Tema 6. Clasificadores Bayesianos

Pedro Larrañaga, Iñaki Inza, Abdelmalik Moujahid
Departamento de Ciencias de la Computación e Inteligencia Artificial
Universidad del País Vasco–Euskal Herriko Unibertsitatea

6.1 Introducción

Tal y como hemos visto en temas anteriores, el problema de *clasificación supervisada* consiste en asignar un vector $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{R}^n$ a una de las r_0 clases de la variable C. La clase verdadera se denota por c y toma valores en $\{1, 2, \ldots, r_0\}$. Se puede contemplar el clasificador como una función γ que asigna etiquetas a observaciones. Es decir:

$$\gamma : (x_1, \ldots, x_n) \rightarrow \{1, 2, \ldots, r_0\}$$

Existe una *matriz de costo* $cos(r, s)$ con $r, s = 1, \ldots, r_0$ en la cual se refleja el costo asociado a las clasificaciones incorrectas. En concreto $cos(r, s)$ indica el costo de clasificar un elemento de la clase r como de la clase s. En el caso especial de la *función de pérdida* 0/1, se tiene:

$$c(r, s) = \begin{cases} 1 & \text{si } r \neq s \\ 0 & \text{si } r = s \end{cases}$$

Subyacente a las observaciones suponemos la existencia de una distribución de probabilidad conjunta:

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

la cual es desconocida. El objetivo es construir un clasificador que minimiza el coste total de los errores cometidos, y esto se consigue (Duda y Hart, 1973) por medio del *clasificador de Bayes*:

$$\gamma(\boldsymbol{x}) = \arg \min_k \sum_{c=1}^{r_0} cos(k, c)p(c|x_1, \ldots, x_n)$$

En el caso de que la función de pérdida sea la 0/1, el clasificador de Bayes se convierte en asignar al ejemplo $\boldsymbol{x} = (x_1, \ldots, x_n)$ la *clase con mayor probabilidad a posteriori*. Es decir:

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

En la práctica la función de distribución conjunta $p(x_1, \ldots, x_n, c)$ es desconocida, y puede ser estimada a partir de una muestra aleatoria simple $\left\{(\boldsymbol{x}^{(1)}, c^{(1)}), \ldots, (\boldsymbol{x}^{(N)}, c^{(N)})\right\}$ extraida de dicha función de distribución conjunta.

6.2 Naïve Bayes

El paradigma clasificatorio en el que se utiliza el teorema de Bayes en conjunción con la hipótesis de independencia condicional de las variables predictoras dada la

clase se conoce bajo diversos nombres que incluyen los de *idiota Bayes* (Ohmann y col. 1988), *naïve Bayes* (Kononenko, 1990), *simple Bayes* (Gammerman y Thatcher, 1991) y *Bayes independiente* (Todd y Stamper, 1994).

A pesar de tener una larga tradición en la comunidad de reconocimiento de patrones (Duda y Hart, 1973) el clasificador naïve Bayes aparece por primera vez en la literatura del aprendizaje automático a finales de los ochenta (Cestnik y col. (1987)) con el objetivo de comparar su capacidad predictiva con la de métodos más sofisticados. De manera gradual los investigadores de esta comunidad de aprendizaje automático se han dado cuenta de su potencialidad y robustez en problemas de clasificación supervisada.

En esta sección se va a efectuar una revisión del *paradigma naïve Bayes*, el cual debe su nombre a las hipótesis tan simplificadoras –independencia condicional de las variables predictoras dada la variable clase– sobre las que se construye dicho clasificador. Partiremos del paradigma clásico de diagnóstico para, una vez comprobado que necesita de la estimación de un número de parámetros ingente, ir simplificando paulatinamente las hipótesis sobre las que se construye hasta llegar al modelo naïve Bayes. Veremos a continuación un resultado teórico que nos servirá para entender mejor las características del clasificador naïve Bayes.

6.2.1 Del Paradigma Clásico de Diagnóstico al Clasificador Naïve Bayes

Vamos a comenzar recordando el teorema de Bayes con una formulación de sucesos, para posteriormente formularlo en términos de variables aleatorias. Una vez visto el teorema de Bayes, se presenta el paradigma clásico de diagnóstico, viéndose la necesidad de ir simplificando las premisas sobre las que se construye en aras de obtener paradigmas que puedan ser de aplicación para la resolución de problemas reales. El contenido de este apartado resulta ser una adaptación del material que Díez y Nell (1998) dedican al mismo.

Teorema 6.1 *(Bayes, 1764)* [1] *Sean A y B dos sucesos aleatorios cuyas probabilidades se denotan por $p(A)$ y $p(B)$ respectivamente, verificándose que $p(B) > 0$. Supongamos conocidas las probabilidades a priori de los sucesos A y B, es decir, $p(A)$ y $p(B)$, así como la probabilidad condicionada del suceso B dado el suceso A, es decir $p(B|A)$. La probabilidad a posteriori del suceso A conocido que se verifica el suceso B, es decir $p(A|B)$, puede calcularse a partir de la siguiente fórmula:*

$$p(A|B) = \frac{p(A,B)}{p(B)} = \frac{p(A)p(B|A)}{p(B)} = \frac{p(A)p(B|A)}{\sum_{A'} p(A')p(B|A')}$$

□

La formulación del teorema de Bayes puede efectuarse también para variables aleatorias, tanto unidimensionales como multidimensionales.

[1] Thomas Bayes (1702–1761) fue uno de los seis primeros reverendos protestantes ordenados en Inglaterra. Comenzó como ayudante de su padre hasta que en 1720 fuera nombrado pastor en Kent. Abandonó los hábitos en 1752. Sus controvertidas teorías fueron aceptadas por Laplace, y posteriormente cuestionadas por Boole. Bayes fué elegido miembro de la *Royal Society of London* en 1742.

Comenzando por la formulación para dos variables aleatorias unidimensionales que denotamos por X e Y, tenemos que:

$$p(Y = y|X = x) = \frac{p(Y = y)p(X = x|Y = y)}{\sum_{y'} p(Y = y')p(X = x|Y = y')}$$

El teorema de Bayes tambien puede ser expresado por medio de una notación que usa el número de componentes de cada una de las variables multidimensionales anteriores $\boldsymbol{X}$ e $\boldsymbol{Y}$, de la siguiente manera:

$$p(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x}) = p(Y_1 = y_1, \ldots, Y_m = y_m|X_1 = x_1, \ldots, X_n = x_n)$$
$$= \frac{p(Y_1 = y_1, \ldots, Y_m = y_m)p(X_1 = x_1, \ldots, X_n = x_n|Y_1 = y_1, \ldots, Y_m = y_m)}{\sum_{y'_1,\ldots,y'_m} p(X_1 = x_1, \ldots, X_n = x_n|Y_1 = y'_1, \ldots, Y_m = y'_m)p(Y_1 = y'_1, \ldots, Y_m = y'_m)}$$

En el problema de clasificación supervisada reflejado en la Tabla 6.1 , tenemos que $\boldsymbol{Y} = C$ es una variable unidimensional, mientras que $\boldsymbol{X} = (X_1, \ldots, X_n)$ es una variable n-dimensional.

	X_1	$\ldots$	X_n	Y
$(\boldsymbol{x}^{(1)}, y^{(1)})$	$x_1^{(1)}$	$\ldots$	$x_n^{(1)}$	$y^{(1)}$
$(\boldsymbol{x}^{(2)}, y^{(2)})$	$x_1^{(2)}$	$\ldots$	$x_n^{(2)}$	$y^{(2)}$
$\ldots$		$\ldots$		
$(\boldsymbol{x}^{(N)}, y^{(N)})$	$x_1^{(N)}$	$\ldots$	$x_n^{(N)}$	$y^{(N)}$

Tabla 6.1: Problema de clasificación supervisada.

Vamos a plantear la *formulación clásica de un problema de diagnóstico* utilizando una terminología habitual en medicina. Es evidente que la terminología puede trasladarse a otras ramas de la ciencia y de la técnica, en particular a la ingeniería. La terminología a usar incluye términos como:

- *hallazgo*, con el cual nos referimos a la determinación del valor de una variable predictora X_r. Así por ejemplo x_r (valor de la variable X_r) puede estar representando la existencia de vómitos en un determinado enfermo;
- *evidencia*, denota el conjunto de todos los hallazgos para un determinado individuo. Es decir $\boldsymbol{x} = (x_1, \ldots, x_n)$ puede estar denotando (si $n = 4$) que el individuo en cuestión es joven, hombre, presenta vómitos y además no tiene antecedentes familiares;
- *diagnóstico*, denota el valor que toman las m variables aleatorias $Y_1, \ldots, Y_m$, cada una de las cuales se refiere a una enfermedad;
- *probabilidad a priori del diagnóstico*, $p(\boldsymbol{y})$ o $p(Y_1 = y_1, \ldots, Y_m = y_m)$, se refiere a la probabilidad de un diagnóstico concreto, cuando no se conoce nada acerca de los hallazgos, es decir, cuando se carece de evidencia;
- *probabilidad a posteriori de un diagnóstico*, $p(\boldsymbol{y}|\boldsymbol{x})$ o $p(Y_1 = y_1, \ldots, Y_m = y_m|X_1 = x_1, \ldots, X_n = x_n)$, es decir, la probabilidad de un diagnóstico concreto cuando se conocen n hallazgos (evidencia).

En el planteamiento clásico del diagnóstico (véase Tabla 6.2) se supone que los m diagnósticos posibles son *no excluyentes*, es decir, pueden ocurrir a la vez, siendo cada uno de ellos dicotómico. Para fijar ideas en relación con el ámbito médico, podemos pensar que cada uno de los m posibles diagnósticos no excluyentes se relaciona con una enfermedad, pudiendo tomar dos valores: 0 (no existencia) y 1 (existencia). Por lo que se refiere a los n hallazgos o síntomas, se representarán por medio de las n variables aleatorias $X_1, \ldots, X_n$ y también asumiremos que cada variable predictora es dicotómica, con valores 0 y 1. El valor 0 en la variable X_i indica la ausencia del i-ésimo hallazgo o síntoma mientras que el valor 1 indica la presencia del hallazgo o síntoma correspondiente.

	X_1	$\ldots$	X_n	Y_1	$\ldots$	Y_m
$(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)})$	$x_1^{(1)}$	$\ldots$	$x_n^{(1)}$	$y_1^{(1)}$	$\ldots$	$y_m^{(1)}$
$(\boldsymbol{x}^{(2)}, \boldsymbol{y}^{(2)})$	$x_1^{(2)}$	$\ldots$	$x_n^{(2)}$	$y_1^{(2)}$	$\ldots$	$y_m^{(2)}$
$\ldots$		$\ldots$			$\ldots$	
$(\boldsymbol{x}^{(N)}, \boldsymbol{y}^{(N)})$	$x_1^{(N)}$	$\ldots$	$x_n^{(N)}$	$y_1^{(N)}$	$\ldots$	$y_m^{(N)}$

Tabla 6.2: Problema clásico de diagnóstico.

El problema del diagnóstico consiste en encontrar el diagnóstico más probable a posteriori, una vez conocido el valor de la evidencia. En notación matemática el diagnóstico óptimo, $(y_1^*, \ldots, y_m^*)$ será aquel que verifique:

$$(y_1^*, \ldots, y_m^*) = arg \max_{(y_1,\ldots,y_m)} p(Y_1 = y_1, \ldots, Y_m = y_m | X_1 = x_1, \ldots, X_n = x_n)$$

Aplicando el teorema de Bayes para calcular $p(Y_1 = y_1, \ldots, Y_m = y_m | X_1 = x_1, \ldots, X_n = x_n)$, obtenemos:

$$p(Y_1 = y_1, \ldots, Y_m = y_m | X_1 = x_1, \ldots, X_n = x_n)$$

$$= \frac{p(Y_1 = y_1, \ldots, Y_m = y_m) p(X_1 = x_1, \ldots, X_n = x_n | Y_1 = y_1, \ldots, Y_m = y_m)}{\sum_{y_1', \ldots, y_m'} p(Y_1 = y_1', \ldots, Y_m = y_m') p(X_1 = x_1, \ldots, X_n = x_n | Y_1 = y_1', \ldots, Y_m = y_m')}$$

Veamos a continuación el número de parámetros que se deben estimar para poder especificar el paradigma anterior y de esa forma obtener el valor de $(y_1^*, \ldots, y_m^*)$. Es importante tener en cuenta que la estimación de cada uno de los parámetros anteriores se deberá efectuar a partir del fichero de N casos, reflejado en la Tabla 6.2.

Para estimar $p(Y_1 = y_1, \ldots, Y_m = y_m)$, y teniendo en cuenta que cada variable Y_i es dicotómica, necesitaremos un total de $2^m - 1$ parámetros. De igual forma, por cada una de las distribuciones de probabilidad condicionadas, $p(X_1 = x_1, \ldots, X_n = x_n | Y_1 = y_1, \ldots, Y_m = y_m)$, se necesitan estimar $2^n - 1$ parámetros. Al tener un total de 2^m de tales distribuciones de probabilidad condicionadas, debemos estimar $(2^n - 1)2^m$ parámetros. Es decir, que el número total de parámetros necesarios para determinar un modelo concreto del paradigma clásico de diagnóstico es: $2^m - 1 + 2^m(2^n - 1)$. Para hacernos una idea del número de parámetros a estimar podemos consultar la Tabla 6.3, en la cual vemos de manera aproximada el número de parámetros a estimar para distintos valores de m (número de enfermedades) y n (número de hallazgos).

m	n		parámetros
3	10	$\simeq$	$8 \cdot 10^3$
5	20	$\simeq$	$33 \cdot 10^6$
10	50	$\simeq$	$11 \cdot 10^{17}$

Tabla 6.3: Número de parámetros a estimar, en función de m (número de enfermedades) y n (número de síntomas), en el paradigma clásico de diagnóstico.

Ante la imposibilidad de poder estimar el elevado número de parámetros que se necesitan en el paradigma clásico de diagnóstico, en lo que sigue se simplificarán las premisas sobre las que se ha construído dicho paradigma.

En primer lugar vamos a considerar que los *diagnósticos son excluyentes*, es decir, que dos diagnósticos no pueden darse al unísono. Esto trae como consecuencia que en lugar de considerar el diagnóstico como una variable aleatoria m–dimensional, este caso pueda verse como una única variable aleatoria unidimensional siguiendo una distribución polinomial con m valores posibles.

Vamos a denotar por $X_1, \ldots, X_n$ a las n variables predictorias. Supongamos que todas ellas sean binarias. Denotamos por C la variable de diagnóstico, que suponemos puede tomar m posibles valores. La Tabla 6.1 refleja la situación anterior. La búsqueda del diagnóstico más probable a posteriori, c^*, una vez conocidos los síntomas de un determinado paciente, $\boldsymbol{x} = (x_1, \ldots, x_n)$, puede plantearse como la búsqueda del estado de la variable C con mayor probabilidad a posteriori. Es decir

$$c^* = arg \max_c p(C = c | X_1 = x_1, \ldots, X_n = x_n)$$

El cálculo de $p(C = c|X_1 = x_1, \ldots, X_n = x_n)$ puede llevarse a cabo utilizando el teorema de Bayes, y ya que el objetivo es calcular el estado de C, c^*, con mayor probabilidad a posteriori, no es necesario calcular el denominador del teorema de Bayes. Es decir,

$$p(C = c|X_1 = x_1, \ldots, X_n = x_n) \ \propto \ p(C = c)p(X_1 = x_1, \ldots, X_n = x_n|C = c)$$

Por tanto, en el paradigma en el que los distintos diagnósticos son excluyentes, y considerando que el número de posibles diagnósticos es m, y que cada variable predictoria X_i es dicotómica, tenemos que el número de parámetros a estimar es $(m-1) + m(2^n - 1)$, de los cuales:

- $m - 1$ se refiere a las probabilidades a priori de la variable C;
- $m(2^n - 1)$ se relacionan con las probabilidades condicionadas de cada posible combinación de las variables predictoras dado cada posible valor de la variable C.

La Tabla 6.4 nos da una idea del número de parámetros a estimar para distintos valores de m y n.

Vemos de nuevo que el número de parámetros a estimar sigue siendo elevado, de ahí que necesitamos imponer suposiciones más restrictivas para que los paradigmas

m	n		parámetros
3	10	$\simeq$	$3 \cdot 10^3$
5	20	$\simeq$	$5 \cdot 10^6$
10	50	$\simeq$	$11 \cdot 10^{15}$

Tabla 6.4: Número de parámetros a estimar, en función de m (número de enfermedades) y n (número de síntomas), en el paradigma clásico de diagnóstico con diagnósticos excluyentes.

puedan convertirse en modelos implementables.

Vamos finalmente a introducir el paradigma *naïve Bayes: diagnósticos excluyentes y hallazgos condicionalmente independientes dado el diagnóstico*. El paradigma naïve Bayes se basa en dos premisas establecidas sobre las variables predictoras (hallazgos, síntomas) y la variable a predecir (diagnóstico). Dichas premisas son:

- los diagnósticos son excluyentes, es decir, la variable C a predecir toma uno de sus m posibles valores: $c_1, \ldots, c_m$;
- los hallazgos son condicionalmente independientes dado el diagnóstico, es decir, que si uno conoce el valor de la variable diagnóstico, el conocimiento del valor de cualquiera de los hallazgos es irrelevante para el resto de los hallazgos. Esta condición se expresa matemáticamente por medio de la fórmula:

$$p(X_1 = x_1, \ldots, X_n = x_n | C = c) = \prod_{i=1}^{n} p(X_i = x_i | C = c) \tag{1}$$

ya que por medio de la regla de la cadena se obtiene:

$$\begin{aligned} p(X_1 = x_1, \ldots, X_n = x_n | C = c) \quad = \quad & p(X_1 = x_1 \mid X_2 = x_2, \ldots, X_n = x_n, C = c) \\ & p(X_2 = x_2 \mid X_3 = x_3, \ldots, X_n = x_n, C = c) \\ & \ldots p(X_n = x_n \mid C = c) \end{aligned}$$

Por otra parte teniendo en cuenta la independencia condicional entre las variables predictoras dada la variable clase, se tiene que:

$$p(X_i = x_i \mid X_{i+1} = x_{i+1}, \ldots, X_n = x_n, C = c) = p(X_i = x_i \mid C = c)$$

para todo $i = 1, \ldots, n$. De ahí que se verifique la ecuación 1.

Por tanto, en el paradigma naïve Bayes, la búsqueda del diagnóstico más probable, c^*, una vez conocidos los síntomas $(x_1, \ldots, x_n)$ de un determinado paciente, se reduce a:

$$c^* = arg \max_c p(C = c | X_1 = x_1, \ldots, X_n = x_n)$$

$$= arg \max_c p(C = c) \prod_{i=1}^{n} p(X_i = x_i | C = c)$$

Suponiendo que todas las variables predictoras son dicotómicas, el número de parámetros necesarios para especificar un modelo naïve Bayes resulta ser $(m-1)+mn$, ya que

- se necesitan $(m-1)$ parámetros para especificar la probabilidad a priori de la variable C;
- para cada variable predictora X_i se necesitan m parámetros para determinar las distribuciones de probabilidad condicionadas.

Con los números reflejados en la Tabla 6.5, nos podemos hacer una idea del número de parámetros necesarios en función del número de posibles diagnósticos y del número de síntomas necesarios para especificar el paradigma naïve Bayes.

m	n	parámetros
3	10	32
5	20	104
10	50	509

Tabla 6.5: Número de parámetros a estimar en el paradigma naïve Bayes en función del número de diagnósticos posibles (m) y del número de síntomas (n).

En el caso de que las n variables predictoras $X_1, \ldots, X_n$ sean continuas, se tiene que el paradigma naïve Bayes se convierte en buscar el valor de la variable C, que denotamos por c^*, que maximiza la probabilidad a posteriori de la variable C, dada la evidencia expresada como una instanciación de las variables $X_1, \ldots, X_n$, esto es, $\boldsymbol{x} = (x_1, \ldots, x_n)$.

Es decir, el paradigma *naïve Bayes con variables continuas* trata de encontrar c^* verificando:

$$c^* = arg\,\max_c p(C = c|X_1 = x_1, \ldots, X_n = x_n)$$

$$= arg\,\max_c p(C = c) \prod_{i=1}^{n} f_{X_i|C=c}(x_i|c)$$

donde $f_{X_i|C=c}(x_i|c)$ denota, para todo $i = 1, \ldots, n$, la función de densidad de la variable X_i condicionada a que el valor del diagnóstico sea c.

Suele ser habitual utilizar una variable aleatoria normal (para cada valor de C) para modelar el comportamiento de la variable X_i. Es decir, para todo c, y para todo $i \in \{1, \ldots, n\}$, asumimos

$$f_{X_i|C=c}(x_i|c) \rightsquigarrow \mathcal{N}(x_i; \mu_i^c, (\sigma_i^c)^2)$$

En tal caso el paradigma naïve Bayes obtiene c^*, como:

$$c^* = arg\,\max_c p(C = c) \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma_i^c} e^{\frac{-1}{2}\left(\frac{x_i - \mu_i^c}{\sigma_i^c}\right)^2} \right]$$

En este caso el número de parámetros a estimar es $(m-1) + 2nm$:

- $m-1$ en relación con las probabilidades a priori $p(C=c)$;
- $2nm$ en relación con las funciones de densidad condicionadas.

Finalmente puede ocurrir que algunos de los hallazgos se recojan en variables discretas mientras que otros hallazgos sean continuos. En tal caso hablaremos del paradigma *naïve Bayes con variables predictoras continuas y discretas.*

Supongamos que de las n variables predictoras, n_1 de ellas, $X_1, \ldots, X_{n_1}$, sean discretas, mientras que el resto $n - n_1 = n_2$, $Y_1, \ldots, Y_{n_2}$, sean continuas. En principio al aplicar directamente la fórmula del paradigma naïve Bayes correspondiente a esta situación se obtiene:

$$p(c|x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}) \propto p(c) \prod_{i=1}^{n_1} p(x_i|c) \prod_{j=1}^{n_2} f(y_j|c)$$

Esta expresión puede propiciar el conceder una mayor importancia a las variables continuas, ya que mientras que $p(x_i|c)$ verifica $0 \leq p(x_i|c) \leq 1$, puede ocurrir que $f(y_j|c) > 1$. Con objeto de evitar esta situación, proponemos la normalización de la aportación de las variables continuas, dividiendo cada uno de los factores correspondientes por el $\text{máx}_{y_j} f(y_j|c)$. Obtenemos por tanto:

$$p(c|x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}) \propto p(c) \prod_{i=1}^{n_1} p(x_i|c) \prod_{j=1}^{n_2} \frac{f(y_j|c)}{\text{máx}_{y_j} f(y_j|c)} \tag{2}$$

En el caso en que las funciones de densidad de las variables continuas condicionadas a cada posible valor de la variable clase sigan distribuciones normales, es decir si $Y_j|C = c \rightsquigarrow \mathcal{N}(y_j; \mu_j^c, (\sigma_j^c)^2)$, se tiene que

$$\frac{f(y_j|c)}{\text{máx}_{y_j} f(y_j|c)} = \frac{\frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{1}{2}\left(\frac{y_j - \mu_j^c}{\sigma_j^c}\right)^2}}{\frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{1}{2}\left(\frac{\mu_j^c - \mu_j^c}{\sigma_j^c}\right)^2}} = e^{-\frac{1}{2}\left(\frac{y_j - \mu_j^c}{\sigma_j^c}\right)^2}$$

y la fórmula 2 se expresa de la manera siguiente:

$$p(c|x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}) \propto p(c) \prod_{i=1}^{n_1} p(x_i|c) \prod_{j=1}^{n_2} e^{-\frac{1}{2}\left(\frac{y_j - \mu_j^c}{\sigma_j^c}\right)^2}$$

La Figura 6.1 refleja la estructura gráfica de un modelo naïve Bayes.

6.2.2 Resultados Teóricos

Minsky (1961) demuestra que si las variables aleatorias predictoras, al igual que la variable clase, son binarias, la superficie de decisión que se deriva de un modelo naïve Bayes es un hiperplano.

Definicion 6.1 *En un problema de decisión binario con dos posibles decisiones (d_1 frente a d_2), una función de decisión es una función continua*

$$r : \Re^n \to \Re$$

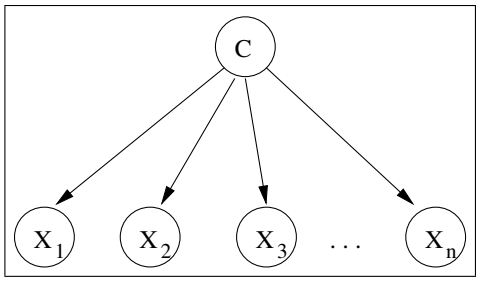


Figura 6.1: Estructura gráfica de un modelo naïve Bayes.

verificando que si $r(\boldsymbol{x}) > 0$ *(*$r(\boldsymbol{x}) < 0$*) se prefiere* d_1 *a* d_2 *(*d_2 *a* d_1*).*

La función $r(\boldsymbol{x}) = 0$ *define una superficie de decisión.*

□

Teorema 6.2 *(Minsky, 1961). Las superficies de decisión de un clasificador naïve Bayes con variables predictoras binarias son hiperplanos.*

DEMOSTRACIÓN: En el modelo naïve Bayes la probabilidad a posteriori de la clase c dado el vector de variables predictoras $\boldsymbol{x} = (x_1, \ldots, x_i, \ldots, x_n)$ viene dada por:

$$p(c|x_1, \ldots, x_n) = \frac{p(c)}{p(x_1, \ldots, x_n)} \prod_{i=1}^{n} p(x_i|c) \tag{3}$$

Escribiendo $p(x_i|c)$ de la siguiente manera:

$$p(x_i|c) = p(X_i = 0|C = c) \left[\frac{p(X_i = 1|C = c)}{p(X_i = 0|C = c)}\right]^{x_i}$$

con $x_i = 0, 1$, sustituyendo en la ecuación 3 y tomando logaritmos, se tiene:

$$\begin{aligned} \log p(c|x_1, \ldots, x_n) &= \log \left[\frac{p(c)}{p(x_1, \ldots, x_n)} \prod_{i=1}^{n} p(X_i = 0|C = c)\right] \\ &\quad + \sum_{i=1}^{n} x_i \log \left[\frac{p(X_i = 1|C = c)}{p(X_i = 0|C = c)}\right] \end{aligned}$$

Denotando por $w_{c0} = \log \left[p(c) \prod_{i=1}^{n} p(X_i = 0|C = c)\right]$ y por

$$w_{ci} = \log \left(\frac{p(X_i = 1|C = c)}{p(X_i = 0|C = c)}\right)$$

se obtiene

$$\log p(c|x_1, \ldots, x_n) = w_{c0} + \sum_{i=1}^{n} x_i w_{ci} - \log p(x_1, \ldots, x_n)$$

Teniendo en cuenta que la variable clase C es dicotómica con posibles valores c_0 y c_1, podemos definir la función de decisión siguiente:

$$\begin{aligned} r_{01}(x_1, \ldots, x_n) &= \log p(c_0|x_1, \ldots, x_n) - \log p(c_1|x_1, \ldots, x_n) \\ &= (w_{00} - w_{10}) + \left(\sum_{i=1}^{n} (w_{0i} - w_{1i}) x_i\right) \end{aligned}$$

De ahí que las superficies de decisión sean hiperplanos.

□

6.3 Seminaïve Bayes

Kononenko (1991) introduce el denominado *seminaïve Bayesian classifier*. En el mismo se trata de evitar las estrictas premisas sobre las que se construye el paradigma naïve Bayes por medio de la consideración de nuevas variables en las cuales no necesariamente tenga que aparecer el producto cartesiano de dos variables, sino tan sólo aquellos valores de dicho producto cartesiano que verifiquen una determinada condición que surge al considerar el concepto de independencia junto con el de la fiabilidad en la estimación de las probabilidades condicionadas, cuestión esta última que es resuelta a partir del teorema de Chebyshev.

Pazzani (1996) introduce el concepto de inducción constructiva con el que a partir del producto cartesiano entre variables y usando el algoritmo voraz BSEJ (ver párrafo siguiente) de una manera de envoltura, desarrolla modelos de clasificadores naïve Bayes así como K-NN. Se trata del trabajo inicial que sirvió de base al famoso trabajo posterior del mismo autor.

Pazzani (1997) presenta una aproximación en la que de manera voraz se va construyendo un modelo naïve Bayes en el que se detectan aquellas variables irrelevantes así como aquellas variables dependientes entre sí. Cuando se detectan variables dependientes, se crea una nueva variable a partir del producto cartesiano de las mismas. El algoritmo está guiado por un score que resulta ser la validación honesta por medio de un 10-*fold cross–validation* (o por medio de un *leave one out* dependiendo del tamaño de la base de datos) del porcentaje de bien clasificados. Se presentan dos algoritmos voraces, uno hacia adelante denominado *FSSJ (Forward Sequential Selection and Joining)* y otro hacia atrás *BSEJ (Backward Sequential Elimination and Joining)*, cuyos pseudocódigos pueden consultarse en Figura 6.2 y Figura 6.4 respectivamente.

Tal y como puede verse en la Figura 6.2 el algoritmo *FSSJ* efectúa una modelización voraz hacia adelante guiado por la estimación del porcentaje de casos bien clasificados. Comienza considerando como modelo inicial la regla simple que consiste en clasificar todos los ejemplos, independientemente de sus características, como pertenecientes a la clase más numerosa. A continuación, mientras se vaya mejorando la estimación del porcentaje de bien clasificados, se va efectuando en cada paso la mejor opción entre incluir en el modelo una variable de las que todavía no formaban parte del mismo, u obtener una nueva variable como producto cartesiano entre alguna de las variables (o supervariables[2]) ya incluidas en el modelo y la que se acaba de incluir. La Figura 6.3 presenta un ejemplo de aplicación del algoritmo *FSSJ*. El algoritmo *BSEJ (Backward Sequential Elimination and Joining)* actúa de manera dual al *FSSJ*, tal y como puede apreciarse en el pseudocódigo de la Figura 6.4.

[2]La denominación supervariable hace alusión a la variable resultante del producto cartesiano entre dos o más variables originales.

Paso 1. Inicializar el conjunto de variables a utilizar a vacío. Clasificar todos los ejemplos como pertenecientes a la clase más frecuente

Paso 2. **Repetir** en cada paso la mejor opción entre:

(a) Considerar cada variable que no está en el modelo como una nueva variable a incluir en el modelo. Dicha variable debe incluirse condicionalmente independiente de las variables presentes en el modelo, dada la variable clase

(b) Juntar cada variable no presente en el modelo con una variable que ya forme parte del mismo

Evaluar cada posible opción por medio de la estimación del porcentaje de bien clasificados

Hasta que ninguna opción produzca mejoras

Figura 6.2 Pseudocódigo del algoritmo *FSSJ* (Pazzani, 1997).

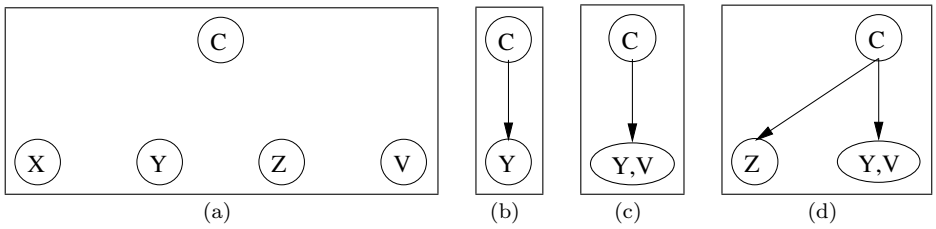


Figura 6.3: Ejemplo de aplicación del algoritmo FSSJ. X, Y, Z y V denotan las variables predictoras, C, la variable a clasificar. La subfigura (a) muestra la situación inicial, donde el ejemplo se clasifica como c^*, siendo $p(c^*) = \arg\max_c p(c)$. La subfigura (b) muestra que, después de comparar todos los modelos naïve Bayes con una única variable predictora, la variable Y ha sido seleccionada. La subfigura (c) muestra el modelo ganador de entre los que tienen como variables predictoras los siguientes subconjuntos de variables: $\{Y, X\}, \{Y, Z\}, \{Y, V\}, \{(Y, X)\}, \{(Y, Z)\}, \{(Y, V)\}$. La subfigura (d) indica que el mostrado ha resultado vencedor entre: $\{X, (Y, V)\}, \{Z, (Y, V)\}, \{(X, Y, V)\}, \{(Z, Y, V)\}$. Al no tener continuidad el algoritmo indica que los modelos $\{X, Z, (Y, V)\}, \{(X, Z), (Y, V)\}, \{Z, (Y, V, X)\}$ son peores al mostrado en la subfigura (d).

6.4 Naïve Bayes Aumentado a Árbol

En esta sección vamos a presentar algunos trabajos que construyen clasificadores con estructura naïve Bayes aumentada a árbol (*Tree Augmented Network (TAN)*). Para obtener este tipo de estructura se comienza por una estructura de árbol con las variables predictoras, para posteriormente conectar la variable clase con cada una de las variables predictoras. La Figura 6.5 ilustra un ejemplo de estructura naïve Bayes aumentada a árbol.

Friedman y col. (1997) presentan un algoritmo denominado *Tree Augmented Network (TAN)* el cual consiste básicamente en una adaptación del algoritmo de Chow–Liu (1968). En dicho algoritmo se tiene en cuenta la cantidad de información mutua condicionada a la variable clase, en lugar de la cantidad de información mutua en la que se basa el algoritmo de Chow-Liu. La cantidad de información mutua entre las variables discretas X e Y condicionada a la variable C se define como:

$$I(X, Y|C) = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{r=1}^{w} p(x_i, y_j, c_r) \log \frac{p(x_i, y_j|c_r)}{p(x_i|c_r)p(y_j|c_r)}$$

Paso 1. Inicializar con el modelo naïve Bayes con todas las variables predictoras
Paso 2. **Repetir** en cada paso la mejor opción entre:
(a) Considerar reemplazar dos de las variables usadas por el clasificador por una nueva variable producto cartesiano de ambas
(b) Considerar eliminar una variable usada por el clasificador
Evaluar cada posible opción por medio de la estimación del porcentaje de bien clasificados
Hasta que ninguna opción produzca mejoras

Figura 6.4: Pseudocódigo del algoritmo *BSEJ* (Pazzani, 1997).

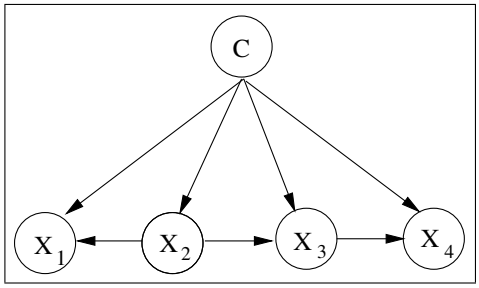


Figura 6.5: Ejemplo de estructura naïve Bayes aumentada a árbol (*Tree Augmented Network (TAN)*).

Tal y como puede verse en el pseudocódigo de la Figura 6.6, *TAN* consta de cinco pasos. En el primer paso se calculan las cantidades de información mutua para cada par de variables (X_i, X_j) condicionadas a la variable C. A continuación se debe construir un grafo no dirigido completo con n nodos, uno por cada una de las variables predictoras, en el cual el peso de cada arista viene dado por la cantidad de información mutua entre las dos variables unidas por la arista condicionada a la variable clase. El algoritmo de Kruskall parte de los $n(n-1)/2$ pesos obtenidos en el paso anterior para construir el árbol expandido de máximo peso de la siguiente manera:

1. Asignar las dos aristas de mayor peso al árbol a construir.
2. Examinar la siguiente arista de mayor peso, y añadirla al árbol a no ser que forme un ciclo, en cuyo caso se descarta y se examina la siguiente arista de mayor peso.
3. Repetir el paso 2 hasta que se hayan seleccionado $n-1$ aristas.

Las propiedades teóricas de este algoritmo de construcción de *TAN* son análogas a las del algoritmo de Chow–Liu (1968). Es decir, si los datos han sido generados por una estructura *Tree Augmented Network*, el algoritmo *TAN* es asintóticamente correcto, en el sentido de que si la muestra de casos es suficientemente grande, recuperará la estructura que generó el fichero de casos. En la Figura 6.7 se muestra un ejemplo de aplicación del algoritmo.

Keogh y Pazzani (1999) proponen un algoritmo voraz que va añadiendo arcos a una estructura naïve Bayes. En cada paso se añade el arco que, manteniendo la condición de que en la estructura final cada variable no tenga más de un padre, mejore en mayor medida el porcentaje de bien clasificados obtenido mediante el mismo.

Paso 1. Calcular $I(X_i, X_j \mid C)$ con $i < j, i, j = 1, \ldots, n$
Paso 2. Construir un grafo no dirigido completo cuyos nodos corresponden a las variables predictoras: $X_1, \ldots, X_n$. Asignar a cada arista conectando las variables X_i y X_j un peso dado por $I(X_i, X_j \mid C)$
Paso 3. A partir del grafo completo anterior y siguiendo el algoritmo de Kruskall construir un árbol expandido de máximo peso
Paso 4. Transformar el árbol no dirigido resultante en uno dirigido, escogiendo una variable como raiz, para a continuación direccionar el resto de aristas
Paso 5. Construir un modelo *TAN* añadiendo un nodo etiquetado como C y posteriormente un arco desde C a cada variable predictora X_i

Figura 6.6: Pseudocódigo del algoritmo *TAN* (Friedman y col. 1997).

6.5 Clasificadores Bayesianos k Dependientes

Sahami (1996) presenta un algoritmo denominado *k Dependence Bayesian classifier* (kDB) el cual posibilita atravesar el amplio espectro de dependencias disponibles entre el modelo naïve Bayes y el modelo correspondiente a una red Bayesiana completa –ver Sección 6.6–. El algoritmo se fundamenta en el concepto de clasificador Bayesiano k-dependiente, el cual contiene la estructura del clasificador naïve Bayes y permite a cada variable predictora tener un máximo de k variables padres sin contar a la variable clase. De esta manera, el modelo naïve Bayes se corresponde con un clasificador Bayesiano 0-dependiente, el modelo TAN sería un clasificador Bayesiano 1-dependiente y el clasificador Bayesiano completo (en la estructura no se refleja ninguna independencia) correspondería a un clasificador Bayesiano $(n-1)$-dependiente. El pseudocódigo del algoritmo kDB puede consultarse en la Figura 6.8.

La idea básica del algoritmo consiste en generalizar el algoritmo propuesto por Friedman y col. (1997) permitiendo que cada variable tenga un número de padres, sin contar la variable clase C, acotado por k. El autor comenta una posible mejora del algoritmo flexibilizando la determinación de k por medio de la obtención de un umbral de cantidad de información mutua, el cual debería de ser sobrepasado para que el correspondiente arco fuese incluido. Se presentan resultados experimentales con cinco bases de datos del repositorio UCI así como con una parte de la base de datos Reuters Text.

6.6 Clasificadores Bayesianos Basados en Redes Bayesianas

Las redes Bayesianas (Jensen, 2001) constituyen un paradigma de amplio uso dentro de la Inteligencia Artificial. En las mismas se efectúa –basándose en una semántica de independencia condicional entre tripletas de variables– una factorización de la función de probabilidad conjunta definida sobre la variable aleatoria n dimensional, tal y como puede verse en la Figura 6.9.

Queda fuera del alcance de estos apuntes el exponer el uso de las redes Bayesianas como paradigmas de clasificación supervisada, remitiéndosele al lector interesado a

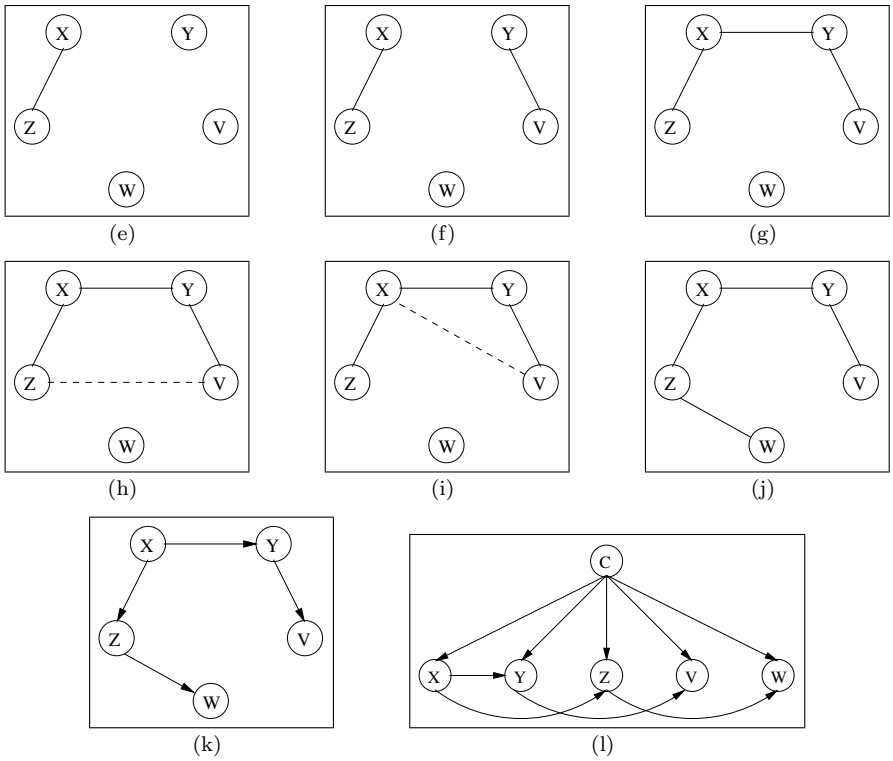


Figura 6.7: Ilustración del algoritmo TAN con cinco variables predictoras X, Y, Z, V y W. Se supone que el orden de las cantidades de información mutuas condicionadas ha sido: $I(X,Z|C) > I(Y,V|C) > I(X,Y|C) > I(Z,V|C) > I(X,V|C) > I(Z,W|C) > I(X,W|C) > I(Y,Z|C) > I(Y,W|C) > I(V,W|C)$. Las subfiguras (a) a (f) corresponden a la aplicación del algoritmo de Kruskall. La subfigura (g) corresponde al Paso 4 del algoritmo TAN y finalmente en la subfigura (h) se realiza el Paso 5 de TAN. El modelo clasificatorio obtenido es: $p(c|x,y,z,v,w) \propto p(c)p(x|c)p(y|x,c)p(z|x,c)p(v|y,c)p(w|z,c)$.

la referencia anterior.

Paso 1. Para cada variable predictora X_i, $i = 1, \ldots, n$, calcular la cantidad de información mútua con respecto a la clase C, $I(X_i, C)$

Paso 2. Para cada par de variables predictoras calcular la cantidad de información mútua condicionada a la clase, $I(X_i, X_j|C)$, con $i \neq j$, $i, j = 1, \ldots, n$

Paso 3. Inicializar a vacío la lista de variables usada $\aleph$

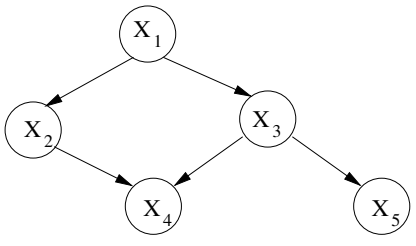
Paso 4. Inicializar la red Bayesiana a construir, BN, con un único nodo, el correspondiente a la variable C

Paso 5. Repetir hasta que $\aleph$ incluya a todas las variables del dominio:

- Paso 5.1. Seleccionar de entre las variables que no están en $\aleph$, aquella X_{max} con mayor cantidad de información mútua respecto a C, $I(X_{max}, C) = \text{máx}_{X \notin \aleph} I(X, C)$
- Paso 5.2. Añadir un nodo a BN, $X \notin \aleph$ representando X_{max}
- Paso 5.3. Añadir un arco de C a X_{max} en BN
- Paso 5.4. Añadir $m = \text{mín}(|\aleph|, k)$ arcos de las m variables distintas X_j en $\aleph$ que tengan los mayores valores $I(X_{max}, X_j|C)$
- Paso 5.5. Añadir X_{max} a $\aleph$

Paso 6. Computar las probabilidades condicionadas necesarias para especificar la red Bayesiana BN

Figura 6.8: Pseudocódigo del algoritmo kDB (Sahami, 1996).



$p(X_1 = 0) = 0{,}20$
$p(X_2 = 0|X_1 = 0) = 0{,}80$
$p(X_2 = 0|X_1 = 1) = 0{,}80$
$p(X_3 = 0|X_1 = 0) = 0{,}20$
$p(X_3 = 0|X_1 = 1) = 0{,}05$
$p(X_4 = 0|X_2 = 0, X_3 = 0) = 0{,}80$
$p(X_4 = 0|X_2 = 1, X_3 = 0) = 0{,}80$
$p(X_4 = 0|X_2 = 0, X_3 = 1) = 0{,}80$
$p(X_4 = 0|X_2 = 1, X_3 = 1) = 0{,}05$
$p(X_5 = 0|X_3 = 0) = 0{,}80$
$p(X_5 = 0|X_3 = 1) = 0{,}60$

Figura 6.9: Factorización de la distribución de probabilidad conjunta obtenida con la red Bayesiana adjunta.

Referencias

1. T. Bayes (1764). Essay towards solving a problem in the doctrine of chances. *The Phylosophical Transactions of the Royal Society of London.*

2. B. Cestnik, I. Kononenko, I. Bratko (1987). ASSISTANT-86: A knowledge elicitation tool for sophisticated users. *Progress in Machine Learning*, 31–45, Sigma Press.

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

4. F. J. Díez, E. Nell (1998). *Introducción al Razonamiento Aproximado.* Departamento de Inteligencia Artificial. UNED.

5. R. Duda, P. Hart (1973). *Pattern Classification and Scene Analysis.* John Wiley and Sons.

6. N. Friedman, D. Geiger, M. Goldszmidt (1997). Bayesian network classifiers. *Machine Learning*, 29, 131–163.

7. A. Gammerman, A. R. Thatcher (1991). Bayesian diagnostic probabilities without assuming independence of symptoms. *Methods of Information in Medicine*, 30, 15–22.

8. F. V. Jensen (2001). *Bayesian Networks and Decision Graphs.* Springer Verlag.

9. E. J. Keogh. M. Pazzani (1999). Learning augmented Bayesian classifiers: a comparison of distribution-based and non distribution-based approaches. *Proceedings of the 7th International Workshop on Artificial Intelligence and Statistics*, 225–230.

10. I. Kononenko (1990).Comparison of inductive and naïve Bayesian learning approaches to automatic knowledge acquisition. *Current Trends in Knowledge Acquisition.*

11. I. Kononenko (1991). Semi-naïve Bayesian classifiers. *Proceedings of the 6th European Working Session on Learning*, 206–219.

12. M. Minsky (1961). Steps toward artificial intelligence. *Transactions on Institute of Radio Engineers*, 49, 8–30.

13. C. Ohmann, Q. Yang, M. Kunneke, H. Stolzing, K. Thon, W. Lorenz (1988). Bayes theorem and conditional dependence of symptoms: different models applied to data of upper gastrointestinal bleeding. *Methods of Information in Medicine*, 27, 73–83.

14. M. Pazzani (1996). Constructive induction of cartesian product attributes. *Information, Statistics and Induction in Science*, 66–77.

15. M. Pazzani (1997). Searching for dependencies in Bayesian classifiers. *Learning from Data: Artificial Intelligence and Statistics V*, 239–248, Springer Verlag.

16. M. Sahami (1996). Learning limited dependence Bayesian classifiers. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 335–338.

17. B. S. Todd, R. Stamper (1994). The relative accuracy of a variety of medical diagnostic programs. *Methods of Information in Medicine*, 33, 402–416.