MVA, Projet PGM: Rapport initial
Factorial HMM

Théis Bazin Valentin De Bortoli Élie Michel

3 janvier 2017

Le but de ce rapport est de présenter l'avancement de nos travaux à la date du 9 décembre 2016. Nous avons divisé notre présentation en trois parties : compréhension mathématique du modèle, implémentation et choix des données.

1 Modèle mathématique : Factorial HMM

La plus grande partie de notre temps a été dédiée à la compréhension des mathématiques sous-jacentes au modèle. Nous commençons par une courte justification. Nous aurions pu modéliser des données observables dépendant de M variables cachées pouvant prendre K valeurs comme un modèle probabiliste graphique de type HMM, *Hidden Markov Model*, avec une variable cachée pouvant prendre K^M valeurs.

Néanmoins cela est très coûteux et ne prend pas en compte le découplage supposé de nos M variables cachées. Pour cela les auteurs de l'article proposent une variante de HMM : *Factorial HMM*. Nous présentons en Figure 1 et Figure 2 les graphes dirigés dans lesquels se factorisent les deux modèles.

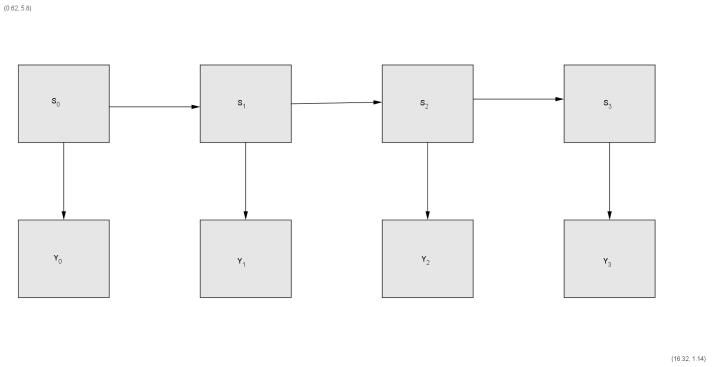

Figure 1 – Hidden Markov Model

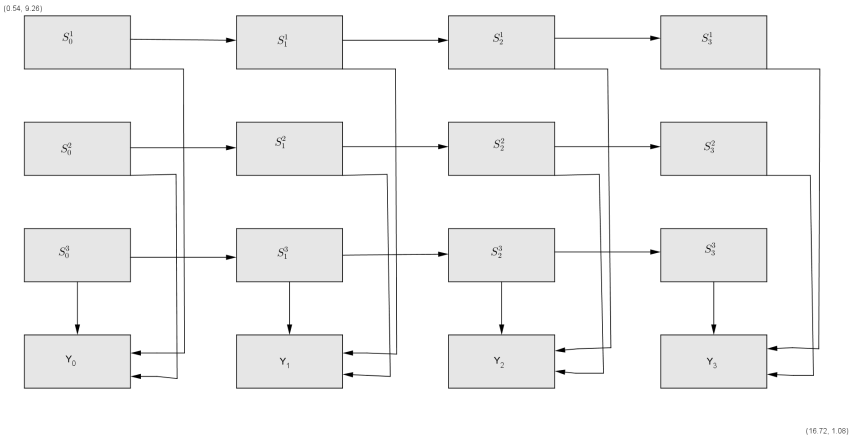

Figure 2 – Factorial Hidden Markov Model

Il s'agit alors, comme dans le modèle HMM, de faire de l'inférence sur les paramètres de notre modèle. Pour cela, nous utilisons l'algorithme EM. L'étape de maximisation, *M-step*, est bien comprise et très similaire à celle de HMM. L'étape de calcul d'espérance, *E-step*, est quant à elle différente puisque nous ne pouvons a priori pas utiliser l'algorithme *sum-product*, car il n'y a pas ici de structure d'arbre.

On note cependant que l'on a tout de même une structure particulière (M couches de chaînes de Markov), ce qui permet d'établir un algorithme de calcul exact des probabilités. Malheureusement, ce calcul exact est très coûteux. Nous considérons donc des approximations de l'espérance, basées sur trois méthodes :

1. L'échantillonnage de Gibbs,
2. *Mean-Field*,
3. *Structured Mean-Field*.

Les particularités mathématiques de chacun des modèles ont été étudiées. L'échantillonnage de Gibbs est une implémentation classique d'un algorithme de type *Markov Chain Monte Carlo* (MCMC). *Mean-Field* rend tous les états des variables cachées indépendants et actualise les paramètres en conséquence. Dans *Structured Mean-Field*, on conserve la structure de chaîne de Markov pour chaque couche, tout en retirant la dépendance entre chaque couche.

Un problème mathématique se pose : les paramètres du modèle sont actualisés comme étant des points fixes d'une certaine fonction compliquée à calculer. Les auteurs ne proposent pas de solution pour trouver ces points fixes, y a-t-il une méthode exacte ou devons-nous utiliser des algorithmes itératifs ?

2 Implémentation

Concernant l'implémentation nous allons nous concentrer sur l'inférence de paramètres du modèle. Les algorithmes seront écrits en Python. L'implémentation n'a pas encore été travaillée ; il nous a semblé plus judicieux de s'intéresser au devoir maison qui traite du cas où $M = 1$: HMM. La structure de notre code s'organisera comme indiqué en Figure 3.

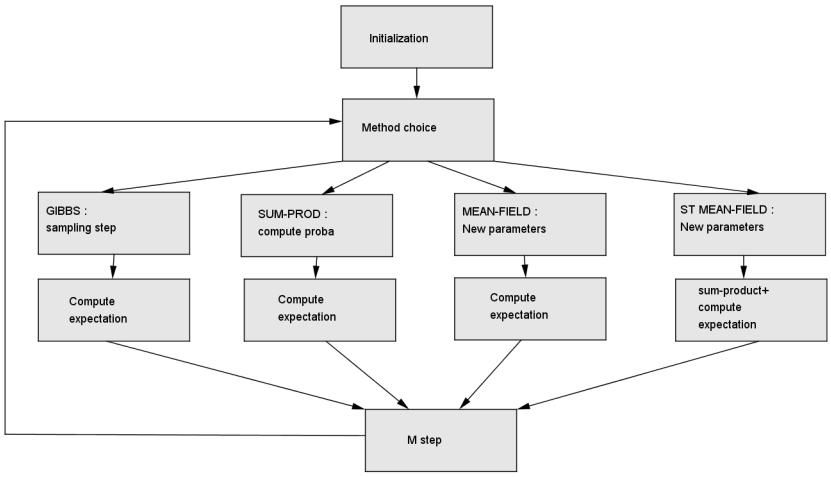

FIGURE 3 – Organisation de l'implémentation

3 Choix des données

Tout d'abord, notre algorithme d'inférence de paramètres sera testé sur des données simulées, puisqu'il est facile de faire un modèle génératif de FACTORIAL HMM. Nous tenterons également de reproduire les résultats obtenus par les auteurs sur les chorales de Bach. Un autre but est également de tester notre algorithme sur des données réelles issues de la sociologie (enquête 2013, "Génération 2010" sur l'évolution en début de carrière de jeunes ayant quitté le système scolaire). Les axes d'étude pour ce dernier jeu de données n'ont pas encore été établis.