Statistique pour la bio-informatique

Séance 9-10 - Decembre 2003

Chaînes de Markov cachées

1 Chaînes de Markov cachées et applications

Les modèles à données latentes (ou manquantes ou cachées) constituent des outils puissants pour modéliser des systèmes dont la dynamique effectue des transitions entre différents états impossible à observer directement. Dans une chaîne de Markov cachée, les différents états d'un système peuvent être caractérisés par un nombre fini de valeurs. On passe alors de l'état s_i à l'état s_j avec la probabilité p_{s_i,s_j} lors d'une transition. Dans chaque état, le système est susceptible émettre un symbole o pris dans un alphabet $\mathcal{O}$ fini ($\mathcal{O}$ pour observable). La probabilité d'émission du symbole o peut dépendre de l'état s. Nous la notons $q_{s,o}$.

Les algorithmes dédiés aux chaînes de Markov cachées sont des algorithmes d'estimation statistique. Etant donnée une suite d'observations de longueur T, $o_1, \ldots, o_T$, ils ont pour objectif typique d'estimer la suite d'états $s_1, \ldots, s_n$ la plus probable. Pour cela, il faudra ajuster correctement les paramètres du modèles $P = (p_{s_i,s_j})$ et $Q = (q_{so})$ à partir d'un ensemble de n séquences dont les états sont connus.

Le premier objectif est généralement rempli par l'algorithme de **Viterbi**. Le second objectif est rempli par l'algorithme EM, dont la version spécifique aux CMC s'appelle algorithme de **Baum-Welch**.

1.1 Applications

Les applications des CMC (ou d'autres modèles à structure latente comme les réseaux de neurones) sont très nombreuses en bio-informatique. Nous illustrons cette approche à l'aide de l'exemple classique la recherche de gènes que nous simplifierons à l'extrême (cf logiciel `genscan` de Burge et Karlin, 1997).

1.2 Algorithmique des chaînes de Markov cachées

Dans cette section, nous notons $\mathcal{S}$ l'ensemble des états cachés et S_t la chaîne associée

$$\forall s_1, s_2 \in \mathcal{S}, \quad p_{s_1,s_2} = \mathrm{P}(S_{t+1} = s_2 \,|\, S_t = s_1).$$

Nous notons π la loi initiale de la chaîne

$$\pi_s = \mathrm{P}(S_1 = s).$$

Nous notons $\mathcal{O}$ l'ensemble des états observables. Conditionnellement à $S_t = s$, la donnée X_t est donc issue de la loi

$$\forall o \in \mathcal{O}, \quad \mathrm{P}(X_t = o \,|\, S_t = s) = q_{s,o}.$$

Ayant observé une séquence de longueur T, $o_1, \ldots, o_T$, la vraisemblance du paramètre multidimensionnel

$$\theta = (\pi, P, Q)$$

est égale à

$$L(\theta) = \mathrm{P}(o_1, \ldots, o_T; \theta).$$

1.2.1 Algorithme forward

La vraisemblance $L(\theta)$ correspond à la vraisemblance incomplète d'un modèle à données manquantes

$$L(\theta) = \sum_{s_1,\ldots,s_T} \mathrm{P}(o_1, \ldots, o_T \,|\, s_1, \ldots, s_T; Q)\, \mathrm{P}(s_1, \ldots, s_T; (\pi, P)).$$

Précisement, nous avons

$$L(\theta) = \sum_{s_1,\ldots,s_T} \pi_{s_1} q_{s_1,o_1} p_{s_1,s_2} q_{s_2,o_2} \cdots p_{s_{T-1},s_T} q_{s_T,o_T} \ .$$

Cette formule suggère un algorithme de calcul naïf, dont la complexité de l'ordre $O(T(\#\mathcal{S})^T)$ rendrait le coût rapidement prohibitif. La solution provient d'un algorithme de programmation dynamique. Il repose sur le calcul de la grandeur

$$\forall s \in \mathcal{S}, \quad \alpha_t(s) = \mathrm{P}(o_1, \ldots, o_t, S_t = s).$$

Cette grandeur représente la probabilité d'observer $o_1, \ldots, o_t$ avec l'état au temps t, $S_t = s$.

Proposition 1.1 Algorithme forward. *Soit $o_1, \ldots, o_T$ une suite d'observations provenant d'une CMC. Posons*

$$\forall s \in \mathcal{S}, \quad \alpha_1(s) = \pi_s q_{s,o_1}$$

et, pour tout $t = 2, \ldots, T$,

$$\forall s_t \in \mathcal{S}, \quad \alpha_t(s_t) = \sum_s \alpha_{t-1}(s) p_{s,s_t} q_{s_t,o_t}.$$

Nous avons

$$L(\theta) = \sum_{s \in \mathcal{S}} \alpha_T(s)$$

L'algorithme de calcul associé est de complexité de l'ordre de $O(T(\#\mathcal{S})^2)$.

Démonstration.

□

De manière un peu moins naturelle, mais complètement équivalente, nous pouvons considérer une variable qui remonte le sens du temps. Cette variable est appelée **variable backward**

$$\forall s \in \mathcal{S}, \quad \beta_t(s) = \mathrm{P}(o_{t+1}, \ldots, o_T \,|\, S_t = s).$$

Cette grandeur représente la probabilité d'observer $o_T, \ldots, o_{t+1}$ conditionnellement à $S_t = s$.

Proposition 1.2 Algorithme backward. *Soit* $o_1, \ldots, o_T$ *une suite d'observations provenant d'une CMC. Posons*

$$\forall s \in \mathcal{S}, \quad \beta_T(s) = 1$$

et, pour tout $t = 1, \ldots, T-1$,

$$\forall s_t \in \mathcal{S}, \quad \beta_t(s_t) = \sum_s \beta_{t+1}(s) p_{s_t,s} q_{s,o_{t+1}}.$$

Nous avons

$$L(\theta) = \sum_{s \in \mathcal{S}} \pi_s \beta_1(s) q_{s,o_1}$$

L'algorithme de calcul associé est de complexité de l'ordre de $O(T(\#\mathcal{S})^2)$.

Démonstration.

□

1.2.2 Algorithme de Viterbi

L'algorithme de Viterbi permet de calculer la suite d'états cachés la plus probable vu les observations $o_1, \ldots, o_T$

$$s_1^*, \ldots, s_T^* = \arg\max \mathrm{P}(o_1, \ldots, o_T \cap s_1, \ldots, s_T; \theta)$$

Notons que la valeur max est appelée **score de Viterbi**. On l'obtient formellement en remplaçant la somme par le maximum dans l'expression de la vraisemblance incomplète $L(\theta)$. Rechercher le maximum de manière naïvement énumérative conduit à un algorithme de complexité exponentiellement croissante en la longueur des observations ($O(T(\#\mathcal{S})^T)$). Comme dans la section précédente, nous pouvons construire un algorithme complexité quadratique $O(T(\#\mathcal{S})^2)$. Cet algorithme, dit **algorithme de Viterbi** s'obtient simplement en remplaçant la somme par le max.

Proposition 1.3 Algorithme de Viterbi. *Soit $o_1, \ldots, o_T$ une suite d'observations provenant d'une CMC. Posons*

$$\forall s \in \mathcal{S}, \quad v_1(s) = \pi_s q_{s,o_1}$$

et, pour tout $t = 2, \ldots, n$,

$$\forall s_t \in \mathcal{S}, \quad v_t(s_t) = \max_s \{v_{t-1}(s) p_{s,s_t} q_{s_t,o_t}\}.$$

Nous avons

$$s_T^* = \arg\max_s v_T(s)$$

et, pour tout $t = T-1, \ldots, 1$,

$$s_t^* = \arg\max_s \{v_t(s) p_{s,s_{t+1}^*}\}.$$

Démonstration.

□

1.2.3 Exercices

Exercice 1. On pose

$$\forall s \in \mathcal{S}, \quad \gamma_t(s) = \mathrm{P}(S_t = s \,|\, o_1, \ldots, o_T; \theta).$$

Montrer que

$$\forall s \in \mathcal{S}, \quad \gamma_t(s) = \frac{\alpha_t(s)\beta_t(s)}{L(\theta)}.$$

Exercice 2. On pose

$$\forall s_1, s_2 \in \mathcal{S}, \quad n_t(s_1, s_2) = \mathrm{P}(S_t = s_1; S_{t+1} = s_2 \,|\, o_1, \ldots, o_T; \theta).$$

Montrer que

$$\forall s_1, s_2 \in \mathcal{S}, \quad n_t(s_1, s_2) = \frac{\alpha_t(s_1) p_{s_1,s_2} q_{s_2,o_{t+1}} \beta_{t+1}(s)}{L(\theta)},$$

et

$$\forall s \in \mathcal{S}, \quad \gamma_t(s) = \sum_{s_2} n_t(s, s_2).$$

1.2.4 Algorithme de Baum-Welch

L'algorithme de Baum-Welch est un algorithme d'estimation itératif dérivé de l'algorithme EM. Plutôt que de détailler les calculs (fastidieux) conduisant aux itérations, nous tentons d'en expliquer les aspects intuitifs. Afin d'estimer le paramètre du modèle

$$\theta = (\pi, P, Q)$$

nous disposons pour l'estimation de n séquences de longueur T. Notons O^i les séquences observées, et $S^{(i)}$ les séquences cachées. Les valeurs initiales S_1 sont choisies de manière arbitraire. Etant données des valeurs initiales de π, P, et Q, nous pouvons calculer une première estimation de π

$$\bar{\pi}(s) = \frac{1}{n} E[\#\{i\,;\, S_1^i = s\} \,|\, \{O^{(i)}\}],$$

une première estimation de P

$$\bar{p}_{s,s'} = \frac{E[N_{s,s'} \,|\, \{O^{(i)}\}]}{E[N_s \,|\, \{O^{(i)}\}]}$$

et une première estimation de Q

$$\bar{q}_{s,o} = \frac{E[N_s(o) \,|\, \{O^{(i)}\}]}{E[N_s \,|\, \{O^{(i)}\}]}.$$

Dans ces équations, $N_{s,s'}$ est le nombre de fois où l'état s est suivi de l'état s'

$$N_{s,s'} = \#\{i, t\,;\, S_t^i = s; S_{t+1}^i = s'\},$$

N_s est le nombre de fois où l'état s apparaît

$$N_s = \#\{i, t\,;\, S_t^i = s\},$$

et $N_s(o)$ est le nombre de fois où l'état s génère l'observation o

$$N_s(o) = \#\{i, t\,;\, S_t^i = s; O_t^i = o\}.$$

Les espérances précédentes peuvent être facilement calculées en terme des variables forward et backward lorsque l'on introduit les variables indicatrices des événements que l'on cherche à compter. On obtient

$$\bar{\pi}(s) = \frac{1}{n}\sum_{i=1}^{n} \gamma_1^{(i)}(s),$$

et

$$\bar{p}_{s,s'} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T-1} n_t^{(i)}(s,s')}{\sum_{s'}\sum_{i=1}^{n}\sum_{t=1}^{T-1} n_t^{(i)}(s,s')}.$$

Pour calculer l'espérance $E[N_s(o) \mid \{O^{(i)}\}]$, il suffit de sommer sur les observations ayant donné un symbole o

$$\bar{q}_{s,o} = \frac{\sum_{i=1}^{n}\sum_{t;O_t^{(i)}=o}\sum_{s'} n_t^{(i)}(s,s')}{\sum_{s'}\sum_{i=1}^{n}\sum_{t=1}^{T-1} n_t^{(i)}(s,s')}.$$