Accès à l'information - Cours sur les modèles structurés

François Yvon

7 décembre 2016

0.1 EM : une application simple

On considère un réseau bayésien très simple comprenant trois variables A, B, C, la loi jointe se factorisant par :

$$\mathrm{P}(A, B, C) = \mathrm{P}(A)\,\mathrm{P}(B)\,\mathrm{P}(C \mid A, B)$$

Les paramètres de ce modèle : π_A, π_B, π_{ABC}.

On observe $\{(B^i, C^i), i = 1 \dots N\}$, ($A$ n'est jamais observé).

B	1	0	1	1	1	0	0
C	1	0	1	0	1	0	1

Écrivez :

1. La vraisemblance
2. La fonction auxiliaire de l'algorithme EM
3. Les équations de mise à jour de l'algorithme EM

0.2 Quelle est la langue de ce mot ?

On observe des séquences $\{T^n_{[1:x]}, n = 1 \dots N\}$ (par exemple des noms de personnes). On suppose que ces séquences sont issues d'un mélange de modèles de Markov (par exemple un modèle de Markov par langue). La probabilité d'une séquence est donnée par :

$$\mathrm{P}(x_{[1:T]}) = \sum_h \mathrm{P}(x_{[1:T]} \mid h; \boldsymbol{\theta})\,\mathrm{P}(h \mid \boldsymbol{\theta}) = \sum_h \prod_t \mathrm{P}(x_t \mid x_{t-1}, h, \boldsymbol{\theta})\,\mathrm{P}(h \mid \boldsymbol{\theta})$$

1. Dessiner la représentation graphique de ce modèle.
2. Quels sont les paramètres de ce modèle ?
3. La règle qui permet de trouver la langue la plus probable quand on connaît une séquence et les paramètres.
4. Écrivez :
   — La vraisemblance
   — La fonction auxiliaire de l'algorithme EM
   — Les équations de mise à jour de l'algorithme EM

Déchiffrer avec EM

1. Implémentez la méthode de (Knight et al, 2006) [1] [Figure 2] pour déchiffrer des codes de substitution simples.
2. Testez-la en utilisant les ressources disponibles sur la site du cours. Vous pourrez comparer vos résultats
   — avec des textes de taille croissante : test10.cy, test100.cy, test500.cy
   — en utilisant des modèles bigramme / trigramme
   — appris uniquement sur des textes anglais, uniquement en français, ou les deux.
3. Comment pourriez-vous étendre la méthode pour également prendre en compte des suppressions de lettres ?

Hint : la liste des symboles qui sont touchés par le codage :
— chiffres : "0123456789"
— minuscules : "abcdefghijklmnopqrstuvwxyzîâôûêéèàëïüöç"
— majuscules : "ABCDEFGHIJKLMNOPQRSTUVWXYZÎÂÔÛÊÉÈÀËÏÜÖÇ"
— ponctuations : " : ;, ? ! ;-.()"
(en particulier les saut de lignes sont conservés).

Il n'est pas exclu que cet exercice soit transformé en test. Faites le sérieusement.

0.3 Regarder la télé

Des petits films de @HugoLarochelle : `https://www.youtube.com/watch?v=9UFXF5EJ4Ek` sur l'alignement de mots.

Aligner avec IBM2

Les ressources fournies contiennent des textes parallèles (deux romans de Jules Verne). Vous pouvez essayer de les aligner en utilisant le logiciel d'alignement `fast_align`, qui implémente une variante vraiment très efficace du modèle IBM 2. `fast_align` peut être récupéré à cette adresse : `https://github.com/clab/fast_align`.

1. L'article de Kevin Knight et al est ici :
`http://www.aclweb.org/anthology/P/P06/P06-2065.pdf`