Inteligencia Artificial

CC-421

César Lara Avila

Universidad Nacional de Ingeniería

(actualización: 2020-12-13)

Bienvenidos

Modelos Ocultos de Markov

- Máquina de estados
- Cadenas de Markov
- Aplicación: PageRank
- Modelos Ocultos de Markov
- Métodos de evaluación
- Métodos de aprendizaje

Máquina de estados

Es un modelo para procesamiento de información especificado por:

- Un conjunto de estados S
- Un estado inicial S_0
- Un conjunto finito de entradas I
- Un conjunto finito de salidas S
- Una función de transición de $S_i \to S_j$
- Una función de salida de $S_i \to O$

Cadena de Markov

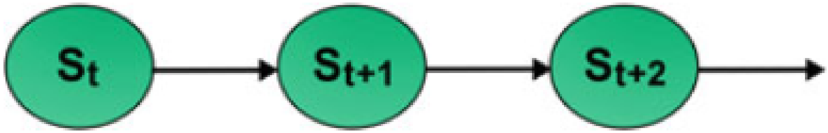
Las cadenas de Markov son otra clase de PGM que representan procesos dinámicos, en particular cómo cambia el estado de un proceso con el tiempo.

Son máquina de estados finitos en la que:

- Las transiciones de un estado a otro no son determinísticas
- La probabilidad de transición de un estado a otro sólo depende del estado actual (y no de los anteriores).

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

La siguiente figura ilustra una cadena de Markov donde cada nodo representa el estado en cierto punto en el tiempo:



Especificación de una Cadena de Markov

Formalmente, una cadena de Markov se define como:

- Conjunto de estados: $Q = \{q_1, q_2, \ldots, q_n\}$
- Vector de probabilidades previas o iniciales: $\Pi = \{\pi_1, \pi_2, \ldots, \pi_n\}$, donde $\pi_i = P(S_0 = q_i)$
- Matriz de probabilidades de transición: $A = \{a_{ij}\}$, donde $a_{ij} = P(S_t = q_j | S_{t-1} = q_i)$ donde n es el número de estados y S_0 es el estado inicial.
- De forma compacta, un MC se representa como $\lambda = \{A, \Pi\}$.

Una cadena de Markov (de primer orden) satisface las siguientes propiedades:

- Axiomas de probabilidad: $\sum_i \pi_i = 1$ y $\sum_j a_{ij} = 1$
- Propiedad de Markov: $P(S_t = q_j | S_{t-1} = q_i, S_{t-2} = q_k, \ldots) = P(S_t = q_j | S_{t-1} = q_i)$.

Ejemplo

Por ejemplo, considere el modelo meteorológico simple anterior con tres estados: $q_1 = soleado, q_2 = nublado, q_3 = lluvioso$.

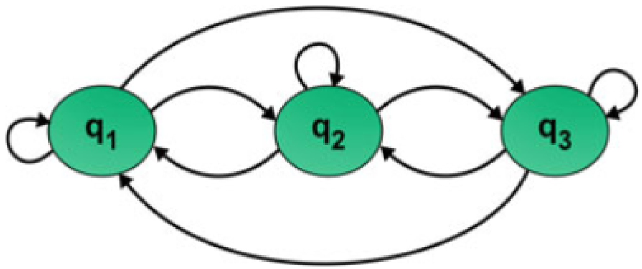
Soleado	Nublado	Lluvioso
0.2	0.5	0.3

En este caso, para especificar un MC, necesitaremos un vector con tres probabilidades previas y una matriz de 3×3 de probabilidades de transición:

	Soleado	Nublado	Lluvioso
Soleado	0.8	0.1	0.1
Nublado	0.2	0.6	0.2
Lluvioso	0.3	0.3	0.4

Diagramas de estado y salidas

La matriz de transición se puede representar gráficamente con un diagrama de transición de estado o simplemente un diagrama de estado. Este diagrama es un grafo dirigido, donde cada nodo es un estado y los arcos representan posibles transiciones entre estados.



A cada estado le corresponde una salida O_i

Una secuencia de observaciones de $t = 1$ a $t = T$ se denota por:
$O = \{o_1 \ldots, o_T\}$.

Preguntas

Dado un modelo de cadena de Markov, hay tres preguntas básicas que podemos hacer:

- Probabilidad de pertenencia - ¿cuál es la probabilidad de una determinada secuencia de estados?

  $$P(q_i, q_j, q_k, \ldots) = a_{0i} a_{ij} a_{jk} \ldots$$

- Probabilidad de permanencia - ¿Cuál es la probabilidad de que la cadena permanezca en cierto estado durante un período de tiempo?

  $$P(d_i) = a_{ii}^{d-1}(1 - a_{ii})$$

- Permanencia de promedio - ¿cuál es el tiempo esperado que la cadena permanecerá en cierto estado?

  $$E(D) = \sum_i d_i P(d_i)$$

  $$E(d_i) = \sum_i d_i a_{ii}^{d-1}(1 - a_{ii}) = 1/(1 - aii)$$

Ejemplo de preguntas

Dado el modelo del clima:

- Probabilidad de pertenencia:
  - Dado que el tiempo inicial es soleado, cual es la P de que sea soleado, soleado, lluvioso, lluvioso, soleado, nublado, soleado
- Probabilidad de permanencia:
  - Probabilidad de que este nublado por 3 días seguidos
- Permanencia promedio:
  - Tiempo esperado que permanezca nublado?

Estimación de parámetros

Para un MC, los parámetros se pueden estimar simplemente contando el número de veces que la secuencia está en un cierto estado i y el número de veces que hay una transición de un estado i a un estado j.

Supongamos que hay N secuencias de observaciones. γ_{0i} es el número de veces que el estado i es el estado inicial en una secuencia, γ_i es el número de veces que observamos el estado i y γ_{ij} es el número de veces que observamos una transición del estado i al estado j.

Los parámetros se pueden estimar con las siguientes ecuaciones:

- Probabilidades iniciales: $\pi \sim \gamma_{0i}/N$
- Probabilidades de transición: $a_{ij} \sim \gamma_{ij}/\gamma_i$

Ejemplo de estimación

Para el ejemplo del clima tenemos las siguientes cuatro secuencias de observación: $q_1 = soleado, q_2 = nublado, q_3 = lluvioso$

$$q_2, q_2, q_3, q_3, q_3, q_3, q_1$$

$$q_1, q_3, q_2, q_3, q_3, q_3, q_3$$

$$q_3, q_3, q_2, q_2$$

$$q_2, q_1, q_2, q_2, q_1, q_3, q_1$$

Los parámetros correspondientes pueden estimarse como se muestra en las tablas:

Soleado	**Nublado**	**Lluvioso**
0.25	0.5	0.25

	Soleado	**Nublado**	**Lluvioso**
Soleado	0	0.33	0.67
Nublado	0.285	0.43	0.285
Lluvioso	0.18	0.18	0.64

Convergencia

- Otra pregunta interesante es: ¿si se 'transita' la cadena un gran número de veces, a la larga cuáles la probabilidad de cada estado (en el límite)?
- Dada una probabilidad inicial Π, la probabilidad después de M iteraciones se obtiene multiplicando Π por $A \times A \times A \ldots$:

$$p = \pi A^M$$

- Después de un cierto número, normalmente el valor de p ya prácticamente no cambia.

Teorema de Perron-Frobenius

Dadas ciertas condiciones, la cadena converge a una distribución invariante p, tal que: $p \times A = p$

Condiciones:

- Irreducible: de cualquier estado hay cierta probabilidad de visitar los demás estados
- Aperiódica: la cadena no cae en ciclos

La rapidez de convergencia está determinada por el segundo eigenvalor de A.

Ejemplo de convergencia

```
A =   0         1.0000     0
      0         0.1000     0.9000
      0.6000    0.4000    0
```

$\pi = 0.5000\ 0.2000\ 0.300$

Si multiplicamos $\pi \times A$:

```
1.0.1800    0.6400    0.1800
2.0.1080    0.3160    0.5760
  ....
10.0.2358    0.4190    0.3452
```

En el límite:

```
p = 0.2    0.40.4
```

Aplicación: PageRank

Podemos representar la Web como una cadena de Markov,donde cada estado es un página y los hipervínculos entre las páginas web corresponden a transiciones de estado.

Cada enlace de salida puede seleccionarse con la misma probabilidad.

Dada la matriz de probabilidad de transición de la WWW, podemos obtener las probabilidades de convergencia para cada estado (página web) de acuerdo con el teorema de Perron-Frobenius.

La probabilidad a la que converge la cadena provee una estimación de que tan probable es que una persona visite una página en cierto tiempo.

Google basa el orden (importancia) de las páginas que encuentra (para cierta búsqueda) en éstas probabilidades.

Modelos Ocultos de Markov(HMM)

Un modelo de Markov oculto (HMM) es una cadena de Markov donde los estados no son directamente observables.

Se puede ver como un doble proceso estocástico:

- Un proceso estocástico oculto que no podemos observar directamente
- Un segundo proceso estocástico que produce la secuencia de observaciones dada el primer proceso.

Ejemplo

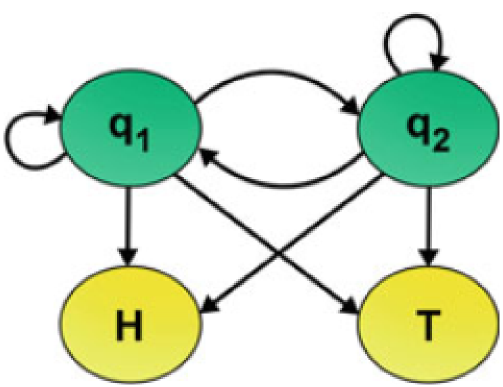
Se tienen dos monedas M_1 y M_2 que se seleccionan en forma aleatoria

Cada moneda está cargada:

- M_1 $P = 0.8$ de salir cara
- M_2 $P = 0.8$ de salir sello

Se tiran en secuencia las moneda y sólo se observa la salida (C o S)

$$\Pi = \begin{array}{cc} M_1 & M_2 \\ \hline 0.5 & 0.5 \end{array} \qquad A = \begin{array}{c|cc} & M_1 & M_2 \\ \hline M_1 & 0.5 & 0.5 \\ M_2 & 0.5 & 0.5 \end{array} \qquad B = \begin{array}{c|cc} & M_1 & M_2 \\ \hline H & 0.8 & 0.2 \\ T & 0.2 & 0.8 \end{array}$$



Especificación de un HMM

- Conjunto de estados: $Q = \{q_1, \ldots, q_n\}$ y de posibles observaciones $O = \{o_1, \ldots, o_m\}$
- Una vector de probabilidades iniciales $\Pi = \{\pi_1, \pi_2, \ldots, \pi_n\}$, donde $\pi_i = P(S_0 = q_i)$
- Un matriz de probabilidades de transición $A = \{a_{ij}\}$, donde $a_{ij} = P(S_t = q_j | S_{t-1} = q_i)$
- Un vector de probabilidades de salida por cada estado (matriz) $B = \{b_{ik}\}$, donde $b_{ik} = P(O_t = o_k | S_t = q_i)$
- En forma compacta: $\lambda = \{A, B, \Pi\}$.

Propiedades de un HMM

Un HMM (de primer orden) satisface las siguientes propiedades:

- Propiedad de Markov:
  $P(S_t = q_j | S_{t-1} = q_i, S_{t-2} = q_k, \ldots) = P(S_t = q_j | S_{t-1} = q_i)$

- Proceso estacionario: $P(S_{t-1} = q_j | S_{t-2} = q_i) = P(S_t = q_j | S_{t-1} = q_i)$ y
  $P(O_{t-1} = o_k | S_{t-1} = q_j) = P(O_t = o_k | S_t = q_i), \forall (t)$

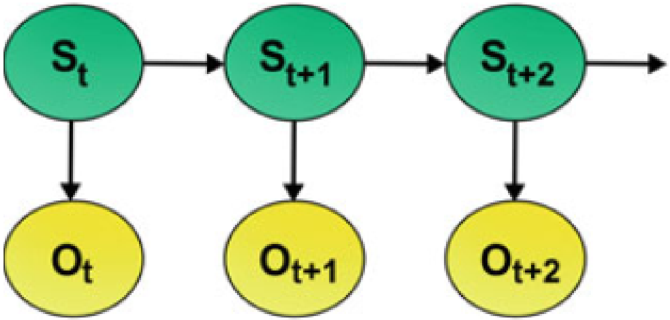
- Independencia de las observaciones:
  $P(O_t = o_k | S_t = q_i, S_{t-1} = q_j, \ldots) = P(O_t = o_k | S_t = q_i)$

  La propiedad de Markov implica que la probabilidad del estado actual solo depende del estado anterior y es independiente del resto de la historia.

  Por proceso estacionario implica que las probabilidades de transición y observación no cambian con el tiempo.

  La independencia de las observaciones especifica que las observaciones solo dependen del estado actual.

Modelo de Grafo de un HMM



Dada una representación HMM de un determinado dominio, hay tres preguntas básicas que son de interés en la mayoría de las aplicaciones :

1. Evaluación: dado un modelo, calcular la probabilidad de una secuencia de observaciones.

2. Secuencia óptima: dado un modelo y una secuencia de observación particular, calcular la secuencia de estado más probable que produjo las observaciones.

3. Aprendizaje de parámetros: dada una serie de secuencias de observaciones, ajustar los parámetros del modelo.

Métodos de evaluación

La evaluación consiste en determinar la probabilidad de una secuencia de observaciónes, $O = \{o_1, o_2, o_3, \ldots\}$, dado un modelo λ, es decir, estimar $P(O|\lambda)$.

Presentamos dos métodos:

- Método directo
- Método interactivo

Método directo

Dada la secuencia de observaciones, $O = \{o_1, o_2, o_3, \ldots o_T\}$, generados por diferentes secuencias de estados, Q_i. Entonces la probabilidad de una secuencia de observaciones se puede estimar como:

$P(O|\lambda) = \sum_i P(O, Q_i|\lambda)$

Para obtener $P(O, Q_i|\lambda)$ se realiza la siguiente operación:

$$P(O, Q_i|\lambda) = \pi_1 b_1(o_1) a_{12} b_2(o_2) \ldots a_{(T-1)T} b_T(o_T)$$

Considerando todas las secuencias: $P(O) = \sum_Q P(O, Q_i)$, que es lo mismo que:

$$P(O|\lambda) = \sum_Q \pi_1 b_1(o_1) a_{12} b_2(o_2) \ldots a_{(T-1)T} b_T(o_T)$$

Para un modelo con N estados y una longitud de observación de T, hay N^T posibles secuencias de estados. Cada término de la suma requiere operaciones $2T$. Como resultado, la evaluación requiere una serie de operaciones del orden de $2T \times N^T$.

Método iterativo

La idea básica del método iterativo, también conocido como *Forward*, es estimar las probabilidades de los estados/observaciones por paso de tiempo.

Es decir, calcula la probabilidad de una secuencia parcial de observaciones hasta el tiempo t (forma inicial $t = 1$) y en base a este resultado parcial, calcula el tiempo $t + 1$ y así sucesivamente.

Definimos una variable auxiliar llamada forward:

$$\alpha_t(i) = P(o_1, o_2, \ldots, o_t, S_t = q_i|\lambda)$$

Es decir, la probabilidad de que la secuencia parcial de observaciones hasta el tiempo t, esté en el estado q_i en el tiempo t.

Algoritmo Forward

El algoritmo iterativo consta de tres partes principales:

- Inicialización: aquí se obtienen las variables α para todos los estados en el tiempo inicial.

$$\alpha_1(i) = P(O_1, S_1 = q_i) = \pi_1 b_i(O_1)$$

- Inducción consiste en calcular $\alpha_{t+1}(i)$ en términos de $\alpha_t(i)$ esto se repite desde $t = 2$ a $t = T$.

$$\alpha_t(j) = [\sum_i \alpha_{t-1}(i) a_{ij}] b_j(O_t)$$

- Terminación : consiste en el cálculo de $P(O|\lambda)$.

$$P(O) = \sum_i \alpha_T(i)$$

Complejidad

- Cada iteración requiere N multiplicaciones y N adiciones (aproximadamente)
- Para T iteraciones, el número de operaciones es del orden de $N^2 \times T$.
    - La complejidad del tiempo se reduce de exponencial en T para el método directo a lineal en T y cuadrática en N para el método iterativo.
- En la mayoría de las aplicaciones $T \ggg N$.

  El procedimiento iterativo que se acaba de describir es la base para resolver las otras dos preguntas para los HMM.

Estimación de estados

Encontrar la secuencia de estados más probable para una secuencia de observación, $O = \{o_1, o_2, o_3, \ldots\}$, se puede interpretar de dos formas:

- Obteniendo el estado más probable S_t en cada paso de tiempo t
- Obteniendo la secuencia de estados más probable $s_0, s_1, \ldots, s_t$.

Obteniendo el estado más probable S_t en cada paso de tiempo t

- Variable backward: $\beta_t(i) = P(o_{t+1}, o_{t+2}, \ldots, o_T, S_t = q_i|\lambda)$
- En forma interativa:

$$\beta_t(i) = \sum_j \beta_{t+1}(j) a_{ij} b_j(o_t)$$

- La variables β para T se definen como $\beta_T(j) = 1$.

Estado óptimo para un cierto tiempo t

- Combinando ambas definiciones e iterar hacia adelante y hacia atrás, coincidimos en un tiempo intermedio t :

  $$P(O, s_t = q_i|\lambda) = \alpha_t(i)\beta_t(i)$$

  La ecuación anterior se puede escribir: $P(O|\lambda) = \sum_i \alpha_t(i)\beta_t(i)$

- Definimos una variable adicional λ es decir, la probabilidad condicional de estar en un cierto estado q_i dada la secuencia de observación:

  $$\gamma_t(i) = P(S_t = q_i|O, \lambda) = P(S_t = q_i, O|\lambda)/P(O)$$

  que se puede escribir en términos de α y β como:

  $$\gamma_t(i) = \alpha_t(i)\beta_t(i)/\sum_i \alpha_t(i)\beta_t(i)$$

  Esta variable γ, proporciona la respuesta al primer problema, el estado más probable (MPS) en un momento t. Encontremos para qué estado tiene el valor máximo:

  $$MPS(t) = Arg \max_i \gamma_t(i)$$

Obteniendo la secuencia de estados más probable Q dada la secuencia de observación O

Secuencia total más probable: $P(Q, O|\lambda)$

El método para obtener la secuencia de estados óptima se conoce como **algoritmo de Viterbi**, que de forma análoga al algoritmo forward, resuelve el problema de forma iterativa.

Variable necesaria δ, la subsecuencia de estados óptimos hasta el tiempo t:

$$\delta_t(i) = \max[P(s_1, s_2, \ldots, s_t = q_i, o_1, o_2, \ldots, o_t|\lambda)]$$

que también se puede obtener de forma iterativa:

$$\delta_{t+1}(i) = [\max \delta_t(i) a_{ij}] b_j(o_{t+1})$$

El algoritmo de Viterbi requiere cuatro fases: inicialización, recursión, terminación y backtracking.

Requiere una variable adicional, $\psi_t(i)$, que almacena para cada estado i en cada paso de tiempo t el estado anterior que dio la máxima probabilidad.

Algoritmo de Viterbi

1. Inicialización

   $$\delta_1(i) = \pi_i b_i(O_1)$$

   $$\psi(i) = 0,\ i = 1 \dots N$$

2. Recursión

   $$\delta_t(j) = \max_i[\delta_{t-1}(i)a_{ij}]b_j(O_t)$$

   $$\psi_t(j) = arg\max_i[\delta_{t-1}(i)a_{ij}],\ t = 2 \dots T,\ j = 1 \dots N$$

3. Terminación

   $$P^* = \max_i[\delta_T(i)]$$

   $$q_T^* = arg\max_i[\delta_T(i)]$$

4. Backtracking

   $$q_{t-1}^* = \psi_t(q_t^*),\ t = T \dots 2$$

Métodos de aprendizaje

Consiste en determinar los parámetros del modelo $\lambda = \{A, B, \Pi\}$, dada una secuencia de observaciones

Para ello, se buscan los parámetros que maximicen $P(\mathbf{O}|\lambda)$.

Para un HMM con N estados y M observaciones necesitamos estimar $N + N^2 + N \times M$ parámetros, para A, B y Π respectivamente.

Algoritmo de Baum-Welch

Necesitamos definir una variable auxiliar ξ, la probabilidad de una transición de un estado i en el tiempo t a un estado j en el tiempo $t+1$ dada una secuencia de observación O:

$$\xi_t(i,j) = P(s_t = q_i, s_{t+1} = q_j | O, \lambda) = P(s_t = q_i, s_{t+1} = q_j, O | \lambda)/P(O)$$

En términos de α y β:

$$\xi_t(i,j) = \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)/P(O)$$

Algoritmo de Baum-Welch

Escribiendo $P(O)$ en términos de α y β:

$$\xi_t(i,j) = \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) / = \sum_i \sum_j \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)$$

La variable γ se puede calcular en términos de ξ como: $\gamma_t(i) = \sum_j \xi_t(i,j)$

Esta variable sumada sobre todos los tiempos da una estimación del número de veces que se pasa por el estado *i*: $\sum_t \gamma_t(i)$

Mientras que la suma sobre t de $\xi_t(i,j)$ da una estimación del número de transiciones de *i* a *j*: $\sum_t \xi_t(i,j)$.

Re-estimación de los parámetros

1 . Probabilidades iniciales-número de veces en el estado *i* en $t = 1$:

$$\pi_i = \gamma_1(i)$$

2 . Probabilidades de transición – número de transiciones dede el estadi *i* al estado *j* entre el número de veces en *i*

$$a_{ij} = \sum_t \xi_t(i, j) / \sum_t \gamma_t(i)$$

3 . Probabilidades de observaciones – número de veces en el estado *j* y observar *k* entre el número de veces en dicho estado:

$$b_{jk} = \sum_{t,O=k} \gamma_t(j) / \sum_t \gamma_t(j)$$

Principio EM

Problema: Necesitamos los parámetros del modelo para el algoritmo de Baum-Welch

Solución: El algoritmo EM

- Se inicia con unos parámetros iniciales para el modelo $\{A, B, \Pi\}$, que pueden inicializarse aleatoriamente o en base a algún conocimiento del dominio.
- Por el algoritmo de Baum-Welch, estos parámetros se vuelven a estimar y se van mejorando iterativamente.
- Este ciclo se repite hasta la convergencia, es decir, hasta que la diferencia entre los parámetros del modelo de un paso al siguiente esté por debajo de cierto umbral.
- El algoritmo EM proporciona un estimador de máxima verosimilitud

- No se garantiza el óptimo global, sin embargo, este estimador tiende a funcionar bien en la práctica.

  Este algoritmo pertenece a la familia de métodos EM.

Puedes revisar el trabajo de Eugene Weinstein, Expectation-Maximization Algorithm and Application sobre este tema.

Fin!