Las probabilidades de Maxent.

Jorge Soberon, Enero 2012

Introducción

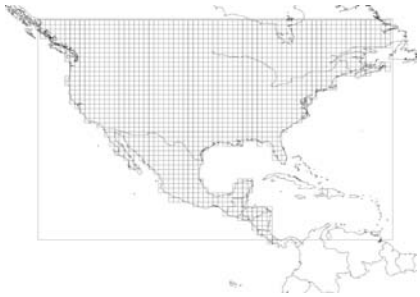
Este breve documento es un resumen de los significados de las dos principales probabilidades que produce el algoritmo Maxent, en la implementación de Phillips: la salida cruda y la salida logística. Al final del documento se incluye una muestra de la forma como se obtiene la ecuación de Gibbs de maxent, por multiplicadores de Lagrange.

La mayor parte de los resultados sobre Maxent se han presentado sobre una doble perspectiva, la geografica (G) y la ambiental (E). En otras palabras, En Maxent la argumentacion a veces esta dada sobre los pixels o celdas de la reticula geografica, y a veces sobre los ambientes. Empezamos con la primera vista:

Vista geografica.

El espacio esta dividido en un reticula de n celdas numeradas consecutivamente x = 1,2,… n, como se ilustra en la figura de abajo. Cada celda tiene un ambiente $z(x)$. La probabilidad que interesa a la gente que quiere estudiar el *area de distribucion* **G**o es:

$$p(Y = 1 | x) \quad . \quad . \quad . \quad (1)$$



Esta es la probabilidad de un ensayo de Bernoulli en el x - pixel. Por ejemplo, si $p(Y = 1| x) = 0.1$, hay un 10% de probabilidad de detectar (¿como? ¿Con que métodos? ¿Durante cuánto tiempo? Todo esto no se dicutira aquí) a la especie en el x - ésimo pixel. Si es 0.9, hay un 90%. Nótese que esta probabilidad no tiene que sumar 1 sobre todos los pixeles. El valor esperado de presencias, sumado

sobre todos los pixeles es la prevalencia π_1 que significa la proporción de pixeles con presencia de la especie.

$$\pi_1 = \sum_{x=1}^{n} p(x) \times p(x \mid Y = 1) \ldots (2)$$

El valor π_1 es la proporción del área total ocupada por la especie. Obviamente, rara vez se conoce este valor. Entonces, *p(Y=1|x)* es una probabilidad definida para la celda *x* y la especie o está presente, o está ausente. O sea, la densidad es sobre solo dos valores en cada celda, como función de la celda.

$$p(Y = 1 \mid x) = 1 - p(Y = 0 \mid x) \ldots (3)$$

Obtener esta probabilidad es el sueño dorado de los que modelan **G**o.
Desafortunadamente hay que aclarar que esta probabilidad NO se puede conocer usando datos de presencias únicas (Phillips & Dudik 2008; Elith et al. 2011; Li et al. 2011), o sin adoptar supuestos simplificadores. Es "indiscernible", si no se cuentan con ausencias estrictas no sesgadas.

¿Que hacer? Uno generalmente no cuenta con datos de ausencias estrictas. Cuando no hay ausencias estrictas, se recurre al truco de utilizar información sobre el "background" para estimar otra probabilidad, bastante distinta pero relacionada a la *p*(*Y=1*|*x*). Aplicando la regla de Bayes a la ecuacion (1):

$$p(Y = 1 \mid x) = \pi_1 \frac{p(x \mid Y = 1)}{p(x)} \ldots (4)$$

Suponiendo un mundo con pixeles equiprobables, *p(x)=1/*|***G***|.
La ecuación (4) expresa resultados fundamentales para los modeladores de nichos o de áreas:

1. La probabilidad de presencia en el *x*-ésimo pixel es una combinación de tres probabilidades: la de presencia no marginal, o prevalencia (π_1), que es la proporción de área |**G**| ocupada por la especie.
2. la probabilidad de visitar el *x*-ésimo pixel, que se supone equiprobable, y *depende del tamaño* de la región |**G**|.
3. y la probabilidad de estar presente en *x*, dado que se cuenta con la información de que la especie esta presente.

La primera probabilidad es en general indiscernible sin datos de ausencias estrictas. La segunda depende de suponer que tanto la especie como el observador visitan de manera equiprobable toda la retícula, lo cual, en general, es falso. La tercera, la cantidad $p(x|Y=1)$, es estimable, y es la que Maxent calcula, como veremos luego, y la expresa como su *raw output.*

La suma de los valores del *raw output* sobre todas las celdas vale 1 y por lo tanto, en retículas grandes, $p(x|Y=1)$ tiene valores muy pequeños. Esta cantidad se debe interpretar como un índice de similitud del ambiente de cada celda con los de aquellas donde se ha observado la especie, y por lo tanto $p(x|Y=1)$, es particularmente útil para los que se interesan en modelar nichos.

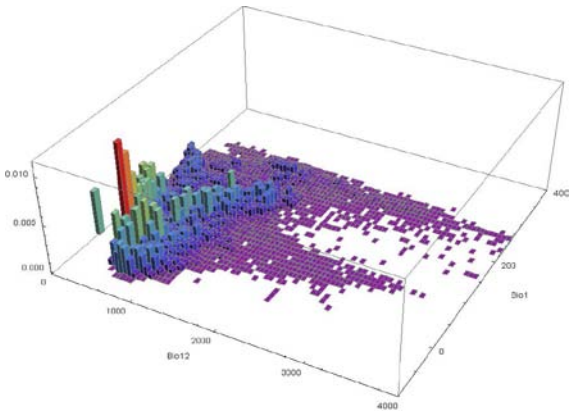
En la siguiente tabla se resumen los puntos anteriores:

$p(Y=1\|x)$	Probabilidad de que la especie este presente dado que se observa la celda *x*. Lo que a los que quieren modelar aéreas ocupadas **G**o les encantaría saber.
$p(x\|Y=1)=f_1(x)$	Probabilidad de estar en la celda *x*, dado que se ha observado a la especie. Es un índice de similitud ambiental respecto a las celdas donde se ha observado a la especie. En otras palabras, relativamente su valor será alto si los ambientes en x se "parecen" a los ambientes donde se ha observado a la especie.
$p(x)$	Probabilidad de que el observador se encuentre en la celda x. Si se muestreara al azar, $p(x)=1/n$. Generalmente no se muestrea al azar, hay toda clase de sesgos y es difícil saber el valor real de $p(x)$.
π_1	"Prevalencia", o la probabilidad no-condicional de observar a la especie. Es la proporción de pixeles o celdas ocupadas, respecto al total de celdas en la retícula.

Si en vez de utilizar la vista geográfica, adoptamos un punto de vista ambiental, empezamos otra vez con el sueño dorado:

$$p(Y=1|\,z)\ \ldots(5)$$

Donde z es la representación de un elemento de volumen en el espacio multidimensional ambiental. En otras palabras, la ecuación anterior se pregunta por cual es la probabilidad de encontrar a la especie en un combinación ambiental *z*. A la derecha se ve una distribución completa de los valores *z* en un espacio de dos variable: precipitación anual y temperatura promedio, para Norteamerica (*sensu lato*). Las combinaciones secas y frias son las mas abundantes.



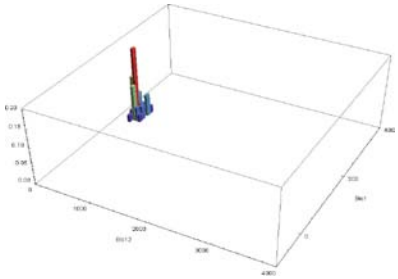
Otra vez, esta probabilidad es "indiscernible" si no se cuenta con información insesgada de ausencias estrictas. Y otra vez, es posible transformarla en algo que se puede estimar:

$$p(Y=1|\,z)=\pi_1\frac{p(z\,|\,Y=1)}{p(z)}\ \ldots(6)$$

Una por una:

1. π_1 es la indiscernible "prevalencia", o sea la proporción del área de referencia **G** donde se encuentra la especie.
2. $p(z)$ es la "densidad" de ambientes en el área de referencia **G**. Se estima directamente de los datos ambientales. Implica, **lo cual es esencialmente importante**, que debe uno especificar la región **G**. Ver figura anterior.

3. $p(z|Y=1)$, que llamaremos de ahora en adelante $f_1(z)$, y que significa la "densidad" de ambientes en las zonas donde la especie ha sido observada, como se ilustra abajo.



Utilizando trucos y supuestos que generalmente el usuario pasa por alto al interpretar (Elith et al. 2011), Maxent tambien intenta estimar la otra probabilidad, $p(Y=1|z)$, que es el *logistic output*.

El *logistic output*.

Sabedores de que muchos usuarios no se quedarían satisfechos con el mero índice de similitud a climas donde la especie ha sido observada, los desarrolladores de Maxent pensaron en el siguiente argumento: imaginemos que se conoce la probabilidad de observar a la especie (¿metodos? ¿tiempo de observación?) en zonas "típicas" para su presencia, y que esa probabilidad se llama τ. Maxent, por default, asume que $\tau = 0.5$. También se define el valor r = la entropía relativa de $f_1(z)$ (que se obtiene de los datos de ocurrencia) respecto a $f(z)$ (que se estima con los puntos de background). Entonces, el logistic output de Maxent proviene de la siguiente ecuación:

$$\hat{p}(Y=1|z) = \frac{\tau e^{\boldsymbol{\beta}\bullet\mathbf{x}-r}}{1-\tau+\tau e^{\boldsymbol{\beta}\bullet\mathbf{x}-r}}$$

El producto vectorial $\boldsymbol{\beta}\bullet\mathbf{x}$ es el resultado de maximizar la entropía sujeta a ciertas restricciones, como se describe abajo. El logistic output de Maxent es esencialmente

un intento de producir una probabilidad sobre cada pixel *x*, o sobre cada combinación ambiental *z*, pero no incluye nueva información sobre la presencia de la especie. Simplemente transforma los valores de $p(x|Y=1)$ a una nueva escala, sin incluir nueva información.

Un ejemplo artificial

Veamos una ilustración numérica de los significados de los términos en la ecuaciones (2) y (6), considerando el siguiente mundo virtual, de 9 celdas y tres tipos de ambientes:

x_1 z=1	x_2 z=1	x_3 z=2
x_4 z=1	x_5 z=1	x_6 z=2
x_7 z=1	x_8 z=2	x_9 z=3

Figura 1. Mundo virtual #1. En las celdas grises el ambiente z=1, en las verdes z=2 y en la amarilla z=3.

Este mundo tiene los siguientes valores y probabilidades:

x	*z*	*p(x)*	$p(Y=1\|x)$ *Logistic Output*	$p(x\|Y=1)$ *Raw Output*
1	1	1/9	0	0
2	1	1/9	0	0
3	2	1/9	½	$(1/2 \times 1/9)/(5/18) = 1/5$
4	1	1/9	0	0
5	1	1/9	0	0

6	2	1/9	½	$(1/2 \times 1/9)/(5/18) = 1/5$
7	1	1/9	0	0
8	2	1/9	½	$(1/2 \times 1/9)/(5/18) = 1/5$
9	3	1/9	1	$(1/9)/(5/18) = 2/5$

Las probabilidades del *Raw Output* y el *Logistic Output* estan relacionadas mediante la ecuación (2). Para estimarlas haría falta conocer los valores de *p(x)* y de π_1, cosa que en la práctica ocurre rarísima vez. El primero se obtuvo de suponer equiprobable la visita a cada celda, y el segundo se obtiene de la esperanza de presencias:

$$\pi_1 = \sum_{x=1}^{9} p(x) \times p(Y = 1 \mid x) = \frac{2.5}{9} = \frac{5}{18}$$

Es fundamental insistir en que la cantidad π_1, (la prevalencia), es indiscernible si no se cuenta con información no sesgada de ausencias estrictas.

Nótese que la especie solamente tiene una probabilidad segura de encontrarse en la novena celda. En las celdas amarillas la probabilidad de observar a la especie (en un periodo de tiempo dado, usando métodos dados) es de ½.

En la práctica, Maxent estima el *raw output* $p(x|Y=1)$ mediante un argumento de máxima entropía que se explica después y que produce una ecuación como la que sigue:

$$p(x \mid Y = 1) = ke^{\beta \cdot v}$$

donde $\boldsymbol{\beta}$ es un vector de parametros ajustados, $\boldsymbol{x}$ otro vector de "features" y *k* una constante de normalizacion (ver Phillips y Dudik, o Elith & Phillips).

En la vista ambiental, la tabla 1 se transforma, acumulando todas las celdas con ambientes idénticos, para obtener:

z	*p(z)*	*p*(*Y=1*|*z*) Logistic Output	*p*(*x*|*Y=1*) Raw Output
1	5/9	0	0
2	3/9	1/2	$(1/2 \times 3/9)/(5/18) = 3/5$
3	1/9	1	$(1 \times 1/9)/(5/18) = 2/5$

Y la prevalencia π_1 se obtiene de la ecuacion (2):

$$\pi_1 = \sum_{z=1}^{3} p(z) \times p(Y=1 \mid z) = \frac{2.5}{9} = \frac{5}{18}$$

Resumen:

- La *raw output* de Maxent es un estimado de la probabilidad *p*(*x*|*Y=1*), o equivalentemente (Elith et al., 2011) de *p*(*z*|*Y=1*), o sea, de la probabilidad de estar presente en el pixel *x* (o el ambiente *z*) dado que sabemos que la especie esta presente. Es un índice de similitud con los pixeles donde se ha observado la especie.
- La *logistic output* de Maxent es una transformación del *raw output* bajo el supuesto de que se conoce el valor de *p*(*Y=1*|*x*) en los pixeles o ambientes típicos.

Obtencion de las ecuaciones de Maxent

Mostraremos ahora como se obtiene la probabilidad de Maxent, para un caso muy simplificado. Se desea estimar una cierta probabilidad sobre las celdas i = 1, 2, … n. Cada una de esas celdas tiene valores ambientales x_i. Se quiere estimar una distribución de probabilidades p_i sobre las celdas. Como es una distribución sobre las celdas, $\sum_i p_i = 1$. Además se cuenta con los datos de las presencias. Esto es, hay una colección de localidades, celdas o puntos k=1, 2,…m donde se ha observado la especie. La media empírica de las variables ambientales en esos puntos es

$$\begin{aligned} \overline{x}_1 &= \frac{1}{m}\sum_{k=1}^{m} x_{k,1} \\ &\vdots \\ \overline{x}_v &= \frac{1}{m}\sum_{k=1}^{m} x_{k,v} \end{aligned}$$

En su caso mas simple, lo que se busca en un *approach* de máxima entropía es encontrar una distribución p_i que tenga medias de los valores ambientales parecidos a los empíricos:

$$\begin{aligned} \hat{x}_1 &= \sum_{i=1}^{n} p_i x_{i,1} \\ &\vdots \qquad\qquad \ldots . (8) \\ \hat{x}_v &= \sum_{i=1}^{n} p_i x_{i,v} \end{aligned}$$

Pero que sea lo mas plana posible (tenga la máxima entropía). La idea, se reduce al siguiente principio: “queremos encontrar una distribución que esté de acuerdo con todo lo que se sabe, pero que no suponga nada que no se sabe”.

La entropía es $H = -\sum_{i=1}^{n} p_i \ln p_i$. Hay que maximizar H sujeta a las restricciones (8) y a que la suma debe ser 1. Esto se hace recurriendo a multiplicadores de Lagrange. Se construye la siguiente función:

$$\Phi = H - \left(\lambda_0 \sum_i p_i + \lambda_1 \sum_i p_i x_{i,1} + \ldots + \lambda_v \sum_i p_i x_{i,v} \right) =$$
$$= \sum_i \left[p_i \ln p_i - \left(\lambda_0 p_i + \lambda_1 p_i x_{i,1} + \ldots + \lambda_v p_i x_{i,v} \right) \right]$$

Cuya diferencial es:

$$d\Phi = \sum_i \frac{\partial}{\partial p_i} \left[p_i \ln p_i - \left(\lambda_0 p_i + \lambda_1 p_i x_{i,1} + \ldots + \lambda_v p_i x_{i,v} \right) \right] dp_i$$
$$= \sum_i \left[1 + \ln p_i - \left(\lambda_0 + \lambda_1 x_{i,1} + \ldots + \lambda_v x_{i,v} \right) \right] dp_i$$

En un extremo de la función Φ cada término en la suma debe igualarse a cero:

$$1 + \ln p_i - \left(\lambda_0 + \lambda_1 x_{i,1} + \ldots + \lambda_v x_{i,v} \right) = 0$$
$$\ln p_i = -\left(1 + \lambda_0 + \lambda_1 x_{i,1} + \ldots + \lambda_v x_{i,v} \right)$$
$$p_i = e^{-(1+\lambda)} e^{-\boldsymbol{\lambda} \cdot \mathbf{x}_i}$$

Y como la suma de las p_i debe ser 1, se obtiene que el valor de $e^{(1+\lambda)} = \sum_i e^{-\boldsymbol{\lambda} \cdot \mathbf{x}_i}$ por lo que se llega al resultado esperado:

$$p_i = \frac{e^{-\boldsymbol{\lambda} \cdot \mathbf{x}_i}}{\sum_i e^{-\boldsymbol{\lambda} \cdot \mathbf{x}_i}}$$

Que es la función de Gibbs que Maxent calcula.

En la práctica, Maxent utiliza una restricción relajada, que toma en cuenta la necesidad de no sobre-ajustar los parámetros (Elith et al. 2011; Phillips & Dudik, 2008), por lo que el esquema real es mas complicado que el que se describe arriba.