Modelos de Markov Escondidos (HMM)

Quantil S.A.S.

Juan Pablo Lozano

Julio - 2014

Definición HMM

- Un HMM está compuesto por dos procesos estocásticos:
  - Un proceso "escondido" que consta de las variables de estado S_t. Las llamaremos estados.
  - Un proceso "observado" de variables Y_t las cuales son generadas por la variable S_t. Las llamaremos observaciones.
- Ambos procesos cumplen la propiedad de Markov.
- Se asume que las variables escondidas son discretas. S_t puede tomar K valores enteros.
- Las variables observadas pueden tomar valores discretos como reales.

HMM Cont...

Basandose en las propiedades Markovianas de los HMM, la probabilidad conjunta de una secuencia de estados y observaciones es la siguiente:

$$P(S_{1:T}, Y_{1:T}) = P(S_1)P(Y_1|S_1)\prod_{t=2}^{T} P(S_t|S_{t-1})P(Y_t|S_t)$$

Esta factorización de la probabilidad conjunta se ilustra de la siguiente manera:

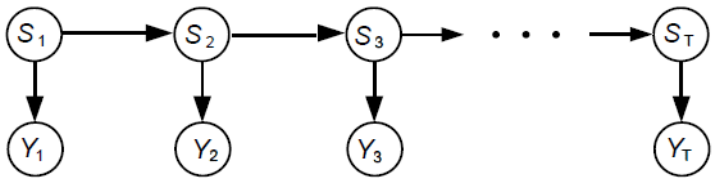


Figure: Representación Grafica HMM

HMM Cont...

¿Qué se necesita para calcular esta probabilidad?

- Distribución de probabilidad inicial $P(S_1)$.
- Matriz $K \times K$ de trancisión.
- Distribución de output para definir $P(Y_t|S_t)$

Se puede incluir un proceso de inputs U_t de las cuales dependa la matriz de transición, $P(S_t|S_{t-1}, U_t)$

Redes Bayesianas

Definición: Una red bayesiana es una representación gráfica de las independencias condicionales de un conjunto de variables aleatorias.

Independencia Condicional: A es condicionalmente independiente de B dado C si $P(A, B|C) = P(A|C)P(B|C)$ para todo A, B, C tal que $P(C) \neq 0$.

De manera más general: Dos conjuntos de nodos A, B son independientes condicionalmente dado C si C *d-separa* a A y B. Es decir, si para todo camino indirecto entre A y B, existe un nodo D tal que: 1) D tiene flechas convergentes y ni D ni sus descendientes están en C o 2) D está en C pero no tiene flechas convergentes.

Ej: $P(W, X, Y, Z) = P(W)P(X)P(Y|W)P(Z|X, Y)$

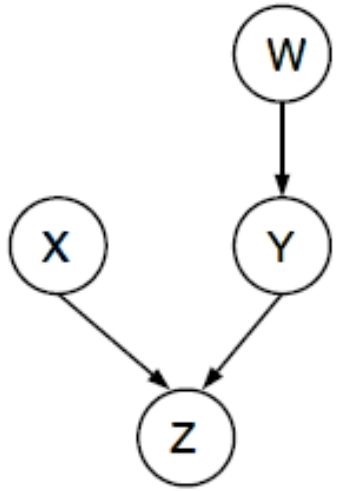
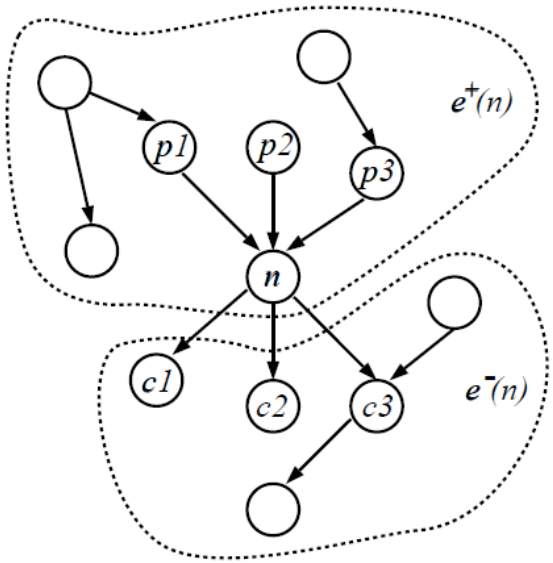


Figure: Ejemplo Red Bayesiana

Con base en este ejemplo se puede ver que W es condicionalmente independiente de X dado $C = \{Z, Y\}$

Evidencia y Belief Propagation

- Algoritmos para calcular probabilidades conjuntas y marginales (Belief Propagation)
- Teniendo los valores de algunas variables en la red, el objetivo es actualizar la probabilidad marginal de todas las variables en la red para incorporar esta evidencia.
- Mensajes locales



$$P(n|e) \propto \left[\sum_{p_1,\ldots,p_k} P(n|p_1,\ldots,p_k) \prod_{i=1}^{k} P(p_i|e^+(p_i) \right] \prod_{j=1}^{l} P(c_j, e^-(c_j)|n)$$

En el ejemplo se tiene, por ejemplo, lo siguiente si se observa que $Z = z, X = x$.

$$P(Y|Z = z, X = x) \propto P(Y)P(Z = z|X = x, Y)P(X = x)$$

Redes Bayesianas Dinámicas y Modelos Espacio-Estado

Redes Bayesianas Dinámicas: Son redes bayesianas para modelar series de tiempo (HMM).

Modelos Espacio-Estado: Son HMM en los cuales las variables observadas son vectores D-dimensionales con valores reales. Lo especial de estos modelos es lo siguiente:

- La probabilidad de transición $P(X_t|X_{t-1})$ se puede descomponer en componentes deterministicos y estocásticos. Es decir: $X_t = f_t(X_{t-1}) + w_t$.
- De igual manera $P(Y_t|X_t)$ se puede descomponer como $Y_t = g_t(X_t) + v_t$

Si f_t y g_t son lineales e invariantes bajo el tiempo, y los ruidos tienen distribución Normal entonces el modelo se conoce como Modelos Espacio-Estado lineal Gaussiano.

$$X_t = AX_{t-1} + w_t; \qquad Y_t = CX_t + v_t$$

Elementos de HMM

- Matriz de transición entre estados A.
- Matriz de probabilidades de observación B.
- Distribución de probabilidad inicial π
- Número de estados N
- El número de valores de los estados K.
- El número de valores de las observaciones M
- Tiempo T.

Preguntas sobre HMM

Dada la estructura de un HMM, surgen tres problemas claves para que estos modelos se implementen en la vida real.

1. Teniendo una secuencia de observaciones $Y = Y_1, \ldots, Y_T$ y un modelo $\mathfrak{M} = (A, B, \pi)$, ¿Cómo se calcula $P(Y|\mathfrak{M})$, la probabilidad que las observaciones se ajusten al modelo?
2. Teniendo una secuencia de observaciones $Y = Y_1, \ldots, Y_T$ ¿Cómo se escoge una secuencia de estados $S_1, \ldots, S_T$ que sea óptima en algún sentido?
3. ¿Cómo ajustar los parámetros del modelo $\mathfrak{M} = (A, B, \pi)$ para maximizar $P(Y|\mathfrak{M})$?

Algoritmo Forward-Backward

Objetivo: Calcular $P(S_t|Y_1, \ldots, Y_T)$

- El algoritmo se divide en dos pasos: el paso Forward y el paso Backward.
- En el paso Forward se define $\alpha_t(S_t) = P(S_t, Y_1, \ldots, Y_t)$ y la idea es encontrar cada $\alpha_t(S_t)$ de forma iterativa.
- En el paso Backward se definen $\beta_t(S_t) = P(Y_{t+1}, \ldots, Y_T|S_t)$ y de igual manera, se quieren encontrar de forma iterativa.
- Al final obtendremos que

$$P(S_t|Y_1, \ldots, Y_T) \propto \alpha_t(S_t)\beta_t(S_t)$$

Luego es un algoritmo que permite hacer inferencia sobre los estados basandose en las variables observables.

Algoritmo Forward

Objetivo: Lograr calcular $P(S_t, Y_1, \ldots, Y_t), \forall t$ sabiendo $\mathfrak{M} = (A, B, \pi)$.

Algoritmo:

1. Se inicializa $\alpha_1(S_1) = \pi_{S_1} P(Y_1|S_1)$
2. Para $t = 1, \ldots, T-1$ y $1 \leq S_t \leq K$

$$\alpha_t(S_t) = \sum_{S_{t-1}} \alpha_{t-1}(S_{t-1}) P(St|S_{t-1}) P(Y_t|S_t)$$

Lo interesante de este algoritmo es que la complejidad computacional es del orden de $O(TK^2)$, en vez de $O((2T-1)K^T$.

Algoritmo Backward

Objetivo: Lograr calcular $P(Y_{t+1}, \ldots, Y_T | S_t), \forall t$ sabiendo $\mathfrak{M} = (A, B, \pi)$.

Algoritmo:

1. Se inicializa $\beta_T(S_T) = 1$
2. Para $t = T-1, \ldots, 1$ y $1 \leq S_t \leq K$

$$\beta_t(S_t) = \sum_{St} \beta_{t+1}(S_{t+1}) P(St+1 | S_t) P(Y_{t+1} | S_{t+1})$$

Este algoritmo también tiene complejidad computacional del orden de $O(TK^2)$.

Algoritmo Forward-Backward Cont...

- De los pasos del algoritmo se pueden llegar a 3 posibles formas de calcular $P(Y|\mathfrak{M})$
  1. $P(Y|\mathfrak{M}) = \sum_{S_T=1}^{K} \alpha_T(S_T)$.
  2. $P(Y|\mathfrak{M}) = \sum_{S_1=1}^{K} \beta_1(S_1)$.
  3. $P(Y|\mathfrak{M}) = \sum_{S_t=1}^{K} \alpha_t(S_t)\beta_t(S_t)$.
- Luego de hacer los pasos Forward y Backward iterativamente para todo t, se derivan las siguientes variables:

$$\gamma_t(S_t) = P(S_t|Y_1, \ldots, Y_T) = \frac{\alpha_t(S_t)\beta_t(S_t)}{\sum_{S_t=1}^{K} \alpha_t(S_t)\beta_t(S_t)}$$

$$\eta_t(S_t) = \frac{\alpha_{t-1}(S_{t-1})P(S_t|S_{t-1})P(Y_t|S_t)\beta_t(S_t)}{\sum_{S_t,S_{t-1}} \alpha_{t-1}(S_{t-1})P(S_t|S_{t-1})P(Y_t|S_t)\beta_t(S_t)}$$

Algoritmo de Viterbi

Objetivo: Encontrar la cadena de estados más probable S^* teniendo en cuenta las observaciones. Es decir, encontrar $S^* = \underset{S_1,\ldots,S_T}{argmax}\ P(S_1,\ldots,S_T|Y_1,\ldots,Y_T)$.

Algoritmo:

- Inicializar un $\mu_1(S_1) = \pi_{S_1} P(Y_1|S_1)$ y $\Psi_1 = 0$.
- Para $2 \leq t \leq T$

$$\mu_t(S_t) = \max_{S_t}\ P(Y_t|S_t)P(S_t|S_{t-1})\mu_{t-1}(S_{t-1})$$

$$\Psi_t(S_t) = \underset{S_t}{argmax}\ P(S_t|S_{t-1})\mu_{t-1}(S_{t-1})$$

Algortimo de Viterbi Cont...

- Se definen $P^* = \max_{S_T} \mu_T(S_T)$ y $i_T^* = \operatorname{argmax}_{S_T} \mu_T(S_T)$.
- Ahora se hace un retroceso para poder calcular la cadena de estados usando el siguiente método: para $t = T-1, \ldots, 1$ tenemos que $i_t^* = \Psi_{t+1}(i_{t+1}^*)$.

Algoritmo EM

El algoritmo EM es el encargado de estimar los parámetros del modelo para que se maximice la probabilidad que las observaciones vengan de un modelo con estos parámetros.

Se basa en usar la logverosimilitud, encontrar una función que sea cota inferior de ella y empezar a estimar parámetros para llegar lo más cerca posible a la logverisimilitud.

La logverosimilitud es entonces: $\mathfrak{L}(\theta) = logP(Y|\theta) = log\sum_S P(Y,S|\theta)$.

Para cualquier distribución $Q(S)$ sobre los estados se tiene:

$$\mathfrak{L}(\theta) \geq \sum_S Q(S) logP(S,Y|\theta) - \sum_S Q(S) logQ(S) = \mathfrak{F}(Q,\theta)$$

Algoritmo EM Cont...

Los pasos del EM son entonces los siguientes en el caso general:

- **Paso E:** $Q_{k+1} \leftarrow \underset{Q}{argmax}\ \mathfrak{F}(Q, \theta_k)$.
- **Paso M:** $\theta_{k+1} \leftarrow \underset{\theta}{argmax}\ \mathfrak{F}(Q_{k+1}, \theta)$

El máximo en el paso E se obtiene cuando $Q_{k+1}(S) = P(S|Y, \theta_k)$, lo que hace que se de la siguiente igualdad $\mathfrak{F}(Q_{k+1}, \theta_k) = \mathfrak{L}(\theta_k)$.

Entonces como antes del paso M $\mathfrak{F}(Q_{k+1}, \theta_k) = \mathfrak{L}(\theta_k)$ y el paso E no cambia a θ, se garantiza que este método no disminuye la verosimilitud luego de cada paso combinado del algoritmo. Es decir, $P(Y|\overline{\mathfrak{M}}) \geq P(Y|\mathfrak{M})$

EM para HMM

Para el caso de HMM el algoritmo EM se simplifica sustancialmente usando los resultados obtenidos del Algoritmo Forward-Backward.

Si se aplica el logaritmo a la probabilidad conjunta de estados y observaciones se obtiene lo siguiente:

$$logP(S_{1:T}, Y_{1:T}) = logP(S_1) + \sum_{t=1}^{T} logP(Y_t|S_t) + \sum_{t=2}^{T} logP(S_t|S_{t-1})$$

Si representamos el estado S_t como un vector unitario K-dimensional. (Ej. $S_t = [0, 0, 1, 0, \ldots, 0]^T$ significa que en el tiempo t el valor de S_t es 3.)

EM para HMM Cont...

Usando estas convenciones se tiene lo siguiente:

- $logP(S_t|S_{t-1}) = S_t^T log\boldsymbol{\phi} S_{t-1}$.
- $logP(S_1) = S_1^T log\pi$.
- $logP(Y_t|S_t) = Y_t^T (logE) S_t$

Ahora, usando resultados anteriores se tiene que:

- $\pi_i = \gamma_{1,i}(S_{1,i})$
- $E_{d,i} = \frac{\sum_{t=1}^{T} Y_{t,d}\gamma_{t,i}}{\sum_{t=1}^{T} \gamma_{t,i}}$
- $\boldsymbol{\phi_{i,j}} = \frac{\sum_{t=2}^{T} \eta_{t,i,j}}{\sum_{t=2}^{T} \gamma_{t,i}}$

Problemas y Generalizaciones

Aunque los HMM son bastante útiles para hacer inferencia y aprendizaje de máquinas, pueden llegar a tener limitaciones en cuanto al número de estados posibles y como el costo computacional que está ligado a esto. Para poder sobrepasar estas limitaciones existen extensiones de HMM.

1. HMM Factoriales.
2. HMM con estructura de árbol.
3. Modelos de espacio con cambio de estado.

Extensiones de HMM

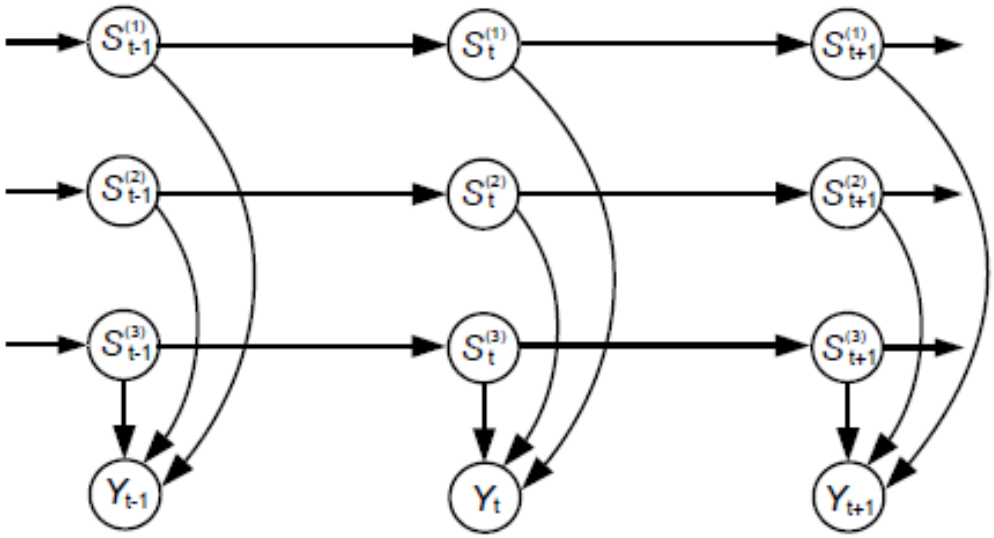


Figure: Representación Grafica HMM Factorial

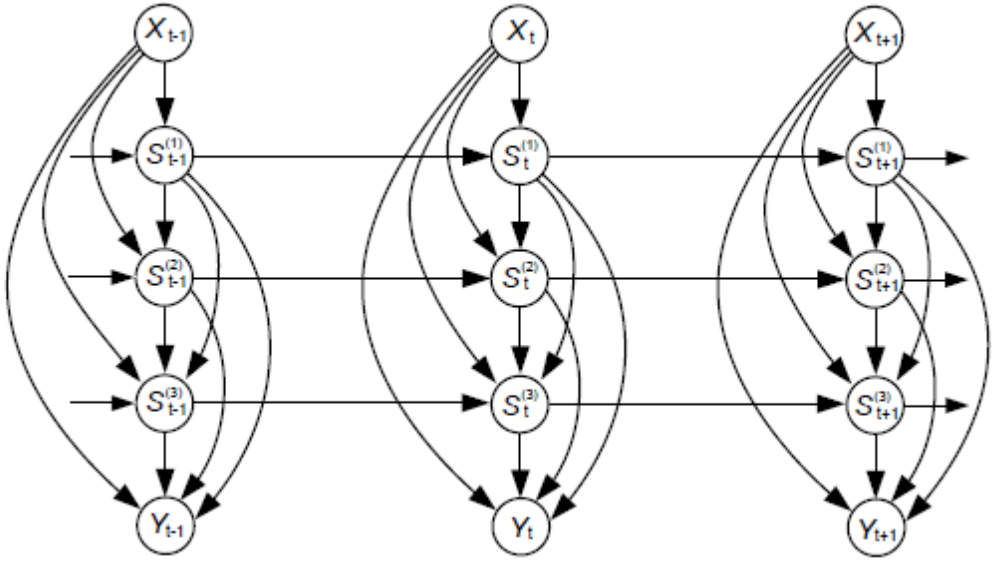


Figure: Representación Grafica HMM Árbol

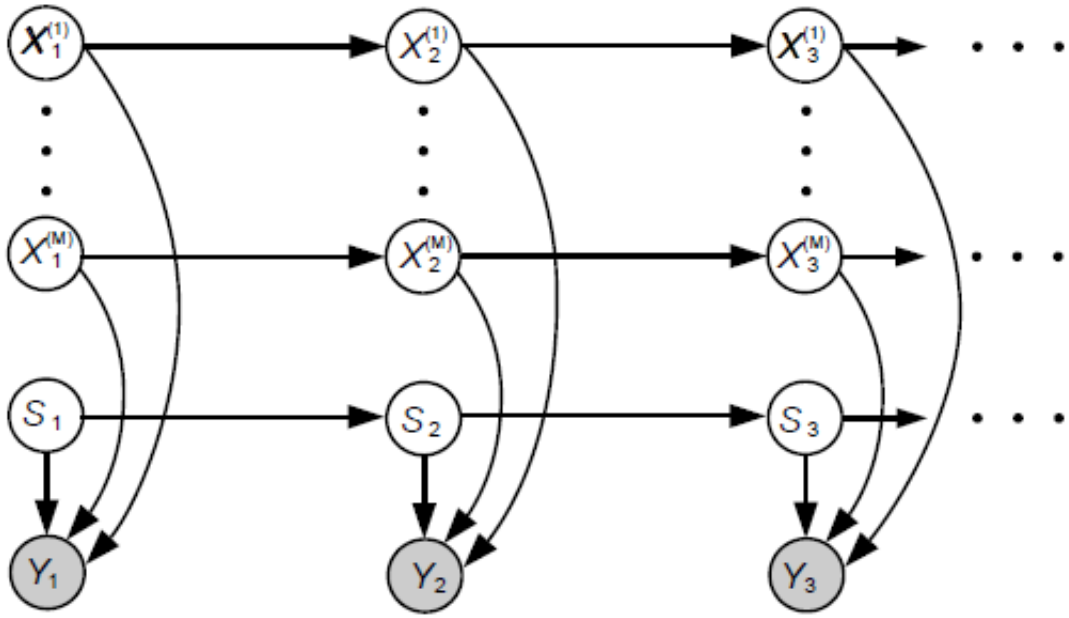


Figure: Representación Grafica HMM Switch

Inferencia Aproximada e Intractabilidad

- El problema con las extensiones de HMM es que algunas probabilidades se vuelven casi imposibles de calcular por medios convencionales por su complejidad computacional(Ej. El Algoritmo Forward-Backward para HMM Factoriales tiene orden $O(TMK^{M+1})$.
- Existen dos métodos para calcular probabilidades usando aproximaciones (Inferencia Aproximada):
  1. Muestreo de Gibbs: Actualización estocástica de las variables de estado muestreando cada una usando la probabilidad condicional del estado condicionado a estados cercanos a este.
  2. Métodos Variacionales: Se define una distribución paramétrica Q y hacer variar los parámetros de esta distibución para aproximar la distribución P.

Estructura de Modelo

Hay dos problemas muy importantes ligados al aprendizaje de HMM: overfitting y la selección y estructura del modelo (número de estados, formas de las matrices de transición y de observación). Existen tres métodos que permiten ayudar a lidiar con estos problemas:

1. Cross-Validation.
2. Regularización.
3. Integración Bayesiana (Monte Carlo, Aproximación de Laplace y el método variacional bayesiano)

Aplicaciones

- Clustering suave (Modelo de mezcla Gaussiana).
- Reconocimiento de voz.
- Posicionamiento de objetos.
- Predicción de texto.
- Motores de busqueda (Google).

Referencias

- Rabiner L.R., Juang B.H., 1986. An Introduction to Hidden Markov Models.
- Gahramani Z., 2001. An Introduction to Hidden Markov Models and Bayesian Networks.
- Dempster A., Laird N., Rubin D., 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm.
- Kim J., Pearl J., 1983. A computational Model for Causal and Diagnostic Reasoning in Inference Systems.