Combinatoria, probabilidad y entropía estadística

Guillem Pérez Nadal

Prefacio

Estas notas son un repaso de los contenidos necesarios para resolver los problemas de la guía 2. Para muchísima más información, les recomiendo consultar el libro *Introduction to probability*, de Grinstead y Snell, del cual hay una copia de distribución libre que pueden consultar acá.

1 Combinatoria

La combinatoria es el arte de contar cosas. Vamos directamente con algunos ejemplos.

- **Permutaciones.** Una *permutación* es una forma de ordenar los elementos de un conjunto. ¿Cuántas permutaciones hay de un conjunto de n elementos? En la primera posición tenemos n posibilidades porque podemos poner cualquiera de los elementos; para cada una de éstas, hay $n-1$ posibilidades en la segunda posición porque uno de los elementos ya fue asignado, y así sucesivamente. Por lo tanto, el número total de permutaciones es $n(n-1)\dots 2 = n!$.

- **k-permutaciones.** Una *k-permutación* es una lista ordenada de k elementos extraídos de un conjunto. ¿Cuántas k-permutaciones hay de un conjunto de $n \geq k$ elementos? Por el mismo argumento de antes vemos que ese número es $n(n-1)\cdots(n-k+1) = n!/(n-k)!$ (igual que en el caso anterior pero nos quedamos sólo con los primeros k factores).

- **Combinaciones.** Una *combinación* es un subconjunto, es decir, lo mismo que en el ítem anterior pero sin importar el orden. ¿Cuántas combinaciones de k elementos hay de un conjunto de $n \geq k$ elementos? El número de k-permutaciones dividido por el número de formas de ordenar la lista,

$$\frac{n!}{k!(n-k)!} = \binom{n}{k}. \tag{1}$$

Muchos problemas de combinatoria se reducen a alguna de estas tres posibilidades, o a una combinación de ellas, o a una combinación de ellas más alguna cosita más. *Relacionar el problema con estas posibilidades es una de las*

estrategias más útiles en combinatoria. En el material complementario sobre combinatoria tienen un análisis más exhaustivo de los problemas básicos de esta disciplina.

Un comentario sobre combinaciones. El número total de combinaciones de un conjunto de n elementos es

$$\sum_{k=0}^{n} \binom{n}{k} = \sum_{k=0}^{n} \binom{n}{k} 1^k 1^{n-k} = (1+1)^n = 2^n. \tag{2}$$

¿Cómo podemos entender esto? Para elegir una combinación tenemos que tomar una decisión binaria sobre cada elemento del conjunto: si lo incluimos o no en el subconjunto. Hay 2 posibilidades para cada elemento, y por lo tanto 2^n en total.

2 Probabilidad

Distribuciones de probabilidad. El *espacio muestral* M de un experimento es el conjunto de todos sus resultados posibles. Por ejemplo, el espacio muestral del experimento de tirar un dado es $M = \{1, 2, 3, 4, 5, 6\}$. Una *distribución de probabilidad* es una función $p : M \to \mathbb{R}$, que a cada resultado m le asigna un número p_m, con las siguientes propiedades:

$$p_m \geq 0 \qquad \sum_{m \in M} p_m = 1. \tag{3}$$

La probabilidad p_m se interpreta como una medida de nuestra confianza en que el resultado del experimento va a ser m. Hay muchas maneras de asignar probabilidades a cada resultado, que dependerán de la información que tengamos. Por ejemplo, si no tenemos ninguna información es natural asignar igual probabilidad a todos los resultados, con lo cual $p_m = 1/|M|$ para todo m, donde $|M|$ es el número de elementos de M. Pero si hemos tenido la oportunidad de hacer estadística repetiendo el experimento muchas veces entonces quizá sea más juicioso asignar $p_m = N_m/N$, donde N_m es el número de veces que salió el resultado m y N el número de veces que repetimos el experimento. Esta segunda forma de asignar probabilidades se llama la *interpretación frecuentista de la probabilidad*.

Eventos. Un *evento* es un subconjunto de M. Decimos que un evento E *ocurre* si el resultado del experimento está en E; por ejemplo, en el caso del dado el evento $\{2, 4, 6\}$ ocurre si el resultado es par. La *probabilidad de* E se define por

$$p(E) = \sum_{m \in E} p_m, \tag{4}$$

y mide nuestro grado de confianza en que ocurra E. Las siguientes propiedades son fáciles de probar:

(i) $p(\varnothing) = 0$;

(ii) $p(M) = 1$;

(iii) $p(A \cup B) = p(A) + p(B) - p(A \cap B)$.

En particular, si A y B son disjuntos (es decir, si $A \cap B = \varnothing$) se tiene

$$p(A \cup B) = p(A) + p(B). \tag{5}$$

Nótese que la ocurrencia de $A \cup B$ es equivalente a la ocurrencia de A o bien de B, mientras que la de $A \cap B$ equivale a la de A y B a la vez; decir que dos eventos son disjuntos es lo mismo que decir que no pueden ocurrir a la vez. Por último, en el caso en que todos los resultados son igualmente probables se tiene $p(E) = |E|/|M|$, es decir, la probabilidad es el número de casos favorables sobre el número de casos posibles.

Probabilidad condicionada. Consideremos un experimento con distribución de probabilidad p, y supongamos que se nos informa de que al hacer el experimento ha ocurrido el evento E. Por ejemplo, tiramos un dado y sabemos que salió par. ¿Cuál es la distribución de probabilidad q que refleja nuestras nuevas expectativas acerca del resultado del experimento, ahora que tenemos este dato extra? La nueva distribución debe cumplir $q_m = 0$ para $m \notin E$, y es natural imponer también $q_m/q_n = p_m/p_n$ para $m, n \in E$ (es decir, si un resultado de E era el doble de probable que otro antes de saber que E ocurre, sigue siéndolo después de saberlo). Estas dos condiciones determinan q completamente. En efecto, la segunda condición implica $q_m = cp_m$ para $m \in E$, donde c es una constante, y reemplazando este resultado y la primera condición en la segunda propiedad de (3) se obtiene

$$1 = c \sum_{m \in E} p_m = cp(E), \tag{6}$$

con lo cual $c = 1/p(E)$ y por lo tanto

$$q_m = \begin{cases} p_m/p(E) & m \in E \\ 0 & m \notin E. \end{cases} \tag{7}$$

La probabilidad de un evento F de acuerdo con esta nueva distribución se denota como $p(F|E)$ y se llama *probabilidad de F condicionada a E*. Tenemos

$$p(F|E) = q(F) = \sum_{m \in F} q_m = \frac{1}{p(E)} \sum_{m \in F \cap E} p_m = \frac{p(F \cap E)}{p(E)}. \tag{8}$$

Por ejemplo, consideremos el caso del dado con todas las caras equiprobables, de manera que $p_m = 1/6$. ¿Cuál es la probabilidad de el resultado sea mayor que 3 sabiendo que es par? Se nos está pidiendo $p(F|E)$ con $E = \{2, 4, 6\}$ y $F = \{4, 5, 6\}$; tenemos $p(E) = 1/2$ y $p(F \cap E) = 1/3$, y por lo tanto la

respuesta es $p(F|E) = 2/3$. También podríamos haber llegado a este resultado dividiendo casos favorables entre casos posibles. Se dice que dos eventos A y B son *independientes* si $p(A|B) = p(A)$, es decir, si saber que ocurre B no altera en nada la probabilidad de que ocurra A. La ecuación (8) implica que, para A y B independientes,

$$p(A \cap B) = p(A)p(B). \tag{9}$$

Valor de expectación y varianza. El *valor de expectación* de una función $f : M \to \mathbb{R}$ se define como

$$\langle f \rangle = \sum_{m \in M} f_m p_m. \tag{10}$$

En la interpretación frecuentista, el valor de expectación de f es simplemente su promedio sobre todas las veces que hemos repetido el experimento,

$$\langle f \rangle = \frac{1}{N} \sum_{m \in M} f_m N_m = \frac{1}{N} \sum_{i=1}^{N} f_{m(i)}, \tag{11}$$

donde $m(i)$ denota el resultado obtenido en la i-ésima repetición. Nótese que el valor de expectación es lineal, $\langle af + bg \rangle = a\langle f \rangle + b\langle g \rangle$, y que si f es constante entonces $\langle f \rangle = f$. La *varianza* de f se define como

$$\Delta f^2 = \langle (f - \langle f \rangle)^2 \rangle = \langle f^2 \rangle - \langle f \rangle^2, \tag{12}$$

y su raíz cuadrada Δf, que se conoce como *desviación típica*, da una medida de cuánto se aleja f de su valor de expectación en un resultado típico del experimento.

3 Entropía estadística

La *entropía* de una distribución de probabilidad $p : M \to \mathbb{R}$ se define como

$$S(p) = -\sum_{m \in M} p_m \ln p_m. \tag{13}$$

Esta cantidad mide cuán incierto es el resultado del experimento de acuerdo con la distribución p. Para ver eso, mostraremos que S es mínima cuando estamos seguros de cuál va a ser el resultado, y máxima cuando no tenemos ni idea.

S es mínima cuando no hay incertidumbre. Notemos primero que $S(p) \geq 0$ para cualquier distribución de probabilidad p, porque $p_m \leq 1$ y por lo tanto $\ln p_m \leq 0$. Ahora, si q es una distribución sin incertidumbre ($q_{\bar{m}} = 1$, $q_m = 0$ para $m \neq \bar{m}$) se tiene $S(q) = 0$ porque $\ln 1 = 0$ y $\lim_{\epsilon \to 0} \epsilon \ln \epsilon = 0$, y por lo tanto $S(q) \leq S(p)$ para cualquier distribución p, como queríamos demostrar.

S es máxima cuando la incertidumbre es máxima. Para ver esto empezamos introduciendo una nueva cantidad llamada entropía relativa. La *entropía relativa* entre dos distribuciones p y q se define como

$$S(p|q) = -\sum_{m\in M} p_m \ln \frac{q_m}{p_m}. \tag{14}$$

La propiedad crucial de la entropía relativa es que es positiva, $S(p|q) \geq 0$. Esto se ve fácilmente teniendo en cuenta que $\ln x \leq x - 1$,

$$S(p|q) \geq -\sum_{m\in M} p_m \left(\frac{q_m}{p_m} - 1\right) = 0. \tag{15}$$

Ahora, sea q la distribución de máxima incertidumbre, $q_m = 1/|M|$. Para esta distribución tenemos $S(q) = \ln |M|$, y por lo tanto

$$0 \leq S(p|q) = \ln |M| - S(p) = S(q) - S(p), \tag{16}$$

así que $S(q) \geq S(p)$ para cualquier distribución p, como queríamos demostrar (en el problema 16 de la guía 2 van a probar esto mismo por otro método).

Como hemos mencionado más arriba, cuando no tenemos ninguna información que privilegie ciertos resultados del experimento sobre otros, lo natural es asignar igual probabilidad a todos los resultados. Como acabamos de ver, ésta es la distribución que maximiza la entropía. Esto sugiere un criterio general para asignar probabilidades: buscamos la distribución que maximiza la entropía de entre todas las que están de acuerdo con nuestros datos. Por ejemplo, supongamos que sabemos que el valor de expectación de cierta función $A : M \to \mathbb{R}$ toma el valor a. Encontrar la distribución que maximiza la entropía y satisface esta propiedad es un problema de optimización que podemos resolver usando el método de los multiplicadores de Lagrange, es decir extremizando la función

$$F(p; \lambda, \mu) = S(p) - \lambda \left(\sum_{m\in M} p_m - 1\right) - \mu \left(\sum_{m\in M} A_m p_m - a\right). \tag{17}$$

Nótese que ésta es una función de $|M| + 2$ variables, las $|M|$ probabilidades p_m y los dos multiplicadores de Lagrange λ y μ. El primer multiplicador de Lagrange está para imponer la condición de normalización, y el segundo para imponer el vínculo sobre el valor medio de A. Para ver cómo se termina de resolver este problema y otros más generales los emplazo a leer las resoluciones de los problemas 18 y 19.