Notas TFG

Juan Tornero Lucas

21 de marzo de 2017

1. Marco Teórico de las Cadenas de Markov

Secuencias

Comenzamos definiendo el conjunto canónico $X = \{0, 1, \ldots, m-1\}$ 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 infinita.Donde 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. ϕ 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 $s = (\phi, (a, a-1))$, entonces $A_s = X^{\mathbb{Z}}$.

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

Procesos

Un proceso Q es una medida 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. Sea $\mathbb{X}$ 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 $(X^{\mathbb{Z}}, \mathbb{X})$ es el espacio medible que asigna probabilidades a los conjuntos A_s fijados por las subsecuencias s. Definiremos entonces $P(s) = P(A_s)$, tenemos en particular:

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

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 $\forall A \in X, P(D(A)) = P(A)$, 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 $l > 0$:

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

Además, para $w = \phi$:

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

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

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

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

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

,

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

DEMOSTRACIÓN PENDIENTE

*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 $\mathbb{X}^+ = \{x^+ = x_0 x_1, \ldots | \forall i \geq 0, x_i \in X\}$, y respectivamente $\mathbb{X}^- = \{x^+ = \ldots x\_2, x\_1 | \forall i \leq 0, x_i \in X\}$. 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 defínción 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

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

Para demostrarlo, usamos el teorema de convergencia de martingale *DEMOSTRACIÓN PENDIENTE*

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 $k-equialentes$ 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 Martingale* $\lim_{l\to\infty} f_l^w(x-) = f^w(x^-)$ *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.

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

Recordamos que un proceso es estacionario si cumple:

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

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

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

Por tanto, aplicando la propiedad 2 de las MC

$$\sum_{z \in S} P(z\phi) = \sum_{z \in S} P(\phi z) = \sum_{z \in S} \pi_z \overbrace{=}^{P2MC} 1 = P(\phi)$$

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

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

En el otro sentido, la unión de z con v no da

$$P(zv) = \pi_z a_{zv_0} \dots a_{v_{l-2}v_{l-1}}$$

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

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