Inteligencia Artificial

CC-421

César Lara Avila

Universidad Nacional de Ingeniería

(actualización: 2020-12-08)

Bienvenidos

Clasificadores Bayesianos

- Introducción a la clasificación
- Evaluación de un clasificador
- Clasificador Bayesiano
- Modelos alternativos: TAN, BAN
- Clasificadores Bayesianos Multidimensionales
- Clasificación Jerárquica

Introducción a la clasificación

La clasificación consiste en asignar clases o etiquetas a los objetos. Hay dos tipos básicos de problemas de clasificación:

- Sin supervisión: en este caso las clases son desconocidas, por lo que el problema consiste en dividir un conjunto de objetos en n grupos o clusters, de manera que se asigne una clase a cada grupo diferente.
- Supervisado: las posibles clases o etiquetas se conocen a priori, y el problema consiste en encontrar una función o regla que asigne cada objeto a una de las clases.

Desde una perspectiva probabilística, el problema de la clasificación supervisada consiste en asignar a un objeto particular descrito por sus atributos, $A_1, A_2, \ldots, A_n$, una de las m clases, $C = \{c_1, c_2, \ldots, c_m\}$, de modo que se maximice la probabilidad de la clase dada los atributos; es decir:

$$Arg_C[\max P(C|A_1, A_2, \ldots, A_n)]$$

Si denotamos el conjunto de atributos como $\mathbf{A} = \{A_1, A_2, \ldots, A_n\}$, la ecuación anterior se puede escribir como: $Arg_C[\max P(C|A)]$.

Evaluación de un clasificador

Los principales aspectos a considerar en la evaluación de un clasificador son:

- Exactitud
- Tiempo de clasificación
- Tiempos de entrenamiento
- Requisitos de memoria
- Claridad

La exactitud en porcentaje se define como: $Acc = (N_C/N) \times 100$ donde N_c es el número de correctas predicciones.

Cuando existe un desequilibrio en los costos de clasificación errónea, debemos minimizar el costo esperado (EC). Para dos clases, esto viene dado por:

$$EC = FN \times P(-)C(-|+) + FP \times P(+)C(+|-)$$

donde: FN es la tasa de falsos negativos, FP es la tasa de falsos positivos, $P(+)$ es la probabilidad de positivo, $P(-)$ es la probabilidad de negativo, $C(-|+)$ es el costo de clasificar un positivo como negativo y $C(+|-)$ es el costo de clasificar un negativo como positivo.

Clasificador Bayesiano

La formulación del Clasificador Bayesiano se basa en la aplicación de la regla de Bayes para estimar la probabilidad de cada clase dados los atributos:

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

que se puede escribir de forma más compacta como:

$$P(C|\mathbf{A}) = P(C)P(\mathbf{A}|C)/P(\mathbf{A})$$

El problema de clasificación (basada en la ecuación anterior), se puede formular como:

$$Arg_C[\max[P(C|\mathbf{A}) = P(C)P(\mathbf{A}|C)/P(\mathbf{A}]]$$

Podemos expresar la ecuación anterior en términos de cualquier función que varíe monótonamente con respecto a $P(C|\mathbf{A})$, por ejemplo:

- $Arg_C[\max[P(C)P(\mathbf{A}|C)]]$
- $Arg_C[\max[\log(P(C)P(\mathbf{A}|C))]]$
- $Arg_C[\max[(\log P(C) + \log P(\mathbf{A}|C)]]$

Clasificador Bayesiano

La probabilidad de los atributos, $P(\mathbf{A})$, no varía con respecto a la clase, por lo que se puede considerar como una constante para la maximización.

Con base en las formulaciones anteriores para resolver un problema de clasificación, necesitaremos una estimación de $P(C)$, conocida como probabilidad previa de las clases, $P(\mathbf{A}|C)$, conocida como likelihood y $P(C|\mathbf{A})$ es la probabilidad posterior. Por lo tanto, para obtener la probabilidad posterior de cada clase, solo necesitamos multiplicar su probabilidad previa por la likelihood que depende de los valores de los atributos.

El clasificador bayesiano sólo puede ser de uso práctico para problemas relativamente pequeños en términos de número de atributos. Una alternativa es considerar algunas propiedades de independencia como en los modelos de grafos , en particular que todos los atributos son independientes dada la clase, lo que da como resultado el clasificador Naive Bayesiano.

Clasificador Naive Bayesiano

Se basa en el supuesto de que todos los atributos son independientes dada la variable de clase, es decir, cada atributo A_i es condicionalmente independiente de todos los demás atributos dada la clase:
$P(A_i|A_j, C) = P(A_i|C), \forall j \neq i.$

Así:

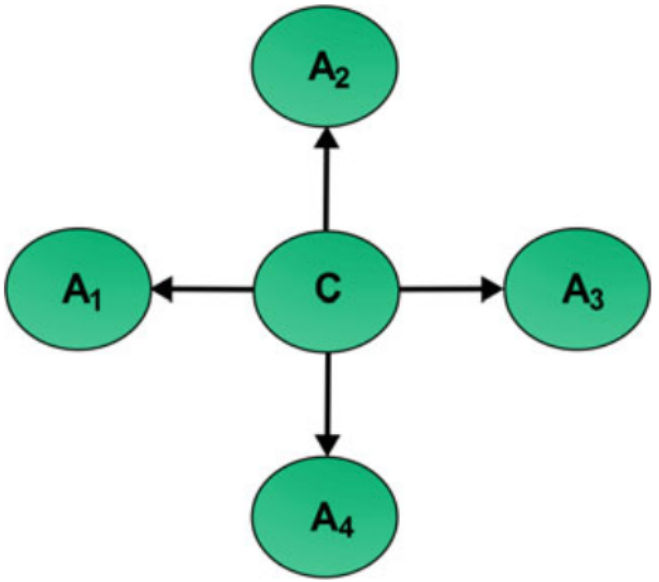
$$P(C|A_1, A_2, \ldots, A_n) = P(C)P(A_1|C)P(A_2|C)\ldots P(A_n|C)/P(\mathbf{A})$$

donde $P(\mathbf{A})$ es una constante de normalización.

La formulación Naive Bayes reduce drásticamente la complejidad del clasificador bayesiano, ya que en este caso solo requerimos la probabilidad previa (vector unidimensional) de la clase, y las n probabilidades condicionales de cada atributo dada la clase (matrices bidimensionales) como los parámetros para el modelo.

Aprender un NBC consiste en estimar la probabilidad previa de la clase, $P(C)$ y la probabilidad condicional de cada atributo dada la clase, $P(A_i|C)$. Estos pueden obtenerse mediante estimaciones subjetivas de un experto o de los datos por máxima verosimilitud.

Clasificador Naive Bayesiano



Las probabilidades se pueden estimar a partir de datos utilizando, por ejemplo, la estimación de máxima verosimilitud. Las probabilidades previas de la variable de clase, C, vienen dadas por:

$$P(c_i) \sim \frac{N_i}{N}$$

donde N_i es el número de veces que c_i aparece en las N muestras.

Clasificador Naive Bayesiano

Las probabilidades condicionales de cada atributo, A_j, se pueden estimar como:

$$P(A_{jk}|c_i) \sim \frac{N_{jki}}{N}$$

donde N_{jki} es el número de veces que el atributo A_j toma el valor k y es de la clase i, y N_i es el número de muestras de la clase c_i en el conjunto de datos.

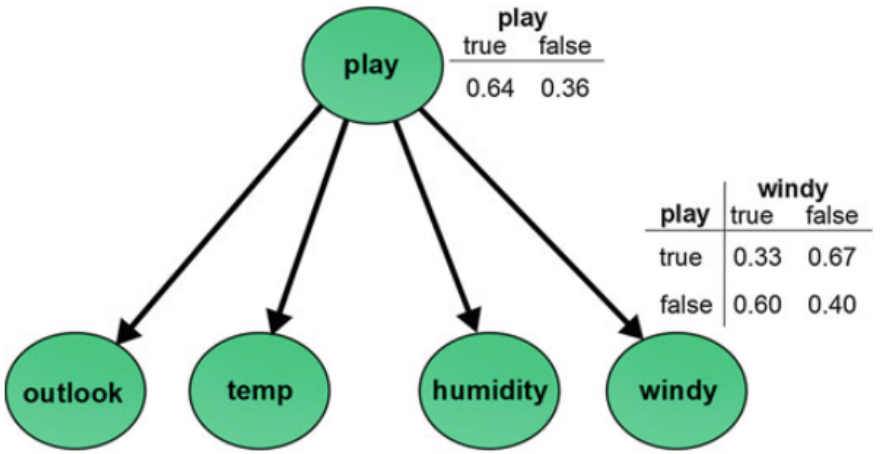
Una vez estimados los parámetros, la probabilidad posterior se puede obtener simplemente multiplicando la probabilidad anterior por el likelihood de cada atributo. Por tanto, dados los valores de m atributos, $a_1, \ldots, a_m$, para cada clase c_i, el posterior es proporcional a:

$$P(c_i|a_1, \ldots, a_m) \sim P(c_i)P(a_1|c_i) \ldots P(a_m|c_i)$$

Se seleccionará la clase c_k que maximice la ecuación anterior.

Ejemplo

Se muestra un NBC para el ejemplo del tenis, incluidas algunas de las tablas de probabilidad requeridas.



Ventajas y desventajas de un Clasificador Naive Bayesiano

Las principales ventajas del clasificador Naive Bayesiano son:

- El bajo número de parámetros requeridos, lo que reduce los requisitos de memoria y facilita su aprendizaje a partir de los datos.
- El bajo costo computacional de la inferencia (estimación de los posteriores) y el aprendizaje.
- El desempeño relativamente bueno (precisión de clasificación) en muchos dominios.
- Es un modelo simple e intuitivo.

Las principales limitaciones del clasificador Naive Bayesiano son las siguientes:

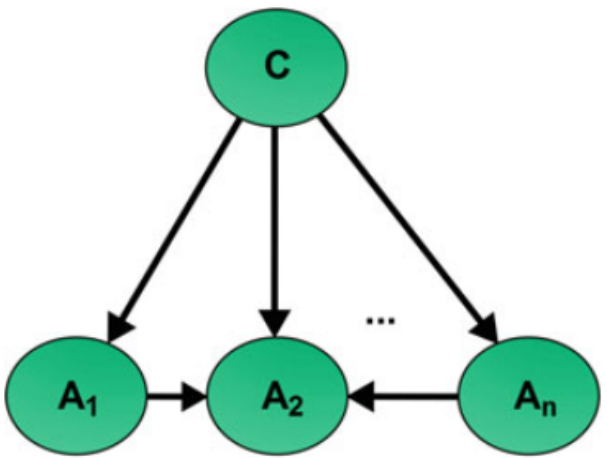
- En algunos dominios, el rendimiento se reduce dado que el supuesto de independencia condicional no es válido.
- Si hay atributos continuos, estos deben discretizarse (o considerar modelos alternativos como el discriminador lineal).

Modelos alternativos: TAN, BAN

TAN (Tree augmented Bayesian Classifier) incorpora algunas dependencias entre los atributos mediante la construcción de un árbol dirigido entre las variables de atributo.

Es decir, los n atributos forman un grafo que está restringido a un árbol dirigido que representa las relaciones de dependencia entre los atributos.

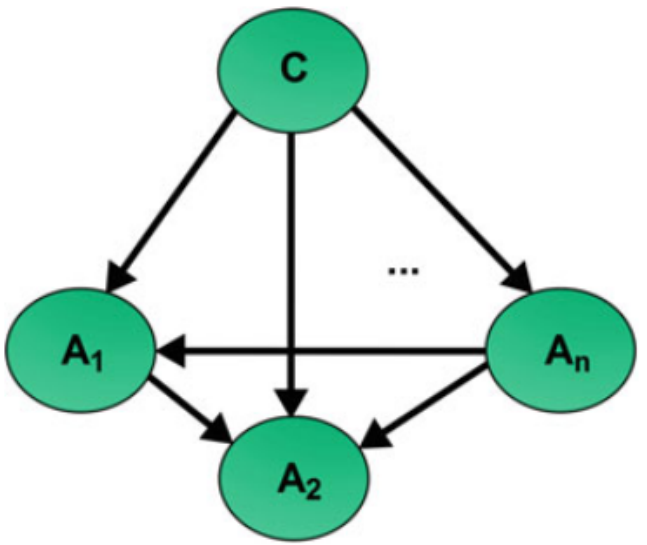
Además, existe un arco entre las variables de clase y cada atributo.



Modelos alternativos: TAN, BAN

Si quitamos la limitación de una estructura de árbol entre atributos, obtenemos el BAN (Bayesian Network augmented Bayesian Classifier), que considera que la estructura de dependencia entre los atributos constituye un grafo acíclico dirigido (DAG).

Al igual que con el clasificador TAN, hay un arco dirigido entre el nodo de clase y cada atributo. La estructura de un clasificador BAN se muestra en la siguienta figura:



Modelos alternativos: TAN, BAN

La probabilidad posterior para la variable de clase dados los atributos se puede obtener de forma similar a la NBC. Por lo tanto, debemos considerar la probabilidad condicional de cada atributo dada la clase y sus atributos principales:

$$
\begin{aligned}
&P(C|A_1, A_2, \ldots, A_n) \\
&= P(C)P(A_1|C, Pa(A_1))P(A_2|C, Pa(A_2)) \ldots P(A_n|C, Pa(A_n))/P(\mathbf{A})
\end{aligned}
$$

donde $Pa(A_i)$ es el conjunto de atributos padres de A_i según la estructura de dependencia de atributos del clasificador TAN o BAN.

Los clasificadores TAN y BAN pueden considerarse como casos particulares de un modelo más general, es decir, redes bayesianas.

Clasificadores bayesianos Semi-Ingenuos

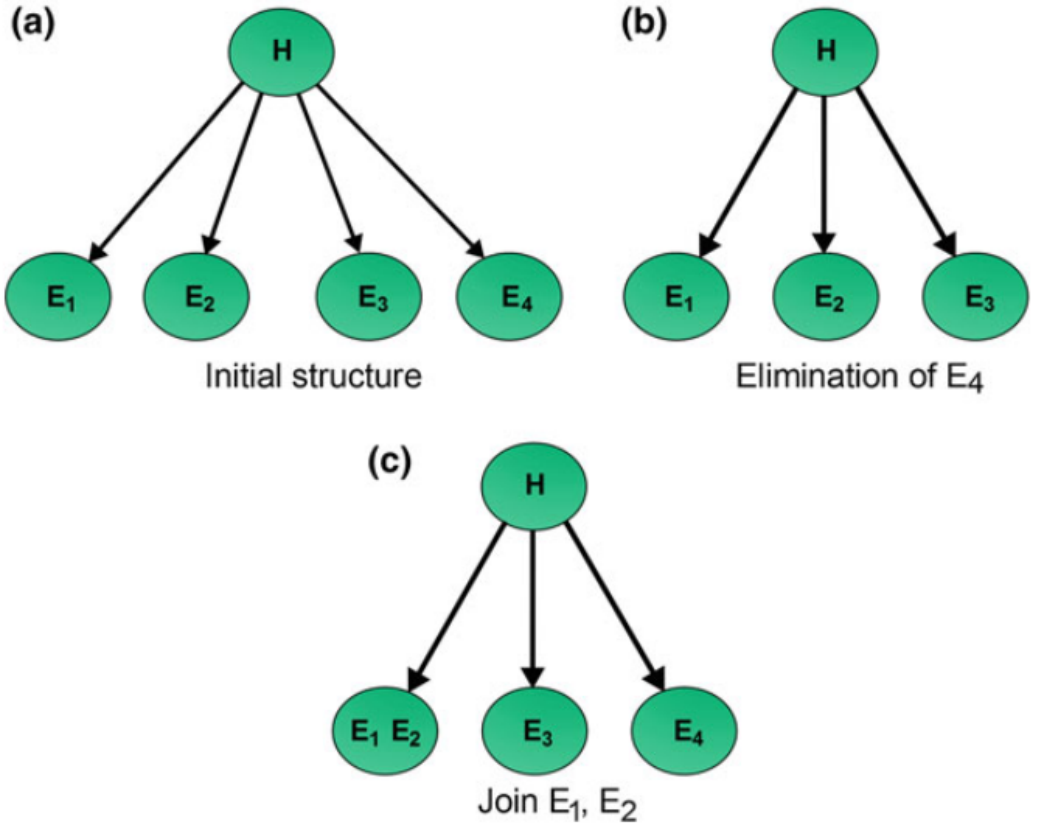
La idea básica del SNBC es eliminar o unir atributos que no sean independientes dada la clase, de manera que mejore el rendimiento del clasificador. Esto es análogo a la selección de características en el aprendizaje automático y existen dos tipos de enfoques:

- Filter: los atributos se seleccionan de acuerdo con una medida local, por ejemplo, la información mutua entre el atributo y la clase.
- Wrapper: los atributos se seleccionan en base a una medida global, generalmente comparando el desempeño del clasificador con y sin el atributo.

Además, el algoritmo de aprendizaje puede comenzar desde una estructura vacía y agregar (o combinar) atributos, o de una estructura completa con todos los atributos, y eliminar (o combinar) atributos.

Clasificadores Bayesianos Semi-Ingenuos

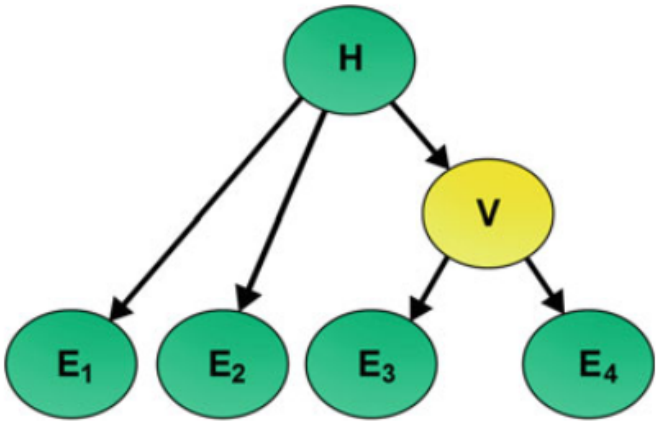
La figura siguiente ilustra las dos operaciones alternativas para modificar la estructura de un NBC: (i) eliminación de nodos y (ii) combinación de nodos, considerando que partimos de una estructura completa.



Clasificadores Bayesianos Semi-Ingenuos

La referencia de *Kwoh, C.K., Gillies, D.F* introducen una operación alternativa para modificar la estructura de un NBC, que consiste en agregar un nuevo atributo que independiza dos atributos dependientes.

Observa la siguiente figura:



Este nuevo atributo es una especie de nodo virtual u oculto en el modelo, para el cual no tenemos ningún dato para conocer sus parámetros.

Una alternativa para estimar los parámetros de variables ocultas en redes bayesianas, como en este caso, se basa en el procedimiento Expectativa-Maximización (EM).

Clasificadores Bayesianos Multidimensionales

Formalmente, el problema de clasificación multidimensional corresponde a la búsqueda de una función h que asigne a cada instancia representada por un vector de m características $\mathbf{X} = (X_1, \ldots, X_m)$ un vector de d valores de clase $\mathbf{C} = (C_1, \ldots, C_D)$.

La función h debería asignar a cada instancia $\mathbf{X}$ la combinación más probable de clases, es decir,

$$ArgMax_{c_1,\ldots,c_d} P(C_1 = c_1, \ldots, C_d = c_d | \mathbf{X})$$

La clasificación de múltiples etiquetas es un caso particular de clasificación multidimensional, donde todas las variables de clase son binarias. En el caso de la clasificación de múltiples etiquetas, existen dos enfoques básicos:

- Relevancia binaria
- Conjunto de potencia de etiqueta.

Clasificadores de Redes Bayesianas Multidimensionales

Un clasificador de red Bayesiano multidimensional (MBC) sobre un conjunto $\mathbf{V} = \{\mathbf{Z_1}, \ldots, \mathbf{Z_n}\}, n \geq 1$, de variables aleatorias discretas es una red Bayesiana con una estructura particular.

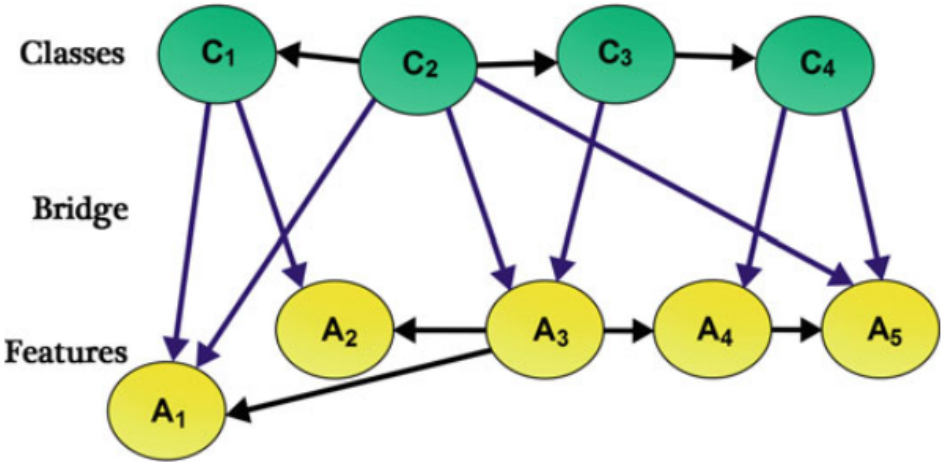
El conjunto $\mathbf{V}$ de variables se divide en dos conjuntos
$\mathbf{V_C} = \{\mathbf{C_1}, \ldots, \mathbf{C_d}\}, d \geq 1$, de variables de clase y
$\mathbf{V_X} = \{\mathbf{X_1}, \ldots, \mathbf{X_m}\}, m \geq 1$, de variables de características $(d + m = n)$.

El conjunto $\mathbf{A}$ de arcos también se divide en tres conjuntos, $\mathbf{A_C}, \mathbf{A_X}, \mathbf{A_{CX}}$, de modo que $\mathbf{A_C} \subseteq \mathbf{V_C} \times \mathbf{V_C}$ está compuesto por los arcos entre las variables de clase, $\mathbf{A_X} \subseteq \mathbf{V_X} \times \mathbf{V_X}$ está compuesto por los arcos entre las variables de características, y finalmente, $\mathbf{A_{CX}} \subseteq \mathbf{V_C} \times \mathbf{V_X}$ se compone de los arcos de las variables de clase a las variables de características.

Diferentes estructuras grafos para los subgrafos de clases y características pueden conducir a diferentes familias de MBC.

Clasificadores de Redes Bayesianas Multidimensionales

Los correspondientes subgrafos inducidos son $\mathbf{G_C} = (\mathbf{V_C}, \mathbf{A_C}), \mathbf{G_X} = (\mathbf{V_X}, \mathbf{A_X})$ y $\mathbf{G_{CX}} = (\mathbf{V}, \mathbf{A_{CX}})$ llamados subgrafos de clase, característica y puente.



El problema de obtener la clasificación de una instancia con un MBC, es decir, la combinación de clases más probable, corresponde al MPE (Most Probable Explanation) o problema de abducción para redes bayesianas. En otras palabras, determinar los valores más probables para las variables de clase $\mathbf{V} = \{\mathbf{C_1}, \ldots, \mathbf{C_n}\}$, dadas las características. Este es un problema complejo con un alto costo computacional.

Clasificadores de Cadena Bayesiana

Los clasificadores de cadena son un método alternativo para la clasificación de múltiples etiquetas que incorporan dependencias de clase, al tiempo que mantienen la eficiencia computacional del enfoque de relevancia binaria.

Un clasificador de cadena consta de d clasificadores binarios base que están enlazados en una cadena, de manera que cada clasificador incorpora las clases predichas por los clasificadores anteriores como atributos adicionales. Por lo tanto, el vector de características para cada clasificador binario, L_i, se extiende con las etiquetas $(0/1)$ para todos los clasificadores anteriores en la cadena.

Cada clasificador de la cadena está entrenado para aprender la asociación de la etiqueta l_i dadas las características aumentadas con todas las etiquetas de clase anteriores en la cadena, $L_1, L_2, \ldots, L_{i-1}$.

Al igual que en el enfoque de relevancia binaria, el vector de clase se determina combinando las salidas de todos los clasificadores binarios de la cadena.

Clasificadores de Cadena Bayesiana

Los clasificadores de cadena bayesiana son un tipo de clasificador de cadena bajo un marco probabilístico. Si aplicamos la regla de la cadena de la teoría de la probabilidad, podemos reescribir la ecuación anterior:

$$Arg \max_{c_1,\ldots,c_d} P(C_1|C_2,\ldots,C_d,\mathbf{X}) = P(C_2|C_3,\ldots,C_d,\mathbf{X})\ldots P(C_d|\mathbf{X})$$

Podemos simplificar la ecuación de la siguiente forma:

$$Arg \max_{C_1,\ldots,C_d} \prod_{i=1}^{d} P(C_i|\mathbf{Pa}(C_i),\mathbf{X})$$

donde $\mathbf{Pa}(C_i)$ son los padres de la clase i en el DAG que representa las dependencias entre las variables de clase. Podemos resolver el siguiente conjunto de ecuaciones como una aproximación de la ecuación anterior:

$$Arg \max_{C_1} P(C_1|\mathbf{Pa}(C_1),\mathbf{X})$$

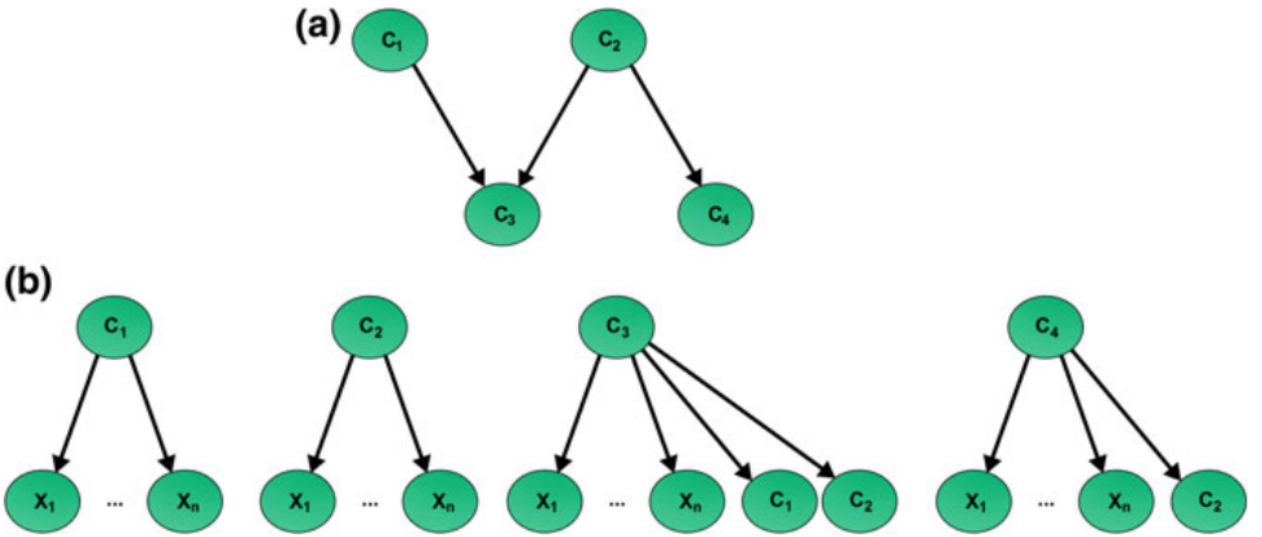
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$Arg \max_{C_d} P(C_d|\mathbf{Pa}(C_d),\mathbf{X})$$

Clasificadores de Cadena Bayesiana

Esta última aproximación corresponde a un clasificador de cadena bayesiana. Por lo tanto, un BCC hace dos supuestos básicos:

- La estructura de dependencia de clases dada las características se puede representar mediante un DAG.
- La combinación conjunta más probable de asignaciones de clases (abducción total) se aproxima mediante la concatenación de las clases individuales más probables.



Clasificación Jerárquica

La clasificación jerárquica es un tipo de clasificación multidimensional en la que las clases se ordenan en una estructura predefinida, normalmente un árbol o, en general, un grafo acíclico dirigido (DAG). Existen dos enfoques básicos:

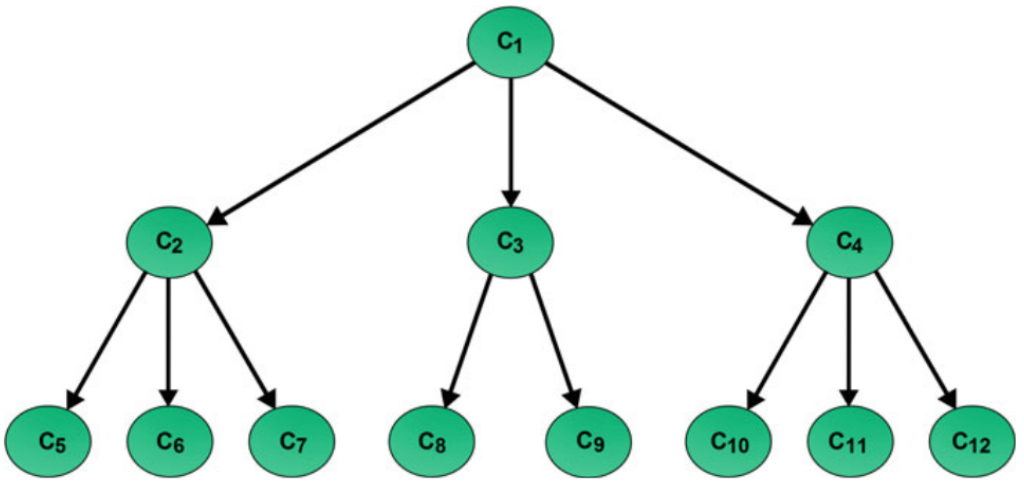
- Clasificadores globales
- Clasificadores locales
    - Clasificador Local por Nivel de Jerarquía
    - Clasificador Binario Local por Nodo
    - Clasificador Local por Nodo Padre (LPCN)

Los métodos locales suelen utilizar un enfoque de arriba hacia abajo para la clasificación. Un enfoque alternativo es analizar las rutas en la jerarquía y seleccionar la mejor ruta de acuerdo con los resultados de los clasificadores locales.

Evaluación de la Ruta Encadenada

La evaluación de la ruta encadenada (CPE) analiza cada ruta posible desde la raíz hasta un nodo hoja en la jerarquía, teniendo en cuenta el nivel de las etiquetas predichas para dar una puntuación a cada ruta y finalmente devolver la que tiene la mejor puntuación.

CPE consta de dos partes, formación y clasificación.



Ejemplo de estructura jerárquica (árbol). Para cada nodo no hoja, se entrena un clasificador local para predecir sus nodos hijos: C_1 clasifica C_2, C_3, C_4, C_2 clasifica C_5, C_6, C_7 y de manera similar para C_3 y C_4.

Entrenamiento y clasificación

- Se entrena un clasificador local para cada nodo en la jerarquía, excepto los nodos hoja, para clasificar sus nodos hijos, es decir, usando el esquema LCPN.

- En la fase de clasificación, las probabilidades de cada clase para todos los clasificadores locales se obtienen con base en los datos de entrada. Después de calcular las probabilidades para cada nodo no hoja en la jerarquía, estas se combinan para obtener una puntuación para cada ruta. La puntuación para cada ruta en la jerarquía se calcula como:

$$puntuación = \sum_{i=0}^{n} w_{C_i} \times \log(P(C_i|X_i, pa(C_i)))$$

donde C_i son las clases para cada LCPN, X_i es el vector de atributos, $pa(C_i)$ es la clase principal predicha y w_{C_i} es un peso.

- Una vez obtenidas las puntuaciones de todas las rutas, se seleccionará la ruta con la puntuación más alta como el conjunto de clases correspondientes a determinada instancia.

Blanket de Markov

- Una alternativa para aprender un MBC es el algoritmo MB-MBC.
- Este algoritmo en particular utiliza el Blanket Markov de cada variable de clase para aligerar la carga computacional de aprender el MBC al filtrar aquellas variables que no mejoran la clasificación.
- Un blanket de Markov de una variable C, denotado como $MB(C)$ es el conjunto mínimo de variables bajo el cual C es condicionalmente independiente de todas las variables restantes.

Fin!