2.1. Modelos ocultos de Markov (MOM)

Un modelo oculto de Markov es un modelo estocástico que es utilizado para modelar fenómenos aleatorios variantes en el tiempo. El modelo de Markov es utilizado para modelar la evolución de fenómenos aleatorios denotados por estados discretos como una función de tiempo, donde la transición de un estado a otro es aleatoria. Supongamos que un sistema puede estar en uno de S estados diferentes, y que en cada instante del tiempo discreto este se puede mover a otro estado de manera aleatoria, con una probabilidad de transición en el tiempo t

dependiente solamente sobre el estado del sistema en el tiempo t (ver Figura 2.1). En la Figura 2.1 se muestran tres estados, desde el estado 1, cualquier transición es posible, se observan cuáles son las probabilidades de pasar de un estado a otro. Tenemos que q_t denota un estado en el tiempo t, y donde q_t puede tomar uno de los valores $Q = \{q_1, q_2, \ldots, q_S\}$; el estado inicial se selecciona de acuerdo a una probabilidad π_i, esto es $\pi_i = P(q_1 = i)$, donde $i = 1, 2, \ldots, S$ [25, 55].

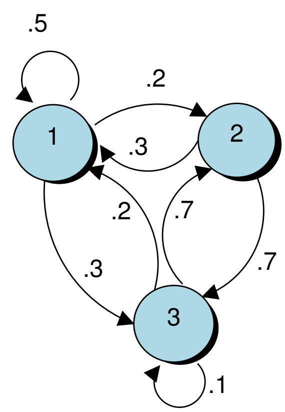

Figura 2.1: Diagrama de estados probabilísticos: modelo de Markov simple

De acuerdo a la descripción anterior, la probabilidad de transición depende solo del estado actual: $P(q_{t+1} = j|q_t = i, q_{t-1} = k, q_{t-2} = l, \ldots) = P(q_{t+1} = j|q_t = i)$. Esta estructura de probabilidades es llamada *propiedad de Markov*, y la secuencia aleatoria de estados $\{q_0, q_1, q_2, \ldots\}$, es llamada *secuencia de Markov o cadena de Markov*. Esta secuencia es la salida del modelo de Markov. Podemos determinar la probabilidad de llegar al siguiente estado sumando todas las probabilidades de los caminos para llegar ahí: $P(q_{t+1} = j) = P(q_{t+1} = j|q_t = 1)P(q_t = 1) + P(q_{t+1} = j|q_t = 2)P(q_2 = 2) + \cdots + P(q_{t+1} = j|q_t = S)P(q_t = S)$. Este cálculo puede ser hecho convenientemente en una matriz, por ejemplo

$$p_t = \begin{bmatrix} P(q_t = 1) & P(q_t = 2) & \cdots & P(q_t = S) \end{bmatrix}.$$

Así tenemos el vector de probabilidades de cada estado, y se tiene la matriz A conteniendo las probabilidades de transición

$$A = \begin{bmatrix} P(1|1) & P(2|1) & \cdots & P(q|1) \\ P(1|2) & P(2|2) & \cdots & P(q|2) \\ . & & & \\ . & & & \\ . & & & \\ P(1|q) & P(2|q) & \cdots & P(q|q) \end{bmatrix},$$

donde $P(j|i)$ es una abreviación de $P(q_{t+1} = j|q_t = i) \mathrel{\hat{=}} a_{ij}$. Por ejemplo para la Figura 2.1, la matriz de probabilidades de transición sería:

$$A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0 & 0.7 \\ 0.2 & 0.7 & 0.1 \end{bmatrix}.$$

Una asignación de probabilidad de estado estacionario es aquella que no cambia de un instante de tiempo al siguiente, de esta manera la probabilidad debe satisfacer la ecuación $p = pA$. Como una ley de probabilidad total, cada renglón de la matriz A debe sumar 1.

La idea detrás de un MOM puede ser ilustrada utilizando los problemas de urnas de probabilidad elemental (ver Figura 2.2). Supongamos que tenemos S diferentes urnas y donde cada una contiene sus propios conjuntos de bolas de colores. En cada instante de tiempo, una urna es seleccionada aleatoriamente de acuerdo al estado en el que esta se encontraba en el instante de tiempo previo (es decir, de acuerdo a un modelo de Markov). De esta manera, una bola se extrae aleatoriamente de una urna seleccionada en el tiempo t. La bola es lo que observamos como salida, y el estado actual es oculto.

En los MOM no se observa directamente el estado, en lugar de eso, cada estado tiene una distribución de probabilidad asociada. Cuando un MOM se mueve del estado q_t en el tiempo t, la salida observada o_t es un resultado de la variable aleatoria O_t que es seleccionada de acuerdo a la distribución $b(o_t|q_t = q)$, que se representa usando la notación $b(o_t|q_t = q) = B_q(o_t)$, o simplemente $b_q(o_t)$ (ver Figura 2.3).

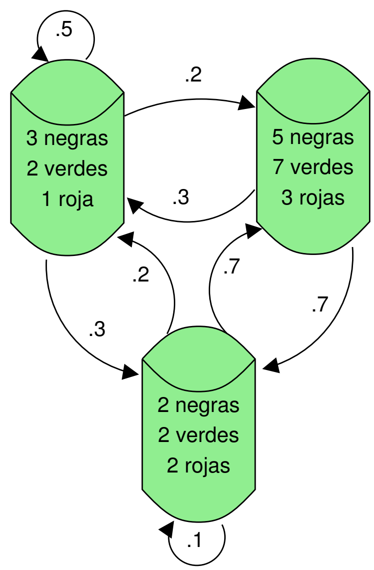

Figura 2.2: Ejemplo de concepto de modelos ocultos de Markov

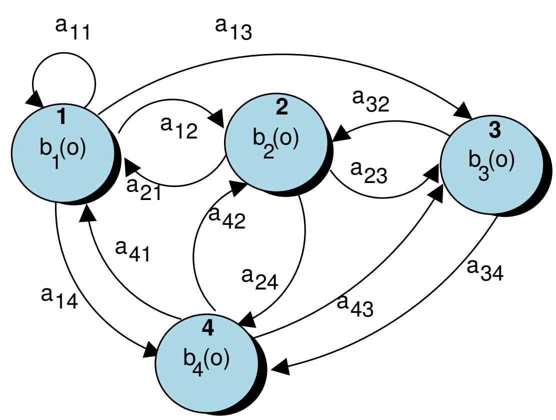

Figura 2.3: Modelo oculto de Markov de cuatro estados

En el ejemplo previo de las urnas, las probabilidades de salida dependían del contenido de las urnas. Una secuencia de salidas de un MOM es $\{o_0, o_1, o_2, \ldots\}$. La información inherente del estado no se observa directamente, está oculta. La distribución de probabilidad de cada estado puede ser de un tipo, y en general, cada estado puede tener su propio tipo de distribución. Sin embargo, en la práctica suelen tener el mismo tipo de distribución pero con parámetros diferentes. Tenemos que M denota el número de posibles resultados de todos los estados y O_t es la variable aleatoria de salida en el tiempo t, con resultado o_t. Determinamos la probabilidad de cada posible salida sumando todas las probabilidades, $P(O_t = i) = P(O_t = i|q_t = 1)P(q_t = 1) + P(O_t = i|q_t = 2)P(q_t = 2) + \cdots + P(O_t = i|q_t = S)P(q_t = S)$. De tal manera que

$$r_t = \begin{bmatrix} P(O_t = 1) & P(O_t = 2) & \cdots & P(O_t = M) \end{bmatrix}$$

y

$$B = \begin{bmatrix} P(O_t = 1|q_t = 1) & \cdots & P(O_t = M|q_t = 1) \\ P(O_t = 1|q_t = 2) & \cdots & P(O_t = M|q_t = 2) \\ . & & \\ . & & \\ . & & \\ P(O_t = 1|q_t = S) & \cdots & P(O_t = M|q_t = S) \end{bmatrix};$$

de tal manera que $b_{ji} = P(O_t = i|q_t = j) = b_j(o_t = i)$. Para el ejemplo de las urnas, con bolas de colores negro, verde y rojo, tenemos una correspondencia con los valores 1, 2 y 3 respectivamente.

$$B = \begin{bmatrix} 1/2 & 1/3 & 1/6 \\ 1/3 & 7/15 & 1/5 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}.$$

Donde cada renglón debe sumar uno. Entonces, las probabilidades de salida pueden ser calculadas por $r_t = p_t B$. Los parámetros de un MOM son descritos por el conjunto $\Lambda = \{A, \pi, B\}$, muy parecido a nuestros modelos de espacio de estados [55].

Por ende, un MOM se compone de un par de procesos estocásticos: una cadena *oculta* de Markov y un proceso *observable*, el cual es una función probabilística. Esto significa que los eventos observables en el mundo real (como las observaciones acústicas) son modelados con distribuciones de probabilidad (posiblemente continuas), y que son la parte observable del modelo, asociados con un proceso Markoviano de primer orden de estados individuales de un tiempo discreto. En general este último no es ergódico. La semántica del modelo (correspondencia conceptual con un fenómeno físico) es usualmente encapsulada en la parte oculta; por ejemplo, en un RAV un MOM puede ser usado para modelar una palabra, donde cada estado de la parte oculta representa un fonema (o una unidad subfonética), mientras que la parte observable cuenta como las características de los eventos acústicos correspondientes en un espacio característico dado (como por ejemplo una señal acústica muestreada, representada de manera adecuada) [56]. De esta manera un MOM se define de la siguiente forma [57]:

1. Un conjunto Q de S estados, $Q = \{q_1, ..., q_S\}$, que son los distintos valores que el proceso estocástico oculto discreto puede tomar.

2. Una probabilidad de distribución del *estado inicial*, por ejemplo $\pi = \{P(q_i|t = 0), q_i \in Q\}$, donde t es un índice de tiempo discreto.

3. Una distribución de probabilidad que caracteriza las transiciones permitidas entre los diferentes estados, es decir $a_{ij} = \{P(q_j$ en el tiempo $t|q_i$ en el tiempo $t-1)$,$q_i \in Q, q_j \in Q\}$, donde las *probabilidades de transición* $a_{ij} \triangleq A$ se asume que son independientes en el tiempo t.

4. Una *observación o espacio característico* O, el cual es un universo continuo o discreto para todos los posibles eventos observables (usualmente un subconjunto de R^d, donde d es la dimensionalidad de las observaciones).

5. Un conjunto de distribuciones de probabilidad (vistas como probabilidades de traslado o de salida), que describen las propiedades estadísticas de las observaciones para cada estado del modelo: $B = \{b_j(o_t) = P(o_t|q_j), q_j \in Q, o_t \in O\}$. Tradicionalmente asociados a modelos de mezclas Gaussianas.

Los MOM representan un paradigma de aprendizaje, en el sentido que los ejemplos del caso que va a ser modelado pueden ser obtenidos y utilizados en conjunto con un algoritmo de entrenamiento con el fin de aprender las estimaciones adecuadas de $\Lambda = \{A, \pi, B\}$. Los algoritmos más populares para estos casos son el algortimo de forward-backward (o Baum-Welch) y el algoritmo de Viterbi [57, 58].

Los MOM pueden ser aplicados a reconocimiento de patrones, donde los patrones ocurren como eventos que se dan secuencialmente en el tiempo. La más exitosa de sus aplicaciones radica en el procesamiento de voz. Donde cada palabra o sonido (fonema) a ser reconocido está representado por un MOM, y donde la salida es un vector con ciertas características que se derivan de los datos de la voz. La variabilidad aleatoria en el vector de características y la cantidad de tiempo que cada característica produce es modelada por el MOM. La variabilidad en la duración de cada palabra es modelada por un modelo de Markov. La variabilidad de las salidas es modelada por una selección aleatoria dentro de cada estado. Por ejemplo, en un sistema con un vocabulario pequeño de N palabras hay N MOM (A_i, π_i, B_i), siendo cada uno entrenado para representar los parámetros de esa palabra. Esta es la fase de entrenamiento del problema de reconocimiento de patrones [55]. Para lograr el reconocimiento de una palabra desconocida, esta secuencia de vectores característicos es calculada, y la probabilidad de que esta secuencia de vectores sea producida por el MOM (A_i, π_i, B_i) es determinada para cada i. El MOM que produce la mayor probabilidad se selecciona como la palabra reconocida [25].

Mientras que los MOM han sido una metodología dominante para el modelado acústico en reconocimiento automático de voz durante décadas, muchas de sus debilidades también han sido bien conocidas y se han convertido en el foco de muchas investigaciones. Una de sus principales debilidades es la imposibilidad de representar la dependencia temporal en las características acústicas de la voz, que sin embargo, es una propiedad esencial de la dinámica de la voz [59].