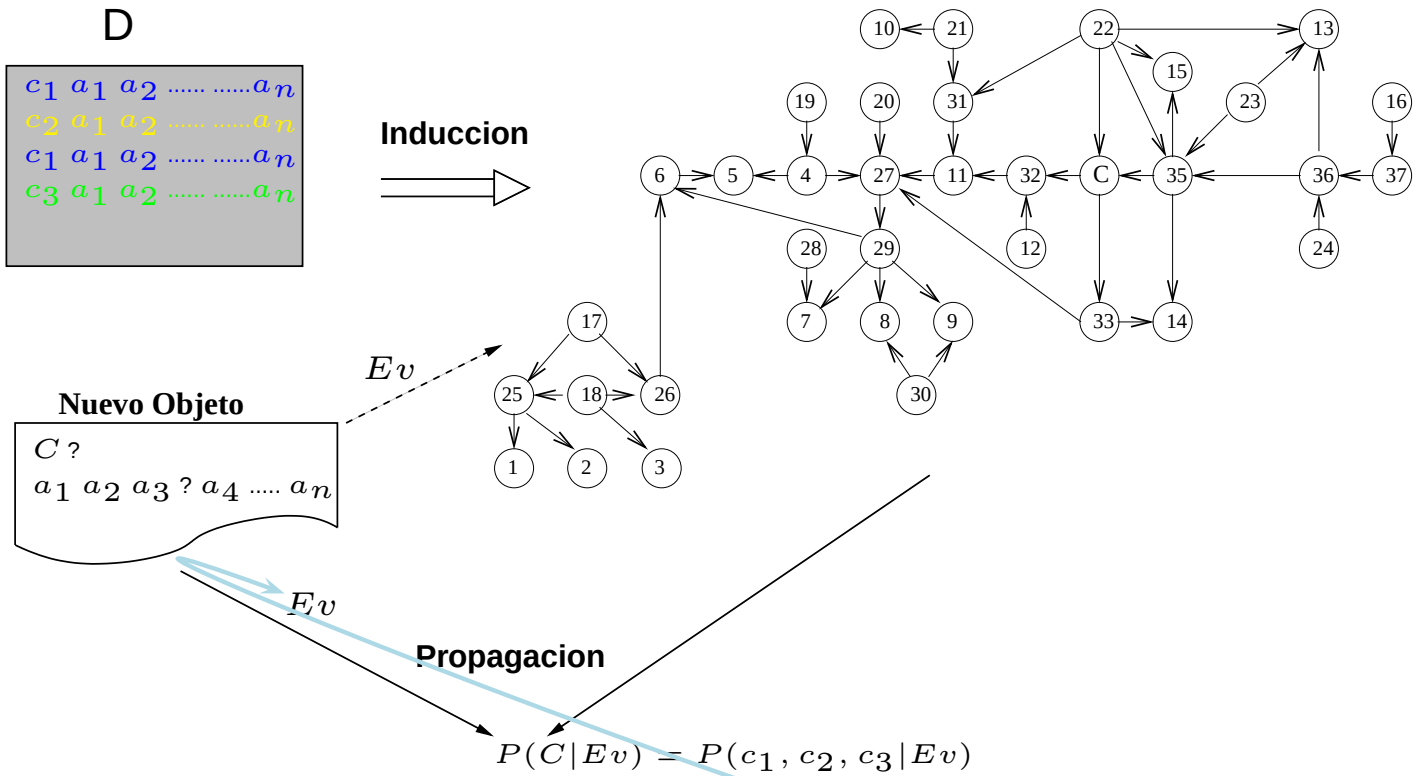


Las RRBB como clasificadores



- Un algoritmo de aprendizaje para construir, a partir de los datos d, una RB s, que represente $P(C, X_1, X_2, ..., X_n)$.

Las RRBB como clasificadores



- Un algoritmo de aprendizaje para construir, a partir de los datos d, una RB s, que represente $P(C, X_1, X_2, ..., X_n)$.
- Un algoritmo de inferencia calcula $P(C|Ev)$ para la evidencia Ev disponible sobre el objeto a clasificar.

Las RRBB como clasificadores

Cualquier problema de clasificación tiene asociado una función de coste de error de clasificación *misclassification cost function* $coste(c_e, c_r)$.
La más simple es 0/1:

$$coste(c_r, c_e) = \begin{cases} 0 & \text{si } c_e = c_r \\ 1 & \text{en otro caso} \end{cases}$$

- Objetivo de un clasificador **minimizar** el coste total de errores en clasificación.

Las RRBB como clasificadores

Cualquier problema de clasificación tiene asociado una función de coste de error de clasificación *misclassification cost function* $coste(c_e, c_r)$.
La más simple es 0/1:

$$coste(c_r, c_e) = \begin{cases} 0 & \text{si } c_e = c_r \\ 1 & \text{en otro caso} \end{cases}$$

- Objetivo de un clasificador **minimizar** el coste total de errores en clasificación.

...sabemos ... las RRBBs, permiten representar eficientemente:

$$p(c, x_1, \ldots, x_n) = p(c|x_1, \ldots, x_n)P(x_1, \ldots, x_n)$$

combinando conjunta y función de coste

$$\gamma(\mathbf{x}) = \arg\min_k \sum_{c=1}^{r_c} coste(k, c)p(c|x_1, \ldots, x_n)$$

Las RRBB como clasificadores

Para coste 0/1,

$$\gamma(\mathbf{x}) = \arg\max_c p(c|x_1, \ldots, x_n) = \arg\max_c p(c, x_1, \ldots, x_n)$$

El clasificador, da el valor r_e, con mayor valor aposteriori para C

Para realizar la clasificación se calcula la condicional

$$p(c|x_1, \ldots, x_n) = \frac{p(c, x_1, \ldots, x_n)}{\sum_{c'} p(c', x_1, \ldots, x_n}$$

La conjunta, que no tenemos, se tiene que estimar apartir de D

Las RRBB como clasificadores

Para coste 0/1,

$$\gamma(\mathbf{x}) = \arg\max_c p(c|x_1, \ldots, x_n) = \arg\max_c p(c, x_1, \ldots, x_n)$$

El clasificador, da el valor r_e, con mayor valor aposteriori para C

Para realizar la clasificación se calcula la condicional

$$p(c|x_1, \ldots, x_n) = \frac{p(c, x_1, \ldots, x_n)}{\sum_{c'} p(c', x_1, \ldots, x_n}$$

La conjunta, que no tenemos, se tiene que estimar apartir de D

La selección de los parámetros del modelo se obtiene maximizando la log-verosimilitud *log-likelihood* (LL)

$$LL = \sum_{d=1}^{N} \log p(c^d, \mathbf{x}^d)$$

Estimación de un clasificador

- Evaluación de clasificador:
  cómo de bien el modelo se comportan ante muestras con la clase desconocid

Lo que se conoce como (TE) tasas de éxito de un clasificador *accuracy*.
Se seleccionan una serie de muestras aleatorias, se clasifican c_e,
c_r y se calcula TE.

Estimación de un clasificador

- Evaluación de clasificador:
  cómo de bien el modelo se comportan ante muestras con la clase desconocid

Lo que se conoce como (TE) tasas de éxito de un clasificador *accuracy*.
Se seleccionan una serie de muestras aleatorias, se clasifican c_e,
c_r y se calcula TE.
Se necesita un estimador de TE de un clasificador.

Estimación de un clasificador

- Evaluación de clasificador:
  cómo de bien el modelo se comportan ante muestras con la clase desconocid

Lo que se conoce como (TE) tasas de éxito de un clasificador *accuracy*.
Se seleccionan una serie de muestras aleatorias, se clasifican c_e,
c_r y se calcula TE.
Se necesita un estimador de TE de un clasificador.

Estimaciones de TE de un clasificador:

- **Error de resustitución**. El mismo conjunto de training se clasifica.
- Estimación **por conjunto de test** independiente *hold-out*. Se parte D
  en **training** y **test** frecuente usar 2/3, 1/3
- Estimación **por validación cruzada** *k-fold cross validation*
  Particionamiento del conjunto en k conjuntos frecuente usar *k=10*
- Estimación por validación cruzada dejando uno fuera ***Leave-one-out***
  llevando k al extremo de $k = N$

Problema de las estimaciones: sesgo *bias* y varianza.

Las RRBB como clasificadores

Experimentalmente las RRBB obtenidas mediante algoritmos de aprendizaje genérico han demostrado ser malos clasificadores, esto es, tener (TE) bajas.

Las RRBB como clasificadores

Experimentalmente las RRBB obtenidas mediante algoritmos de aprendizaje genérico han demostrado ser malos clasificadores, esto es, tener (TE) bajas.

Un algoritmo de aprendizaje tipo *métrica* + *búsqueda* optimiza $Score(s,d) = \sum_{i=1}^{n} score(P_s(X_i, pa(X_i)), P_d(X_i, pa(X_i)_s))$ de forma conjunta, $\forall X_i$.

Las RRBB como clasificadores

Experimentalmente las RRBB obtenidas mediante algoritmos de aprendizaje genérico han demostrado ser malos clasificadores, esto es, tener (TE) bajas.

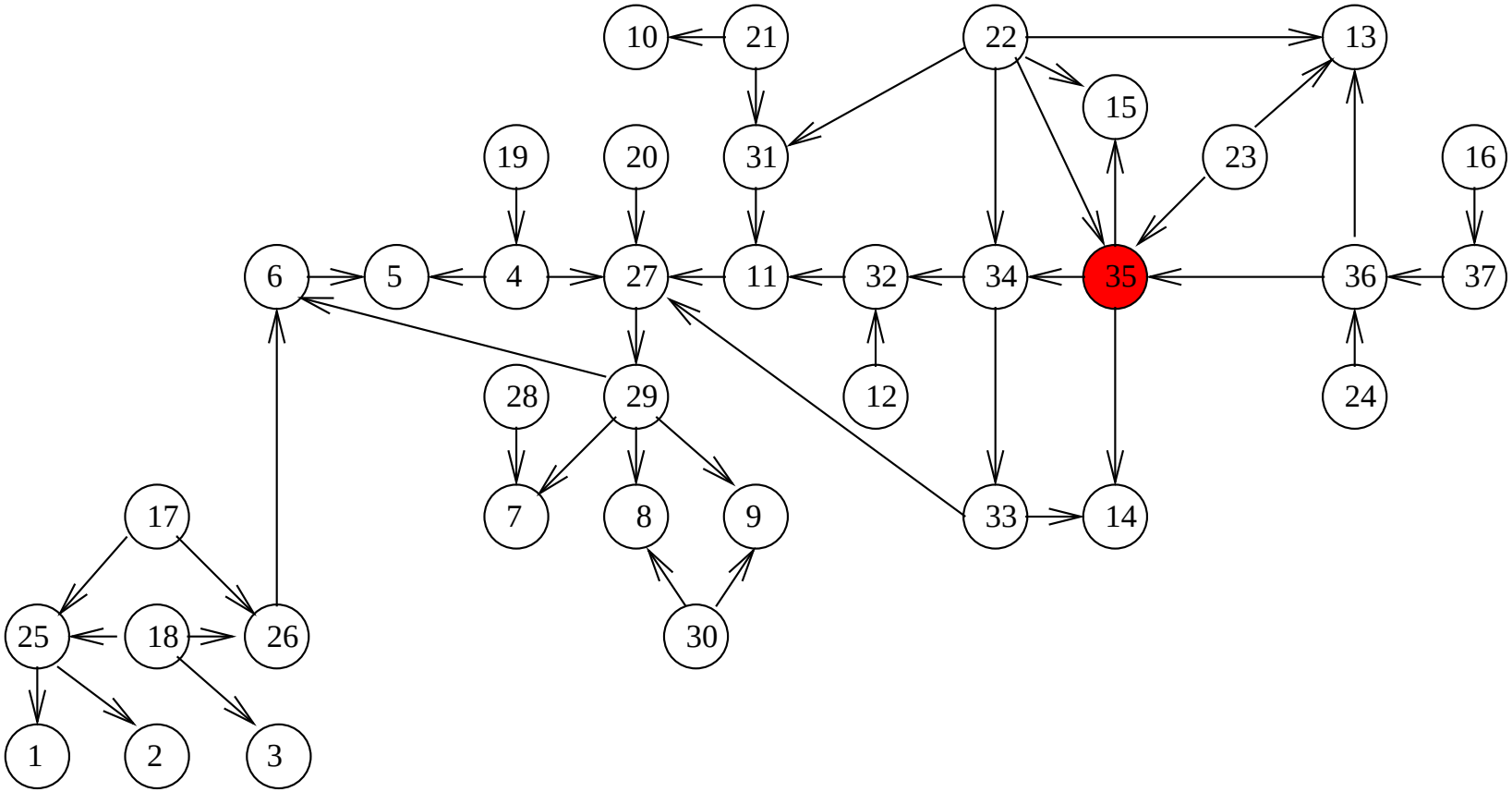
Un algoritmo de aprendizaje tipo *métrica* + *búsqueda* optimiza $Score(s,d) = \sum_{i=1}^{n} score(P_s(X_i, pa(X_i)), P_d(X_i, pa(X_i)_s))$ de forma conjunta, $\forall X_i$.

Pero,... una buena representación de $P(C, X_1, X_2, \ldots X_n)$ no necesariamente lo es de $P(C|X_1, X_2, ..., X_n)$.

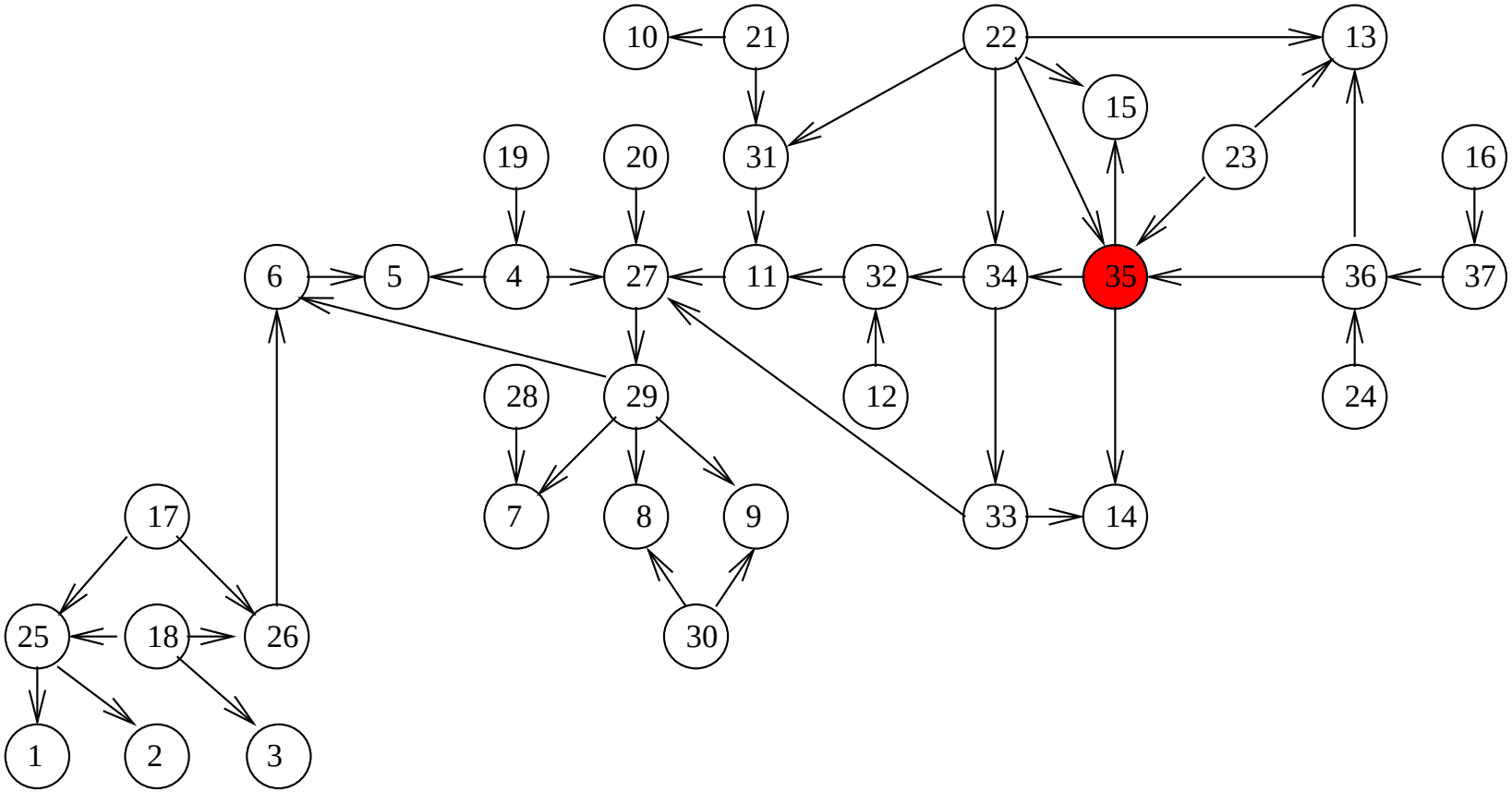
Así ocurre que: dadas s_1, s_2

$$Score(s_1, d) > Score(s_2, d) \not\Longrightarrow TE(s_1, d') < TE(s_2, d')$$

Las RRBB como clasificadores



Las RRBB como clasificadores



Varias alternativas

Métricas específicas para la clasificación

Según paradigma: *métrica + búsqueda*

¿Podría plantearse una métrica condicional?

ejemplo: MDL basado en la medida de *verosimilitud LL(s|d)*

$LL(s|d) = N \sum_{i=1}^{N} \sum_{x_i, pa(x_i)} P_d(X_i, pa(X_i)_s) \log P_s(X_i, pa(X_i))$

se maximiza cuando $P_s(X_i, pa(X_i)) = P_d(X_i, pa(X_i)_s)$ (1)

Esto es, para dos RRBB, $s_1 = (g, \theta_1)$ y $s_1 = (g, \theta_2)$ si

θ_1 cumple (1) entonces $LL(s_1|d) \geq LL(s_2|d)$

Métricas específicas para la clasificación

Según paradigma: *métrica + búsqueda*

¿Podría plantearse una métrica condicional?

ejemplo: MDL basado en la medida de *verosimilitud LL(s|d)*

$LL(s|d) = N \sum_{i=1}^{N} \sum_{x_i, pa(x_i)} P_d(X_i, pa(X_i)_s) \log P_s(X_i, pa(X_i))$

se maximiza cuando $P_s(X_i, pa(X_i)) = P_d(X_i, pa(X_i)_s)$ (1)

Esto es, para dos RRBB, $s_1 = (g, \theta_1)$ y $s_1 = (g, \theta_2)$ si

θ_1 cumple (1) entonces $LL(s_1|d) \geq LL(s_2|d)$

- Verosimilitud condicional $CLL(s|d)$?

dado $CLL(s|d) = N \sum_{i=1}^{N} \log P(C|X_1, X_2 \ldots X_n)$

no hay forma eficiente de maximizar $CLL(s|d)$, esto es

¿cómo seleccionar los parámetros óptimos? solución NO viable

Métricas específicas para la clasificación

Según paradigma: *métrica + búsqueda*

¿Podría plantearse una métrica condicional?

ejemplo: MDL basado en la medida de *verosimilitud LL(s|d)*

$LL(s|d) = N \sum_{i=1}^{N} \sum_{x_i, pa(x_i)} P_d(X_i, pa(X_i)_s) \log P_s(X_i, pa(X_i))$

se maximiza cuando $P_s(X_i, pa(X_i)) = P_d(X_i, pa(X_i)_s)$ (1)

Esto es, para dos RRBB, $s_1 = (g, \theta_1)$ y $s_1 = (g, \theta_2)$ si

θ_1 cumple (1) entonces $LL(s_1|d) \geq LL(s_2|d)$

- Verosimilitud condicional $CLL(s|d)$?

dado $CLL(s|d) = N \sum_{i=1}^{N} \log P(C|X_1, X_2 \ldots X_n)$

no hay forma eficiente de maximizar $CLL(s|d)$, esto es

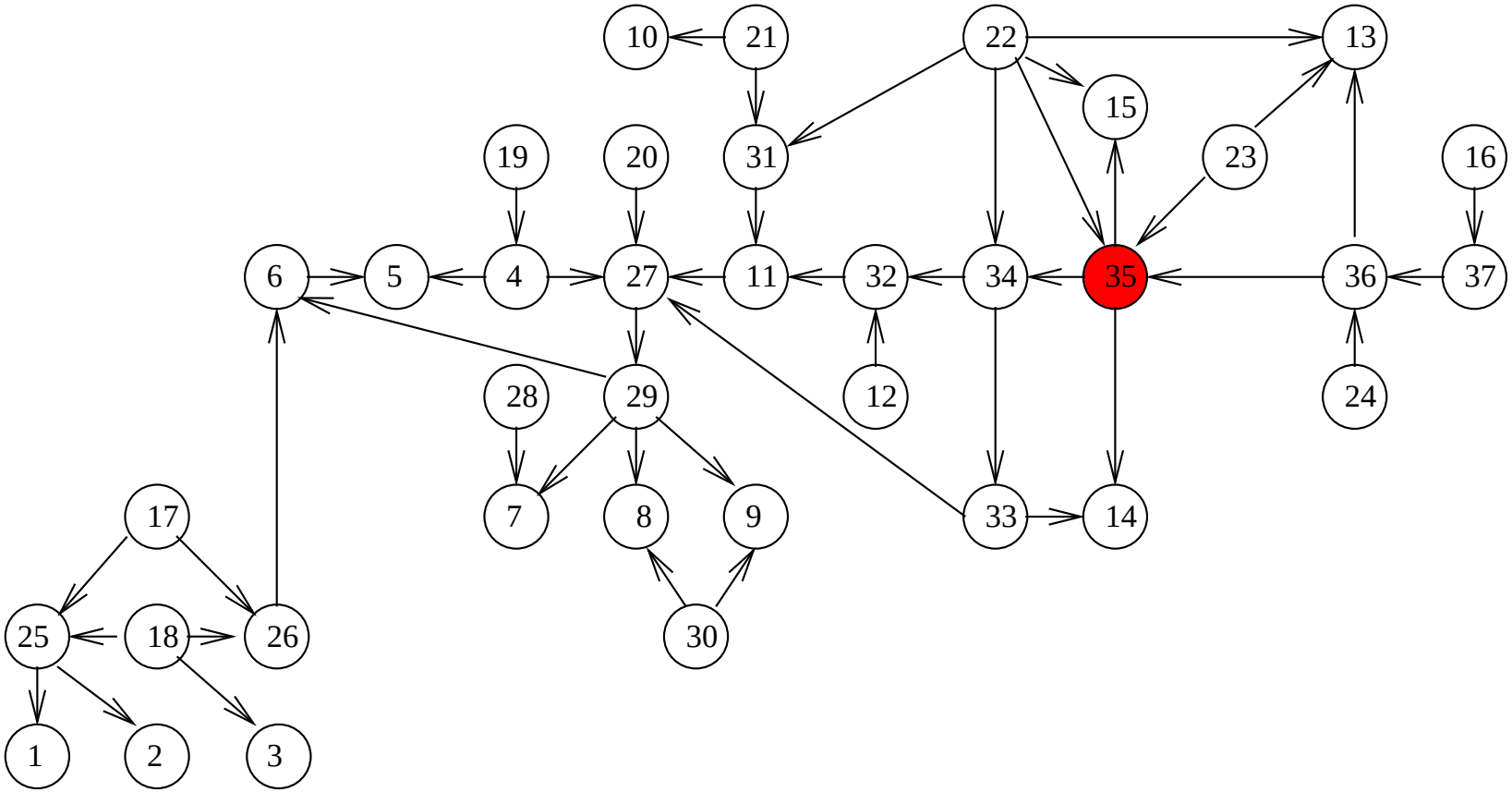
¿cómo seleccionar los parámetros óptimos? solución NO viable

- Uso de métricas que consideren las TE directamente...

(algoritmos de envolturas... veremos después)

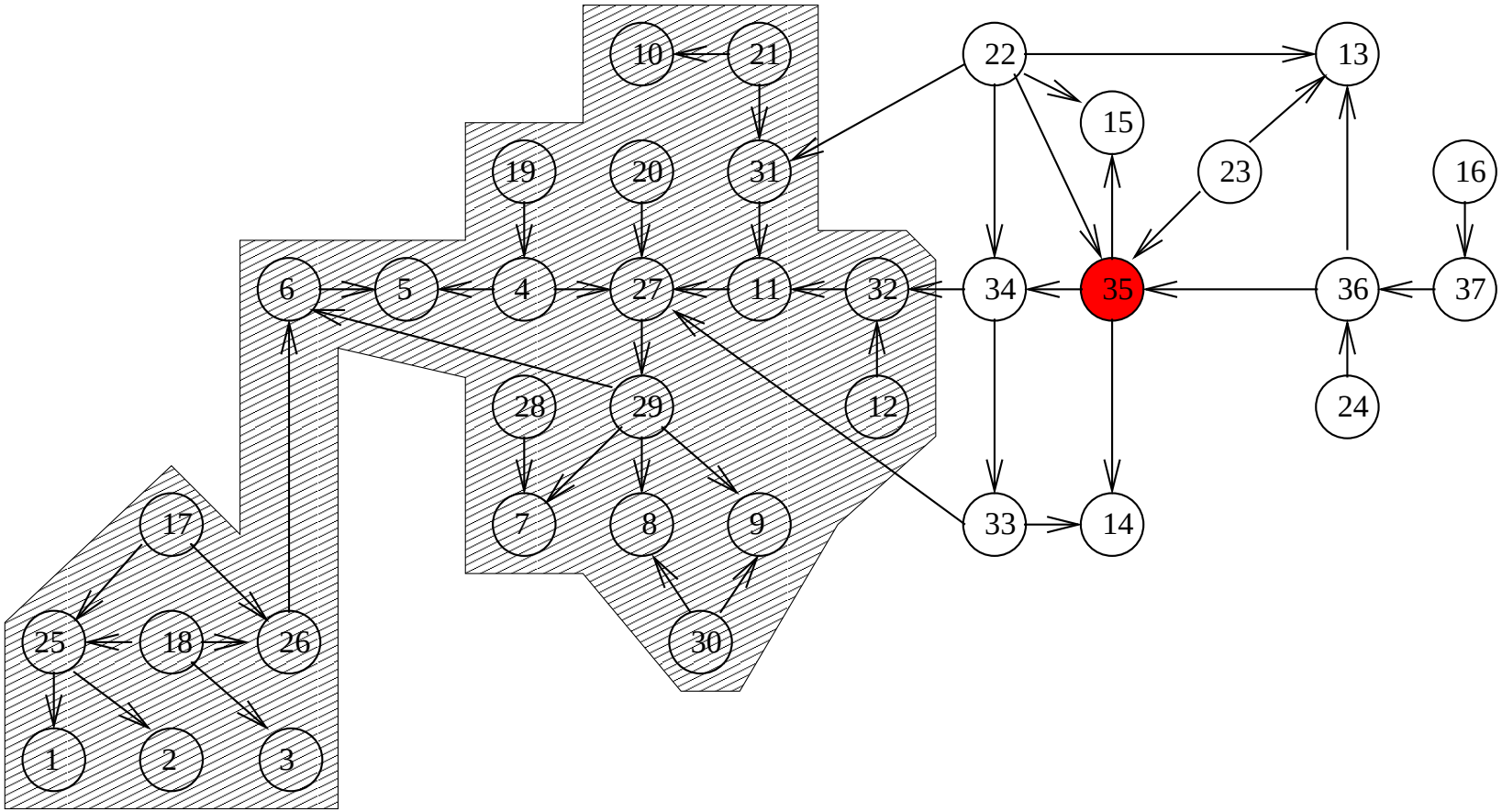
Las RRBB como clasificadores. Estrategias

Otras alternativas



Las RRBB como clasificadores. Estrategias

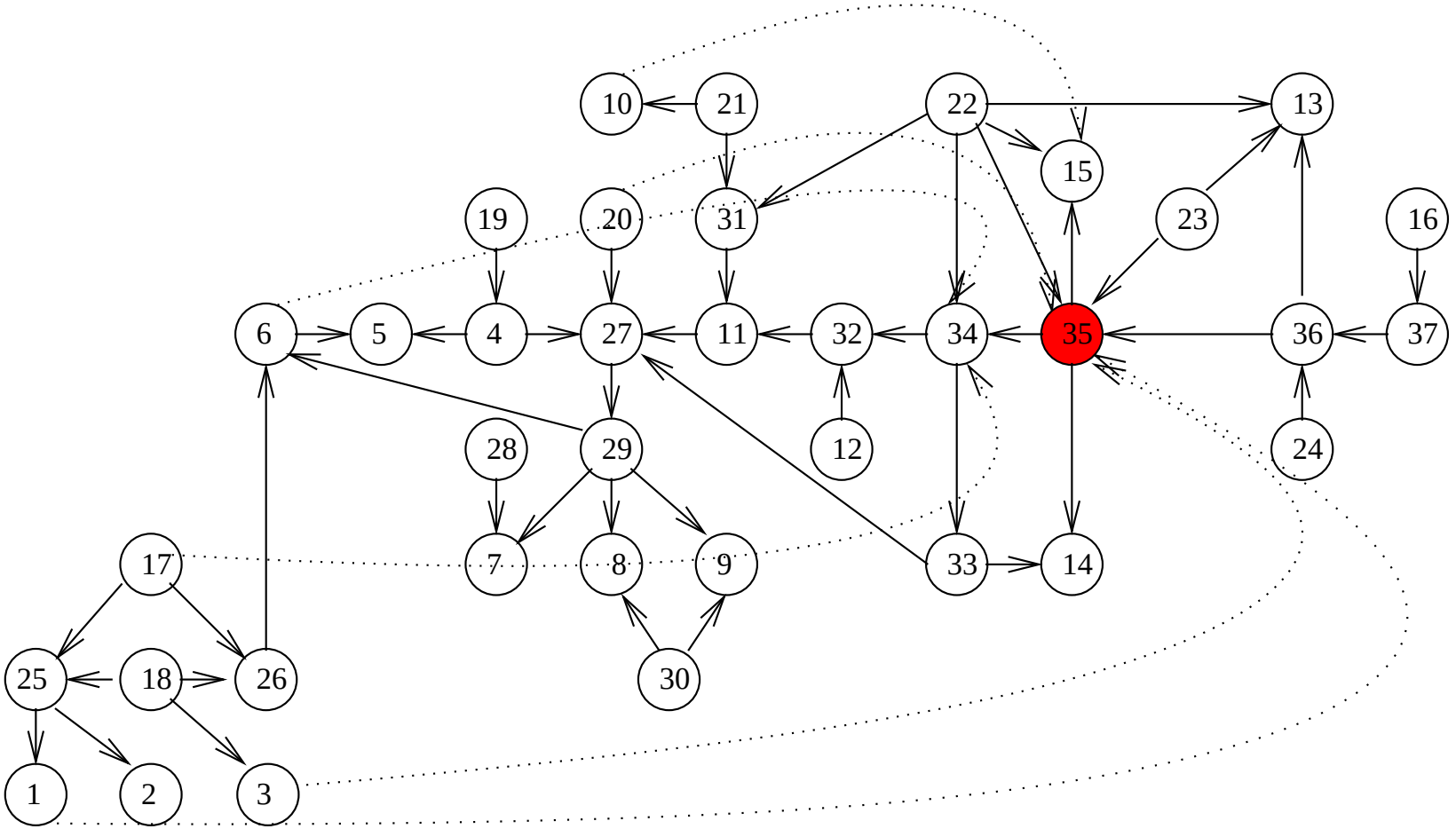
Otras alternativas



Desechar variables

Las RRBB como clasificadores. Estrategias

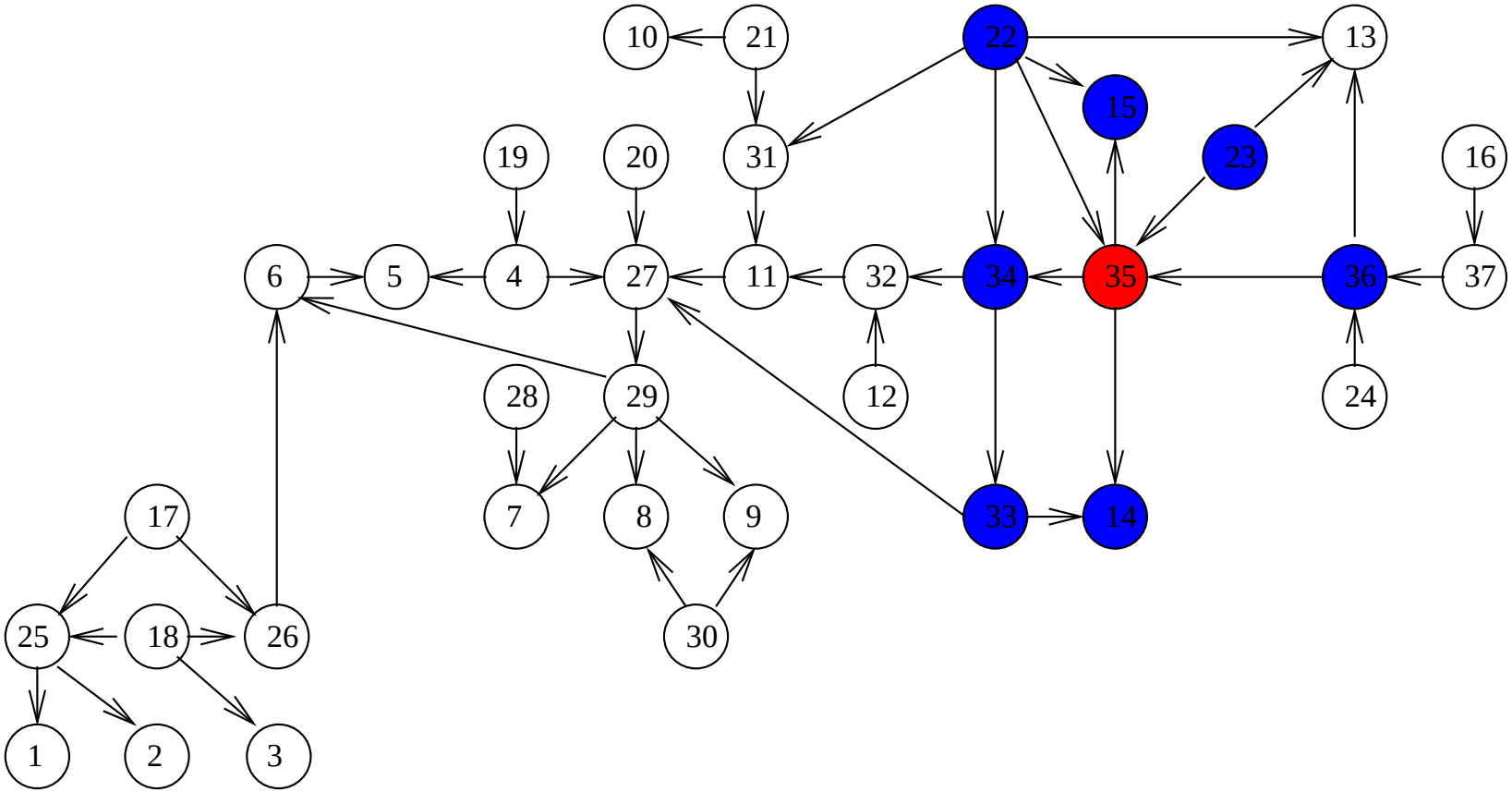
Otras alternativas



Relaciones extras con la clase

Las RRBB como clasificadores. Estrategias

Otras alternativas

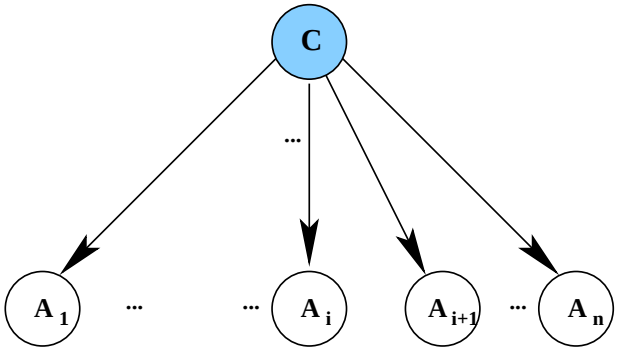


Otras formas de viajar en el espacio...Manto de Markov $MB(C)$

Revisando ... clasificadores

Del estudio de modelos que son *buenos* clasificando

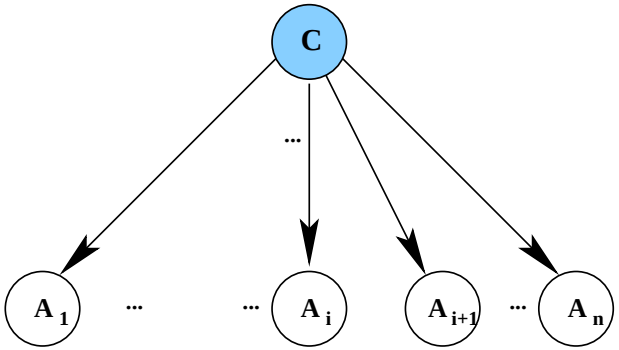
- El punto de partida es el modelo Ingenuo **Naive Bayes**.



Revisando ... clasificadores

Del estudio de modelos que son *buenos* clasificando

- El punto de partida es el modelo Ingenuo **Naive Bayes**.



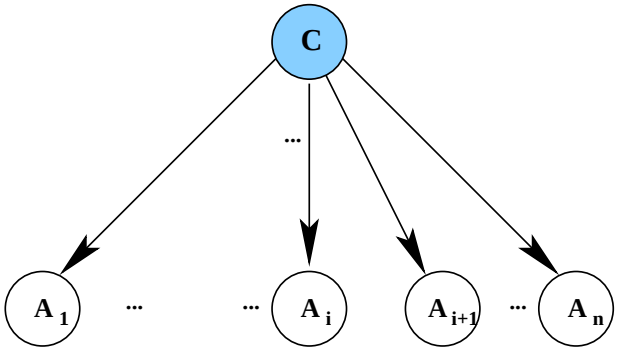
- Suposición: *Todos* A_i *son* **condicionalmente independientes** *dada* C
- Razones del éxito:

(Kononenko, 1990, ... Domingos and Pazzani, 1997)
uso en personalización, clasificación textos ...

Revisando ... clasificadores

Del estudio de modelos que son *buenos* clasificando

- El punto de partida es el modelo Ingenuo **Naive Bayes**.



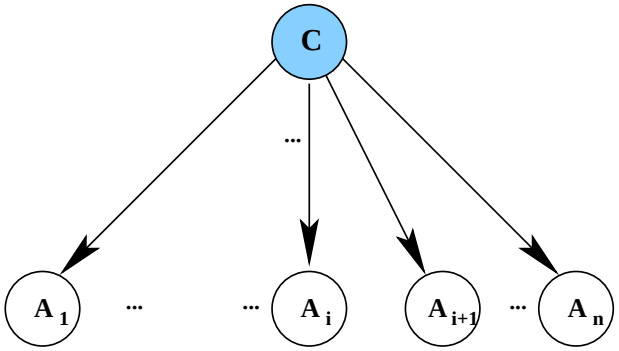
- Suposición: *Todos* A_i *son* **condicionalmente independientes** *dada* C
- Razones del éxito:
  - Se fija directamente la estructura que representa las restricciones impuestas.
  - Estimaciones robustas
  - Es muy competitivo *frente a modelos más sofisticados.*

(`Kononenko, 1990, ...  Domingos and Pazzani, 1997`)
uso en personalización, clasificación textos ...
NB parte de unas suposiciones muy restrictivas y poco realistas

Revisando ... clasificadores

Del estudio de modelos que son *buenos* clasificando

- El punto de partida es el modelo Ingenuo **Naive Bayes**.



- Suposición: *Todos* A_i *son* **condicionalmente independientes** *dada* C
- Razones del éxito:
  - Se fija directamente la estructura que representa las restricciones impuestas.
  - Estimaciones robustas
  - Es muy competitivo *frente a modelos más sofisticados.*

(`Kononenko, 1990, ...  Domingos and Pazzani, 1997`)
uso en personalización, clasificación textos ...
NB parte de unas suposiciones muy restrictivas y poco realistas

Revisando... clasificadores

Modelo Naive Bayes

$$P(C \mid A_1, A_2, \ldots, A_n) = P(A_1, A_2, \ldots, A_n\, C)/P(A_1, A_2, \ldots, A_n)$$

$$P(C \mid A_1, A_2, \ldots, A_n) \propto P(A_1, A_2, \ldots, A_n,\, C) \text{luego,}$$

$$arg\max_C P(C \mid A_1, A_2, \ldots, A_n) = arg\max_C P(A_1, A_2, \ldots, A_n,\, C) \text{pero,}$$

$$P(A_1, A_2, \ldots, A_n,\, C) = P(C)\; P(A_1, A_2, \ldots, A_n \mid C) \text{así}$$

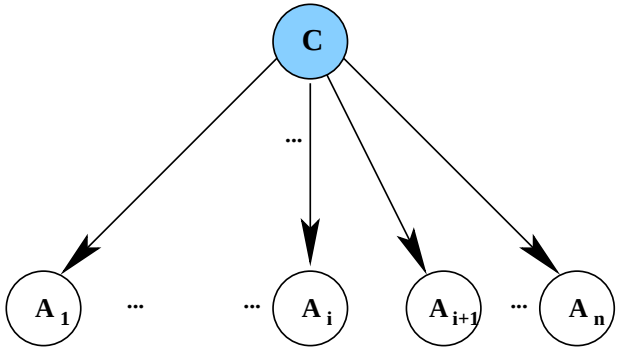
$$P(C \mid A_1, A_2, \ldots, A_n) = \alpha\; P(C)\; P(A_1, A_2, \ldots, A_n \mid C)$$

con las suposiciones se transforma en

$$P(C \mid A_1, A_2, \ldots, A_n) = \alpha\;\; P(C) \prod_{i=1}^{n} P(A_i \mid C)$$

- Sólo hace falta estimar $P(A_i|C)\ \forall A_i$, mediante cálculo de frecuencias, las $\hat{\theta}_{A_i|C}$ son robustas.

Revisando... clasificadores



la construcción del modelo es muy simple y rápida, y

su empleo es muy eficiente.
si n es número de variables $A_1, \ldots, A_n$,
k número de casos de C, $c_1, ..., c_{r_C}$
v es el número de casos promedio por A_i y
t número de muestras $1..N$

Entrenamiento		Clasificación	
Tiempo	Espacio	Tiempo	Espacio
O(nt)	O(knv)	O(kn)	O(knv)

Revisando ... clasificadores

Dada la factorización del modelo Naive Bayes

$P(C \mid A_1, A_2, \ldots, A_n) = \alpha \;\; P(C) \prod_{i=1}^{n} P(A_i \mid C)$

Problemas

1. La información redundante, o fuertemente correladas, degrada la eficacia de clasificador. Suponiendo que $A_1 = A_2$, la misma evidencia toma el doble de relevancia.

Revisando ... clasificadores

Dada la factorización del modelo Naive Bayes

$P(C \mid A_1, A_2, \ldots, A_n) = \alpha \;\; P(C) \prod_{i=1}^{n} P(A_i \mid C)$

Problemas

1. La información redundante, o fuertemente correladas, degrada la eficacia de clasificador. Suponiendo que $A_1 = A_2$, la misma evidencia toma el doble de relevancia.
2. La información irrelevante puede degradar la eficacia del clasificador al introducir ruido.

Revisando ... clasificadores

Dada la factorización del modelo Naive Bayes

$P(C \mid A_1, A_2, \ldots, A_n) = \alpha \;\; P(C) \prod_{i=1}^{n} P(A_i \mid C)$

Problemas

1. La información redundante, o fuertemente correladas, degrada la eficacia de clasificador. Suponiendo que $A_1 = A_2$, la misma evidencia toma el doble de relevancia.
2. La información irrelevante puede degradar la eficacia del clasificador al introducir ruido.
3. La suposición de independencia condicional no es realista
   Dominio: la concesión de préstamos.
   ¿ Podemos suponer que no existe correlación entre *Edad*, *NivelEstudios* e *Ingresos*?

RRBB como clasificadores. Estrategias

Del estudio de modelos que son *buenos* clasificando

RRBB como clasificadores. Estrategias

Del estudio de modelos que son *buenos* clasificando

- El punto de partida es el modelo Ingenuo (Naive) Bayes.

Se plantean diferentes estrategias de especialización

RRBB como clasificadores. Estrategias

Del estudio de modelos que son *buenos* clasificando

El punto de partida es el modelo
Ingenuo (Naive) Bayes.

Se plantean diferentes estrategias de especialización

1. Incorporando técnicas de selección y/o agrupación de características
   para eliminar atributos **redundantes** o **irrelevantes**,
   o para explícitamente tener en cuenta las interacciones entre subconjuntos de atributos.

RRBB como clasificadores. Estrategias

Del estudio de modelos que son *buenos* clasificando

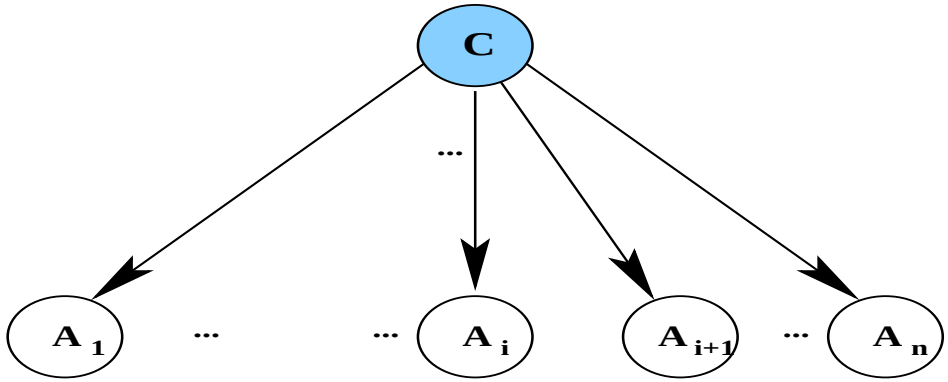
El punto de partida es el modelo
Ingenuo (Naive) Bayes.

Se plantean diferentes estrategias de especialización

1. Incorporando técnicas de selección y/o agrupación de características
   para eliminar atributos **redundantes** o **irrelevantes**,
   o para explícitamente tener en cuenta las interacciones entre subconjuntos de atributos.
2. Modificando o restringiendo el proceso de búsqueda
   para dirigirlo hacia redes que tengan un buen comportamiento como clasificadores.
3. Nuevos formalismos de MGP

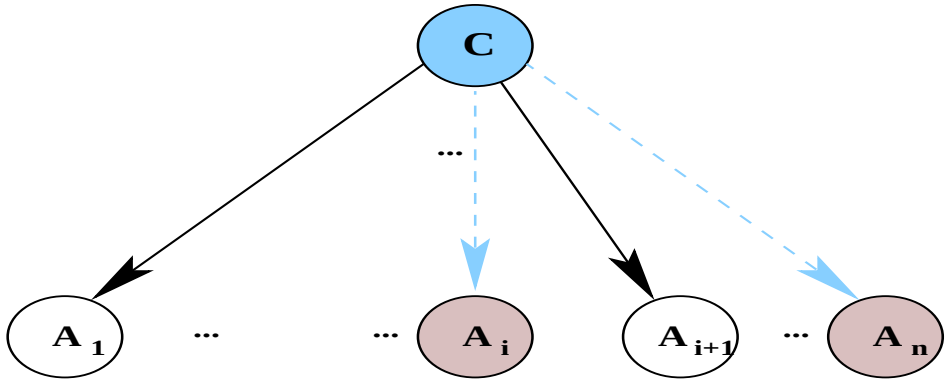
Selección de Características

No todos los atributos contribuyen a la determinación de la clase.



Selección de Características

No todos los atributos contribuyen a la determinación de la clase.



- Se incorporan técnicas de selección de características para eliminar atributos irrelevantes y/o redundantes

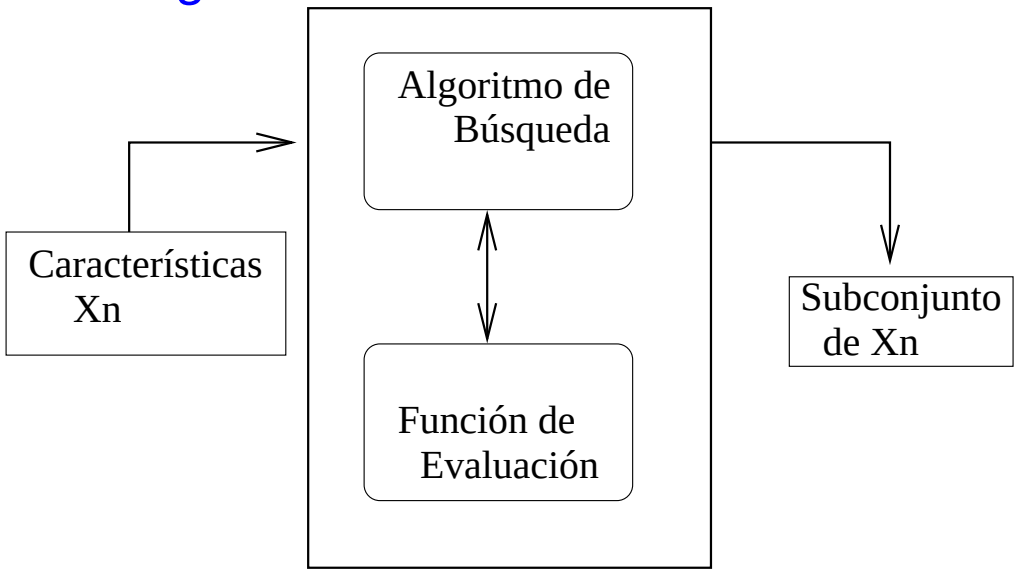
Selección de Características

(Feature selection. FS)

- Objetivo
  Reducir el número de atributos necesarios para caracterizar los datos.
- Consecuencias
  - Se mejora la **eficiencia** de tareas como la clasificación.
  - La estimación de los parámetros es más **robusta**.
- Diferentes estrategias
  - Filtros *(Filters)*
  - Envolturas *(Wrappers)*

Selección de Características

Estrategia de filtrado



- Proceso *previo*, compuesto de:
  - Búsqueda específica en el espacio de variables
  - Función de evaluación específica independiente a la clasificación
- Salida
  Un subconjunto de las variables del problema
- Existen muchas y muy diversas métricas, (*... respecto a* C, por conjuntos)
  información mutua, entropía, distancia Euclídea, divergencia de Kullback-Leibler

Selección de Características

Una propuesta de filtrado

Evaluación utilizando la información mútua

$$I(X,Y) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

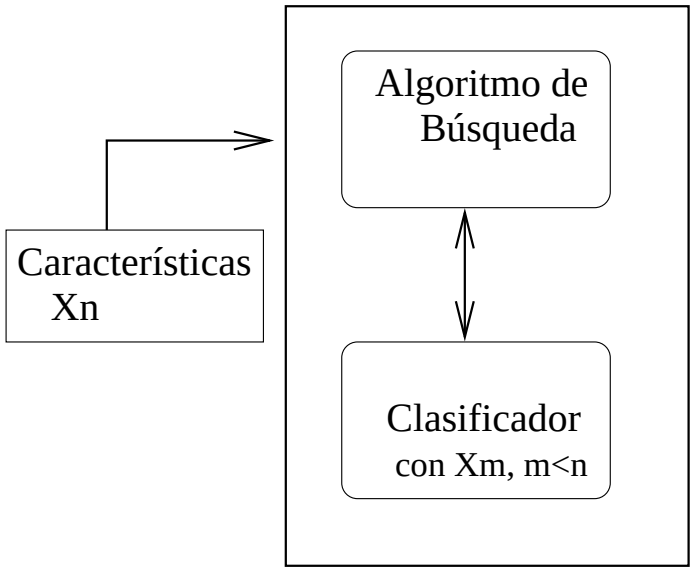
Búsqueda la selección de los atributos con mayores valores respecto a C

```
for i= 1 .. n
    Xvector[i] = (X_i) // se calcula
    Ivector[i] = I(X_i,C) // se calcula
ordenarDesc (Ivector[i], Xvector[i]) // orden decreciente de
Seleccionadas = {}
for i= 1 .. k
  Seleccionadas += {Xvector[i]} // los k atrib. más relevant
NB (Seleccionadas)    // se monta el NB con las X seleccionad
```

Se fija un k, número de variables a seleccionar
o un umbral λ para comparar las medidas de información $I(X,C)$
Otras medidas: *Ganancia de Información condicional,*
para dejar fuera los atributos más correlados.

Los Selective naive bayes

Estrategia con envolturas



La selección de características es una consecuencia de la clasificación.

- Búsqueda en el espacio de variables combinada con la búsqueda en el espacio de configuraciones
- Métrica TE, Función de evaluación del clasificador

Salida: La mejor red bayesiana evaluada

Métricas específicas

Algoritmos de tipo wrapper

```
i=0, G_i={0}  // G_0 estructura inicial
mejor_G=G_0
E(G_0)      // Se estiman sus parámetros
mejor_V=TE(G_0|D)   // Se clasifica D
while not fin {
  siguiente configuración G_i+1 en el espacio
  i++
  E(G_i)      // Se estiman sus parámetros
  V_i=TE(G_i|D)   // Se clasifica D
  if  mejor_V > V_i then {
      mejor_V=V_i
      mejor_G=G_i }
} return mejor_G   // mejor estructura de las exploradas
```

Los Selective naive bayes

Los algoritmos de búsqueda

- espacio de tamaño 2^n,con n atributos predictivos o características
- de tipo exponencial, secuencial o aleatorio
  hill-climbing; simulated annealing; genéticos...
  (Langley and Sage, 1994); (Vinciotti et al., 2006); (Inza et al., 2001)

Los Selective naive bayes

Los algoritmos de búsqueda

- espacio de tamaño 2^n,con n atributos predictivos o características
- de tipo exponencial, secuencial o aleatorio
  hill-climbing; simulated annealing; genéticos...
  (Langley and Sage, 1994); (Vinciotti et al., 2006); (Inza et al., 2001)
  - variantes para el tipo secuencial:

	Forward	Backward
Inicio	$\emptyset$	X
Decisión	Incluir	Eliminar

Los Selective naive bayes

Los algoritmos de búsqueda

- espacio de tamaño 2^n,con n atributos predictivos o características
- de tipo exponencial, secuencial o aleatorio
  hill-climbing; simulated annealing; genéticos...
  (Langley and Sage, 1994); (Vinciotti et al., 2006); (Inza et al., 2001)
  - variantes para el tipo secuencial:

	Forward	Backward
Inicio	$\emptyset$	X
Decisión	Incluir	Eliminar

  - Criterio de parada:
    - Se añaden atributos mientras mejora
    - Se añaden atributos mientras no empeora

Clasificadores Selective

Métodos de clasificación
mediante selección de características combinación de:

Estrategia $\times$ Tipo de Alg. de Búsqueda $\times$ variantes $\times$ función de evaluación

Los clasificadores bayesianos selectivos

Clasificadores Selective

Métodos de clasificación
mediante selección de características combinación de:

Estrategia $\times$ Tipo de Alg. de Búsqueda $\times$ variantes $\times$ función de evaluación

Los clasificadores bayesianos selectivos

Selective naive `LangleySage94`

Wrapper + secuencial + FSS

Clasificadores Selective

Métodos de clasificación
mediante selección de características combinación de:

Estrategia $\times$ Tipo de Alg. de Búsqueda $\times$ variantes $\times$ función de evaluación

Los clasificadores bayesianos selectivos

- Selective naive `LangleySage94`
  Wrapper + secuencial + FSS
- Selective Bayesian network `Singh-Provan96`
  Filter + secuencial + métrica : ganancia de información + K2

Selección de Características

Selective Bayesian network `LangleySage94`

V_i, conjunto de variables seleccionadas.

Se clasifica NB con un núm. variable de atributos $NB(V_i)$

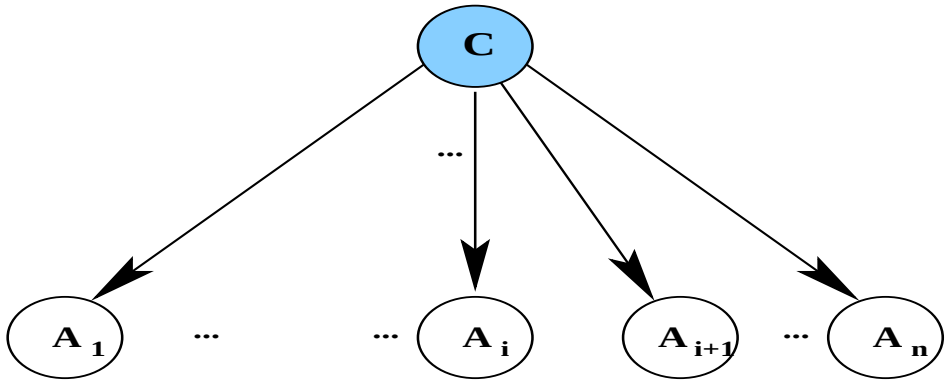
Criterio de evaluación: *estimación de TE mediante Leave one out*

Criterio de parada: No empeora la bondad del clasificador

```
i=0,  V_i={}  // Conjunto vacío
Candidatos={X_1,X_2,X_3, ...X_n}
mejor_TE=TE(NB(V_i),D)   // Se clasifica D
do {
  for X_l in Candidatos {
     current_TE = TE(NB(V_i U X_l),D)...
  } // se obtiene el mejor de una vuelta
  if current_TE >=  mejor_TE  //
     actualiza V_i y Candidatos  // var. candidata + predic
}while current_TE < mejor_TE && Candidatos!={}
```

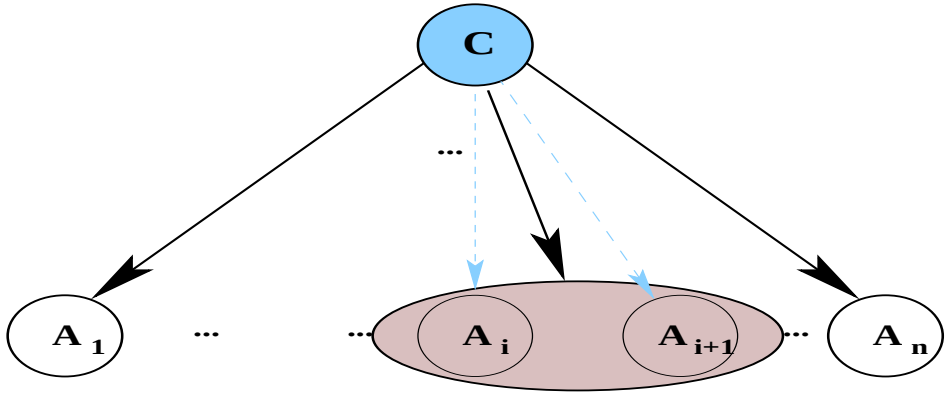
Los semi, por agrupación de características

No todos los atributos son condicionalmente independientes dada la clase.



Los semi, por agrupación de características

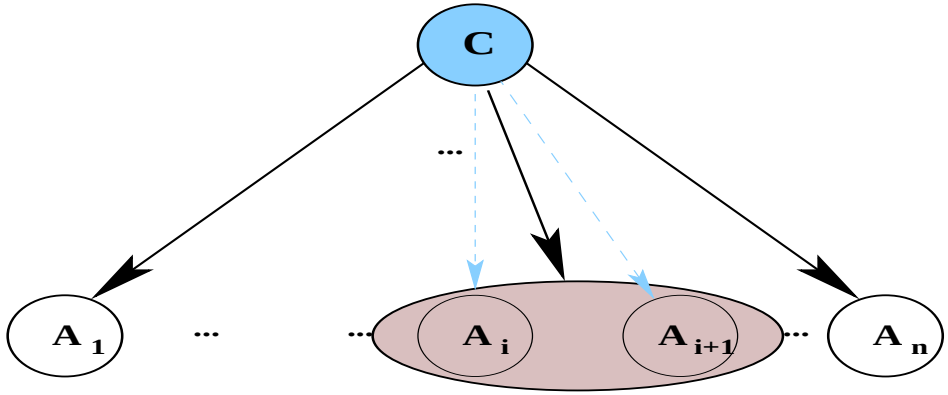
No todos los atributos son condicionalmente independientes dada la clase.



se crea una nueva variable, fusión de atributos fuertemente correlados

Los semi, por agrupación de características

No todos los atributos son condicionalmente independientes dada la clase.



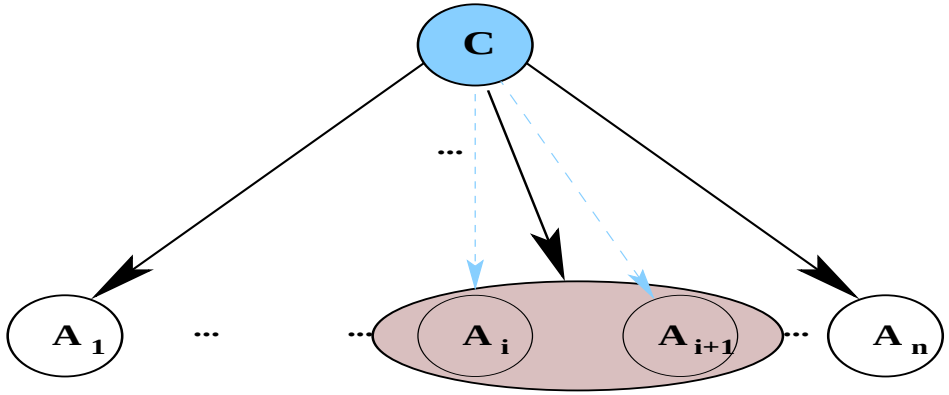
se crea una nueva variable, fusión de atributos fuertemente correlados

$$P(C|A_1, A_2, \ldots, A_n) = \alpha P(C) \times P(A_1|C) \times \ldots \times P(A_i, A_{i+1}|C)$$

$$\times \ldots \times P(A_n|C)$$

Los semi, por agrupación de características

No todos los atributos son condicionalmente independientes dada la clase.



se crea una nueva variable, fusión de atributos fuertemente correlados

$$P(C|A_1, A_2, \ldots, A_n) = \alpha P(C) \times P(A_1|C) \times \ldots \times P(A_i, A_{i+1}|C)$$

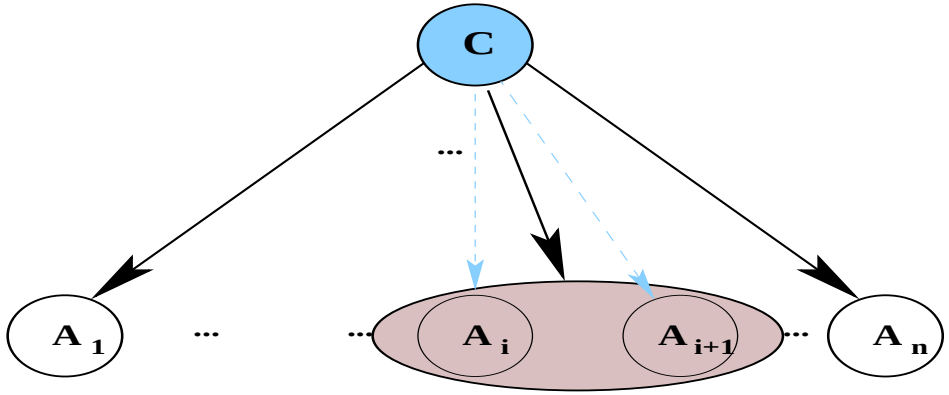
$$\times \ldots \times P(A_n|C)$$

- (Semi-Naive) Kononenko91

  test estadísticos para $Dep(X_i, X_j|C)$ fusión de atributos

Los semi, por agrupación de características

No todos los atributos son condicionalmente independientes dada la clase.



se crea una nueva variable, fusión de atributos fuertemente correlados

$$P(C|A_1, A_2, \ldots, A_n) = \alpha P(C) \times P(A_1|C) \times \ldots \times P(A_i, A_{i+1}|C)$$

$$\times \ldots \times P(A_n|C)$$

- (Semi-Naive) `Kononenko91`
  test estadísticos para $Dep(X_i, X_j|C)$ fusión de atributos
- Semi-Naive `Pazzani95` (FSSJ)

Agrupación de Características

Algoritmo Forward Sequential Selection and Joining (FSSJ)

Se clasifica NB con un núm. variable de variables $NB(V_i)$

V_i contiene X_j o var. nuevas **compuestas**

Criterio de evaluación: TE mediante Leave one out

Criterio de parada: No mejora la bondad del clasificador

```
i=0, V_i={} Candidatos={X_1,X_2,X_3, ...X_n}
mejor_TE=TE(NB(V_i),D)   // Se clasifica D
repeat
  para cada atributo de Candidatos
   la mejor operacion entre
      1. añadirlo condicionalmente independiente de los de V_i
      2. juntarlo con cada X_j de NB(V_i)
    current_TE = TE(NB(V_i),D)
  if current_TE >=  mejor_TE  // mejor de la vuelta
       i++ ; se actualiza V_i y Candidatos
until criterio de parada
```

Nueva estrategia. Restringir el proceso de búsqu

Otras topologías que generalicen el modelo Naive Bayes

Los clasificadores **Augmented**

ej. Augmented Naive Bayesian Networks (ABNs)

Todos los atributos influyen en la clase, pero

se relaja la suposición de independencia condicional.

Nueva estrategia. Restringir el proceso de búsqu

Otras topologías que generalicen el modelo Naive Bayes

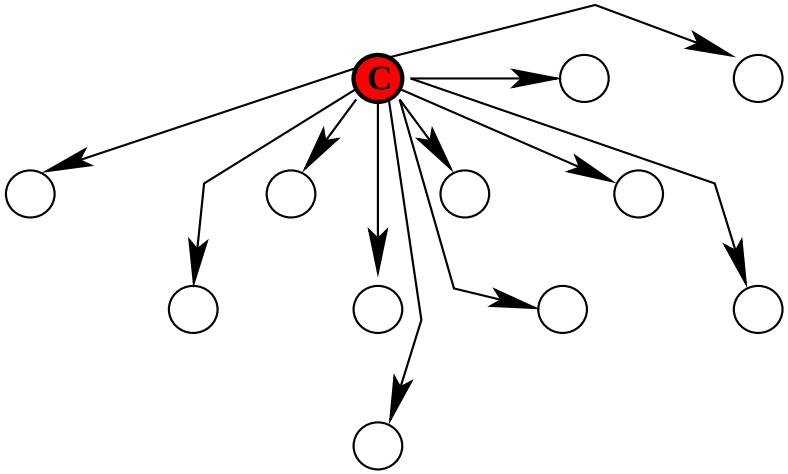
Los clasificadores **Augmented**

ej. Augmented Naive Bayesian Networks (ABNs)

Todos los atributos influyen en la clase, pero

se relaja la suposición de independencia condicional.

Se parte del NB



Nueva estrategia. Restringir el proceso de búsqu

Otras topologías que generalicen el modelo Naive Bayes

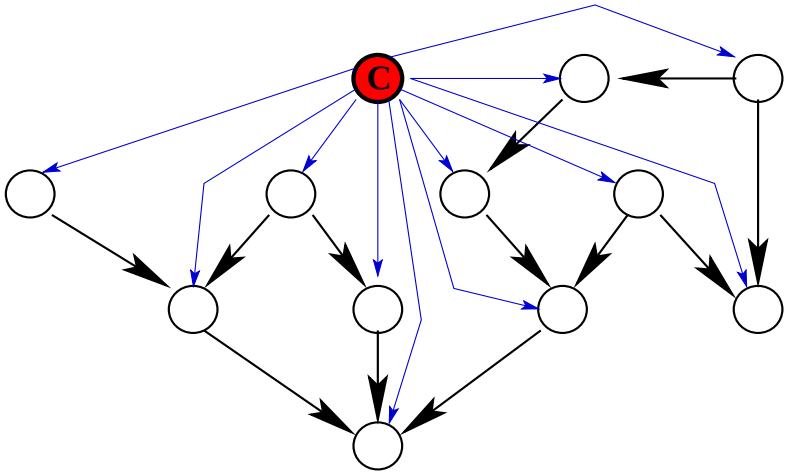
Los clasificadores **Augmented**

ej. Augmented Naive Bayesian Networks (ABNs)

Todos los atributos influyen en la clase, pero

se relaja la suposición de independencia condicional.

Se parte del NB



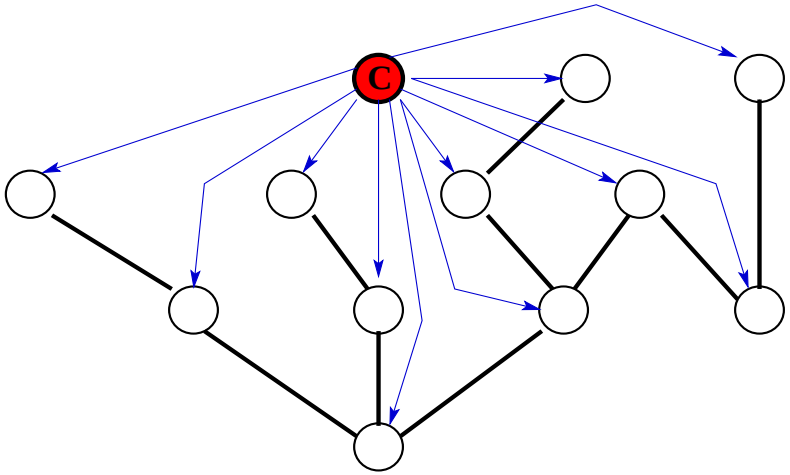
se completa añadiendo arcos entre atributos, *augmented arcs*.

Concepto de *n-dependencia*,

un atributo tiene dependencia con n atributos además de la clase.

`(Sahami, 1996)`

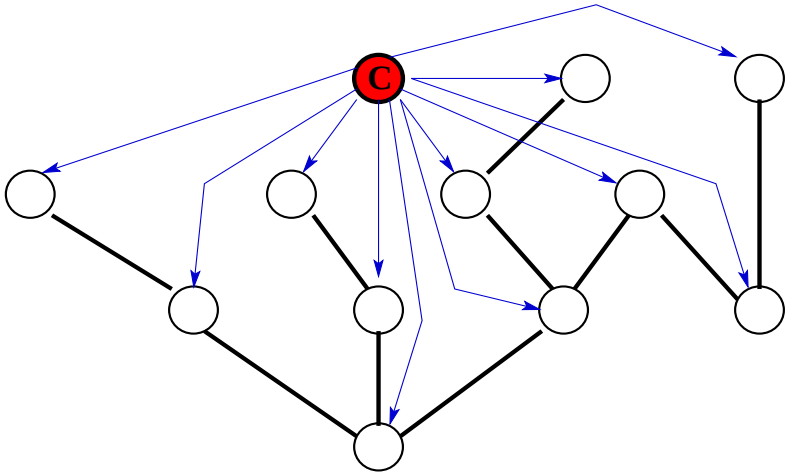
Restringiendo el proceso de búsqueda



tree-augmented naive Bayesian network TAN

un clasificador bayesiano de 1-dependencia

Restringiendo el proceso de búsqueda

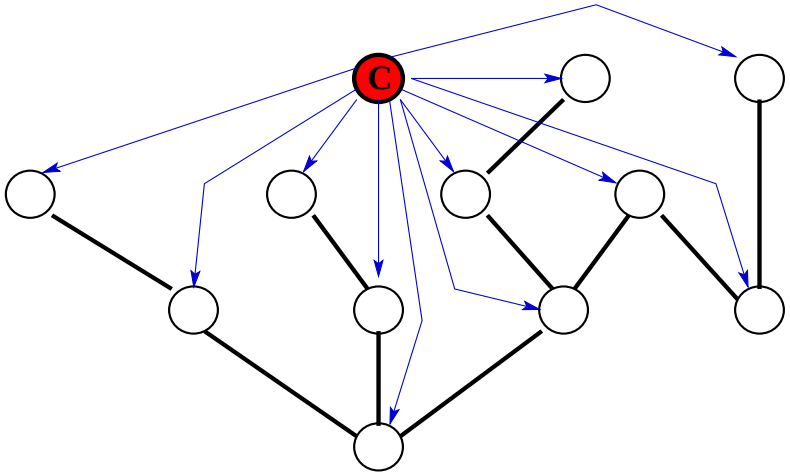


tree-augmented naive Bayesian network TAN

un clasificador bayesiano de 1-dependencia

- (TAN) `Friedman97`
  Es una extensión del método de ChowLiu68, considerando el rol especial de C

Restringiendo el proceso de búsqueda



tree-augmented naive Bayesian network TAN

un clasificador bayesiano de 1-dependencia

- (TAN) `Friedman97`
  Es una extensión del método de ChowLiu68, considerando el rol especial de C
- (TAN) `keogh 2002`
  wrapper que realiza una búsqueda HC sobre estructuras TAN comenzando con el NB y los arco que maximizan TE se añaden *leaving-one-out* para estimar the calidad del clasificador.

Restringiendo el proceso de búsqueda

El clasificador (TAN) de `Friedman97`

Se utiliza la medida de información condicional

$$D(X,Y|C) = \sum_{\mathtt{x},\mathtt{y},\mathtt{c}} P(\mathtt{x},\mathtt{y},\mathtt{c}) \log \frac{P(\mathtt{x},\mathtt{y}|\mathtt{c})}{P(\mathtt{x}|\mathtt{c})P(\mathtt{y}|\mathtt{c})}$$

`Para cada par X, Y`
`Calcula` $D(X,Y|C)$ // peso de los enlaces X,Y
`Construye el árbol expandido de máximo peso` Kruskal
`Orienta el árbol` // se elige un nodo como raíz
`Se introduce` C `y todos los enlaces del NB`

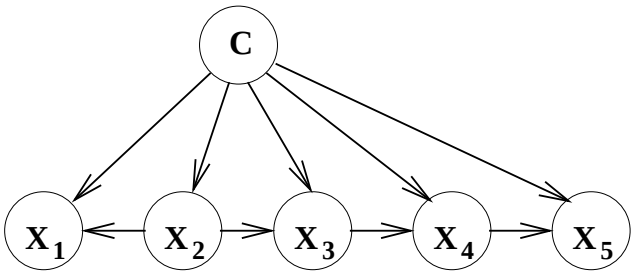
Restringiendo el proceso de búsqueda

El clasificador (TAN) de `Friedman97`

Se utiliza la medida de información condicional

$$D(X,Y|C) = \sum_{\mathtt{x},\mathtt{y},\mathtt{c}} P(\mathtt{x},\mathtt{y},\mathtt{c}) \log \frac{P(\mathtt{x},\mathtt{y}|\mathtt{c})}{P(\mathtt{x}|\mathtt{c})P(\mathtt{y}|\mathtt{c})}$$

`Para cada par X, Y`
`Calcula` $D(X,Y|C)$ // peso de los enlaces X,Y
`Construye el árbol expandido de máximo peso` Kruskal
`Orienta el árbol` // se elige un nodo como raíz
`Se introduce` C `y todos los enlaces del NB`



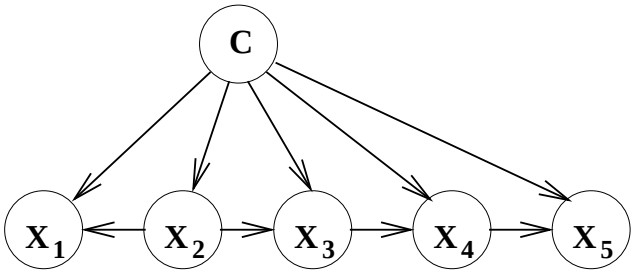
Restringiendo el proceso de búsqueda

El clasificador (TAN) de `Friedman97`

Se utiliza la medida de información condicional

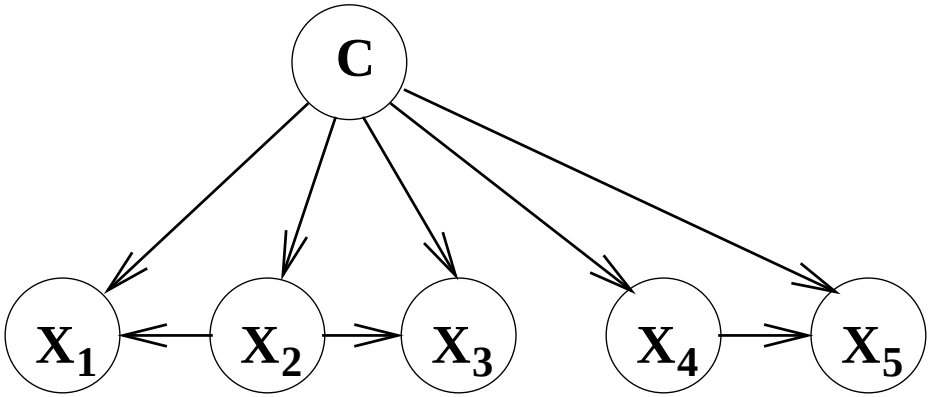
$$D(X,Y|C) = \sum_{\mathtt{x},\mathtt{y},\mathtt{c}} P(\mathtt{x},\mathtt{y},\mathtt{c}) \log \frac{P(\mathtt{x},\mathtt{y}|\mathtt{c})}{P(\mathtt{x}|\mathtt{c})P(\mathtt{y}|\mathtt{c})}$$

`Para cada par X, Y`
`Calcula` $D(X,Y|C)$ // peso de los enlaces X,Y
`Construye el árbol expandido de máximo peso` Kruskal
`Orienta el árbol` // se elige un nodo como raíz
`Se introduce` C `y todos los enlaces del NB`



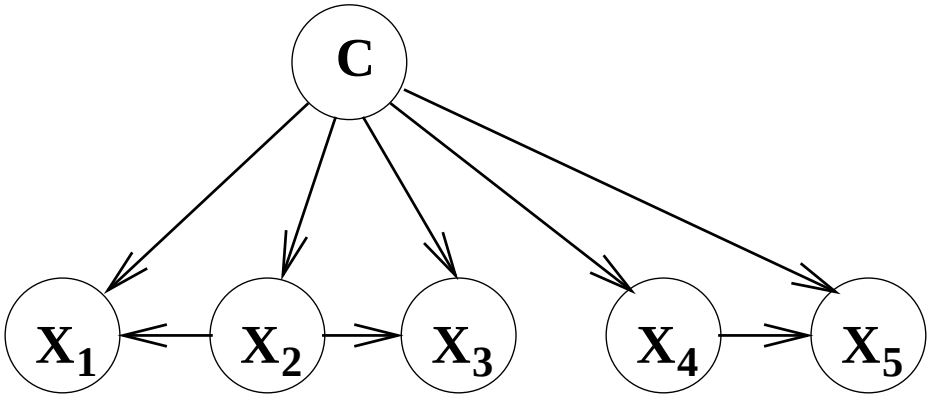
Se demuestra que el TAN así construido maximiza la verosimilitud dado D, $LL(s|D)$

Restringiendo el proceso de búsqueda



forest-augmented naive Bayesian network FAN

Restringiendo el proceso de búsqueda

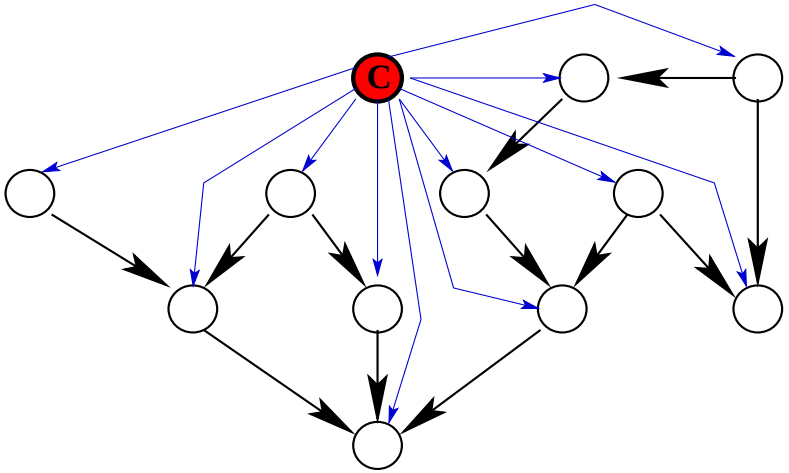


forest-augmented naive Bayesian network FAN

- (FAN) `Lucas 2002`,
  Es una extensión del TAN de `Friedman97`
  una estructura de 1-dependencia con más de un atributo como raíz.

Restringiendo el proceso de búsqueda

bayesian network augmented naive Bayesian clasifier (BAN)



un clasificador bayesiano de n-dependencias

Se fija la estructura NB y
se buscan los arcos entre atributos mediante cualquier algoritmo de aprendizaje d
RRBB.

Restringiendo el proceso de búsqueda

Métodos basados en *métrica+búsqueda*

- BAN de `(Friedman 97)` usa la métrica MDL + HC,
- BAN de `(Ezawa 96)`
  extensión del algoritmo K2 `(Cooper, Herskovits)`

Restringiendo el proceso de búsqueda

Métodos basados en *métrica+búsqueda*

- BAN de `(Friedman 97)` usa la métrica MDL + HC,
- BAN de `(Ezawa 96)`
  extensión del algoritmo K2 `(Cooper, Herskovits)`

Métodos basados en *independencias*

- BAN de `(Cheng,Greiner 99)`

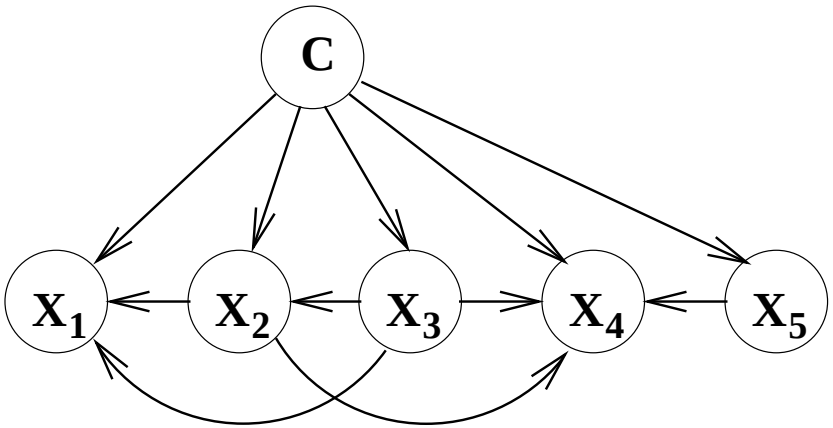
Restringiendo el proceso de búsqueda

Métodos basados en *métrica+búsqueda*

- BAN de `(Friedman 97)` usa la métrica MDL + HC,
- BAN de `(Ezawa 96)`
  extensión del algoritmo K2 `(Cooper, Herskovits)`

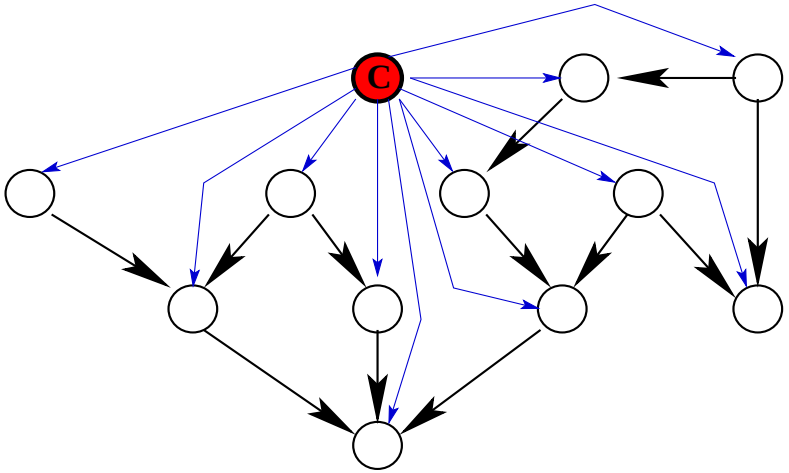
Métodos basados en *independencias*

- BAN de `(Cheng,Greiner 99)`



Restringiendo el proceso de búsqueda

bayesian network augmented naive Bayesian clasifier (KDB)



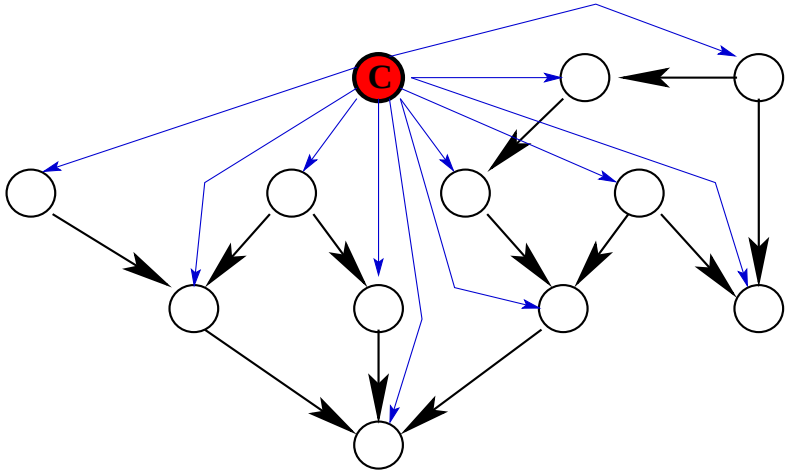
$k = 3$

un clasificador bayesiano de k-dependencias

- `Sahami96`
  con k como parámetro
  Utiliza la información mútua de KL + búsqueda greedy

Restringiendo el proceso de búsqueda

Se relaja la suposición de independencia condicional y
todos los atributos influyen en la clase, (son hijos).



Otras topologías que generalicen aún más el modelo Naive Bayes

- **El Manto de Markov** un concepto clave para la influencia sobre la variable clase C.

Modificando el proceso de búsqueda

El concepto de Manto de Markov

Sean X_i, G;

$MB_G(X_i)$ incluye Π_{X_i}, los padres de X_i;

$Y_j(X_i)$, los hijos de X_i y $F_j(X_i)$,
los padres de los hijos de X_i en G.

El $MB_G(X_i)$ tiene la propiedad de:
$I(X_i, X - MB_G(X_i) - X_i | MB_G(X_i))$

Modificando el proceso de búsqueda

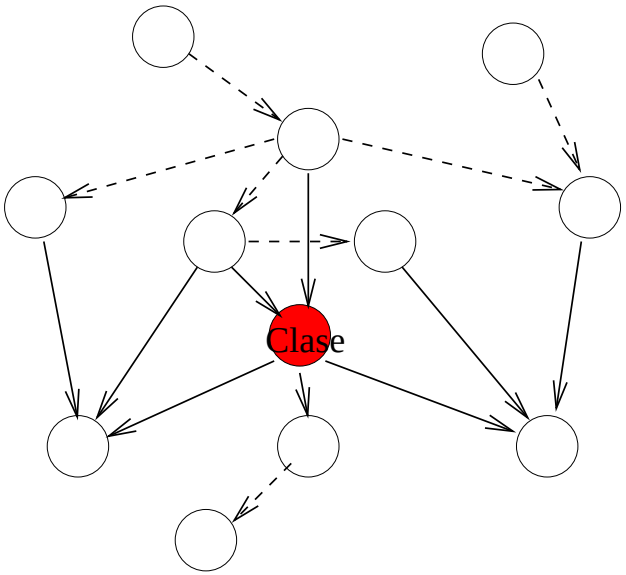
El concepto de Manto de Markov

Sean X_i, G;

$MB_G(X_i)$ incluye Π_{X_i}, los padres de X_i;

$Y_j(X_i)$, los hijos de X_i y $F_j(X_i)$,
los padres de los hijos de X_i en G.

El $MB_G(X_i)$ tiene la propiedad de:
$I(X_i, X - MB_G(X_i) - X_i | MB_G(X_i))$



Modificando el proceso de búsqueda

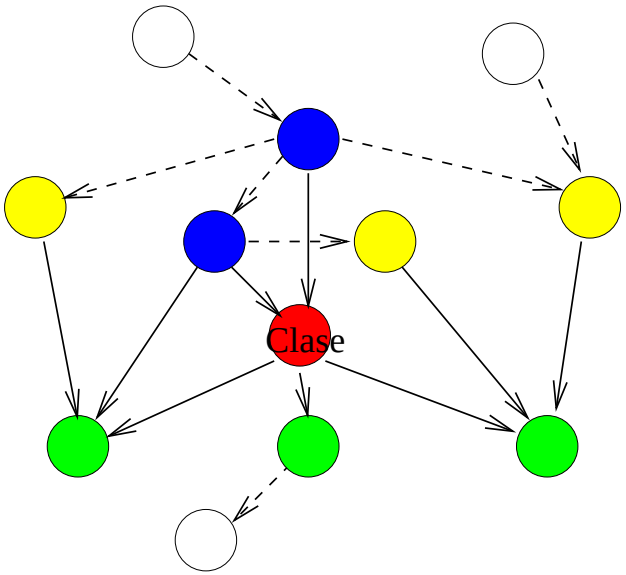
El concepto de Manto de Markov

Sean X_i, G;

$MB_G(X_i)$ incluye Π_{X_i}, los padres de X_i;

$Y_j(X_i)$, los hijos de X_i y $F_j(X_i)$,
los padres de los hijos de X_i en G.

El $MB_G(X_i)$ tiene la propiedad de:
$I(X_i, X - MB_G(X_i) - X_i | MB_G(X_i))$



Modificando el proceso de búsqueda

Los algoritmos MB

Utilizando el $MB_G(c)$ la distribución de probabilidad de C condicionada al estado del resto queda descompuesta

$$P(c|x) = \alpha P(c|\pi_c)\Pi_j P(y_j|f_j(c))$$

α una constante norm, c,x,π_c,y_j y $f_j(c)$ valores para C, $X = X_1 \ldots X_n$, Π_j Y_j, y F_j.
Π_C padres de C, Y_j hijos de C, y F_j padres de los hijos de C.

Nuevo Espacio de búsqueda los posibles $MB(C)$

cualquier conectividad entre las variables que forman parte del $MB(C)$ y la variable clase

- MB-GA (`Sierra and Larrañaga,1998`)
  Wrapper + búsqueda Alg. genéticos
  se imponen restricciones entre la conectividad entre los atributos predictores

Consecuencia,
se realiza una **Selección de Características**

Modificando el proceso de búsqueda

Un algoritmo aproximado hacia atrás para $MB(C)$

Backward Sequential Feature Selection

$x_j = (X_1 = x_{j1}, X_2 = x_{j2}, \ldots, X_n = x_{jn})$

$g_j = (X_i = x_{ji},\ i = 1..n / X_i \in MB_G(C))$

$$\lambda_G(x_j) = D(P(C|x_j), P(C|g_j))$$

Buscamos el subconjunto $G \subset \mathbf{X}$, nuestro $MB_G(C)$ que minimice

$$\Delta_G = \sum_{x_j} P(x_j)\lambda_G(x_j)$$

$G = X_1, X_2, X_3, ...X_n$ // Conjunto completo
repeat
    para cada atributo X_i de G
        eliminar X_i / $arg\ min_{x_i} \Delta_G - \{X_i\}$
until $|G| = k$ // número predefinido de atrib.
o $\Delta_G - \{X_i\}$ próximo a Δ_G

Problema: $MB(C)$ suele ser grande -> costoso

Modificando el proceso de búsqueda

Proceso de búsqueda local

Red bayesiana inicial G_0

```
completamente inconexa
con una configuración fija
con una configuración aleatoria
```

Repetir

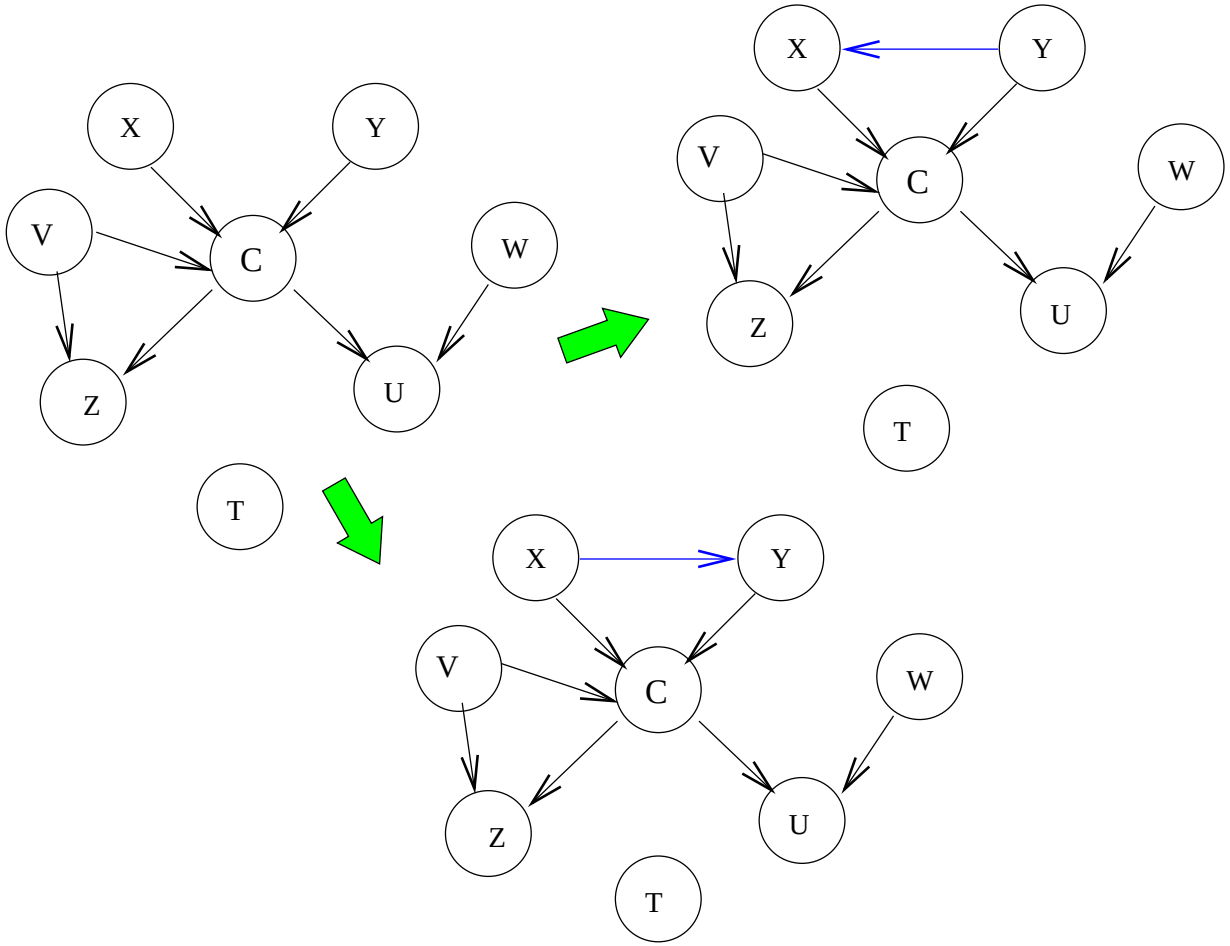
```
evaluar posibles cambios // n cambios
aplicar los cambios que supongan una mejora en el score
repetir
```

hasta que no se pueda mejorar score

Se trata de buscar en el *espacio de clases de equivalencia*
éste espacio es más reducido que el espacio de DAGs

Modificando el proceso de búsqueda

I sea $G_i = (X, E_i)$ ¿Qué es una configuración vecina?



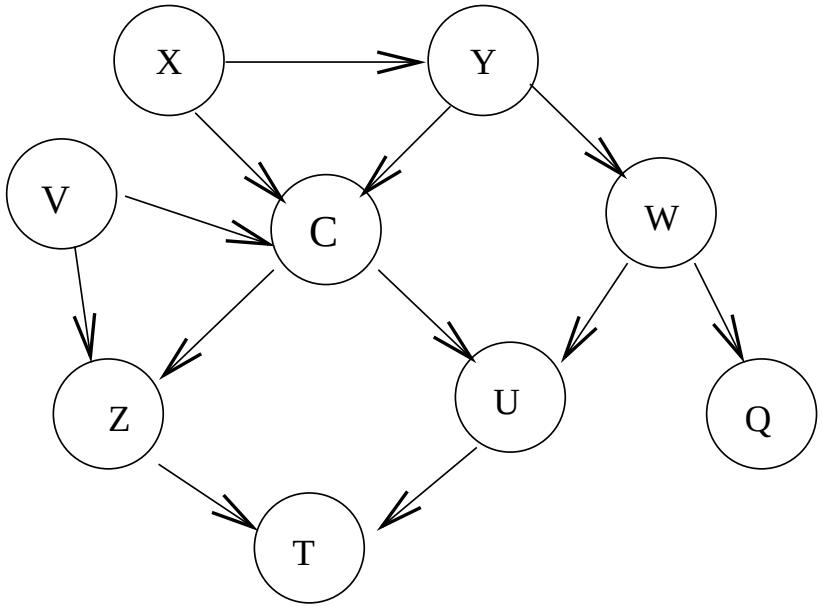
Modificando el proceso de búsqueda

Concepto de equivalencia por clasificación

Modificando el proceso de búsqueda

Concepto de equivalencia por clasificación

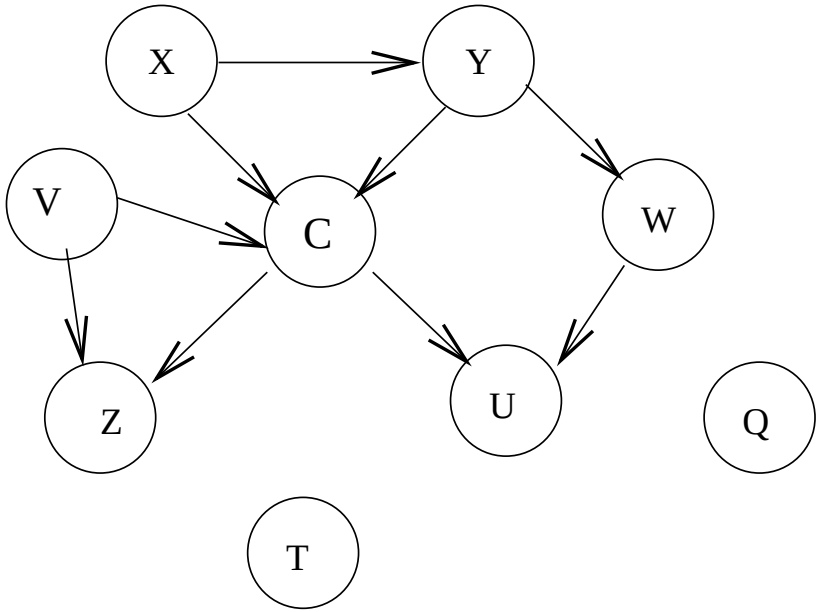
sea $G_1 = (X, E_1)$



Modificando el proceso de búsqueda

Concepto de equivalencia por clasificación

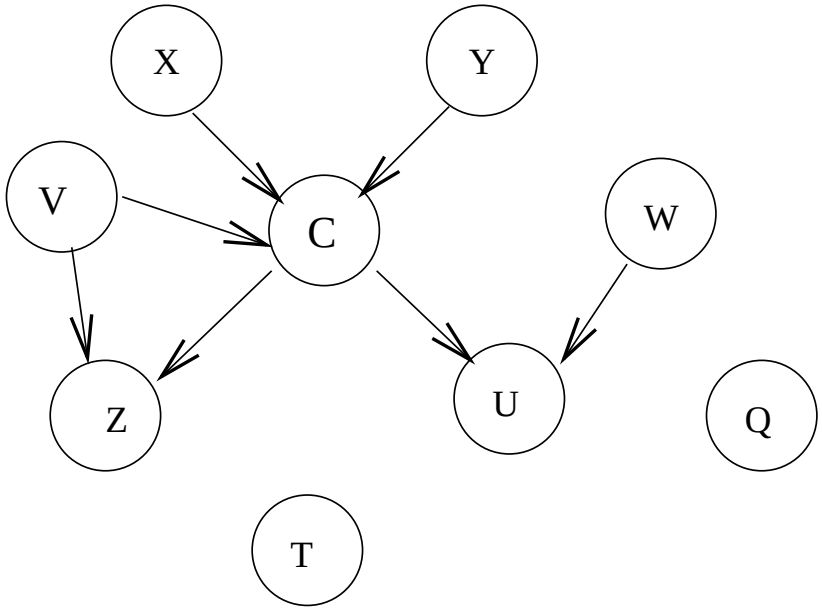
sea $G_2 = (X, E_2)$



Modificando el proceso de búsqueda

Concepto de equivalencia por clasificación

sea $G_3 = (X, E_3)$



el mismo valor para $P_{G_i}(C/x)\ \forall x$,

las 3 RRBB son equivalentes por clasificación.

Modificando el espacio de búsqueda

Un C-DAG $G = (U, E)$ *Class-focused DAG*

$$\forall X, Y \in \mathbf{U}, \text{ si } X \rightarrow Y \in E \text{ entonces}$$

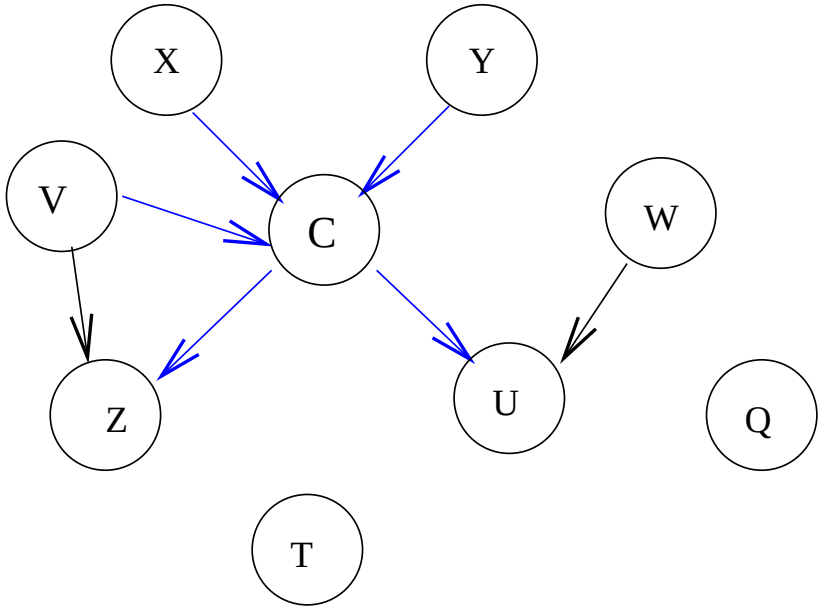
$$\text{o } Y = C \text{ o } X = C \text{ o } C \rightarrow Y \in E$$

Modificando el espacio de búsqueda

Un C-DAG $G = (U, E)$ *Class-focused DAG*

$$\forall X, Y \in \mathbf{U}, \text{ si } X \rightarrow Y \in E \text{ entonces}$$

$$\text{o } Y = C \text{ o } X = C \text{ o } C \rightarrow Y \in E$$

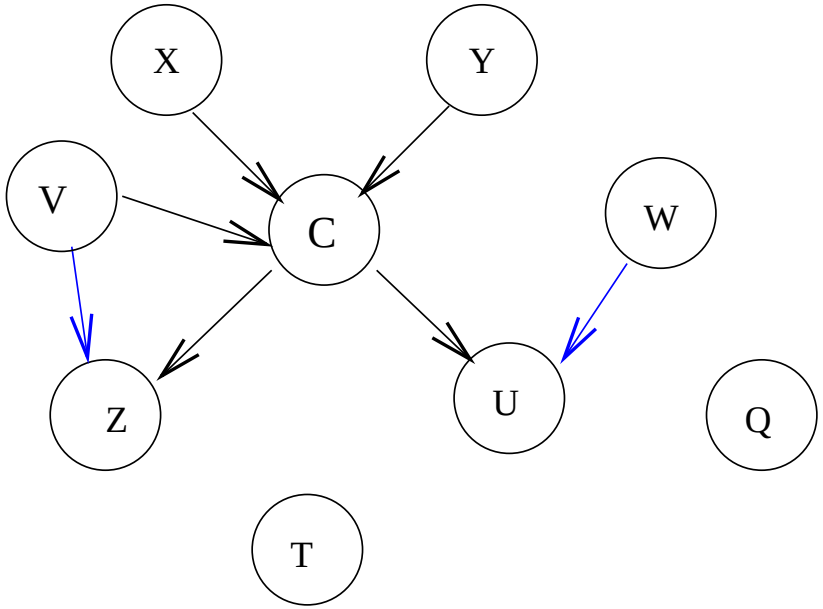


Modificando el espacio de búsqueda

Un C-DAG $G = (U, E)$ *Class-focused DAG*

$$\forall X, Y \in \mathbf{U}, \text{ si } X \rightarrow Y \in E \text{ entonces}$$

$$\text{o } Y = C \text{ o } X = C \text{ o } C \rightarrow Y \in E$$

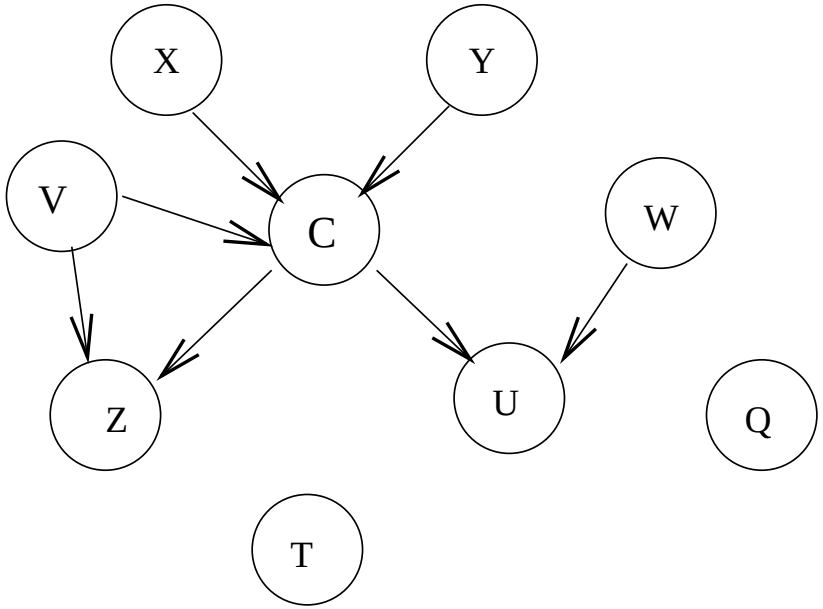


Modificando el espacio de búsqueda

Un C-DAG $G = (U, E)$ *Class-focused DAG*

$$\forall X, Y \in \mathbf{U}, \text{ si } X \rightarrow Y \in E \text{ entonces}$$

$$\text{o } Y = C \text{ o } X = C \text{ o } C \rightarrow Y \in E$$

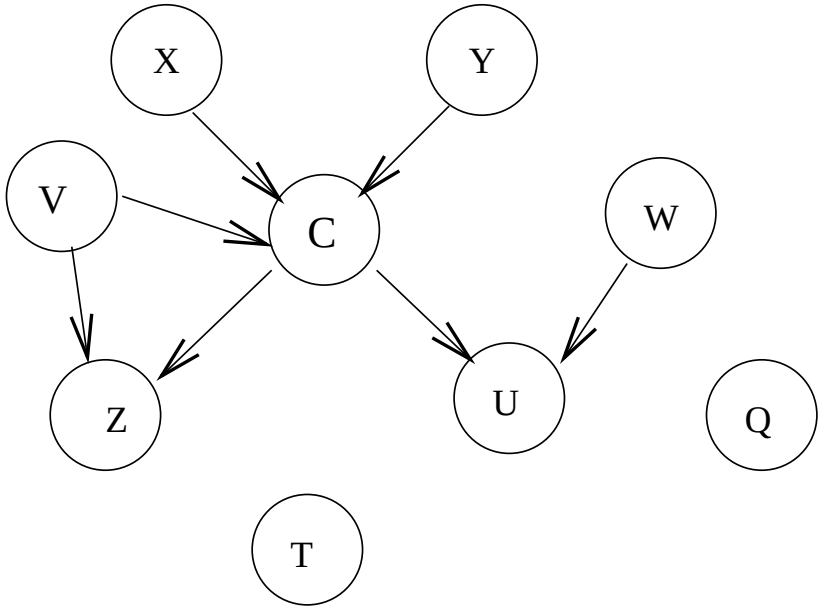


Modificando el espacio de búsqueda

Un C-DAG $G = (U, E)$ *Class-focused DAG*

$$\forall X, Y \in \mathbf{U}, \text{ si } X \rightarrow Y \in E \text{ entonces}$$

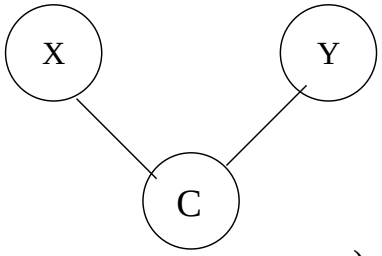
$$\text{o } Y = C \text{ o } X = C \text{ o } C \rightarrow Y \in E$$



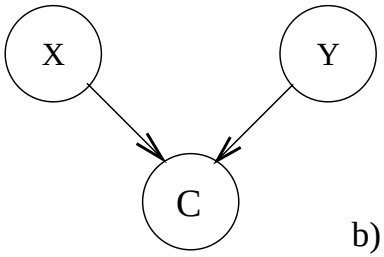
representación canónica de las clases de equivalencia por clasificación

Modificando el proceso de búsqueda

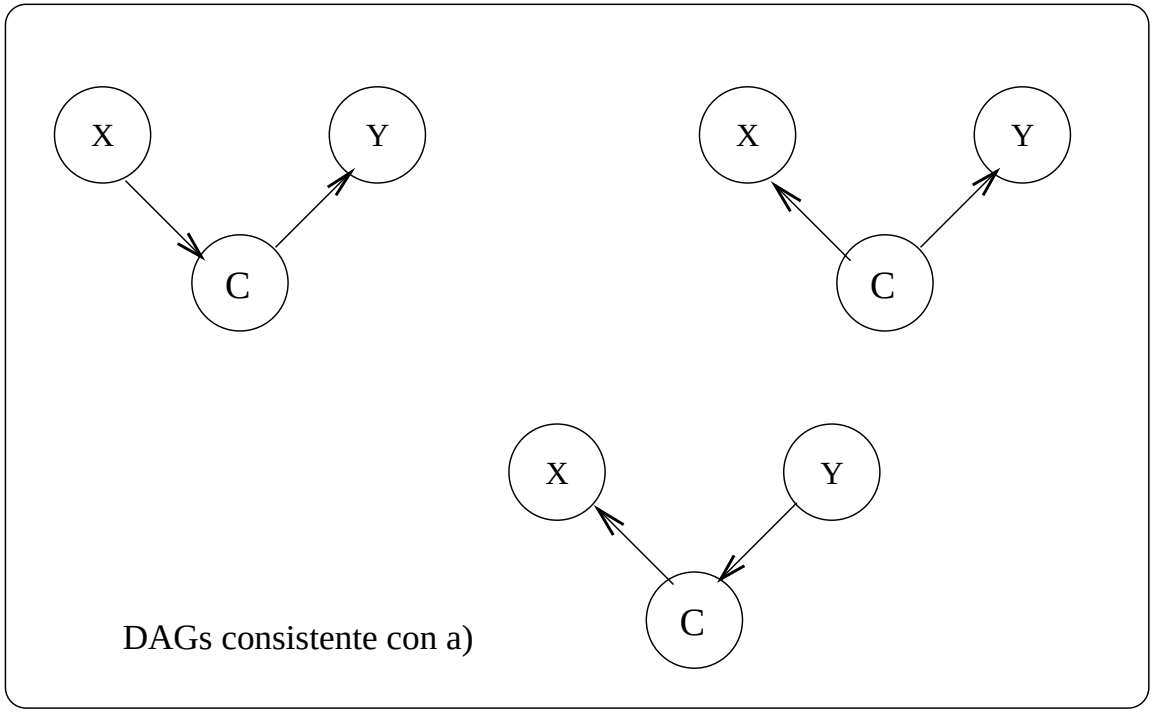
Aprovechar la estructura del dominio



a)



b)



Nuevos modelos

Reduciendo aún más el número de configuraciones del espacio.

Se trata de buscar en el espacio de
clases de equivalencia por independencia

Cada clase de equivalencia es representada por:
RPDAG o PDAG (dag parcialmente orientado (restringido) compuesto de *esqueleto*
+ *estructurasCabezaCabeza*

Espacio más reducido que el espacio de DAGs
Espacio con menos máximos locales y menos mesetas

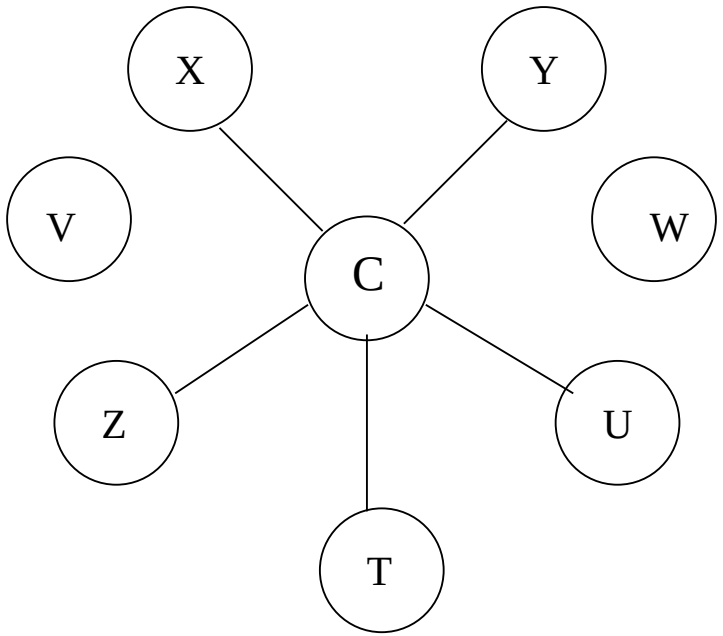
Nuevos modelos

Moviéndonos en el espacio de los C-RPDAGs

Nuevos modelos

Moviéndonos en el espacio de los C-RPDAGs

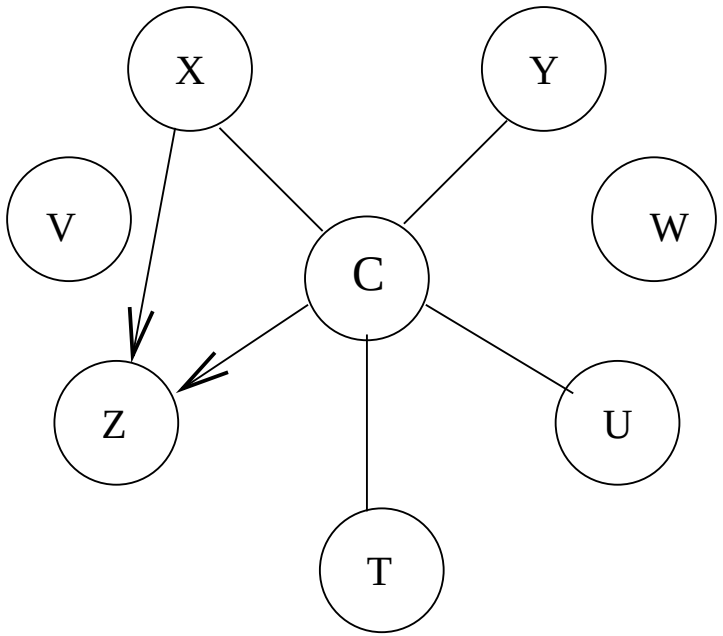
- C-RPDAG (`Acid and de Campos,2005`)
  búsqueda local + métrica general



Nuevos modelos

Moviéndonos en el espacio de los C-RPDAGs

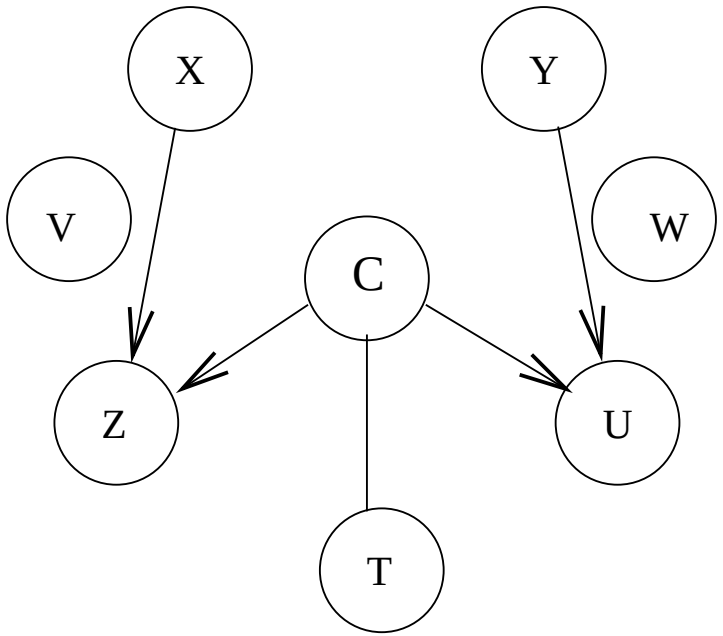
- C-RPDAG (`Acid and de Campos,2005`)
  búsqueda local + métrica general



Nuevos modelos

Moviéndonos en el espacio de los C-RPDAGs

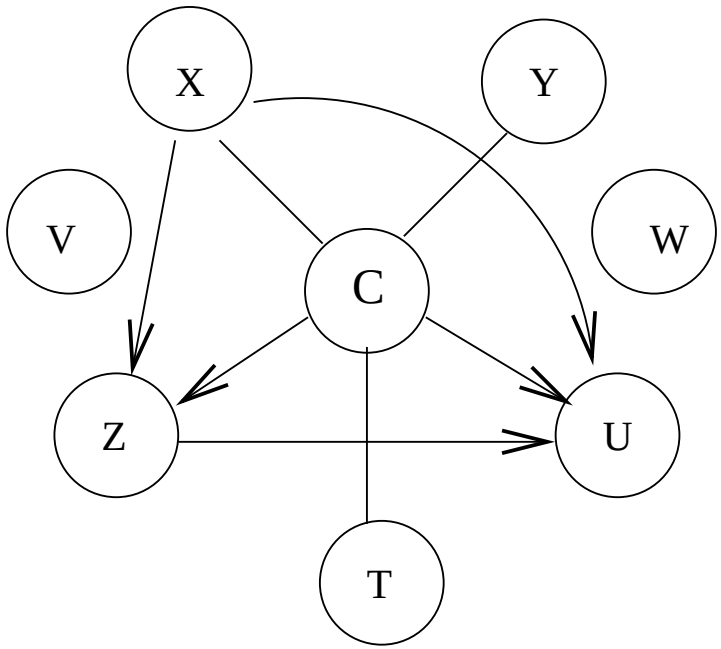
- C-RPDAG (`Acid and de Campos,2005`)
  búsqueda local + métrica general



Nuevos modelos

Moviéndonos en el espacio de los C-RPDAGs

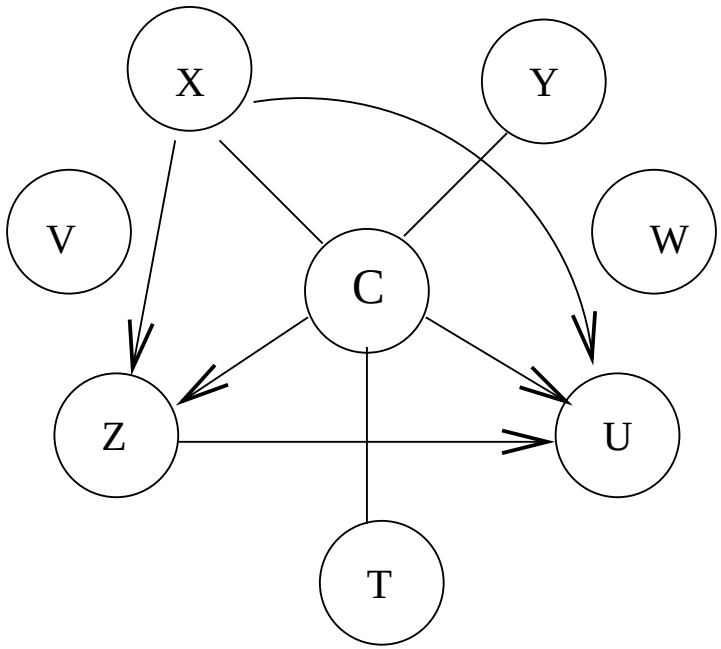
- C-RPDAG (`Acid and de Campos,2005`)
  búsqueda local + métrica general



Nuevos modelos

Moviéndonos en el espacio de los C-RPDAGs

- C-RPDAG (`Acid and de Campos,2005`)
  búsqueda local + métrica general

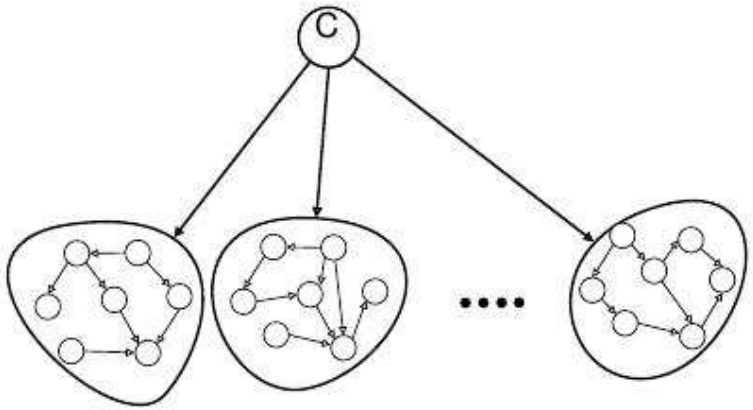


Los RPDAGs son las clases de equivalencia por independencia

otra clases de equivalencia por independencia

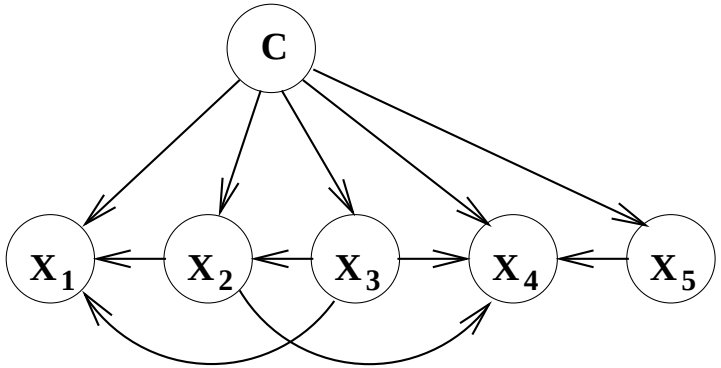
Nuevos modelos

Modelos híbridos, las Multiredes *Bayesian multinets*



Nuevos modelos

Generalización del Naive Bayes Un RRBB establece que las relaciones entre atributos X_i son las mismas para $C = c_i \; \forall i, \;\; i = 1..r_C$



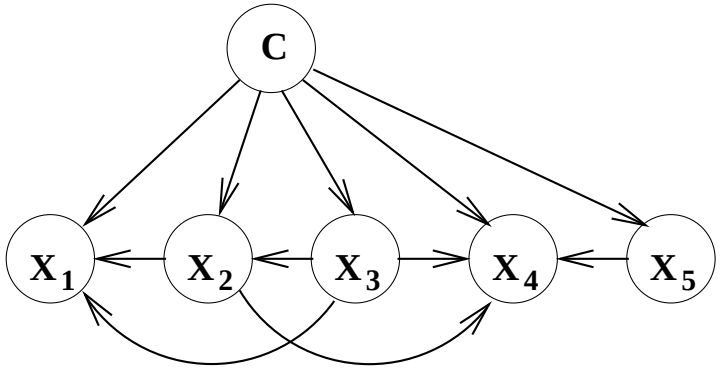
Nueva generalización:

existen relaciones entre $X_i, X_j \;\; i \neq j$ distintas para cada clase.

permite representar independencias asimétricas (`Heckerman 1991`)

Nuevos modelos

Generalización del Naive Bayes Un RRBB establece que las relaciones entre atributos X_i son las mismas para $C = c_i \; \forall i, \;\; i = 1..r_C$

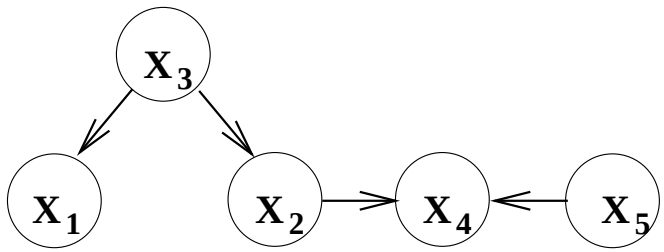


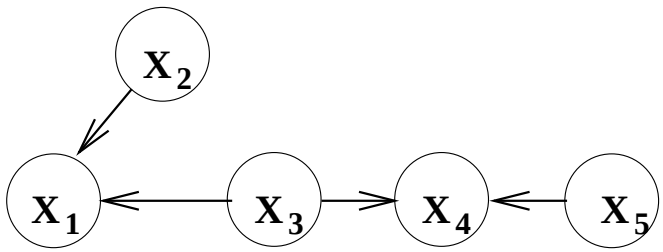
Nueva generalización:

existen relaciones entre $X_i, X_j \;\; i \neq j$ distintas para cada clase.

permite representar independencias asimétricas (`Heckerman 1991`)

C=c1 C=c2





Nuevos modelos

Las multiredes como clasificadores [Friedman et al.,1997]. Sea C la clase con k valores y $P(C)$ a priori

$M = (P_C, B_1, ...B_k)$

Se parte D en k particiones, $D_k = \{(x, c) \in D / c = c_k\}$

Se induce una RB, B_i, apartir de D_i $i = 1..k$ sobre $X_1, \ldots X_n$

Una multired define:

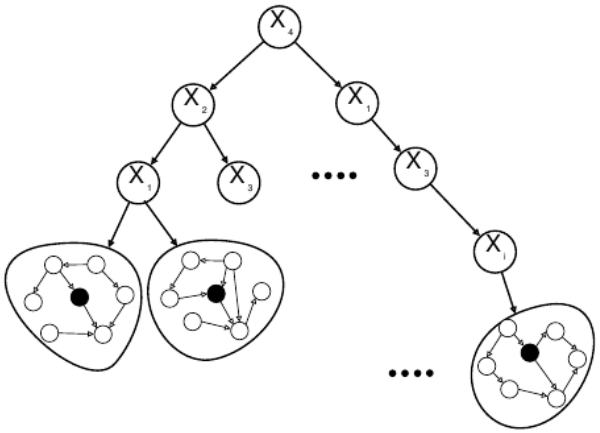
$$P_M(C, X_1, \ldots X_n) = P_C(C) P_{B_i}(X_1, \ldots X_n) \;\; con \;\; C = c_i$$

en realidad $P_{B_i}(X_1, \ldots X_n)$ aproxima $P_D(X_1, \ldots X_n | C = ci)$ Se clasifica en el valo de la clase que maximiza

$$P_M(C | X_1, \ldots X_n)$$

Nuevos modelos

Modelos híbridos, multiredes bayesianas recursivas



Caso particular, los Naive Bayes Tree (Kohavi 1996)

Y

hasta aquí hemos llegado

References

[Acid and de Campos,2005] Acid, S., & de Campos, L.M. & Castellano, J.G (2005). Learning bayesia network classifiers: searching in a Space of Partially directed acyclic graphs. *Machine Learning, 59*, 213–235.

[Cheng and Greiner,1999] Cheng, J., & Greiner, R. (1999). Comparing Bayesian network classifiers. *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (pp. 101–10

[Chickering,2002] Chickering, D.M (1992). Learning equivalence classes of Bayesian network structures. *Journal of Machine Learning Research, 2*, 30445–498.

[Chow and Liu,1968] Chow, C., & Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory, 14*, 462–467.

[Cooper and Herskovits,1992] Cooper, G.F., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9*, 309–348.

[Domingos, P. and Pazzani, M. (1997)] Domingos, P. and Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 29(2-3), 103–130.

[Ezawa et al.,1996] Ezawa, K., Singh, M., & Norton, S. (1996). Learning goal oriented Bayesian networks for telecommunications risk management. In *Proceedings of the Thirteenth International Conference on Machine Learning* (pp. 139–147).

References

[Friedman et al.,1997] Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifier *Machine Learning, 29*, 131–163.

[Inza, I., Larranaga, P., and Sierra, B. (2001)] . Feature subset selection by Bayesian networks: a comparison with genetic and sequential algorithms. International Journal of Approximate Reasoning, 27(2), 143–164.

[Keogh and Pazzani,2002] Keogh, E., & Pazzani, M. (2002). Learning augmented Bayesian classifier *International Journal on Artificial Intelligence Tools, 11*, 587–601.

[Kononenko 90] Kononenko, I. (1990). Current Trends in Knowledge Acquisition, chapter *Comparison inductive and naive Bayesian learning approaches to automatic knowledge acquisition. IOS Press*.

[Kononenko,1991] Kononenko, I. (1991). Semi-naive Bayesian classifier. In *Proceedings of the Secon International Conference on Knowledge Discovery in Databases* (pp. 206–219).

[Langley and Sage,1994] Langley, P., & Sage, S. (1994). Induction of selective Bayesian classifiers. I *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence* (pp. 399–406).

[Lucas,2002] Lucas, P. (2002). Restricted Bayesian network structure learning. In *Proceedings of the First European Workshop on Probabilistic Graphical Models* (pp. 117–126).

[Pazzani,1995] Pazzani, M.J. (1995). Searching for dependencies in Bayesian classifiers. *Lecture Notes in Statistics, 112*, 239–248.

References

[Sahami,1996] Sahami, M. (1996). Learning limited dependence Bayesian classifiers. In *Proceedings the Second International Conference on Knowledge Discovery and Data Mining* (pp. 335–33

[Sierra and Larrañaga,1998] Sierra, B., & Larrañaga, P. (1998). Predicting survival in malignant skin melanoma using Bayesian networks automatically induced by genetic algorithms. An empiri comparison between different approaches. *Artificial Intelligence in Medicine, 14*, 215–230.

[Singh and Provan,1996] Singh, M., & Provan, G.M. (1996). Efficient learning of selective Bayesian network classifiers. In *Proceedings of the Thirteenth International Conference on Machine Learning* (pp. 453–461).

[Vinciotti, V., Tucker, A., Kellam, P., and Liu, X. (2006)] The robust selection of predictive genes via a simple classifier. In *Applied Bioinformatics, 5(1)*,1–11.

References