Bases Formales de la Computación: Sesión 3. Modelos Ocultos de Markov

Prof. Gloria Inés Alvarez V.

Departamento de Ciencias e Ingeniería de la Computación
Pontificia Universidad Javeriana Cali

Periodo 2008-2

Contenido

Introducción

- Un Hidden Markov Model (HMM) es una función de probabilidad de un modelo de Markov.
- Su primer uso fue en lingüística, modelando secuencias de letras en literatura rusa. sin embargo, fueron diseñados como un método general.
- Los HMM han sido la parte del modelamiento estadístico que más se aplica en los sistemas de reconocimiento de habla modernos. Hoy dia siguen siendo la técnica más exitosa y la más usada para varias tareas, por ejemplo, para el problema POS (Parts of speech tagging).

Construcción de Modelos de Señales

- La mayoría de los sistemas producen una salida observable que puede ser caracterizada como una señal.
- Un problema muy importante es el de modelar un sistema a partir de las señales que emite.
  - Porque esto permite obtener una descripción teórica del sistema mediante la cual se puede aprender a procesar las señales para obtener una salida deseada del sistema. Ejemplo: eliminación de ruido y distorsión.
  - El modelo permite aprender sobre el sistema que emite las señales.
  - Porque se comportan supremamente bien en la práctica.

Clases de Modelos de Señales

- Modelos Deterministas: se basan en una propiedad específica y conocida del sistema a modelar. Por ejemplo, el comportamiento se modela mediante una señal seno. Sólo se necesita determinar o estimar unos pocos parámetros: amplitud, frecuencia, fase.
- Modelos Estadísticos: se busca caracterizar únicamente las propiedades estadísticas de la señal. HMM pertenecen a esta clase. La hipótesis de base es que el proceso se puede modelar como un proceso aleatorio paramétrico y que los parámetros se pueden estimar de forma precisa.

Modelos Ocultos de Markov

- Los primeros trabajos fueron de Baum y sus colegas a finales de la década de 1960 y principios de los 70's.
- Se implementaron para aplicaciones de procesamiento de habla por Baker y Jelinek en los laboratorios de IBM en 1970.
- Sólo se entendieron y aplicaron ampliamente hasta finales de los 80's.

Cadenas de Markov

Definición

Un proceso de Markov es un proceso estocástico que sirve para representar secuencias de variables aleatorias no independientes entre sí. Es decir, donde la probabilidad del siguiente estado sobre una secuencia completa depende de estados previos al estado actual.

Por ejemplo, si la variable aleatoria consiste en contar los libros que hay en la biblioteca, entonces saber cuántos libros hay hoy puede servir para calcular cuántos habrá mañana y no es necesario saber cuántos había hace una semana o hace un año.
Es decir, que los elementos futuros en la secuencia son condicionalmente independientes de los pasados, dado el elemento presente.

La Propiedad de Markov

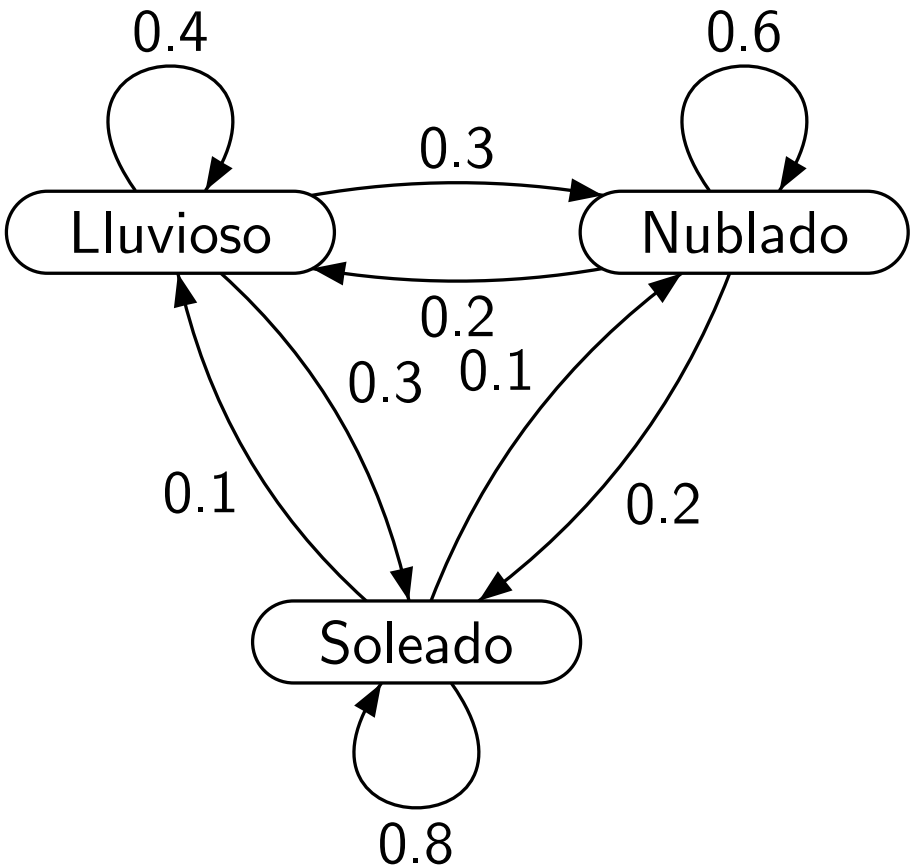
Definición

Sea $X = \{X_1, \ldots, X_T\}$ *una secuencia de variables aleatorias que toman valores en un conjunto finito* $S = \{s_1, \ldots, s_N\}$ *que se llama el espacio de estados. Entonces, las propiedades de Markov son:*

- *Horizonte limitado:*
  $P(X_{t+1} = s_k | X_1, \ldots, X_t) = P(X_{t+1} = s_k | X_t)$
- *Invariante en el tiempo*
  $P(X_{t+1} = s_k | X_1, \ldots, X_t) = P(X_2 = s_k | X_1)$

Si X *cumple estas dos propiedades, se dice que* X *es una cadena de Markov o que tiene la propiedad de Markov.*

Ejemplo de Cadena de Markov



$$\mathbf{A} = \begin{pmatrix} 0{,}4 & 0{,}3 & 0{,}3 \\ 0{,}2 & 0{,}6 & 0{,}2 \\ 0{,}1 & 0{,}1 & 0{,}8 \end{pmatrix}$$

$$\mathbf{\Pi} = \begin{pmatrix} 0{,}25 & 0{,}25 & 0{,}5 \end{pmatrix}$$

Preguntas que se pueden responder

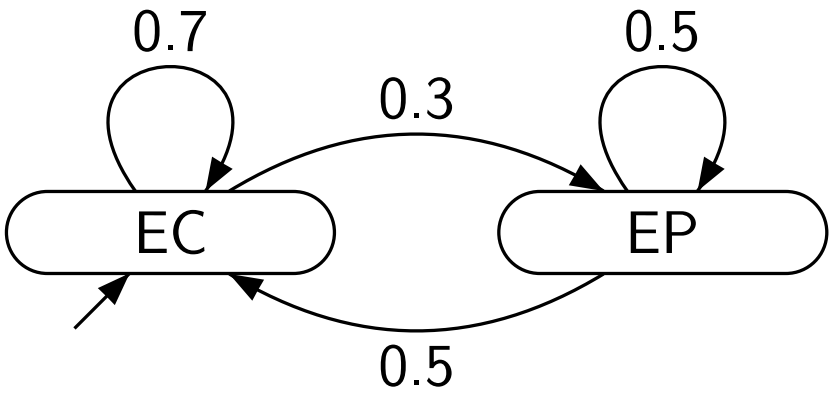
- Cuál es la probabilidad de la secuencia: LSLSLSN ?
  $P(X_1 \ldots, X_T) = \Pi_{X_1} \prod_{t=1}^{T-1} A[X_t, X_t + 1]$
- Cuál es la probabilidad de tener el mismo clima por d días?

$$
\begin{aligned}
P(O|M, q_1 = i) &= P(O, q_1 = i|M)/P(q_1 = i) \\
&= \Pi[i](A[i,i])^{d-1}(1 - A[i,i])/\Pi[i] \\
&= (A[i,i])^{d-1}(1 - A[i,i]) \\
&= p_i(d) \\
\bar{p}_i(d) &= \sum_{d=1}^{\infty} d p_i(d) \\
&= \frac{1}{1 - A[i,i]}
\end{aligned}
$$

Modelos Ocultos de Markov

- En las cadenas de Markov, las señales observadas corresponden a los estados del modelo.
- En los modelos ocultos de Markov no se conoce la secuencia de estados por la que pasa el modelo, sino una función probabilística de ella.

Ejemplo de Modelo Oculto de Markov



La máquina vende: CocaCola,PremioRojo y Nestea.

$$\mathbf{B} = \begin{pmatrix} 0{,}6 & 0{,}1 & 0{,}3 \\ 0{,}1 & 0{,}7 & 0{,}2 \end{pmatrix}$$

Cuando la máquina está en el estado *CocaCola* tiende a entregar una CocaCola, independientemente de lo que el usuario haya solicitado y análogamente cuando está en el estado *PremioRojo*.

Uso del Modelo

Cual es la probabilidad de que la máquina expenda la secuencia *Nestea*, *PremioRojo* si la máquina empieza en el estado *CocaCola*?

- Se deben considerar todos los caminos en el modelo que conducen a esa secuencia.
- Hay cuatro posibilidades:
  1. EC EC $0{,}7 \times 0{,}3 \times 0{,}7 \times 0{,}1$
  2. EC EP $0{,}7 \times 0{,}3 \times 0{,}3 \times 0{,}1$
  3. EP EC $0{,}3 \times 0{,}3 \times 0{,}5 \times 0{,}7$
  4. EP EP $0{,}3 \times 0{,}3 \times 0{,}5 \times 0{,}7$
- Sumando estas opciones se obtiene la probabilidad de la secuencia que es 0,084

Modelos Ocultos de Markov

Definición

Un modelo oculto de Markov es una tupla $M = \{S, \Sigma, A, B, \Pi\}$ *donde:*

- S *es un conjunto finito de estados.*
- Σ *es un conjunto finito de símbolos.*
- A *es la matriz de probabilidades de transición entre estados de* S. $A[i,j]$ *es* $P(X_{t+1} = Sj | X_t = s_i)$
- B *es la matriz de probabilidad de emisión de símbolos de* Σ. $B[j,k]$ *es* $B[j,k] = P(O_t = k | X_t = s_j)$
- Π *es el vector de probabilidades iniciales.* $\Pi[i]$ *es* $P(X_1 = S_i)$

Se denota una secuencia de estados $X = (X_1, \ldots, X_{T+1})$ *donde* $X_t : S \rightarrow \{1, \ldots, N\}$ *y una secuencia de observaciones* $O = (o_1, \ldots, o_T)$ *con* $o_T \in \Sigma$

Los tres problemas fundamentales en HMM

1. Problema de evaluación de la probabilidad (o verosimilitud) de una secuencia de observaciones dado un HMM.
2. Problema de determinación de la secuencia más probable de estados.
3. Problema de ajuste de los parámetros del modelo para que den mejor cuenta de las señales observadas.

Problema de Evaluación

Definición

Dada una secuencia de observaciones $O_1, \ldots, O_T$ y un modelo de Markov $M = (A, B, \Pi)$, cómo calcular eficientemente $P(O|M)$, la probabilidad de la secuencia dado el modelo?

El problema de evaluación tiene particular interés cuando se desea elegir entre varios modelos posibles, ya que se puede elegir aquel que mejor explique una secuencia de observaciones

Solución al Problema de Evaluación

- La solución directa consiste en extender todos los caminos de longitud T, calcular la probabilidad de cada uno y sumarlas. El costo es exponencial, porque en el peor caso desde cada estado se puede ir a N estados, es decir, que habría N^T caminos distintos y en cada camino se hacen $2T$ cálculos para obtener la probabilidad. Este costo no es aceptable en la práctica: con $N = 5$ y $T = 100$ el costo sería del orden de 10^{72}.
- Existe una manera más eficiente de resolver el problema, que se basa en la definición de la variable forward:

$$\alpha_t(i) = P(O_1, \ldots, O_t, X_t = s_i | M)$$

Definición de la Variable Forward

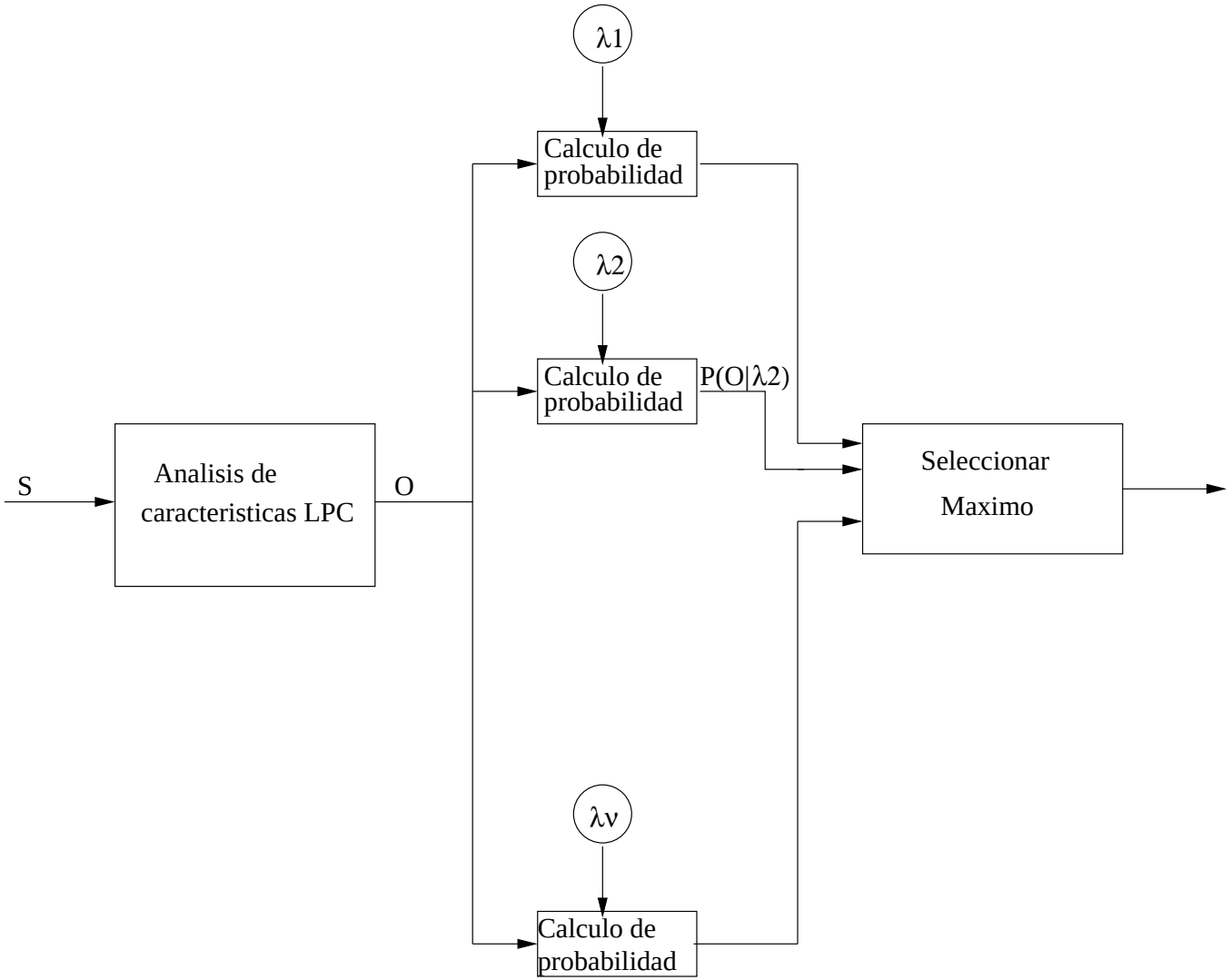
- $\alpha_t(i)$ es la probabilidad de haber observado la secuencia parcial $O_1, \ldots, O_t$ y encontrarse en el tiempo t en el estado s_i, dado el modelo M.
- $\alpha_t(i)$ se puede calcular inductivamente:
  - Caso base: $\alpha_1(i) = \Pi[i]B[i, O_1], 1 \leq i \leq N$
  - Caso general: $\alpha_{t+1}(j) = \left(\sum_{i=1}^{N} \alpha_t(i)A[i,j]\right) B[j, O_{t+1}]$ con $1 \leq t \leq T-1$ y $1 \leq j \leq N$
  - Terminación: $P(O|M) = \sum_{i=1}^{N} \alpha_T(i)$

Aplicación en Reconocimiento de Habla

En el reconocimiento de palabras aisladas se puede aplicar HMM

- Se selecciona el vocabulario de trabajo
- Se eligen las unidades básicas de reconocimiento: fonemas, sílabas, ...
- Se construye un HMM especializado en reconocer cada palabra, que es una secuencia de observaciones, donde cada observación corresponde a una unidad básica
- Al recibir una secuencia de observaciones desconocida, se alimentan con ella todos los HMM y se elige como la palabra reconocida la representada por el modelo que explique la secuencia de observaciones con máxima probabilidad.

Esquema del Modelo Conceptual



Definición de la Variable Backward

- $\beta_t(i)$ es la probabilidad de observar la secuencia parcial $O_{t+1}, \ldots, O_T$ y encontrarse en el tiempo t en el estado s_i, dado el modelo M.
- $\beta_t(i)$ se puede calcular inductivamente:
  - Caso base: $\beta_T(i) = 1, 1 \leq i \leq N$
  - Caso general: $\beta_t(i) = \sum_{j=1}^{N} A[i,j]B[j, O_{t+1}]\beta_{t+1}(j)$ con $t = T-1, T-2, \ldots 1$ y $1 \leq j \leq N$

Problema 2: Determinación de la secuencia de estados más probable

- Dependiendo del criterio de optimalidad elegido pueden haber varias secuencias de estados más probables, dada una secuencia de observaciones y un modelo.
- El criterio más utilizado consiste en elegir la secuencia de estados que maximiza $P(S|O, M)$. Esto da orígen al algoritmo conocido como Algoritmo de Viterbi. Para hacer el cálculo se utiliza la variable $\delta_t(i)$.

$$\delta_t(i) = \max_{X_1,\ldots,X_{t-1}} P(X_1, \ldots, X_t = i, O_1, \ldots, O_t|M)$$

Algoritmo de Viterbi

Donde $\delta_t(i)$ es la más alta probabilidad que se obtiene a lo largo de un camino, hasta el tiempo t, que da cuenta de las primeras t observaciones y termina en el estado s_i
Para poder encontrar la secuencia precisa de estados, es necesario almacenar el estado que maximiza la función $\delta_t(i)$ en cada momento, esto se almacena en la variable $\psi_t(j)$.

Algoritmo de Viterbi

- Inicialización:

$$\begin{aligned} \delta_1(i) &= \Pi[i]B[i, O_i], 1 \leq i \leq N \\ \psi_1(i) &= 0 \end{aligned}$$

- Recursión:

$$\begin{aligned} \delta_t(i) &= \max_{1 \leq i \leq N}(\delta_{t-1}(i)A[i,j])B[j, O_t], \text{con} 2 \leq t \leq T \text{y} 1 \leq i \leq N \\ \psi_t(i) &= argmax_{1 \leq i \leq N}(\delta_{t-1}(i)A[i,j]), \text{con} 2 \leq t \leq T \text{y} 1 \leq i \leq N \end{aligned}$$

- Terminación:

$$\begin{aligned} p^* &= \max_{1 \leq i \leq N}(\delta_T(i)) \\ q_T^* &= argmax_{1 \leq i \leq N}(\delta_T(i)) \end{aligned}$$

- Obtención del camino:

$$q_T^* = \psi_{t+1}(q_{t+1}^*) t = T-1, T-2, \ldots, 1$$

Aplicación en Procesamiento de Lenguaje Natural

El problema de etiquetamiento mediante categorías sintácticas se puede resolver mediante HMMs

- Dado un HMM, es posible asociar a cada palabra de una oración la categoría sintáctica que le corresponde
- Cada palabra es una observación, por lo tanto la oración es una secuencia de observaciones
- Cada estado es una categoría sintáctica
- Mediante el algoritmo de Viterbi se puede establecer la secuencia de estados, es decir la secuencia de categorías sintácticas que con mayor probabilidad explica la secuencia de observaciones