ANALÍTICA AVANZADA DE DATOS: CLASIFICADOR BAYESIANO

A. Alejandra Sánchez Manilla

asanchezm.q@gmail.com

Clasificadores Probabilísticos

- Sucesos deterministas: fenómeno que da lugar a un resultado cierto
  - *Después de las 6:00 son las 7:00.*
  - *Después del día sigue la noche.*

- Suceso aleatorios
  - *Lanzar una moneda al aire.*
  - *Lanzamiento de un dado.*

- Los fenómenos aleatorios exhiben regularidad estadística.

- Para el lanzamiento de una moneda: $$\lim_{n \to \infty} p = \frac{1}{2}$$

Clasificadores Probabilísticos

- *Probabilidad* : es el conjunto de posibilidades de que un evento ocurra o no en un momento determinado.
- Estas posibilidades se miden en una escala del 0 al 1. Donde:
  - *0 indica que es imposible que el evento ocurra*
  - *1 indica que el evento ocurrirá con certeza.*

  La probabilidad de un evento E está dada por (definición clásica de la probabilidad):

$$P(E) = \frac{\#\, resultados\ favorables}{\#\, total\ de\ posibles\ resultados} = \frac{n(E)}{n(S)}$$

Clasificadores Probabilísticos

Ejemplo: Lanzar un dado

Evento: que salga un número par

Definimos el evento de la siguiente manera:

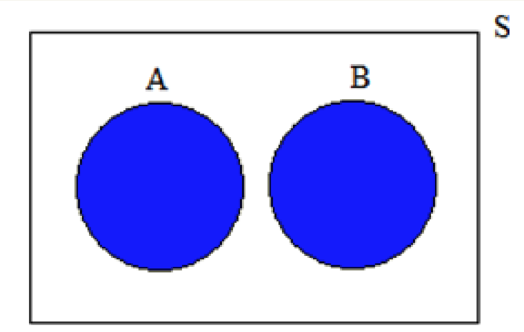
E = sale número $par \;=\; (2,4,6) \rightarrow P(E) = \; 3/6 = 1/2 = 0.5$

Clasificadores Probabilísticos

Propiedades de la Probabilidad

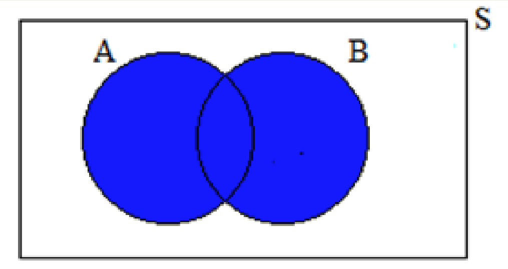
Regla de la adición

Eventos excluyente	Eventos no excluyente
$P(A \text{ o } B) = P(A) + P(B)$ $P(A \cup B) = P(A) + P(B)$	$P(A \text{ o } B) = P(A) + P(B) - P(A \text{ y } B)$ $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



Regla de la multiplicación

Eventos excluyente	Eventos no excluyente
$P(A \text{ y } B) = P(A)P(B)$ $P(A \cap B) = P(A)P(B)$	$P(A \text{ y } B) = P(A)P(B\|A)$ $P(A \cap B) = P(A)P(B\|A)$



Clasificadores Probabilísticos

Probabilidad Condicional

$P(A|B)$ Probabilidad condicional de A, dado B.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$P(B) > 0$, no es posible que exista $P(A)$ dado B, si B no sucede.

$$P(B|A) \neq P(A|B)$$

Clasificadores Probabilísticos

Probabilidad Condicional

Ejemplo: se realizó una encuesta sobre hábitos de lectura que se resumen en la siguiente tabla:

	Si lee	No lee	Total
Hombre	10	20	30
Mujer	90	30	120
Total	100	50	150

$$a)\ P(M) = \frac{120}{150} = 0.8$$

$$b)\ P(L \cap M) = \frac{90}{120} = 0.75$$

$$c)\ P(L|M) = \frac{P(L \cap M)}{P(M)} = \frac{0.75}{0.8} = 0.94$$

¿Cuál es la probabilidad de que una persona elegida al azar le guste leer dado que es una mujer?

Probabilidad Condicional

Teorema de Bayes

Desarrollado por el monje Thomas Bayes en el siglo XVII. El teorema de Bayes <u>se utiliza para revisar probabilidades previamente calculadas cuando se posee nueva información</u>

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$$

$P(A)$ $\rightarrow$ probabilidad a priori de A (prior).

$P(B|A)$ $\rightarrow$ probabilidad de B dado A (verosimilitud).

$P(A|B)$ $\rightarrow$ probabilidad a posteriori de A dado B (posterior).

Probabilidad Condicional

Clasificador Naïve Bayes

Consideremos una fábrica de tapones de corcho, cuya producción esta restringida a dos clases en la calidad de los corchos: promedio y superior

Cantidad de corchos de la clase 1: $(c_1) = n_1 = \quad 901{,}402$
Cantidad de corchos de la clase 2: $(c_2) = n_2 = 1{,}352{,}130$
Cantidad de corchos totales: $n \quad = 2{,}253{,}550$

Con esta información se puede obtener la probabilidades a priori:

$$P(c_1) = n_1/n = 0.4 \qquad\qquad P(c_2) = n_2/n = 0.6$$

Probabilidad Condicional

Clasificador Naïve Bayes

Supongamos que nos piden adivinar a que clase pertenece un corcho sin haberlo visto y la única información disponible es la probabilidad a priori

La opción lógica, es decir que el corcho pertenece a la **clase 2** ya que con esta decisión esperaríamos estar equivocados solo un 40%

Probabilidad Condicional

Clasificador Naïve Bayes

Supongamos que ahora se nos permite conocer los valores del vector x de características del corcho en cuestión. Sea $P(c_k|x)$ la probabilidad condicional de que el corcho representado por x pertenezca a la clase c_k

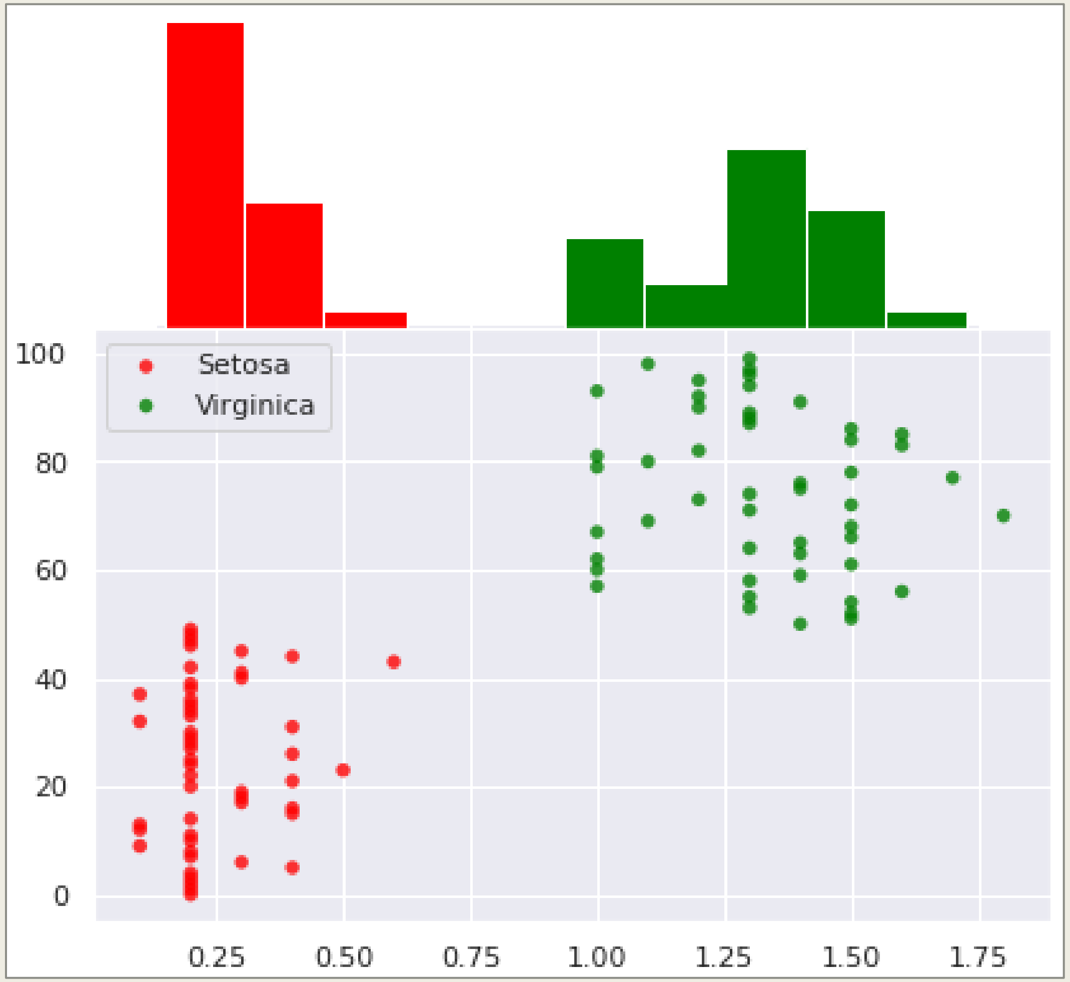
Si somos capaces de estimar $P(c_1|x)$ y $P(c_2|x)$ podríamos determinar la frontera de decisión con la siguiente regla:

$$Si\ P(c_1|x) > P(c_2|x)\ entonces\ \ x \in c_1\ \ sino\ \ x \in c_2$$

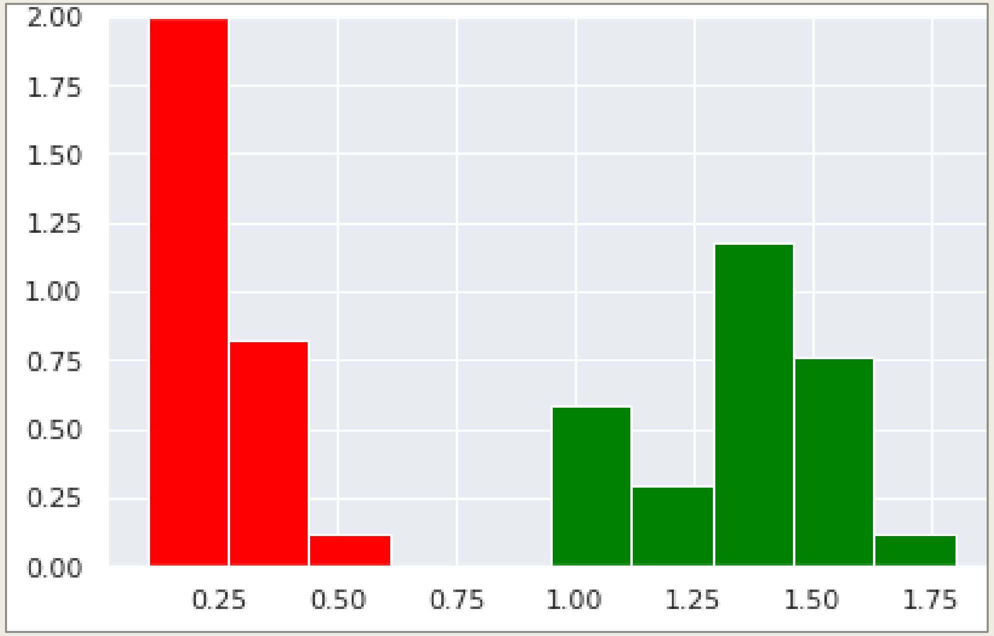
Probabilidad Condicional

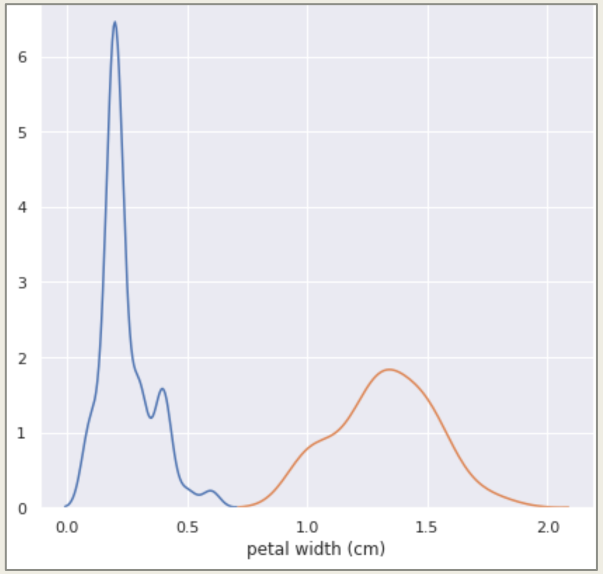
Clasificador Naïve Bayes

Ejemplo: al graficar el ancho del pétalo para la iris setosa e iris virginica, podemos observar la distribución de los datos.



Clasificador Naïve Bayes



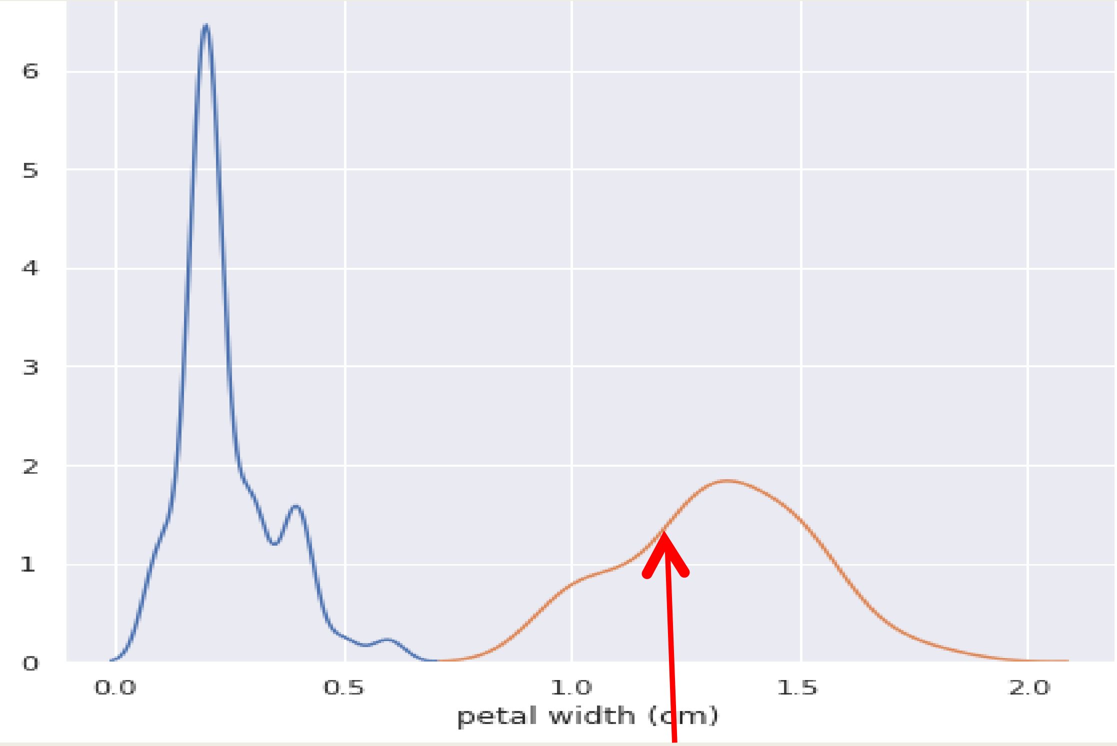


Los histogramas pueden ser resumidos usando una curva que indique la distribución de los datos

Probabilidad Condicional

Clasificador Naïve Bayes

Si nos dieran el ancho del pétalo de una iris. ¿Es posible clasificar que tipo de iris es?



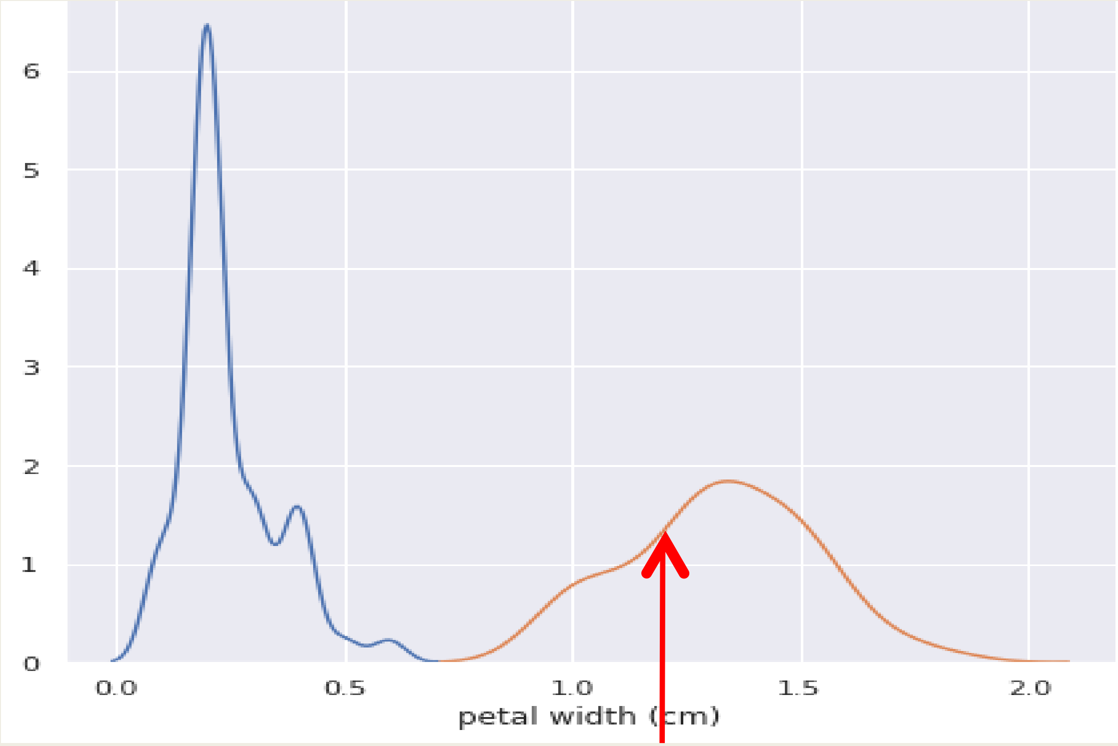
ap = 1.3

Dada la distribución de los datos los más probables es que sea una Iris Virginica.

Probabilidad Condicional

Clasificador Naïve Bayes

De manera más formal podríamos plantear usando probabilidad condicional:



ap = 1.3

$$p(c_i|d)$$

la probabilidad de la clase c_i dado que ya observamos d.

$$p(iris\_setosa|1.3)$$

$$p(iris\_virginica|1.3)$$

Probabilidad Condicional

Clasificador Naïve Bayes

Supongamos que tenemos un conjunto de entrenamiento de 100 muestras, donde 50 elementos son de la clase Setosa y 50 elementos son de la clase Virginica, con una distribución como la que se muestra en la tabla:

Ancho Pétalo	Cantidad	Clase
1.3	5	Setosa
1.3	40	Virginica
2.2	45	Setosa
2.2	10	Virginica

Probabilidad a priori

$$P(iris_{setosa}) = \frac{50}{100} = 0.5$$

$$P(iris_{virginica}) = \frac{50}{100} = 0.5$$

Verosimilitud

$$P(1.3|iris_{setosa}) = \frac{5}{50} = 0.1$$

$$P(1.3|iris_{virginica}) = \frac{40}{50} = 0.8$$

Probabilidad Condicional

Clasificador Naïve Bayes

Los cálculos para el otro posible valor del ancho de pétalo. La probabilidad de clase (a priori) no cambia

Ancho Pétalo	Cantidad	Clase
1.3	5	Setosa
1.3	40	Virginica
2.2	45	Setosa
2.2	10	Virginica

Probabilidad a priori

$$P(iris_{setosa}) = \frac{50}{100} = 0.5$$

$$P(iris_{virginica}) = \frac{50}{100} = 0.5$$

Verosimilitud

$$P(2.2|iris_{setosa}) = \frac{45}{50} = 0.9$$

$$P(2.2|iris_{virginica}) = \frac{10}{50} = 0.2$$

Probabilidad Condicional

Clasificador Naïve Bayes

Una vez calculadas las probabilidades a priori y verosimilitud, se puede calcular la probabilidad de que una nueva muestra pertenezca a una clase dado que se conoce el ancho del pétalo. ¿Qué pasa si el ancho del pétalo = 2.2?

$$p(iris_{setosa}) = 0.5$$
$$p(iris_{virginica}) = 0.5$$

$$p(1.3|iris_{setosa}) = 0.1$$
$$p(1.3|iris_{virginica}) = 0.8$$

$$p(2.2|iris_{setosa}) = 0.9$$
$$p(2.2|iris_{virginica}) = 0.2$$

$$p(iris_{setosa}|2.2) = p(iris_{setosa}) \times p(1.3|iris_{setosa})$$
$$p(iris_{setosa}|2.2) = 0.5 \times 0.9 = 0.45$$

$$p(iris_{virginica}|2.2) = p(iris_{virginica}) \times p(1.3|iris_{virginica})$$
$$p(iris_{virginica}|2.2) = 0.5 \times 0.2 = 0.1$$

Es más probable que pertenezca a la **clase setosa**

Probabilidad Condicional

Clasificador Naïve Bayes

De acuerdo con Kuncheva [1], un error mínimo se garantiza al seleccionar la clase con mayor probabilidad a posteriori.

Este clasificador es un modelo *basado en probabilidad condicional*: dada una instancia a clasificar representada por un vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ que representan n rasgos (variables independientes). El clasificador asigna un valor de probabilidad a cada instancia $P(c_k|x_1, \ldots, x_n)$.

Obtener estimaciones precisas de las funciones de distribución de probabilidad $P(c_k|\boldsymbol{x})$ es difícil, especialmente cuando la dimensionalidad del espacio de características es alto o los rasgos toman una gran cantidad de valores.

Clasificador Naïve Bayes

$$P(c_k|\boldsymbol{x}) = \frac{P(c_k) \cdot P(\boldsymbol{x}|c_k)}{P(\boldsymbol{x})} \qquad posterior = \frac{prior \times likelihood}{evidence}$$

De este ecuación lo importante es el numerador, ya que los valores de $\boldsymbol{x}$ son conocidos, por lo que se le considera una constante [2].

El numerador es equivalente a la distribución conjunta de probabilidades:

$$P(c_k, x_1, \dots, x_n)$$

Que puede ser re-escrita:

$$P(c_k)\, P(x_1, \dots, x_n|c_k) = P(c_k)\, P(x_1|c_k)\, P(x_2, \dots, x_n|c_k, x_1)$$

$$= P(c_k)\, P(x_1|c_k)\, P(x_2|c_k, x_1)\, P(x_3, \dots, x_n|c_k, x_1, x_2)$$

$$= P(c_k)\, P(x_1|c_k)\, P(x_2|c_k, x_1)\, P(x_3|c_k, x_1, x_2) P(x_4, \dots, x_n|c_k, x_1, x_2, x_3)$$

Probabilidad Condicional

Clasificador Naïve Bayes

Usando la asunción ingenua de la independencia condicional, cada rasgo x_i es condicionalmente independiente de cualquier otro rasgo x_j para $i \neq j$, lo que significa:

$$P(x_i | c_k, x_j) = P(x_i | c_k)$$

Por tanto la probabilidad conjunta puede ser escrita:

$$P(c_k, x_1, \dots, x_n) = P(c_k)\, P(x_1|c_k)\, P(x_2|c_k)\, P(x_3|c_k) \dots P(x_n|c_k)$$

$$= P(c_k) \prod_{i=1}^{n} P(x_i|c_k)$$

Probabilidad Condicional

Clasificador Naïve Bayes

Ejemplo: consideramos el siguiente conjunto de entrenamiento

Example training data set

Cook	Mood	Cuisine	Tasty
Sita	Bad	Indian	Yes
Sita	Good	Continental	Yes
Asha	Bad	Indian	No
Asha	Good	Indian	Yes
Usha	Bad	Indian	Yes
Usha	Bad	Continental	No
Asha	Bad	Continental	No
Asha	Good	Continental	Yes
Usha	Good	Indian	Yes
Usha	Good	Continental	No

Consideremos un nuevo patrón a clasificar:

(cook=sita, modo=bar, cuisine=continental)

Probabilidad Condicional

Clasificador Naïve Bayes

Probabilidad a priori de las clases:

Probabilidad a priori de Tasty = yes → $P(\text{Tasty=yes}) = \frac{6}{10} = 0.6$

Probabilidad a priori de Tasty = no → $P(\text{Tasty=}no) = \frac{4}{10} = 0.4$

Probabilidad versosimilitud de los rasgos:

$$P(\text{Cook=sita}|\text{Tasty=yes}) = \frac{2}{6} = 0.33$$

$$P(\text{Cook=sita}|\text{Tasty=no}) = 0 = 0.01$$

Se asigna un valor pequeño para no anular todo el cálculo de la probabilidad.

$$P(\text{Mood=bad}|\text{Tasty=yes}) = \frac{2}{6} = 0.33$$

$$P(\text{Mood=bad}|\text{Tasty=no}) = \frac{3}{4} = 0.75$$

Probabilidad Condicional

Clasificador Naïve Bayes

Probabilidad verosimilitud de los rasgos:

$$P(\text{Cuisine=continental}|\text{Tasty=yes}) = \frac{2}{6} = 0.33$$

$$P(\text{Cuisine=continental}|\text{Tasty=no}) = \frac{3}{4} = 0.75$$

Por tanto las probabilidades a posteriori son:

$$P(\text{Tasty=yes}|\boldsymbol{X}) = 0.6 \times 0.33 \times 0.33 \times 0.33 = 0.0216$$

$$P(\text{Tasty=no}|\boldsymbol{X}) = 0.4 \times 0.01 \times 0.75 \times 0.75 = 0.00225$$

El nuevo patrón es clasificado como perteneciente a la clase *Tasty* = *yes*

Probabilidad Condicional

Ventajas:

- Rápido de entrenar y al clasificar
- No es sensitivo a rasgos irrelevantes
- Puede manejar valores discretos y continuos

Desventajas:

- Asume independencia en los datos

Referencias

- [1] Leondes, C.T. (2018). *Image Processing and Pattern Recognition*. California: Academic Press.
- [2] Duda, R.O., Hart, P.E. & Stork, D.G. (2001). *Pattern Classification*. 2nd edition. Wiley-Interscience.
- [3] Marques de Sá, J:P. (2001). *Pattern Recognition: Concepts, Methods and Applications*. Berlin: Springer-Verlag.
- [4] Kuncheva, L. (2014). *Combining Pattern Classifiers: Methods and Algorithms*. 2nd edition. USA: Wiley.
- [5] Witten, I.H., Frank, E. & Hall, M.A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd edition. USA: Elsevier.
- [6] Murty, N.M. & Devi, V.S. (2011). *Pattern Recognition: An Algorithmic Approach*. Springer.
- [7] Zaki, M.J. & Meira, W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.
- [8] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H. & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems With Applications*, 73, 220-239.