Teoría de Probabilidades

Dr. Luis Gerardo de la Fraga

Cinvestav, Departamento de Computación
fraga@cs.cinvestav.mx

14 de abril, 2020

Resumen

Este documento es una traducción propia de la sección 1.2 *Probability Theory*, del libro *Pattern Recognition and Machine Learning* de C.M. Bishop.

Un concepto clave en el campo de reconocimiento de patrones es sobre incertidumbre. Esto resulta tanto del ruido en las mediciones, como a través del tamaño finito de los conjuntos de datos. La teoría de probabilidades provee un marco consistente para la cuantificación y manipulación de la incertidumbre y forma uno de los fundamentos centrales para el reconocimiento de patrones. Cuando se combina con la teoría de decisiones, nos permite realizar predicciones óptimas dada toda la información disponible que tengamos, aún cuando esa información pueda ser incompleta o ambigua.

Se introducirán los conceptos básicos de probabilidad considerando un ejemplo simple. Imagine que tiene dos cajas, una roja y otra azul, la caja roja contiene 2 manzanas y 6 naranjas, y la caja azul contiene 3 manzanas y 1 naranja. Esto se ilustra en la figura 1. Supongamos ahora que se toma aleatoriamente una de las cajas y de esa caja aleatoriamente seleccionamos una fruta, y observando que tipo de fruta es, la reemplazamos en la caja de donde provino. Podríamos imaginar que repetimos este proceso muchas veces. Supongamos que tomamos la caja roja un 40 % de las veces y que tomamos la caja azul un 60 % de las veces, y cuando quitamos una fruta de una caja tenemos la misma probabilidad de seleccionar cualquiera de las piezas de fruta dentro de la caja.

En este ejemplo, la identidad de la caja que será escogida es una variable aleatoria, la cual se denotará por C. Esta variable aleatoria puede tomar uno de dos valores posibles, es decir r (correspondiente a la caja roja) o a (correspondiente a la caja azul). De forma similar, la identidad de la fruta es también una variable aleatoria y se denotará por F. Esta puede tomar los valores de m (para manzana) o n (para naranja).

Para comenzar, se definirá la probabilidad de un evento como la fracción de veces que ese evento ocurra entre el número total de intentos, en el límite de que el número total de eventos tienda a infinito. Así, la probabilidad de seleccionar la caja verde es 4/10 (de la suposición inicial) y la probabilidad de seleccionar

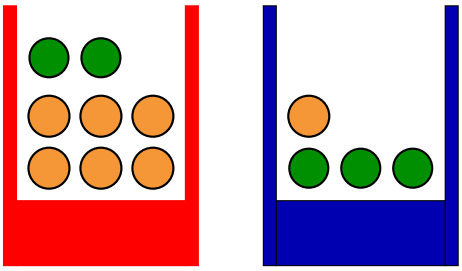

Figura 1: El ejemplo simple de dos cajas coloreadas cada una conteniendo fruta (las manzanas se muestran en verde y las naranjas en anaranjado) para introducir las ideas básicas de probabilidad.

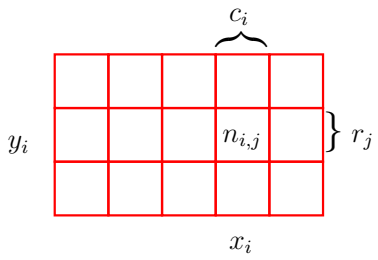

Figura 2: Se pueden derivar las reglas de suma y de producto de la probabilidad considerando dos variables aleatorias, X, la cual toma valores $\{x_i\}$ donde $i = 1, \ldots, M$, y Y, la cual toma valores $\{y_j\}$, para $j = 1, \ldots, L$. Es esta ilustración se tiene $M = 5$ y $L = 3$. Si se considera un total de N instancias de estas variables, se denota el número de instancias donde $X = x_i$ y $Y = y_j$ como $n_{i,j}$, la cual es el número de puntos en la celda correspondiente del arreglo. El número de puntos en la columna i, correspondiente a $X = x_i$, se denota por c_i, y el número de puntos en el renglón j, correspondiente a $Y = y_j$ se denota como r_j.

la caja azul es 6/10. Se escribirán estas probabilidades como $p(C = r) = 4/10$ y $p(C = a) = 6/10$. Nótese que por definición, las probabilidades deben yacer en el intervalo $[0, 1]$. También, si los eventos son mutuamente exclusivos e incluyen todos los posibles resultados (por ejemplo, la caja debe ser o roja o azul), entonces la suma de los probabilidades de esos eventos deben de sumar uno.

Se pueden hacer ahora preguntas como: "¿cuál es la probabilidad total que el procedimiento de selección tome una manzana?", o "dado que ya se ha tomado una naranja, ¿cuál es la probabilidad de que en la caja que escogimos sea la azul?" Se pueden responder preguntas de este tipo, y de hecho preguntas mucho más complejas asociados con problemas en reconocimiento de patrones, una vez que estemos equipados con las reglas elementales de probabilidad, conocidas como la *regla de suma* y la *regla de producto*. Habiendo deducido estas reglas, se regresará al ejemplo de las cajas de frutas.

Para derivar las reglas de probabilidad, consideremos el ejemplo más general mostrado en la figura 2 que envuelve dos variables aleatorias X y Y (las cuales podrían ser las variables de la caja y la fruta consideradas anteriormente). Supondremos que X pude tomar valores x_i donde $i = 1, \ldots, M$, y Y puede tomar valores y_j donde $j = 1, \ldots, L$. Considérese un total de N intentos en los cuales se muestrea ambas variables X y Y, y sea n_{ij} el número de tales intentos en los cuales $X = x_i$ y $Y = y_j$. También, sea c_i el número de intentos en los cuales X toma los valores x_i (independientemente del valor que Y tome), y similarmente sea r_j el número de intentos en los cuales Y toma valores y_j.

La probabilidad de que X tomará el valor x_i y Y tomará el valor y_j se escribe como $p(X = x_i, Y = y_j)$ y se le llama probabilidad *conjunta* de $X = x_i$ y $Y = y_i$. Esta probabilidad está dada por el número de puntos que yacen en la celda i, j como una fracción del número total de puntos, y es

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}. \tag{1}$$

Aquí se está considerando implícitamente el límite $N \to \infty$. Se forma similar, la probabilidad de que X tome el valor x_i independientemente del valor de Y se escribe como $p(X = x_i)$ y está dada por la fracción del número total de puntos que yacen en la columna i, de forma que

$$p(X = x_i) = \frac{c_i}{N}. \tag{2}$$

Debido a que el número de instancias en la columna i en la figura 2 es justo la suma del número de instancias en cada calda en esa columna, se tiene $c_i = \sum_j n_{ij}$ y por lo tanto, de (1) y (2), tenemos

$$p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j) \tag{3}$$

la cual es la *regla de suma* de la probabilidad. Nótese que a $p(X = x_i)$ se le llama a veces la probabilidad *marginal*, porque se obtiene marginalizando, o sumando, las otras variables (en esta caso Y).

Si consideramos aquellas instancias para las cuales $X = x_i$, entonces la fracción de tales instancias para las cuales $Y = y_j$ se escribe como $p(Y = y_i|X = x_i)$ y se le llama la probabilidad *condicional* de $Y = y_j$ dada $X = x_i$. Esta se obtiene encontrando la fracción de aquellos puntos en la columna i que yacen en la celda i, j y entonces es dado por

$$p(Y = y_i|X = x_i) = \frac{n_{i,j}}{c_i}. \tag{4}$$

De las expresiones (1), (2) y (4), se puede derivar la siguiente relación

$$\begin{aligned} p(X = x_i, Y = y_j) = \frac{n_{i,j}}{N} = \frac{n_{i,j}}{c_i} \cdot \frac{c_i}{N} = \\ = p(Y = y_i|X = x_i)p(X = x_i) \end{aligned} \tag{5}$$

la cual es la *regla de producto* de la probabilidad.

Hasta ahora se ha tenido mucho cuidado en hacer la distinción entre una variable aleatoria, tal como la caja C en el ejemplo con la fruta, y los valores que la variable aleatoria puede tomar, por ejemplo r si la caja es roja. Así la probabilidad que C tome el valor r se denota $p(C = r)$. Aunque esto ayuda a evitar ambigüedad, resulta en una notación engorrosa, y en muchos casos no hay necesidad para tal pedantería. En vez de ello, se podría escribir simplemente $p(C)$ para denotar una distribución sobre la variable aleatoria C, o $p(r)$ para denotar la distribución evaluada para el particular valor de r, previniendo que la interpretación es clara a partir del contexto.

Con esta notación más compacta, las dos reglas fundamentales de la teoría de probabilidad se podrían escribir en la siguiente forma:

$$\textbf{regla de suma} \qquad p(X) = \sum_{Y} p(X, Y) \tag{6}$$

$$\textbf{regla de producto} \qquad p(X, Y) = p(Y|X)\, p(X). \tag{7}$$

Aquí $p(X, Y)$ es una probabilidad conjunta y se pronuncia como "la probabilidad de X *y* Y". De forma similar, la cantidad $p(Y|X)$ es una probabilidad condicional y se pronuncia como "la probabilidad de Y *dada* X", mientras la cantidad $p(X)$ es una probabilidad marginal y es simplemente "la probabilidad de X". Estas dos reglas simples forman la base para toda la maquinaria probabilística que se usa en el libro.

A partir de la regla de producto, junto con la propiedad de simetría $p(X, Y) = p(Y, X)$, se puede obtener inmediatamente la siguiente relación entre probabilidades condicionales

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \tag{8}$$

la cual se llama *teorema de Bayes* y juega un papel central en reconocimiento de patrones y aprendizaje automático. Usando la regla de suma, el denominador en el teorema de Bayes puede expresarse en términos de la cantidades que aparecen en el numerador

$$p(X) = \sum_{Y} p(X|Y)p(Y). \tag{9}$$

Se puede ver el denominador en el teorema de Bayes como la constante de normalización requerida para asegurar que la suma de la probabilidad condicional en el lado izquierdo de (8) sobre todos los valores de Y sea igual a uno.

En la figura 3 se muestra un ejemplo simple que envuelve una distribución conjunta sobre dos variables para ilustrar el concepto de distribuciones marginales y condicionales. Aquí una muestra finita de $N = 60$ puntos se ha dibujado a partir de la distribución conjunta y se muestra arriba a la izquierda. Arriba a la derecha está el histograma de las fracciones de puntos teniendo cada uno de los dos valores de Y. De las definiciones de probabilidad, estas fracciones deberían ser iguales a las probabilidades correspondiente $p(Y)$ en el límite $N \to \infty$. Se puede ver el histograma como una forma fácil para modelar una distribución de probabilidades dados solamente un número finito de puntos dibujados de esa

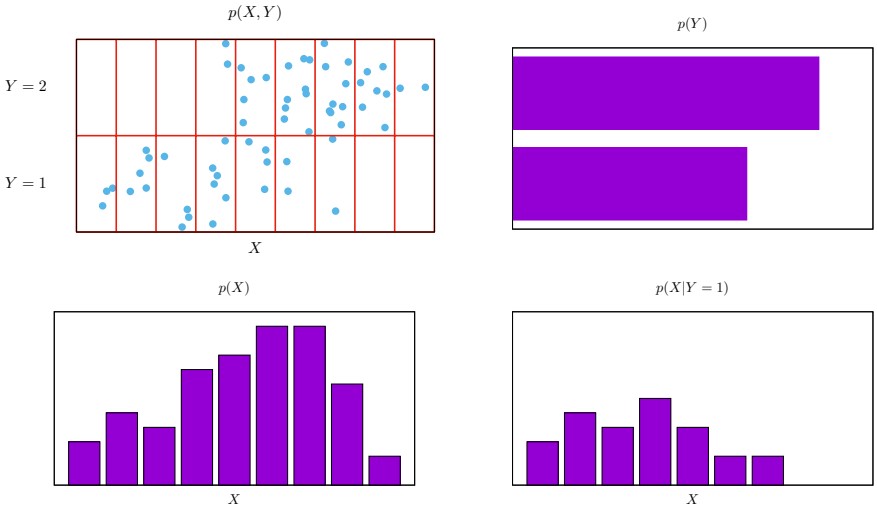

Figura 3: Ilustración de una distribución sobre dos variables, X, la cual toma 9 valores posiblesm y Ym la cual toma 2 valores posibles. Arriba a la izquierda se ve una muestra de 60 puntos dibujados a partir de una distribuciób de probabilidad conjunta sobre esas variables. Las subsiguientes figuras muestran histogramas de las distribuciones marginales $p(X)$ y $p(Y)$, así como también la distribución condicional $p(X|Y = 1)$ correspondiente a la gráfica abajo a la derecha.

distribución. Modelar distribuciones a partir de datos yace en el corazón del reconocimiento de patrones estadístico. Las dos gráficas restantes de la figura 3 muestran los histogramas correspondientes estimados de $p(X)$ y $p(X|Y = 1)$.

Regresando ahora al ejemplo de las cajas de fruta. Por el momento, regresemos de nuevo a la distinción explícita entre variables aleatorias y sus instanciaciones. Hemos visto que las probabilidades de seleccionar ya sea la caja roja o azul está dadas por

$$p(C = r) = 4/10 \tag{10}$$
$$p(C = a) = 6/10 \tag{11}$$
$$\tag{12}$$

respectivamente. Nótese que estas satisfacen $p(C = r) + p(C = a) = 1$.

Ahora suponga que se toma una caja aleatoriamente, y se ve que es la caja azul. Entonces la probabilidad de seleccionar una manzana es justo la fracción de manzanas en la caja azul, la cual es $3/4$, así $p(F = m|C = a) = 3/4$. De hecho, se pueden escribir todas las cuatro probabilidades conjuntas para el tipo

de fruta, dada la caja seleccionada

$$p(F = m|C = r) = 1/4 \tag{13}$$
$$p(F = n|C = r) = 3/4 \tag{14}$$
$$p(F = m|C = a) = 3/4 \tag{15}$$
$$p(F = n|C = a) = 1/4. \tag{16}$$

De nuevo, nótese que estas probabilidades están normalizadas de forma que

$$p(F = m|C = r) + p(F = n|C = r) = 1 \tag{17}$$

y de forma similar

$$p(F = m|C = a) + p(F = n|C = a) = 1. \tag{18}$$

Ahora podemos usar las reglas de suma y producto de probabilidad para evaluar todas las probabilidades de escoger una manzana

$$\begin{aligned} p(F = m) &= p(F = m|C = r)p(C = r) + p(F = m|C = a)p(C = a) \\ &= \frac{1}{4} \times \frac{4}{10} + \frac{3}{4} \times \frac{6}{10} = \frac{11}{20} \end{aligned} \tag{19}$$

a partir del cual sigue, usando la regla de suma, que $p(F = n) = 1 - 11/20 = 9/20$.

Supongamos ahora que decimos que la pieza de fruta ha sido seleccionada y es una naranja, y quisiésemos saber de cuál caja fue tomada. Esto requiere que se evalúe la distribución de probabilidad sobre las cajas condicionada sobre la identidad de la fruta, mientras que las probabilidades en (14)-(16) dan la distribución de probabilidades sobre la fruta condicionadas sobre la identidad de la caja. Se puede resolver el problema de volver al revés la probabilidad condicional usando el teorema de Bayes para dar

$$p(C = r|F = n) = \frac{p(F = n|C = r)p(C = r)}{p(F = n)} = \frac{3}{4} \times \frac{4}{10} \times \frac{20}{9} = \frac{2}{3}. \tag{20}$$

De la regla de suma, sigue que $p(C = a|F = n) = 1 - 2/3 = 1/3$.

Se puede proveer una interpretación importante del teorema de Bayes como sigue. Si se ha preguntado cuál caja ha sido escogida antes de decir la identidad de la fruta seleccionada, entonces la información más completa disponible está proveída por la probabilidad $p(C)$. A esta se le llama *probabilidad anterior* porque esta es la probabilidad disponible *antes* de observar la identidad de la fruta. Una vez que se dice que la fruta es una naranja, se puede usar el teorema de Bayes para calcular la probabilidad $p(C|F)$, la cual se llamará la *probabilidad posterior* porque esta es la probabilidad obtenida *después* de que haya observado F. Nótese que en este ejemplo, la probabilidad anterior de seleccionar la caja roja fue 4/10, de forma que es más probable seleccionar la caja azul que la roja. Sin embargo, una vez que se ha observado que la pieza de fruta seleccionada

es una naranja, encontramos que la probabilidad posterior de la caja roja es ahora 2/3, de forma que ahora es más probable que la caja se seleccionemos fue de hecho la roja. Este resultado está en acuerdo con nuestra intuición, como la proporción de naranjas es más alta en la caja roja que en la caja azul, y la observación de que la fruta fue una naranja provee evidencia significativa a favor de la caja roja. De hecho, la evidencia es lo suficientemente fuerte que sobrepesa la anterior y hace que se más probable que la caja roja haya sudo seleccionada en vez de la azul.

Finalmente, nótese que si la distribución conjunta de las dos variables factoriza en el producto de las marginales, de forma que $p(X, Y) = p(X)p(Y)$, entonces X y Y se dicen que son *independientes*. De la regla de producto, vemos que $p(Y|X) = p(Y)$, y de esta forma la distribución condicional de Y dado X es de hecho independiente del valor de X. Por ejemplo, en las cajas de fruta, si cada caja contiene la misma fracción de manzanas y naranjas, entonces $p(F|C) = p(F)$, de forma que la probabilidad de seleccionar, digamos, una manzana es independiente de cual caja se escoja.