Métodos Bayesianos I

García Prado, Sergio
sergio@garciparedes.me

26 de abril de 2017

Resumen

En este documento se resuelven distintos ejercicios relacionados con el Teorema de Bayes*, todos ellos desde una perspectiva práctica relacionada con el aprendizaje automático. Además, se realiza una clasificación manual basada en el algoritmo* Naive Bayes *[JL95]. Las 3 últimas secciones se corresponden con la realización de experimentos sobre distintas variaciones del conjunto de datos* weather-nominal *[dat] y* Naive Bayes *para después discutir los resultados.*

1. Se sabe que un 1 % de las mujeres de 40 años que participan en un examen rutinario tienen cáncer de mama. También se sabe que un 80 % de las mujeres que tienen cáncer de mama, darán positivo al hacerse una mamografía. Sin embargo, un 9,6 % de las mujeres que no tienen cáncer de mama darán positivo en una mamografía. En este contexto una mujer de 40 años se somete a un examen rutinario y su mamografía da positivo. ¿Cuál es la probabilidad de que realmente tenga cáncer de mama?

Para resolver este problema de probabilidad condicionada se han definido las variables X e Y tal y como se muestra a continuación. Seguidamente se ha definido la probabilidad de los sucesos descritos en el enunciado del ejercicio. Mediante el teorema de bayes se ha obtenido la probabilidad del que una mujer de 40 años tenga cáncer realmente tras haberse realizado un examen rutinario.

Se ha obtenido una probabilidad del 0,8 % de que tras haber dado positivo en el examen finalmente tenga cancer de mama.

$$X = \text{Tener Cancer de mama} \rightarrow \{0,1\} \tag{1}$$
$$Y = \text{Dar positivo al hacerse una mamografía} \rightarrow \{0,1\} \tag{2}$$

$$Pr(Y = 1|X = 1) = 0{,}800 \tag{3}$$
$$Pr(Y = 1|X = 0) = 0{,}096 \tag{4}$$
$$Pr(X = 1) = 0{,}010 \tag{5}$$

$$Pr(X = 1|Y = 1) = Pr(X = 1) \cdot Pr(Y = 1|X = 1) = 0{,}010 \cdot 0{,}800 = 0{,}008 \tag{6}$$

Nótese que en el enunciado del ejercicio se habla de casos en los cuales las probabilidades son referidas a mujeres de 40 años mientras que hay otros casos en que se habla en general. Sin embargo esto no afecta al resultado del problema debido a la cuestión que se pregunta ya que se presupone que el suceso referido a todas las mujeres se da con la misma probabilidad en mujeres de 40 años.

	x_1	x_2	x_3	x_4	$Pr(Y)$
y_1	2/16	1/16	1/16	1/16	5/16
y_2	1/16	2/16	2/16	1/16	6/16
y_3	1/16	1/16	1/16	0	3/16
y_4	0	2/16	0	0	2/16
$Pr(X)$	4/16	6/16	4/16	2/16	16/16

Tabla 1: *Frecuencias relativas de la distribución de probabilidad conjunta de X e Y*

2. Dadas dos variables aleatorias discretas, X e Y, y dada su distribución de probabilidad conjunta que aparece en la tabla, se pide:

2.1. ¿Cumple la distribución conjunta las propiedades de una distribución de probabilidades?

Si que cumple la distribución de probabilidad conjunta debido a que la suma de frecuencias de las variables X e Y da como resultado la unidad ($= 1$). Esa es la restricción necesaria para que la tabla 1 se refiera a la distribución conjunta de frecuencias relativas.

2.2. ¿Cuál es la probabilidad de $Pr(X = x_1)$?

La probabilidad que la variable X tome el valor x_1 se puede obtener mediante el sumatorio de las frecuencias de la variable Y condicionadas por $X = x_1$. Esto se muestra en la ecuación (7)

$$\begin{aligned} Pr(X = x_1) &= \sum_{i=1}^{4} fr(Y = y_i, X = x_1) \\ &= fr(Y = y_1, X = x_1) + fr(Y = y_2, X = x_1) + fr(Y = y_3, X = x_1) + fr(Y = y_4, X = x_1) \\ &= 2/16 + 1/16 + 1/16 + 0 \\ &= 4/16 \end{aligned} \tag{7}$$

2.3. ¿Cuáles son las distribuciones marginales de $Pr(X = x)$ y $Pr(Y = y)$?

Las distribuciones marginales de frecuencias relativas para las variables X e Y se muestran en la última fila y columna de la tabla 1 respectivamente. Dicha construcción se lleva a cabo de manera sencilla mediante la fórmula descrita en la ecuación (8).

$$Pr(A = a_i) = \sum_{j} fr(B = b_j, X = a_i) \tag{8}$$

2.4. ¿Verifican las distribuciones marginales las propiedades de una distribución de probabilidades?

Las distribuciones marginales de las variables X e Y si que verifican las propiedades necesarias para ser condicionadas distribuciones de probabilidad debido a que la suma de ellas da como resultado la unidad ($\sum_i fr(A = a_i) = 1$) para las dos variables. La razón de ello es que la tabla representa la distribución conjunta de frecuencias relativas de dichas variables.

3. Utilizando el conjunto de datos *weather-nominal-practica* que se proporciona, determinar la clasificación Naive Bayes de las siguientes instancias, utilizando la estimación de máxima verosimilitud (frecuencial)

$$x_1 =< sunny, cool, normal, false > \tag{9}$$
$$x_2 =< overcast, mild, high, true > \tag{10}$$

En este ejercicio se pide clasificar las instancias x_1 y x_2 utilizando el algoritmo *Naive Bayes* de manera manual. Por lo tanto, para ello la primera tarea es calcular las distribuciones de probabilidad de los valores posibles para la clase de destino. A continuación es necesario calcular las probabilidades condicionadas del conjunto de valores que pueden tomar los atributos del conjunto de datos condicionados al valor de la clase. Para realizar dichas tareas se utiliza el conjunto de entrenamiento.

$$\begin{aligned}
Pr(play = yes) &= 9/14 \\
Pr(play = no) &= 5/14 \\
Pr(outlook = sunny|\ play = yes) &= 2/9 \\
Pr(outlook = overcast|\ play = yes) &= 2/9 \\
Pr(outlook = rainy|\ play = yes) &= 5/9 \\
Pr(outlook = sunny|\ play = no) &= 3/5 \\
Pr(outlook = overcast|\ play = no) &= 0/5 \\
Pr(outlook = rainy|\ play = no) &= 2/5 \\
Pr(temperature = hot|\ play = yes) &= 1/9 \\
Pr(temperature = mild|\ play = yes) &= 4/9 \\
Pr(temperature = cool|\ play = yes) &= 4/9 \\
Pr(temperature = hot|\ play = no) &= 2/5 \\
Pr(temperature = mild|\ play = no) &= 2/5 \\
Pr(temperature = cool|\ play = no) &= 1/5 \\
Pr(humidity = high|\ play = yes) &= 3/9 \\
Pr(humidity = normal|\ play = yes) &= 6/9 \\
Pr(humidity = high|\ play = no) &= 4/5 \\
Pr(humidity = normal|\ play = no) &= 1/5 \\
Pr(windy = true|\ play = yes) &= 4/9 \\
Pr(windy = false|\ play = yes) &= 5/9 \\
Pr(windy = true|\ play = no) &= 3/5 \\
Pr(windy = false|\ play = no) &= 2/5
\end{aligned}$$

El siguiente paso es calcular la probabilidad de que se den los valores de los atributos de cada instancia de manera conjunta en cada una de las clases para después clasificar la instancia en la clase que mayor probabilidad presente. Los resultados se muestran en la tabla 2.

$$\begin{aligned}
&Pr(outlook = sunny, temperature = cool, humidity = normal, windy = false|\ play = yes) = \\
&Pr(play = yes) \cdot Pr(outlook = sunny|\ play = yes) \cdot Pr(temperature = hot|\ play = yes) \cdot \\
&Pr(humidity = normal|\ play = yes) \cdot Pr(windy = false|\ play = yes) = \\
&9/14 \cdot 2/9\ \cdot 4/9 \cdot 6/9 \cdot 5/9 = 40/1701 = 0{,}02351557
\end{aligned} \tag{11}$$

$$\begin{aligned} Pr(outlook = sunny, temperature = cool, humidity = normal, windy = false|\ play = no) = \\ Pr(play = no) \cdot Pr(outlook = sunny|\ play = no) \cdot Pr(temperature = hot|\ play = no) \cdot \\ Pr(humidity = normal|\ play = no) \cdot Pr(windy = false|\ play = no) = \\ 5/14 \cdot 3/5 \cdot 1/5 \cdot 1/5 \cdot 2/5 = 3/875 = 0{,}003428571 \end{aligned} \tag{12}$$

En las ecuaciones (11) y (12) se calculan las probabilidades de pertenencia a las correspondientes clases para la instancia x_1. Nótese que se omite el denominador, la razón de ello se debe a que no se pretende conocer la probabilidad exacta de que la instancia pertenezca a dicha clase, sino utilizarlo como punto de comparación respecto del resto de clases. Dado que dicho denominador es constante para todos los cálculos entonces se puede omitir. Tal y como lleva a cabo el algortimo *Naive Bayes*, la clase en la cual se debe etiquetar la instancia es la que maximice dichos cálculos, por tanto x_1 **se clasifica como yes**.

$$\begin{aligned} Pr(outlook = overcast, temperature = mild, humidity = high, windy = true|\ play = yes) = \\ Pr(play = yes) \cdot Pr(outlook = overcast|\ play = yes) \cdot Pr(temperature = mild|\ play = yes) \cdot \\ Pr(humidity = high|\ play = yes) \cdot Pr(windy = true|\ play = yes) = \\ 9/14 \cdot 2/9\ \cdot 4/9 \cdot 3/9 \cdot 4/9 = 16/1701 = 0{,}00940623 \end{aligned} \tag{13}$$

$$\begin{aligned} Pr(outlook = overcast, temperature = cool, humidity = high, windy = true|\ play = no) = \\ Pr(play = no) \cdot Pr(outlook = overcast|\ play = no) \cdot Pr(temperature = mild|\ play = no) \cdot \\ Pr(humidity = high|\ play = no) \cdot Pr(windy = true|\ play = no) = \\ 5/14 \cdot 0/5 \cdot 2/5 \cdot 4/5 \cdot 4/5 = 0 \end{aligned} \tag{14}$$

En las ecuaciones (11) y (12) se calculan las probabilidades de pertenencia a las correspondientes clases para la instancia x_2. Al igual que en el caso anterior, no es necesario calcular el denominador de la igualdad del *Teorema de Bayes*. Tal y como lleva a cabo el algortimo *Naive Bayes*, la clase en la cual se debe etiquetar la instancia es la que maximice dichos cálculos, por tanto x_2 **se clasifica como yes**.

Naive Bayes - Manual	
Instancia	Clase
$< sunny, cool, normal, false >$	yes
$< overcast, mild, high, true >$	yes

Tabla 2: *Clasificación obtenida de manera manual*

4. Utilizando Weka y el clasificador NaiveBayes determinar la clasificación de los ejemplos anteriores, ¿Coindice con la clasificación calculada en el ejercicio anterior?

Los resultados obtenidos al clasificar las instancias x_1 y x_2 del ejercicio anterior mediante Weka se mestran en la tabla 3. Dichos resultados son equivalente a los calculados de manera manual. Sin embargo, podría no haber sido así debido a que la implementación de *Naive Bayes* en *Weka* utiliza la estimación Bayesiana (m-estima), la cual se muestra en la ecuación (15) donde n_c representa el número de ejemplos de entrenamiento con la clase b_j, n el número de ejemplos de la de entrenamiento con la clase b_j, y el atributo a_i, $p = Pr(A = a_i|B = b_j)$, es decir, la estimación a priori y m un determinado peso para la estimación a priori. A partir de dicha estrategia se tratan los casos con probabilidad nula, que provocan que la probabilidad al aplicar el *Teorema de Bayes* sobre nuevas instancias sea nula, y por tanto, las instancias no se etiqueten en dicha clase debido a un único atributo con probabilidad 0 ($Pr(A = a_i|B = b_j) = 0$). La estimación mediante la ecuación (15) resuelve esta problemática.

$$Pr'(A = a_i|B = b_j) = \frac{n_c + mp}{n + m} \tag{15}$$

Naive Bayes - Weka	
Instancia	Clase
$< sunny, cool, normal, false >$	yes
$< overcast, mild, high, true >$	yes

Tabla 3: *Clasificación obtenida mediante WEKA*

5. Entrenar con Weka, un clasificador Naive Bayes para el conjunto de datos *weather-nominal*

5.1. Estimar la tasa de error cometida por el clasificador utilizando validación cruzada de 10 particiones

En la tabla 4 se muetra la tasa de erro cometida tras realizar un experimento de validación cruzada de 10 particiones sobre el conjunto de datos *weather-nominal*.

Naive Bayes - Weka	
Datos	Tasa de Error
Weather Nominal	42,8571 %

Tabla 4: *Validación Cruzada de 10 particiones con Naive Bayes*

5.2. Examinar la salida proporcionada por el Explorer y determinar cómo está estimando esta implementación de *Naive Bayes* los parámetros del clasificador

El algoritmo *Naive Bayes Simple* implementado en Weka funciona de la misma manera que la implementación manual realizada en este documento. Primero se calcula la probabilidad a priori de cada uno de los valores que puede tomar la clase de destino sobre el conjunto de datos de entrenamiento. Seguidamente calcula las probabilidades condicionadas de cada uno de los valores de los atributos del conjunto de datos respecto de las distintas clases de destino.

Para clasificar las instancias supone independencia entre atributos y calcula la probabilidad conjunta de todos los atributos de la instancia condicionada a las distintas clases para despues clasificar dicha instancia en la clase que maximice el valor de probabilidad de la misma.

6. El conjunto de datos *weather-nominal-X6* se ha generado repitiendo cada instancia del conjunto *weather-nominal* seis veces. Entrenar con Weka un clasificador Naive Bayes para este conjunto de datos:

6.1. Estimar la tasa de error cometida por el clasificador utilizando validación cruzada de 10 particiones

En la tabla 5 se muetra la tasa de erro cometida tras realizar un experimento de validación cruzada de 10 particiones sobre el conjunto de datos *weather-nominal-X6*.

Naive Bayes - Weka	
Datos	Tasa de Error
Weather Nominal x6	9,5238 %

Tabla 5: *Validación Cruzada de 10 particiones con Naive Bayes*

6.2. Compare esta tasa de error con la estimada en el ejercicio anterior y discuta los resultados

Los resultados de los experimentos realizados sobre los dos conjuntos de datos se muestran en las tablas 4 y 5. A pesar de corresponderse con el mismo conjunto de datos (solo que con instancias duplicadas 6 veces en el caso del último) los resultados obtenidos presentan una gran diferencia en cuanto a la tasa de error.

La razón por la cual ha sucedido este fenómeno se puede deber al tamaño tan reducido del conjunto de datos referido tan solo a 14 instancias en el caso en que no se repiten, y al tipo de experimento realizado, que divide dicho conjunto en 10 particiones. Esto hace que en algunos casos de test no haya instancias que añadan probabilidad a algunos de los posibles casos, lo cual se deriva en probabilidad nula durante la clasificación, lo que produce un error. Dicho fenómeno es más difícil que suceda cuando se presentan 6 veces más instancias (84 instancias) por lo que los resultados son mejores.

Referencias

[CCAG17] Teodoro Calonge Cano and Carlos Javier Alonso GonzáLez. Técnicas de Aprendizaje Autómatico, 2016/17.

[dat] Weather Nominal Data Set. `http://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/weather.nominal.arff`.

[GP17] Sergio García Prado. Métodos bayesianos 1. `https://github.com/garciparedes/machine-learning-bayesian-1`, 2017.

[JL95] George H John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc., 1995.

[too] Weka. `http://www.cs.waikato.ac.nz/ml/weka/`.