Notas TFG

Juan Tornero Lucas

21 de marzo de 2017

1. Marco Teórico de las Cadenas de Markov

En este primer capítulo desarrollaremos toda la teoría que involucra a las Cadenas de Markov. Empezaremos viendo qué son las secuencias y, a continuación, los espacios de probabilidad conocidos como *Procesos*. Además, mediante una versión adaptada del *Teorema de Extensión de Kolmogorov* veremos cómo estos procesos están definidos de forma única. A continuación definiremos los estados de un proceso, para lo cuál necesitaremos un par de resultados relacionados con los procesos estocásticos de martingala.

Finalmente, veremos la definición de una Cadena de Markov y demostraremos que se tratan de *procesos*.

Secuencias

Comenzamos definiendo el conjunto canónico $X = \{0, 1, \dots, m-1\}$ con $m \in \mathbb{N}$ llamado *alfabeto*, cuyos elementos se denominan *símbolos*, los cuales no tienen por qué ser necesariamente números naturales. Si $x \in X^{\mathbb{Z}}$, entonces $x = \cdots x_{-1}x_0x_1 \cdots$ es una secuencia numerable bi-infinita, los índices $i < 0$ denotan el pasado de la secuencia, y los $i \geq 0$ el futuro, particularmente el índice $i = 0$ es el primer símbolo desconocido de la secuencia.

En estos términos definimos una palabra $w \in X^l$ de longitud l, como una l-tupla de X. $\emptyset$ denotará la palabra vacía de longitud 0. Una subsecuencia s es una estructura $s = (w, (a, b))$, donde w es una palabra y $a, b \in \mathbb{Z}$ tq $|w| = b - a + 1$. Así, s puede también escribirse $s = s_a \cdots s_b$, y a representaría el tiempo inicial, y b el tiempo final. Una secuencia x contendrá a la subsecuencia s si $\forall t \in [a, b], s_t = x_t$.

El conjunto $A_s = \{x \in X^{\mathbb{Z}} | x_i = s_i \forall i \in [a, b]\}$ es el conjunto de las secuencias de $X^{\mathbb{Z}}$ que contienen a s. Si por ejemplo $s = (\emptyset, (a, a-1))$, entonces $A_s = X^{\mathbb{Z}}$.

El conjunto X^* denotará el de todas las palabras.

Procesos

Un proceso Q es una espacio de probabilidad estacionaria en un espacio de secuencias.

Una medida de probabilidad es una función que asigna probabilidades a conjuntos (el espacio medible de probabilidades), en este caso a conjuntos de secuencias $\mathbb{X}$, que es la σ-álgebra definida como la menor colección de subconjuntos de $X^{\mathbb{Z}}$ tq:

1. Para toda secuencia $s, A_s \in \mathbb{X}$.
2. $\mathbb{X}$ es cerrado bajo complementos y uniones contables.

El par $(X^{\mathbb{Z}}, \mathbb{X})$ asigna probabilidades a los conjuntos A_s fijados por las subsecuencias s. Definiremos entonces $P(s) = P(A_s)$, tenemos en particular:

$$P(\emptyset) = P(X^{\mathbb{Z}})$$

En nuestra definición de proceso nos referíamos también a ellos como estacionarios. Sea D la función desplazamiento $D : X^{\mathbb{Z}} \to X^{\mathbb{Z}}$, que actúa sobre todo $x \in X^{\mathbb{Z}}$ de manera que $D(x_t) = x_{t+1}$, es decir desplaza el tiempo de origen.

Decimos que P es una medida de probabilidad estacionaria si $\forall A \in \mathbb{X}, P(D(A)) = P(A)$, como D es de hecho un automorfismo sobre $X^{\mathbb{Z}}$, entonces $P(D^{-1}(A)) = P(D^{-1}(D(A))) = P(A)$

Finalmente podemos definir de forma más formal un proceso Q como el espacio de probabilidades estacionario $(X^{\mathbb{Z}}, \mathbb{X}, P)$.

Sea w una palabra y $s = (w, (a, b))$, entonces si P estacionario $P(w) = P(s)$. Además de manera trivial podemos obtener que, si W_l es el conjunto de las palabras de longitud $l > 0$:

$$\sum_{z \in W_l} P(wz) = \sum_{z \in W_l} P(zw) = P(w) \ (1)$$

Además, para $w = \emptyset$:

$$P(\emptyset) = 1 \ (2)$$

El recíproco también es cierto, y cualquier función de X^* que satisfaga (1) y (2) define un proceso.

A continuación, demostraremos la unicidad de los procesos.

Teorema Dado $f : X^* \to [0, 1]$ que satisface:

$$1. f(\emptyset) = 1$$

$$2. \forall w \in X^*, f(w) = \sum_{z \in X^*} f(zw) = \sum_{z \in X^*} f(wz)$$

,

existe un único proceso $Q = (X^{\mathbb{Z}}, \mathbb{X}, P)$ tq $\forall w \in X^*, P(w) = f(w)$.

Kolmogorov

Este resultado deriva del Teorema de Extensión de Kolmogorov (TEK). Usando las notaciones R^n y $R^{\mathbb{N}}$ para referirnos a las σ-álgebras en $\mathbb{R}^n$ y $\mathbb{R}^{\mathbb{N}}$ respectivamente.

Teorema de Extensión de Kolmogorov. Suponemos que nos dan un conjunto de medidas de probabilidad μ_n consitentes en $(\mathbb{R}^n, R^n)$, es decir que cumplen:

$$\mu_{n+1}((a_1, b_1] \times \ldots \times (a_n, b_n]x\mathbb{R}) = \mu_n((a_1, b_1] \times \ldots \times (a_n, b_n]).$$

Entonces existe una única medida de probabilidad P' en $(\mathbb{R}^{\mathbb{N}}, R^{\mathbb{N}})$, con

$$P'(x|x \in (a_i, b_i], i \in \{1, \ldots, n\}) = \mu_n((a_1, b_1] \times \ldots \times (a_n, b_n])).$$

No demostraremos el TEK directamente, sino que lo usaremos para demostrar una variante del mismo que se adapte a nuestras necesidades, teniendo en cuenta las diferencias que existen entre el espacio de probabilidad $(\mathbb{R}^{\mathbb{N}}, R^n, P')$ y el de un proceso estacionario $(X^{\mathbb{Z}}, \mathbb{X}, P)$

1. $\mathbb{R}^{\mathbb{N}}$ es producto de copias de números reales, mientras que $X^{\mathbb{Z}}$ es producto de copias de un conjunto finito X. Para resolver esta diferencia utilizaremos un mapeado inyectivo $g : X \to \mathbb{R}$.

2. Los elementos de $\mathbb{R}^{\mathbb{N}}$ son secuencias semi-infinitas mientras que los de $X^{\mathbb{Z}}$ son, de hecho, secuencias bi-infinitas. Utilizaremos de nuevo un mapeado, esta vez biyectivo $h : \mathbb{N} \to \mathbb{Z}$. Consideraremos los dígitos en el orden $0, 1, -1, 2, -2, \ldots$.

3. Por último, P' no necesita ser estacionario, así que usaremos el TEK para provar que P existe, y luego veremos que es estacionario.

Introducimos primero los mapas g y h.

Para el caso $g : X \to \mathbb{R}$ solo ha de cumplir que sea inyectivo, da igual la imagen de los símbolos $x \in X$.

El mapa biyectivo $h : \mathbb{N} \to \mathbb{Z}$ será

$$h(n) = \begin{cases} n/2 & n\text{par} \\ (1-n)/2 & n\text{impar} \end{cases}$$

Su inversa, por lo tanto:

$$h^{-1}(y) = \begin{cases} 2y & y > 0 \\ 1 - 2y & y \leq 0. \end{cases}$$

Vemos como h envía $1, 2, 3, 4, 5$ a $0, 1, -1, 2, -2$ respectivamente, y viceversa h^{-1}. Definiremos el conjunto $J(n) = \{y \in \mathbb{Z}|h^{-1}(y) \leq\}$, y los valores $d(n)$ y $c(n)$ como los mayores y menores de $J(n)$ respectivamente, caracterizadas de la siguiente manera:

$$c(n) = min(h(n), h(n-1)) = \begin{cases} \frac{2-n}{2} & n \text{ par} \\ \frac{1-n}{2} & n \text{ impar} \end{cases}$$

$$d(n) = max(h(n), h(n-1)) = \begin{cases} \frac{n}{2} & n \text{ par} \\ \frac{n-1}{2} & n \text{ impar} \end{cases}$$

$$J(n) = \{c(n), \ldots, d(n)\}.$$

Observamos que c y d cumplen las siguientes propiedades:

P1. Si n par, entonces $h(n+1) < 0, c(n+1) = c(n) - 1, d(n+1) = d(n)$. P2. Si n impar, entonces $h(n+1) > 0, c(n+1) = c(n), d(n+1) = d(n) + 1$.

Demostración

P1. Si n par, entonces $n = 2 * i$, para algún $i \in \mathbb{N}\backslash 0$, y:

$$h(n+1) \overbrace{=}^{n+1 \text{ impar}} \frac{1-(n+1)}{2} = \frac{-2i}{2} = -i < 0$$

$$\begin{cases} c(n+1) \overbrace{=}^{n+1 \text{ impar}} \frac{1-(2i+1)}{2} = -i \\ c(n) \overbrace{=}^{n \text{ par}} \frac{2-2i}{2} = -i+1 \end{cases}$$

Luego $c(n+1) = c(n) - 1$

$$\begin{cases} d(n) \overbrace{=}^{n \text{ par}} \frac{2i}{2} = i \\ d(n+1) \overbrace{=}^{n+1 \text{ impar}} \frac{(2i+1)-1}{2} = i \end{cases}$$

Luego $d(n+1) = d(n)$.

P2. Si n impar, entonces $n = 2i + 1$, para algún $i \in \mathbb{N}\backslash\{0\}$, y:

$$h(n+1) \overbrace{=}^{n+1 \text{ par}} \frac{2i+2}{2} = i + 1 > 0$$

$$\begin{cases} c(n+1) \overbrace{=}^{n+1 \text{ par}} \frac{2-(2i+2)}{2} = -i \\ c(n) \overbrace{=}^{n \text{ impar}} \frac{1-(2i+1)}{2} = -i \end{cases}$$

Luego $c(n+1) = c(n)$

$$\begin{cases} d(n) \overbrace{=}^{n \text{ impar}} \frac{(2i+1)-1}{2} = i \\ d(n+1) \overbrace{=}^{n+1 \text{ par}} \frac{2i+2}{2} = i+1 \end{cases}$$

Y finalmente $d(n+1) = d(n) + 1$

Necesitaremos también definir una serie de conjuntos de subsecuencias, una función para mapear esos conjuntos y, finalmente, un conjunto producto.

Empezamos definiendo estos conjuntos de subsecuencias:

1. $X^{[a,b]} \subset X^*$ el conjunto de todas las subsecuencias $s = (w, (a, b))$ que empiezan en el tiempo a y terminan en el tiempo b.

2. De forma similar $X^{J(n)} \subset X^*$ denotará el conjunto de todas las subsecuencias $s = (w, (c(n), d(n)))$ que empiezan en $c(n)$ y terminan en $d(n)$. Como $|J(n)| = n$, $X^{J(n)}$ es un producto de n copias de X. La diferencia entre X^n y $X^{J(n)}$ es cómo están etiquetadas las coordenadas, en el primer caso el orden es $1, 2, \ldots, n$ mientras que el segundo lo hace como $c(n), \ldots, d(n)$. Por las propiedades P1 y P2 demostradas anteriormente, tenemos que:

Si n es impar:

$$X^{J(n+1)} = X^{J(n)} \times X$$

si n es par:

$$X^{J(n+1)} = X \times X^{J(n)}$$

Ahora definimos la función H que envía subsecuencias en $X^{J(n)}$ a subsecuencias en $\mathbb{R}^n$, en el caso que $n = \infty$ H enviará secuencias bi-infinitas en $X^{\mathbb{Z}}\$$ a secuencias semi-infinitas en $\mathbb{R}^{\mathbb{N}}$. El objetivos de esta función es reordenar los argumentos de las coordenadas y enviarlas a $\mathbb{R}$. Si $x \in X^{J(n)}$, entonces hay una subsecuencia $v \in \mathbb{R}^n$ que satisface $v_i = g(x_{h(i)}), \forall i \in \{1, \ldots, n\}$. Recordamos que g es una función a la que solo imponemos que sea inyectiva, por lo tanto no es invertible. Definimos $H : X^{J(n)} \to \mathbb{R}^n$ como $H(x) = v$. Si $x = x_{c(n)}, \ldots, x_{d(n)} \in X^{J(n)}$, entonces:

$$H(x_{c(n)} x_{c_n+1} \ldots x_{d(n)}) = g(x_0)g(x_1)\ldots g(x_{c(n)} + d(n) + 1)$$

Esta función puede ser igualmente definida para el caso $x \in X^{\mathbb{Z}}$. Y aunque H no es invertible, porque g no es invertible. Sí que tiene conjuntos inversos. Por ejemplo, si $A \subset \mathbb{R}^n$, entonces:

$$H^{-1}(A) = \{x \in X^{J(n)} | H(x) \in A\}$$

Finalemte, definimos el conjunto producto S MCo el subconjunto de $X^{[a,b]}$ de la siguiente forma $S = S_a \times \ldots \times S_b$, donde $S_i \subset X$. Si $S' \subset X$, entonces $S \times S'$ es un producto contenido en $X^{[a,b+1]}$. De forma natural definimos el cilindro $C(S)$ de la siguiente manera:

$$C(S) = \{x \in X^{\mathbb{Z}} | x_i \in S_i, i \in [a, b]\}$$

De forma parecida a como definíamos A_s, tenemos que $C(S)$ es el conjunto de todas las secuencias de $X^{\mathbb{Z}}$ que coinciden con la subsecuencia en S.

A partir de $C(S)$ definimos también el cilindro desplazado $D(S)$ como:

$$D(S) = \{x \in X^{[a-1,b-1]} | x_i \in S_{i+1}, i + 1 \in [a, b]\}.$$

Lema. Los cilindros C y T verifican:

1. $C(S) = C(S \times X) = C(X \times S)$
2. $D(S \times X) = D(S) \times X = X \times D(S)$
3. $D(C(S)) = C(D(S))$

Demostración.

Sea $S = S_a \times \ldots \times S_b$.

1. Por definición, $\forall i \in \mathbb{Z}, x_i \in X$:

   $$C(S) = \{x \in X^{\mathbb{Z}} | x_i \in S_i, i \in [a,b]\} = \{x \in X^{\mathbb{Z}} | x_{a-1} \in X \text{ y } x_i \in S_i, i \in [a,b]\} = C(X \times S)$$

   De la misma manera:

   $$C(S) = \{x \in X^{\mathbb{Z}} | x_i \in S_i, i \in [a,b]\} = \{x \in X^{\mathbb{Z}} | x_{b+1} \in X \text{ y } x_i \in S_i, i \in [a,b]\} = C(S \times X)$$

2. Tenemos en cuenta $S \times X \subset X^{[a,b+1]}$:

   $$\begin{aligned} D(S \times X) &= \{x \in X^{[a-1,b]} | x_b + 1 \in X \text{ y } x_i \in S_{i+1}, i+1 \in [a,b]\} \\ &= \{x \in X^{[a-1,b-1]} | x_i \in S_{i+1}, i+1 \in [a,b]\} \times X \end{aligned}$$

   De forma semejante a lo visto anteriormente también tenemos:

   $$\begin{aligned} D(S \times X) &= \{x \in X^{[a-2,b-1]} | x_{a-2} \in X \text{ y } x_i \in S_{i+1}, i+1 \in [a,b]\} \\ &= X \times \{x \in X^{[a-1,b-1]} | x_i \in S_{i+1}, i+1 \in [a,b]\} \end{aligned}$$

3. $$\begin{aligned} D(C(S)) &= \{x \in X^{\mathbb{Z}-1} | x_i \in S_{i+1}, \forall i+1 \in [a,b] \text{ y } x_j \in X, \forall j \notin [a,b]\} \\ &= \{x \in X^{\mathbb{Z}} | x_i \in S_{i+1}, \forall i+1 \in [a,b]\} = C(D(S)) \end{aligned}$$

   cqd

Corolario. Sea $S = S_{c(n)} \times \ldots \times S_{d(n)} \subset X^{J(n)}$ un conjunto producto, y sea

$$A = C(S) = \{x \in X^{\mathbb{Z}} | x_i \in S_i, i \in [c(n), d(n)]\}$$

Entonces tenemos

$$D(S) = \{x \in X^{[c(n)-1, d(n)-1]} | x_i \in S_{i+1}, i+1 \in [c(n), d(n)]\}$$

y

$$D(A) = \{x \in X^{\mathbb{Z}} | x_i \in S_{i+1}, i+1 \in [c(n), d(n)]\}$$

Además $D(A) = C(D(A))$ y, por el lema $D(A) = C(D(S) \times X)$.

Finalmente, observamos que para cada $S \in X^{[a,b]}$ existe un conjunto $S' \in X^{J(n)}$, donde $n = min(-2a, 2b+1)$, tal que $C(S) = C(S')$. Como cada conjunto cilindrico puede ser escrito

como $C(S)$ para algún $S \in X^{[a,b]}$, esto significa que cada conjunto cilíndrico puede ser escrito como $C(S')$ para algún $S' \in X^{J(n)}$.

Finalmente, ya estamos en disposición de demostrar nuestra variante del TEK.

Teorema. Suponemos que tenemos una secuencia de medidas v_n en $X^{J(n)}$ que satisfacen que $\forall n$ y para todo $S \subset X^{(J_n)}$, si n es impar,

$$v_n(S) = v_{n+1}(S \times X) \quad \text{(T.1)}$$

y si n es par,

$$v_n(S) = v_{n+1}(X \times S) \quad \text{(T.2)} v_n(S) = v_{n+1}(D(S) \times X), \quad \text{(T.3)}$$

Entonce existe un único proceso estacionario $Q = (X^{\mathbb{Z}}, \mathbb{X}, P)$ tal que, $\forall n, S \subset X^{J_n}$,

$$P(C(S)) = v_n(S)$$

La demostración la haremos en dos partes: primero demostraremos que P existe y, posteriormente comprobaremos que P es estacionario.

Demostración. Definimos una secuencia de medidas μ_n rn $\mathbb{R}^{\mathbb{N}}$ de la siguiente manera: Si $A \subset \mathbb{R}^{\mathbb{N}}$, entonces definimos $\mu_n(A) = v_n(H^{-1}(A)) = v_n\{x \in X^{J(n)} | H(x) \in A\}$, por la definición de anticonjuntos que establece H.

La consistencia de estas μ's se puede comprobar:

$$\mu_{n+1}(A \times \mathbb{R}) = v_{n+1}(H^{-1}(A \times \mathbb{R}))$$

Dado que

$$H^{-1}(A \times \mathbb{R}) = \begin{cases} H^{-1}(A) \times X & n\text{impar} \\ X \times H^{-1}(A) & n\text{par} \end{cases}$$

Obtenemos para μ_{n+1}

$$\mu_{n+1}(A \times \mathbb{R}) = \begin{cases} v_{n+1}(H^{-1}(A) \times X) & n \text{ impar} \\ v_{n+1}(X \times H^{-1}(A)) & n \text{ par} \end{cases}$$

Ahora, aplicando T.1 y T.2 a la derecha de las igualdades

$$\mu_{n+1}(A \times \mathbb{R}) = v_n(H^{-1}(A)) = \mu_n(A).$$

Aplicando TEK sobre las μ_n's, obtenemos una única P' en $(\mathbb{R}^{\mathbb{N}}, R^{\mathbb{N}})$ que concuerda con μ_n. Por lo tanto, si A_i es una σ-álgebra $\forall i \in \{1, \ldots, n\}$ y $A = A_1 \times \ldots \times A_n$, entonces

$$P'(x | x_i \in A_i, i \in \{1, \ldots, n\}).$$

Ahora pasamos a demostrar la segunda parte del Teorema. Definimos P tq $\forall S \in \mathbb{X}$,

$$P(S) = P'(H(x)|x \in S).$$

Probamos primero que P existe. Sea $S \in X^{J(n)}$, tenemos

$$v_n(S) = \mu_n(H(S)) = P'\Big(a \in \mathbb{R}^{\mathbb{N}} | a_i \dots a_n \in H(S)) = P(x \in X^{\mathbb{Z}} | x_{c(n)} \dots x_{d(n)} \in S) = P(C(S))\Big)$$

Por lo tanto, esta P es en realidad una extensión de las μ_n's y existe.

Para ver que P es estacionaria, necesitamos demostrar que $P(D(A)) = P(A)$ para cualquier $A \in \mathbb{X}$ que contenga al cilindro $C(S)$. Llamaremos a esta colección Θ

$$\Theta = \{C(S) | S \subset X^{J(n)} \text{ para algún } n\}.$$

Aunque cada $A \in \Theta$ pueden tener asociado más de un par n, S siempre escogeremos el de menor n -que tiene asociado una única S-.

Si n es par, T.3 y el lema demostrado anteriormente nos da

$$\begin{aligned} P(A) = v_n(S) = v_{n+1}(D(S) \times X) = P(C(D(S)) \times X) \\ = P(C(D(S))) = P(D(A)) \end{aligned}$$

Y por último, vemos que ocurre lo mismo con n impar

$$\begin{aligned} P(A) = v_n(S) = v_{n+1}(S \times X) = v_{n+1}(D(S \times X) \times X) = P(C(D(S \times X) \times X)) \\ = P(C(D(S \times X) \times X)) = P(C(D(S) \times X \times X)) = P(C(D(S) \times X) \times X) \\ = P(D(A) \times X) = P(D(A)) \end{aligned}$$

Por lo tanto, $P(A) = P(D(A))$. *cqd*

Finalmente estamos en disposición de demostrar el nuestro nuestro teorema de partida.

Teorema Dado $f : X^* \to [0, 1]$ que satisface:

$$1. f(\emptyset) = 1$$

$$2. \forall w \in X^*, f(w) = \sum_{z \in X^*} f(zw) = \sum_{z \in X^*} f(wz)$$

,

existe un único proceso $Q = (X^{\mathbb{Z}}, \mathbb{X}, P)$ tq $\forall w \in X^*, P(w) = f(w)$.

Demostración.

*Ejemplo de Proceso.*Para ilustrar el concepto de proceso, pondremos un pequeño ejemplo basado en un dado 'ideal' de seis caras, $X = \{1, 2, 3, 4, 5, 6\}$ y cualquier palabra de longitud l tendrá asociada una probabilidad de $P(w) = \frac{1}{6}^l$. En particular, si la palabra es de longitud $l = 0$, entonces $P(w) = \frac{1}{6}^0 = 1$. Si además tenemos otra palabra $z \in X$ entonces, a partir de (1) vemos que $f(wz) = f(zw) = \frac{1}{6}^l (\sum_{i=1}^{6} \frac{1}{6}) = \frac{1}{6} = f(w)$. Por lo tanto cumple tanto (1) como (2) y es un proceso. *cqd*

Estados de un Proceso

Supongamos que tenemos un proceso $Q = (X^{\mathbb{Z}}, \mathbb{X}, P)$ del que conocemos algunos símbolos recientes. Si $s = (w, (a, b))$ es una subsecuencia conocida, con $b = -1$, s induce una distribución condicional en un espacio futuro (DCF) del tipo $P(\cdot|s) = P(\cdot|A_s)$. Esta DCF condensa todo nuestro conocimiento sobre las posibilidades futuras del proceso Q, por lo que al concer esta distribución, en cierto sentido, podemos olvidar todo lo acontecido en el proceso anteriormente. Esta es la base con la que decimos que $P(\cdot|s)$ es un estado.

Ahora ampliaremos esta noción. Primero definimos $\mathbb{X}^+ = \{x^+ = x_0x_1, \ldots | \forall i \geq 0, x_i \in X\}$, y respectivamente $\mathbb{X}^- = \{x^+ = \ldots x_-2, x_-1 | \forall i \leq 0, x_i \in X\}$. Los campos σ generados por $\mathbb{X}^+$ y $\mathbb{X}^-$ son $\mathbb{F}$ y $\mathbb{H}$, respectivamente. Las DCF están condicionadas por secuencias $s = (w, (a, b))$ con w palabra y $w \in \mathbb{X}^-$, por lo que estas subsecuencias pueden o no acabar en la última observación conocida (también conocidas como *historias*), pero nos centraremos únicamente en este último caso.

Si dos historias, z e y acaban determinando la misma DCF $P(\cdot|z) = P(\cdot|y)$, entonces decimos que $y \sim z$ son equivalentes, es decir podemos definir la clase de equivalencia $\sim$ y $\overline{z} = \{y \in X^* | y \sim z\}$.

Sea π la función proyección sobre las clases de equivalencia de $\mathbb{X}^-$, es decir $\pi(z) = \overline{z}$. Para toda palabra futura w se cumplirá entonces $P(s|z) = P(s|\overline{z})$.

Utilizando las clases de equivalencia podemos olvidarnos de la historia concreta que ha generado la DCF. Así, imaginemos la historia w, cuya clase de equivalencia $\overline{w}$ induce una DCF sobre x_0. Así, si escogemos un símbolo k adecuado para x_0, tenemos la nueva palabra $z = wk$, y tenemos entonces de nuevo una clase de equivalencia $\overline{z}$ que inducirá una DCF sobre x_1, a las diferentes clases de equivalencia $\overline{z}$ que se obtienen a partir de las DCF inducidas por $\overline{w}$ las llamamos *estados del proceso.*

Usando el ejemplo del dado, en este caso todas las historias son equivalentes, luego hay un solo estado del proceso, el cual es visitado recurrentemente. Daremos a continuación una defineión más formal de los estados de un proceso.

Sea $s_l = x_{-l} \ldots x_{-1}$, con $l > 0$ y $R_l = \{s_l | P(s_l) > 0\}$. Dada la historia s_l, para cualquier evento futuro A:

$$P(A|s_l) = \frac{P(A \cap A_{s_l})}{P(A_{s_l})},$$

En concreto, si $x^- = \lim_{l\to\infty} s_l$, entonces tenemos

$$P(A|x^-) = \lim_{l\to-\infty} P(A|s_l) \quad (1)$$

Si definimos:

$$1_A = \begin{cases} 1 & \text{if si} x \in A \\ 0 & \text{en caso contrario,} \end{cases}$$

$P(A|w_l)$ es la esperanza condicional $E(1_A)|\mathbb{H})(x)$ para cualquier secuencia x que coincide con w. Por lo tanto podemos escribir (1) como $\lim_{l\to\infty} E(1_A|\mathbb{H})$. Este límite converge casi seguramente a $E(1_A|H)(x)$, mediante el Teorema de convergencia de Martingale, el cual demostramos al final del capítulo, para no entorpecer la lectura. De esta manera tendremos que (1)=$P(\cdot|x^-)$.

Si x^- tiene una historia s_l con probabilidad nula, entonces $P(A|x^-) = 0$. Denominaremos al conjunto de historias con probabilidad nula $N = \{x^- \in X^{\mathbb{Z}} | P(x^-) = 0\}$.

Un estado del proceso es un DCF que se da al condicionar en una historia de probabilidad no nula. El conjunto de todos los estados del proceso para un proceso dado es:

$$PS = \{P(\cdot|w)|w \in R\} \cup \{P(\cdot|x^-) \in \mathbb{X}^-\backslash N)\}$$

Además, definimos la función $G : \mathbb{X}^- \to PS$ que envía historias no nulas al DCF que inducen, $G(x^-) = P(\cdot|x^-)$. El conjunto inducido de historias de un estado del proceso A es un conjunto que particuliza G mediante $G^{-1}(A)$, y que resulta en todas las historias que induce A.

Proposición Para cualquier estado del proceso A, $G^{-1}(A) \times \mathbb{X}^+$ es la intersección de un conjunto medible y de un conjunto de medida completa (complementario de medida nula)

Usaremos l para las longitudes de las historias y k para longitudes futuras. Además diremos que $A(w)$ s la probabilidad que asigna A a la palabra w.

Además, diremos que dos estados del proceso B y c son $k - equivalentes$ si, para todas palabras w tq $|w| \leq k$, B(w)=C(w)

Demostración Para cada natural l y palabra w, y para cada historia x^-, definimos:

$$f_l^w(x-) = P(w|x_{-l} \dots x_{-1})$$

y

$$f^w(x^-) = P(w|x^-),$$

Por el *Teorema de Convergencia de Martingalas* $\lim_{l\to\infty} f_l^w(x-) = f^w(x^-)$ *casi seguramente.* Como f_l^w es medible, su límite también lo será.

Sea A un estado del proceso y $J_k(A)$ el conjunto de historias x^- que inducen estados del proceso $G(x^-)$ que son k-equivalentes a A, $J_k(A) = \{x^- \text{ tq } \forall w \text{ tq } |w| \leq k, P(w|x^-) = A(w)\}$. Es decir:

$$J_k(A) = \bigcap_{w:|w|=k} \{x^- \text{ tq } P(w|x^-) = A(w)\} = \bigcap_{w:|w|=k} (f^w)^-1(A(w)).$$

$\{A(w)\}$ es un conjunto medible y f^w es una función medible, así que su la antiimagen $(f^w)^-1(A)$ ha de ser medible. Por lo que $J_k(A)$ es intersección finita de conjuntos medibles y por lo tanto medible. Sea $J(A) = \bigcap_{k=1}^{\infty} J_k(A)$, $J(A)$ es intersección contable de conjuntos medibles, por lo que es medible. Ahora, cada historia $J(A) \cap (X^{\mathbb{Z}} \backslash N)$ induce un estado A, con $X^{\mathbb{Z}} \backslash N$ un conjunto demedida completa, y no hay ninguna historia que no esté en $J(A)$ que induzca A. Por lo tanto $J(A) = \bigcap_{k=1}^{\infty} J_k(A) = G^{-1}(A)$. *cqd*

La idea central es que un estado es una predicción basada en lo sabido sobre el pasado del proceso. En la siguiente sección clasificaremos los diferentes tipos de estados que podemos encontrar.

Martingales

Tipos de Estado

La definición del conjunto de estados de un proceso nos indica que estos están inducidos por historias, ya sean finitas o semi-infinitas.

Definimos antes de describir la tipología de estados, y algunas de sus propiedades, una propiedad asociada a los estados. Consideramos por ejemplo un estado del proceso $A \in PS$, y consideramos $G^{-1}(A)$, el conjunto de historias que induce A. Podemos asignar a este conjunto una probabilidad $P(A) = P(G^{-1}(A))$. En particulas, ante la historia w, $P(A)$ será la probabilidad de que s induzca el estado del proceso A.

Un estado del proceso está *infinitamente precedido* si la historia w que lo genera es de longitud infinita, $|w| = \infty$ y $P(w) > 0$.

Un estado del proceso es *alcanzable* si al menos existe una historia, de longitud cualquiera, w con $P(w) > 0$ que induce este estado.

Un estado del proceso $A \in PS$ es recurrente si $P(A) > 0$.

Proposición.

1. Cada estado del proceso inalcanzable está infinitamente precedido.
2. Cada estado del proceso recurrente está infinitamente precedido.

Demostración.

1. Por definición de estado de un proceso, ya que cada uno de ellos está inducido por alguna historia.

2. Sea A un estado recurrente. Entonces $G^{-1}(A)$ tiene probabilidad no nula y por lo tanto no es vacío. Como $G^{-1}(A)$ contiene solo historias con probabilidad no nula, entonces contiene alguna historia que induce A de probabilidad no nula. *cqd*

A partir de esta proposición podemos asegurar que tres tipos de estado no ocurren:

- Estados inalcanzables que no están infinitamente precedidos.
- Estado alcanzables recurrentes que no están infinitamente precedidos.
- Estados inalcanzables que no son recurrentes ni infinitmanete precedidos.

Nos quedan por definir dos tipos de estado.

Un estado del proceso es *transitivo* si es alcanzable pero no recurrente, y si además no está infinitamente precedido se dice que *estrictemente transitivo*.

Un estado del proceso es *elusivo* si es inalcanzable y no recurrente. Este tipo de estado con frecuencia puede ser ignorado, ya que únicamente pueden ser inducidos por eventos de probabilidad nula, y por lo tanto un conjunto medible de estados elusivos pueden obviarse sin cambiar la medida de probabilidad del proceso. Sin embargo, hay procesos cuyos estados elusivos vienen inducidos por los estados recurrentes, o procesos en los que estos estados elusivos son incontables y tienen probabilidad positiva. En estos casos tenemos que tenerlos en cuenta.

Incluir tabla de estados

Ahora imaginemos que hay dos observadores: el observador 1 conoce menos información histórica que el observador 2. Puede ocurrir que esta diferencia de conocimiento permita al observador 2 conocer mejor el futuro del proceso, pero también puede ocurrir que esta información extra sea irrelevante. Si eso ocurre decimos que estamos *sincronizados*.

Una palabra w es una *palabra sincronizante* si toda historia de probabilidad no nula que termina en w induce el mismo estado del proceso.

Proposición. Si w es una palabra sincronizante, entonces induce un estado del proceso que es alcanzable y recurrente.

Demostración. Sea A este estado del proceso inducido por w. Como A es inducido por una historia con probabilidad no nula, entonces es alcanzable. w es en sí misma una historia que termina en w, y A es el estado del proceso inducido por todas las historias que terminan en w. Además, como w es una palabra no nula, y en conjunto de historias que contienen w es un subconjunto de $G^{-1}(A)$, tenemos:

$$P(A) \geq P(\text{historias que terminan en w}) \geq P(w) > 0.$$

Por lo tanto A es recurrente. *cqd*

Cadenas de Markov.

Una *Cadena de Markov* de n-estados es un triplete (S, A, π), donde S es un conjunto de orden n, A es una matriz cuadrada de dimensión n, y π es un vector de longitud n. Y cumplen las siguientes propiedades:

1. Cada columna de A tiene suma 1.
2. $\sum_i \pi_i = 1$.
3. $\pi A = \pi$.

Los elementos de S son los estados, A es la matriz de transición y π es la distribución estacionaria sobre los estados S.

Si $\mathbb{S}$ es el campo definido por el conjunto de los subconjuntos de $S^{\mathbb{Z}}$, entonces $(\mathbb{S}^{\mathbb{Z}}, S)$ será nuestro espacio medible, y definimos nuestra medida de probabilidades P de manera que si $v = v_0 v_1 \dots v_{l-1}$, con $v_i \in S$,

$$P(v) = \pi_{v_0} a_{v_0 v_1} \dots a_{v_{l-2} v_{l-1}},$$

Definiendo $P(\emptyset) = 0$, vemos que una MC es un proceso estacionario.

Recordamos que un proceso es estacionario si cumple:

$$\sum_{z \in S} P(wz) = \sum_{z \in S} P(zw) = P(w) \ (1)$$
$$P(\emptyset) = 1 \ (2),$$

La condición (2) se cumple trivialmente. Para ver la (1) sea $z \in S$, entonces

$$P(\emptyset z) = P(z\emptyset) = P(z) = \pi_z.$$

Por tanto, aplicando la propiedad 2 de las MC

$$\sum_{z \in S} P(z\emptyset) = \sum_{z \in S} P(\emptyset z) = \sum_{z \in S} \pi_z \overset{P2MC}{=} 1 = P(\emptyset)$$

Ahora sea la probabilidad de la unión de v y z, y teniendo en cuenta la primera propiedad de las MC

$$P(vz) = P(v) a_{v_{l-1} v} \to \sum z \in S P(vz) = P(v) \sum_{z \in S} a_{v_{l-1} z} \overset{P1MC}{=} P(v) \cdot 1 = P(v)$$

En el otro sentido, la unión de z con v nos da

$$P(zv) = \pi_z a_{z v_0} \dots a_{v_{l-2} v_{l-1}}$$

Por lo tanto

$$\sum_{z\in S} P(zv) = \left(\sum_{z\in S} \pi_z a_{zv_0}\right) a_{zv_0} \dots a_{v_{l-2}v_{l-1}}$$

Finalmente, aplicando la tercera propiedad de las MC, si tenemos en cuenta que $\sum_{z\in S} \pi_z a_{zv_0}$ es el elemento v_0 del vector πA

$$\sum_{z\in S} \pi_z a_{zv_0} \overbrace{=}^{P3MC} \pi_{v_0}$$

Finalmente vemos que

$$\sum_{z\in S} P(zv) = \pi_{v_0} a_{v_0v_1} \dots a_{v_{l-2}v_{l-1}} = P(v)$$

concluimos que (S, A, π) define un proceso en $(S^{\mathbb{Z}}, \mathbb{S}, P)$, *cqd*

Nota: hasta ahora hemos utilizado una notación particular referida a la teoría de secuencias y procesos, a partir de ahora introduciremos notación estándard de estadística para términos que hasta ahora denominábamos *palabras*, *estados del proceso...*

Las MC's pueden considerarse procesos "sin memoria", es decir, que la distribución de probabilidad futra para $X = (X_1, \dots, X_n)$ secuencia de variables aleatorias, depende únicamente del valor actual, más formalmente:

$$P(X_{n+1} = x_{n+1} | X_n = x_n \dots X_1 = x_1) = P(X_{n+1} = x_{n+1} | X_n = x_n)$$

Esta identidad es la denominada *propiedad de Márkov*.

2. Modelos Ocultos de Markov (HMM)

Introducción

En los Modelos ocultos de Markov (HMM) la MC subyace tras las observaciones. Los estados del HMM solo pueden ser inferidos a partir de los símbolos registrados. Correlacionando las observaciones y las transiciones de estado lo que se busca es encontrar el estado de secuencias más probable.

Un HMM pertenece a los mecanismos de *Machine Learning*, puede ser entendido como una especie de doble proceso estocástico:

- El primer proceso estocástico es un conjunto finito de estados, cada uno de ellos generalmente asociado a una distribución de probabilidad multidimensional. Mediante la denominada matriz de transición se controla las transiciones entre estados.
- El segundo proceso estocástico es aquel en que cualquier estado puede ser observado, es decir, analizaremos lo observado sin ver en qué estado está ocurriendo, de aquí el epíteto de *oculto* que define a este modelo.

Componentes de un HMM

Para definir completamente un HMM, se necesitan cinco elementos:

1. Los N estados del modelo $S = \{S_1, \ldots, S_n\}$
2. Las M observaciones de los diferentes símbolos por estado $v = \{V_1, \ldots, V_M\}$. Si las observaciones son contínuas, obviamente M es infinito.
3. La matriz de transición $A = \{a_{ij}\}$, donde $a_{ij} = P(q_{t+1} = j | q_t = i)$, siendo q_t el estado actual. Esta matriz es equivalente a la vista en la definición de las MC. Hay que obserar que si uno de los a_{ij} es definido cero, permanecerá cero durante todo el proceso de entrenamiento.
4. La probabilidad de distribución de observación de los símbolos en cada estado, $B = \{b_j(k)\}$, donde $b_j(k)$ es la probabilidad de observar el símbolo v_k en el estado S_j.

$$b_j(k) = P(o_t = v_k | q_t = j), \forall j \in \{1, \ldots, N\}, \forall k \in \{1, \ldots, M\}$$

Donde v_k denota el k-ésimo símbolo observado en el alfabeto, y o_t el actual vector de parámetros.

Se deben satisfacer ciertas restricciones:

$$b_j(k) \geq 0, \forall j \in \{1, \ldots, N\}, \forall k \in \{1, \ldots, M\} \text{ y además } \sum_{k=1}^{M} b_j(k) = 1, \forall j \in \{1, \ldots, N\}$$

En el caso contínuo tendremos una función de densidad contínua, en vez de un conjunto de probabilidades discretas. En este caso lo que especificamos es el conjunto de parámetros de la función de densidad, aproximada como una suma ponderada de M distribuciones Gaussianas (GHMM),

$$b_j(o_t) = \sum_{m=1}^{M} c_{jm} N(\mu_{jm}, \Sigma_{jm}, o_t)$$

donde c_{jm} es son los pesos, μ_{jm} es el vector de medias, y Σ_{jm} las matrices de covarianza. Observar que c_{jm} debe satisfacer las condiciones estocásticas $c_{jm} \geq 0, \forall m \in \{1, \ldots, M\}$ y

$$\sum_{m=1}^{M} c_{jm} = 1, \forall j \in \{1, \ldots, N\}$$

5. La distribución inicial de estados $\pi = \{\pi_i\}$, donde π_i es la probabilidad de que el modelo está en el estado S_i en el tiempo inicial $t = 0$, con

$$\pi_i = P(q_1 = i), \forall i \in \{1, \ldots, N\}$$

Para denotar los parámetros de un HMM con frecuencia se usa:

$$\lambda = (A, B, \pi)$$

Para denotar distribuciones discretas, o bien:

$$\lambda = (A, c_{jm}, \mu_{jm}, \Sigma_{jm}, \pi)$$

Cuando se trata de distribuciones contínuas asociadas a funciones de densidad.

Arquitectura de los HMM

????

Algoritmos Relacionados con los HMM's

El objetivo de un HMM, a grosso modo, es aprender a clasificar los datos proporcionados. Es decir, un HMM tiene que ser capaz de discernir las diferencias de comportamiento que posean los diferentes estados del *stream* de observaciones con el que lo alimentamos.

En la historia de los HMM han destacado los estudios de tres problemas:

1. Problema de Evaluación

Dada la secuencia de observaciones $O = \{o_1, \ldots, o_m\}$, ¿Cuál es la probabilidad de que O haya sido generada por el modelo $P(O|\lambda)$? Dado un λ.

2. Problema de Decodificación

Dada la secuencia de observaciones $O = \{o_1, \ldots, o_m\}$, ¿Cuál es la secuencia de estados más problemas dado el modelo λ?

3. Problema de Aprendizaje

Dada la secuencia de observaciones $O = \{o_1, \ldots, o_m\}$, ¿Cómo debemos ajustar los parámetros de λ para maximiazar $P(O|\lambda)$?

El problema de evaluación es la piedra angular en muchos estudios de reconocimiento de voz. El problema de decodificación resulta útil a la hora de segmentar, y el problema de aprendizaje debe ser resuelto si queremos entrenar un HMM para su uso en labores de reconocimiento.