Aprendizaje de Sucesiones
Modelos ocultos de Markov

Lección 28

Dr. Pablo Alvarado Moya

CE5506 Introducción al reconocimiento de patrones
Área de Ingeniería en Computadores
Tecnológico de Costa Rica

II Semestre, 2019

Contenido

Introducción

- En aprendizaje supervisado y no supervisado supusimos datos independientes e idénticamente distribuidos.
- Esta suposición es inválida en gran cantidad de aplicaciones, particularmente aquellas que tratan con **sucesiones** de datos.
- Esto ocurre cuando se analizan
  - Fenómenos naturales (biológicos, meteorológicos, astronómicos)
  - Tendencias de mercados
  - Señales acústicas (habla, música) o de vídeo
  - Lenguaje (escrito, hablado, señas, etc.)

Modelos de Markov

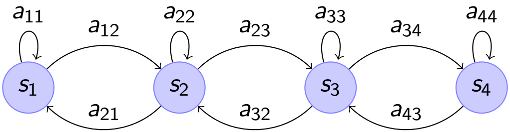
- Ya presentamos MDP (procesos de decisión de Markov) y POMDP (MDP parcialmente observables).
- Ahora revisaremos otros **modelos de Markov**.
- Primero revisaremos el concepto más sencillo de todos: Cadenas de Markov (*Markov Chains*).
- Luego extenderemos el concepto a los Modelos Ocultos de Markov (*Hidden Markov Models*)

Algunos Modelos de Markov

Sistema	**Estado observable**	**Estado oculto**
Autónomo	Cadena de Markov **MC**	Modelo Oculto de Markov **HMM**
Controlado	Proc. Decisión de Markov **MDP**	MDP parcial. observable **POMDP**

Modelo de Markov

- La cadena de Markov modela el estado (observable) de un sistema con una variable aleatoria.
- La Cadena de Markov la definimos como la tupla $\langle \mathcal{S}, \mathbf{A}, \underline{\boldsymbol{\pi}} \rangle$ con $\mathcal{S}$ el conjunto de estados, $\mathbf{A}$ las probabilidades de transición entre estados, y $\underline{\boldsymbol{\pi}}$ las probabilidades de estado inicial.



$$\mathcal{S} = \{s_1, s_2, s_3, s_4\} \qquad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \qquad \mathbf{A}\underline{\mathbf{1}} = \underline{\mathbf{1}}$$

Propiedades

- Sea S_t la variable aleatoria que contiene el estado en el instante t
- La cadena de Markov tiene horizonte limitado (propiedad de Markov)

$$P(S_{t+1} = s_k | S_1, \ldots, S_t) = P(S_{t+1} = s_k | S_t)$$

  es decir, el estado siguiente solo depende del estado actual.
- La MC es estacionaria (invariante en el tiempo)

Probabilidad de una sucesión

- La probabilidad de una sucesión de estados se calcula con la regla de la cadena:

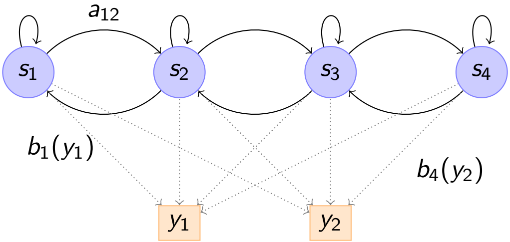
$$
\begin{aligned}
&P(S_1,\ldots,S_T)\\
&\qquad = P(S_1)P(S_2|S_1)P(S_3|S_1,S_2)\cdots P(S_T|S_1,\ldots,S_{T-1})
\end{aligned}
$$

pero usando la propiedad de Markov

$$
\begin{aligned}
&= P(S_1)P(S_2|S_1)P(S_3|S_2)\cdots P(S_T|S_{T-1})\\
&= \pi_{S_1}\prod_{t=1}^{T-1} a_{S_t S_{t+1}}
\end{aligned}
$$

Modelos ocultos de Markov
Hidden Markov Models

- En los modelos ocultos no podemos observar el estado directamente, sino solo alguna observación que está relacionada con el estado actual.



- Observamos los *símbolos* y_i y los estados s_i están **ocultos**
- Si requerimos un clasificador de secuencias, usamos HMM como clasificador generativo, esto es, tendremos un modelo para cada clase.

Aplicaciones

- Los HMM tienen años en la comunidad y sus usos se encuentran en gran variedad de áreas:
  - Reconocimiento de habla
  - Síntesis de habla
  - Reconocimiento de señales acústicas
  - Reconocimiento de gestos
  - Reconocimiento de escritura manual
  - Análisis criptográfico
  - Alineamiento de biosecuencias
  - Predicción genética
  - Reconocimiento de actividades
  - ...

Definición de HMM

- Un HMM se define como la tupla $\langle \mathcal{S}, \mathbf{A}, \mathcal{Y}, \mathbf{B}, \underline{\boldsymbol{\pi}} \rangle$ con
  - $\mathcal{S} = \{s_1, \ldots, s_n\}$ el conjunto de n estados,
  - $\mathbf{A} \in \mathbb{R}^{n \times n}$ las probabilidades de transición entre estados,

$$a_{ij} = P(S_{t+1} = s_j | S_t = s_i)$$

  - $\mathcal{Y} = \{y_1, \ldots, y_m\}$ los m símbolos observables,
  - $\mathbf{B} \in \mathbb{R}^{n \times m}$ las probabilidades de emisión de los m símbolos, y

$$b_{ik} = P(Y_t = y_k | S_t = s_i)$$

  - $\underline{\boldsymbol{\pi}} \in \mathbb{R}^n$ las probabilidades de estado inicial.

$$\pi_i = P(S_1 = s_i)$$

- Con frecuencia se abrevia al HMM con $\lambda = \langle \mathbf{A}, \mathbf{B}, \underline{\boldsymbol{\pi}} \rangle$

Generación de sucesiones de símbolos

- Así generaría un HMM sucesión de símbolos $Y = Y_1 Y_2 \ldots Y_T$:
  1. Haga $t = 1$ y elija un estado inicial $S_1 = s_i$ de acuerdo a la distribución inicial de estados $\underline{\boldsymbol{\pi}}$.
  2. Elija $Y_t = y_k$ de acuerdo a la probabilidad de emisión b_{ik}.
  3. Cambie a nuevo estado $S_{t+1} = s_j$ de acuerdo a la probabilidad de transición a_{ij}
  4. Haga $t = t + 1$ y si $t \leq T$ salte al paso 2

Tres tareas básicas de los HMM

- Evaluación Dada la sucesión de observaciones $Y = Y_1 Y_2 \ldots Y_T$ y el modelo $\lambda = \langle \mathbf{A}, \mathbf{B}, \underline{\boldsymbol{\pi}} \rangle$ calcule $P(Y|\lambda)$
- Reconocimiento Dada la sucesión de observaciones $Y = Y_1 Y_2 \ldots Y_T$ y el modelo $\lambda = \langle \mathbf{A}, \mathbf{B}, \underline{\boldsymbol{\pi}} \rangle$ encuentre la sucesión de estados $S = S_1 S_2 \ldots S_T$ que explica las observaciones
- Entrenamiento Dada la sucesión de observaciones $Y = Y_1 Y_2 \ldots Y_T$ ajuste los parámetros del modelo $\lambda = \langle \mathbf{A}, \mathbf{B}, \underline{\boldsymbol{\pi}} \rangle$ para maximizar $P(Y|\lambda)$.

1. Evaluación

Evaluación

- Vamos a revisar dos estrategias de solución del problema de evaluación:
  1. Por fuerza bruta, solo como motivación para el segundo método
  2. Por procedimiento hacia adelante y hacia atrás
- Para clasificación de sucesiones de observaciones requerimos un modelo por clase y seleccionamos clase más probable como la ganadora.

Evaluación: fuerza bruta (1)

- Queremos encontrar $P(Y|\lambda)$, es decir, la probabilidad de la sucesión de observaciones $Y = Y_1 Y_2 \ldots Y_T$ dado el HMM λ.
- En principio podemos enumerar todas las posibles sucesiones de estados de longitud T $S = S_1 S_2 \ldots S_T$
- Para una sucesión de estados S, la probabilidad de la sucesión Y de observaciones es

$$P(Y|S,\lambda) = \prod_{t=1}^{T} P(Y_t|S_t,\lambda) = \prod_{t=1}^{T} b_{S_t Y_t}$$

- La probabilidad de la sucesión S de estados es

$$P(S|\lambda) = P(S_1) \prod_{t=2}^{T} P(S_t|S_{t-1}) = \pi_{S_1} \prod_{t=2}^{T} a_{S_{t-1} S_t}$$

Evaluación: fuerza bruta (2)

- Por lo tanto, la probabilidad conjunta

$$P(Y,S|\lambda) = P(S|\lambda)P(Y|S,\lambda) = \pi_{S_1} \prod_{t=2}^{T} a_{S_{t-1}S_t} \prod_{t=1}^{T} b_{S_t Y_t}$$

- Considerando todas las posibles secuencias de estados marginalizamos S

$$P(Y|\lambda) = \sum_{S} \pi_{S_1} b_{S_1 Y_1} \prod_{t=2}^{T} a_{S_{t-1}S_t} b_{S_t Y_t}$$

- Total de cálculos $\mathcal{O}(Tn^T)$ (¡exponencial en la longitud!)
- **No es práctico**

Procedimiento hacia adelante

- Definimos la variable **hacia adelante** $\alpha_j(t)$ como la probabilidad de la sucesión parcial de observaciones **hasta** el tiempo t, con estado s_j en el instante t

$$\alpha_j(t) = P(Y_1 Y_2 \ldots Y_t, S_t = s_j|\lambda)$$

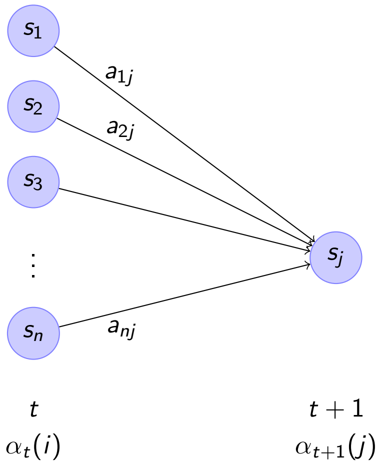
- Esto se calcula inductivamente con

$$\begin{aligned} \alpha_j(1) &= \pi_j b_{jY_1} & 1 \leq j \leq n \\ \alpha_j(t+1) &= \left(\sum_{i=1}^{n} \alpha_i(t) a_{ij} \right) b_{jY_{t+1}} & 1 \leq t \leq T-1 \end{aligned}$$

- Con solo $n^2 T$ operaciones

$$P(Y|\lambda) = \sum_{i=1}^{n} P(Y, S_T = s_i|\lambda) = \sum_{i=1}^{N} \alpha_i(T)$$

Procedimiento hacia adelante



Procedimiento hacia atrás

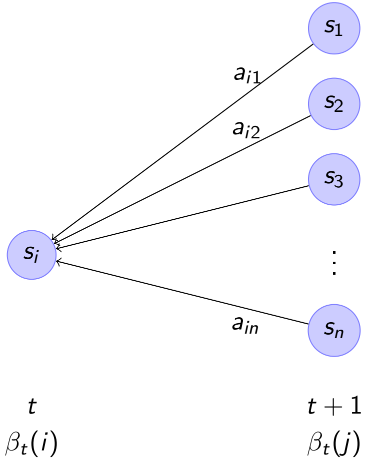
- Similarmente, definimos la variable **hacia atrás** $\beta_i(t)$ como la probabilidad de la sucesión de observaciones parcial **después** del tiempo t, dado el estado s_i en el instante t

$$\beta_i(t) = P(Y_{t+1}Y_{t+2}\dots Y_T | S_t = s_i, \lambda)$$

- También se calcula de forma inductiva:

$$\begin{aligned} \beta_i(T) &= 1 && 1 \leq i \leq N \\ \beta_i(t-1) &= \sum_{j=1}^{n} a_{ij} b_{jY_t} \beta_j(t) && 2 \leq t \leq T \end{aligned}$$

Procedimiento hacia atrás



2. Reconocimiento

Reconocimiento

- En la tarea de **reconocimiento** queremos descubrir la sucesión de estados, dadas las observaciones.
- Contrario a la tarea de evaluación, en reconocimiento no existe **una** sucesión óptima.
  1. Elija estados que individualmente son más probables (maximiza el número de estados correctos)
  2. Encuentre la mejor sucesión de estados (garantiza que la sucesión sea válida)
- La primera opción encuentra $\arg\max_i \gamma_i(t)$ para todo t con

$$\begin{aligned}\gamma_i(t) &= P(S_t = s_i | Y, \lambda) \\ &= \frac{P(Y_1 \ldots Y_t, S_t = s_i | \lambda) P(Y_{t+1} \ldots Y_T | S_t = s_i, \lambda)}{P(Y|\lambda)} \\ &= \frac{\alpha_i \beta_i}{\sum_{j=1}^{n} \alpha_j(t) \beta_j(t)}\end{aligned}$$

Algoritmo de Viterbi

- Para la segunda opción, encontrar la mejor sucesión implica calcular $\arg\max_S P(S|Y,\lambda)$, que equivale a $P(S,Y|\lambda)$.
- El **algoritmo de Viterbi** (programación dinámica) define $\delta_j(t)$ como la mayor probabilidad de una ruta de longitud t que explica las observaciones y termina en el estado s_j

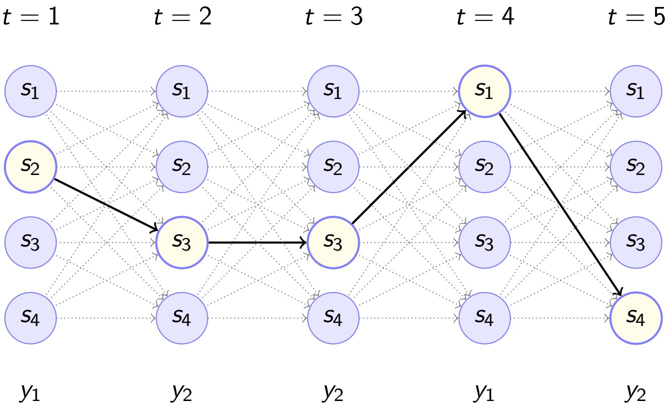
$$\delta_j(t) = \max_{S_1,S_2\ldots,S_{t-1}} P(S_1 S_2 \ldots S_t = j, Y_1 Y_2 \ldots Y_t|\lambda)$$

- Por inducción

$$\begin{aligned} \delta_j(1) &= \pi_j b_{jY_1} & 1 \leq j \leq n \\ \delta_j(t+1) &= \left(\max_i \delta_i(t) a_{ij}\right) b_{jY_{t+1}} & 1 \leq t \leq T-1 \end{aligned}$$

- Tomando el j que maximiza la probabilidad para cada t produce el resultado final (*backtracking*)

Diagrama de Trellis



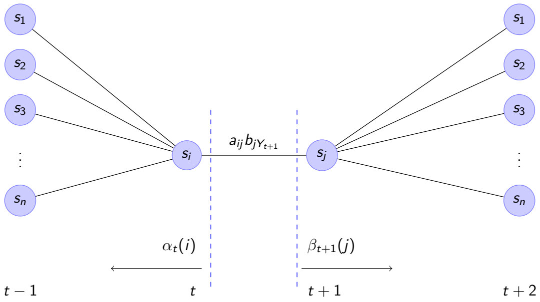
3. Entrenamiento

Entrenamiento

- No se conoce ningún algoritmo que permita encontrar analíticamente los parámetros del HMM para maximizar la probabilidad de la sucesión observada.
- Se utilizan métodos de optimización que corren el riesgo de pegarse en un óptimo local:
  - técnicas basadas en gradientes
  - reestimación Baum-Welch (equivalente a EM)
- Definimos $\xi_{ij}(t)$ como la probabilidad de estar en el estado s_i en t y en el estado s_j en $t+1$:

$$\begin{aligned}\xi_{ij}(t) &= P(S_t = s_i, S_{t+1} = s_j | Y, \lambda) \\ &= \frac{\alpha_i(t) a_{ij} b_{jY_{t+1}} \beta_j(t+1)}{P(Y|\lambda)} \\ &= \frac{\alpha_i(t) a_{ij} b_{jY_{t+1}} \beta_j(t+1)}{\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i(t) a_{ij} b_{jY_{t+1}} \beta_j(t+1)}\end{aligned}$$

Estimación de parámetros del HMM (1)



Estimación de parámetros del HMM (2)

- Debido a que $\gamma_i(t)$ y $\xi_{ij}(t)$ son probabilidades, podemos marginalizar

$$\gamma_i(t) = \sum_{j=1}^{n} \xi_{ij}(t)$$

- Ahora, si sumamos a través del tiempo t
  - $\sum_{t=1}^{T-1} \gamma_i(t)$ = número esperado de veces que se visitó s_i = número esperado de transiciones desde s_i
  - $\sum_{t=1}^{T-1} \xi_{ij}(t)$ = núm. esperado de transiciones desde s_i hasta s_j

Reestimación de Baum-Welch

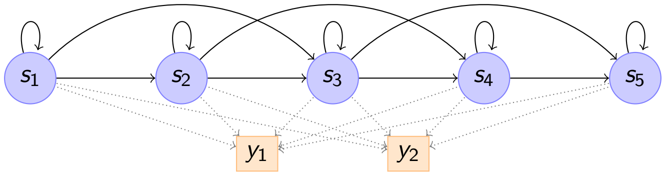
- Dado un modelo preliminar λ, podemos reestimar los parámetros con:

$$\bar{\pi}_i = \gamma_i(1) \qquad \bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_{ij}(t)}{\sum_{t=1}^{T-1} \gamma_i(t)} \qquad \bar{b}_{jk} = \frac{\sum_{Y_t=y_k} \gamma_j(t)}{\sum_{t=1}^{T} \gamma_j(t)}$$

- Baum et al. demostraron que si actualizamos $\lambda = \langle \mathbf{A}, \mathbf{B}, \underline{\boldsymbol{\pi}} \rangle$ con lo anterior para obtener $\bar{\lambda} = \langle \bar{\mathbf{A}}, \bar{\mathbf{B}}, \underline{\bar{\boldsymbol{\pi}}} \rangle$ entonces
  - Si $\bar{\lambda} = \lambda$ estamos en un máximo local de la verosimilitud
  - $P(Y|\bar{\lambda}) > P(Y|\lambda)$: el modelo $\bar{\lambda}$ es más probable.
- Si iteramos obtenemos un estimado de máxima verosimilitud para el HMM
- Desafortunadamente la verosimilitud no es convexa y es probable que se converja a un máximo local.

HMM no ergódicos

- Hasta ahora consideramos HMM ergódicos, es decir, cada estado es alcanzable desde cualquier otro estado en un número finito de pasos.
- En aplicaciones de reconocimiento de voz se restringe más el modelo.
- Se usan particularmente esquemas de Bakis (izquierda derecha), en donde estados "anteriores" no son alcanzables.



Resumen

Este documento ha sido elaborado con software libre incluyendo LATEX, Beamer, GNUPlot, GNU/Octave, XFig, Inkscape, GNU-Make y Subversion en GNU/Linux