Algorithmes d'apprentissage pour les chanes de Markov caches

Alexis Jacq

July 5, 2015

1 Un peu de philo...

Pour proposer une explication un phnomne, dans le but de le prdire ou bien de l'exploiter, l'Homme a recours des modles simplifis de ce phnomne. Ces modles consistent projeter le phnomne dans un formalisme accessible afin de le rendre comprhensible et descriptible. Ainsi, les reprsentations cognitives du monde qui se forment dans nos cerveaux, nos langages, nos religions et enfin nos modles scientifiques sont autant de modles abords pour s'entendre sur une comprhension – d'abord personnelle puis commune – de notre environnement.

L'apprentissage consiste adapter o bien affiner les modles choisis afin d'optimiser leur capacit dcrire les phnomnes observs. Toujours dans les mmes exemples : au niveau du cerveaux, l'apprentissage neuronal nous permet de cerner des concepts toujours plus pertinents et de mmoriser leur dynamique. Au niveau des langages, nous nous efforons dacqurir, par les changes verbaux, un vocabulaire toujours plus adapt a nos socits et nos meurs, puis l'on suppose ce vocabulaire assez efficace pour l'utiliser et ainsi le tester. Dans les religions, nous cherchons "grandir dans la foi", ce qui consiste en un perptuel effort introspectif d'adaptation aux convictions spirituelles et d'acceptation des idologies. Enfin, en science, la dmarche est plus simple dcrire : tablir et supposer des modles mcanistiques puis affiner leur paramtrages pour les adapter la description du phnomne tudi. L'apprentissage se fait donc en deux tapes : l'**acceptation** du modle, puis son **adaptation**.

Ces deux tapes ne sont absolument pas disjointes : en adaptant le modle d'une certaine faon, on le modifie, et on en accepte un nouveau, etc... L'apprentissage peut donc être vu comme un permanent saute-mouton entre acceptation et adaptation : on accepte le modle pour tablir une prdiction, on adapte le modle selon l'cart entre ce qui a t prdit et ce qui est advenu, puis on accepte le modle adapt et on recommence etc... Il est important de voir que tout les algorithmes ou autres dmarches visant un apprentissage reprennent ce schma d'acceptation/adaptation.

2 Un peu de formalisme...

Nous nous plaons ici dans le cadre d'un formalisme mathmatique. Comme dans la sance prtendante [**?**], nous nous intressons une *classe* de modles trs simple et trs maniables : tant donn le phnomne observ $X \rightarrow Y^{\text{obs}}$ o l'on cherche expliquer la variable d'intrt Y^{obs} avec la variable X, on tablit le modle suivant :

$$\mathcal{M}_P : X \mapsto Y$$

o $\mathcal{M}_P$ dsigne le modle (dpendent du jeu de paramtre P), vu comme une application qui X (variable connue) associe Y (la prdiction). Ici, adapter le modle, c'est optimiser les paramtres P dans le but de minimiser l'erreur de prdiction, savoir l'“cart” entre Y prdit et Y^{obs} observ. Comme nous l'avons vu prcdemment, la dfinition de cet cart va beaucoup dpendre du contexte. Dans cette sance, nous allons nous intresser un contexte stochastique. La mesure d'erreur de prdiction sera donc dfinie par la vraisemblance. Cette quantit a dj t introduite dans la sance prcdente ([**?**] section 2.3.1). Rappelons juste qu'il s'agit de la capacit du modle *expliquer* l'observation Y^{obs}. On la note $\mathcal{L}$ (pour *likelihood* = vraisemblance) :

$$\mathcal{L}(\mathcal{M}_P, X, Y^{\text{obs}}) = \mathbb{P}[Y^{\text{obs}} | \mathcal{M}_P, X]$$

3 ...Maintenant on peut y aller

Dans cette sance, nous tudions un modle trs particulier : les chaînes de Markov caches (HMM pour Heidden Markov Model). Mais avant tout : qu'est-ce qu'une chane de Markov ?

3.1 Approche non rigoureuse

Prenons un processus purement alatoire, par exemple un ”pile ou face”. Pourquoi un *processus* ? Parce qu'on va faire plusieurs jets de suite, ce qui apporte une dimension temporelle. De ce fait, on va pouvoir les nommer par indice, par ex j_1 le premier jet, j_2 le deuxime etc... A chaque fois j_i vaut pile ou face selon le rsultat obtenu. Pourquoi *purement* alatoire ? En fait, un jet de pice n'est absolument pas alatoire : il dpend de la force et du point d'application du coup de pouce, de la position initiale dans l'espace, de la topographie du sol, peut-être même de la temprature... Mais notre chelle, tant incapable d'tablir ces liens de causalit, on prfre modliser le jet de pice comme un vnement ne dpendant d'absolument rien. Rien ne permet de prdire si le rsultat sera pile ou face.

Maintenant, regardons un autre processus. On continue jeter successivement des pices. Sauf que cette fois-ci, on compte un score. Supposons que l'on joue pile : si le rsultat d'un lancer donne pile, on gagne un point, sinon on perd un point. Indions l'tat de ce score pour tudier son avancement : s_1 pour le score aprs le premier jet, s_2 aprs le second etc... Ainsi, s_n reprsente le score au bout de n jets :

$$\begin{gathered}
s_1 = \mathbf{1}_{[j_1=\text{pile}]} - \mathbf{1}_{[j_1=\text{face}]} \\
s_2 = s_1 + \mathbf{1}_{[j_2=\text{pile}]} - \mathbf{1}_{[j_2=\text{face}]} \\
\ldots \\
s_n = s_{n-1} + \mathbf{1}_{[j_n=\text{pile}]} - \mathbf{1}_{[j_n=\text{face}]}
\end{gathered}$$

Cette fois-ci, on voit que s_n dpend de s_{n-1} qui est connu l'instant n. En effet si l'instant $n-1$, le score est $s_{n-1} = 10$ on sait que s_n vaudra 11 ou 9 selon j_n. En outre, sans la connaissance de s_{n-1} on ne pourrai rien prdire de plus que “s_n sera compris entre -n et $+n$ (inclus) et aura la mme parit que n".

Remarquons autre chose : si on connaît s_{n-1}, a ne sert rien de connaître s_{n-2} pour obtenir des informations propos de s_n. En effet, les informations apportes par s_{n-2} sont inclues dans les informations apportes par s_{n-1}. Par ex, si on connaît $s_{n-2} = 9$ on sait que s_n vaudra 7, 9, ou 11, tandis que si on connaît $s_{n-1} = 10$, on sait bien que $s_n \in \{7, 9, 11\}$ mais plus encore : $s_n \in \{9, 11\}$. On appelle cela "proprit d'oubli" ou bien "proprit de Markov". C'est dire que si on connaît s_{n-1} et s_{n-2}, on peut *oublier* s_{n-2} qui ne sert plus rien.

Plus formellement, on dit que la **loi** (c'est dire la nature du modle stochastique) de s_n **sachant** s_{n-1}, s_{n-2}, ... , s_2, s_1 est la même que celle de s_n **sachant** seulement s_{n-1}.

Il est trs frquent de rencontrer des modles stochastiques vrifiant cette proprit. Le premier exemple, et sans doute le plus simple, est celui des **marches alatoires**. Il faut imaginer un homme complêtement soûl errant dans une ville. Tantt il fait un pas vers la gauche, puis vers l'avant, puis un autre vers la gauche suivit d'un pas en arrire. chaque instant (les instants ici reprsentent les pas successifs) la position courante seule donne autant d'information sur la position future que l'intgralit de l'errance depuis la sortie du pub. En fait, les marches alatoires se gnralisent comme des accumulations additives de ralisation d'une variable alatoire. On peut donc les dfinir par rcurrence comme suit :

$$z_n = z_{n-1} + x_n$$

o z_n est la n-me position de la marche, et x_n la n-me ralisation d'une variable alatoire. Vous l'avez sans doûte devin, le processus dfini par les scores de jets de pice est donc une marche alatoire. Mais attention : toutes les chaînes de Markov ne sont pas des marches alatoires ! En effet, on pourrait trs bien dfinir la chaîe suivante (cette rcurrence s'appelle le *critre fondamental des chaînes de Markov*, toute chaîne de Markov peut s'crire sous cette forme et n'importe quelle suite vrifiant cette proprit dfinie une chaîne de Markov):

$$z_n = f(z_{n-1}, x_n, n)$$

o f est mesurable de $\mathbb{R}^2 \times \mathbb{N}$ dans $\mathbb{R}$, mais a (les mesures et tout) on le verra plus loin, si a vous intresse. Pour l'instant, nous allons nous contenter de la dfinition suivante (je vous rassure, elle est on ne peux plus rigoureuse et suffisante) :

3.2 Approche rigoureuse des chaînes de Markov

Je vais copier-coller (Ha, si seulement on pouvait rellement copier-coller du LateX !) l'introduction aux chaînes de Markov du Pr. Jean Jacod que l'on retrouve en dbut de ses poly de cours Jussieu ([**?**],[**?**]) :

L'ide des chaînes de Markov est trs simple : il s'agit d'une suite $(X_n)_{n\in\mathbb{N}}$ de variables alatoires valeurs dans un espace mesurable $(E,\mathcal{E})$, dfinies sur un espace de probabilit $(\Omega,\mathcal{F},\mathbb{P})$, et telle que pour tout n on ait la proprit suivante :

$$\left.\begin{array}{r}\text{Conditionnellement la valeur de } X_n\text{, les variables } (X_0,...,X_{n-1})\\ \text{d'une part, } (X_{n+1},X_{n+1},...)\text{ d'autre part, sont indpendantes.}\end{array}\right\}\tag{1}$$

Ce modle permet de rendre compte d'un trs grand nombre de situations concrtes.

Encore une fois, nous ne nous intresserons pas pour l'instant la thorie de la mesure (qui pourrait trs bien – et ce serait une superbe ide – tre la thmatique d'un GT part entire). Donc pour vous expliquer la phrase ... valeurs dans un espace mesurable $(E,\mathcal{E})$, dfinies sur un espace de probabilit $(\Omega,\mathcal{F},\mathbb{P})$..., je vous dirais tout simplement : "*cela veut dire que tout est bien dfinie et utilisable dans le contexte voulu*". C'est une phrase qui permet d'carter tout phnomne pathologique exotique qui pourrait se glisser dans notre dfinition. Un peu comme une formule de marabou [**?**] pour repousser les mauvais esprits.

Cette dfinition se rcrit en termes plus symboliques :

Definition/ Proprit [*chaîne de Markov*] : Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables alatoires dfinies *blablabla* et valeur dans l'espace mesurable *blablabla*. $(X_n)_{n\in\mathbb{N}}$ dfinie une chaîne de Markov si et seulement si $\forall\ n$:

$$\mathbb{P}[X_{n+1},X_{n+2},...|X_0,...X_{n-1},X_n]=\mathbb{P}[X_{n+1},X_{n+2},...|X_n]\tag{2}$$

Cette dernire proprit (2) est tout bonnement la **proprit de Markov faible "gnralise"** (*faible* car il y a une version forte qui se gnralise un domaine d'application plus tendue (elle est super complique donc je ne prendrais pas le risque de l'introduire) et *gnralis* car elle conserne "tout l'avenir" sachant "tout le pass"). Dans ce chapitre, nous ne nous intressons qu'aux chaînes de Markov **homognes**. C'est dire que la loi d'un vnement futur X_{n+1} ou de n'importe quelle runion dvnements futurs $(X_{n+1},X_{n+2},...)$ conditionnellement un vnement de dpart X_n reste la mme quelque soit l'instant n. En d'autres termes :

Proprit [*homognit*] :

$$\forall n,\quad \mathbb{P}[X_{n+1},X_{n+2},...|X_n]=\mathbb{P}[X_1,X_2,...|X_0]\tag{3}$$

On tablie donc la **rgle d'oubli** des chaînes de Markov homognes :

proprit [*rgle d'oubli*] :

$$\mathbb{P}[X_{n+1}, X_{n+2}, ...|X_0, ...X_{n-1}, X_n] = \mathbb{P}[X_1, X_2, ...|X_0] \tag{4}$$

Cette dernire proprit sera trs importante par la suite. Notons que l'homognit peut aussi se voir comme la simplification du critre fondamentale : $(Z_n)_{n\in\mathbb{N}}$ est une chaîne de Markov homogne si et seulement si ses ralisations $(z_n)_{n\in\mathbb{N}}$ vrifient la rcurrence suivante :

$$z_n = f(z_{n-1}, x_n)$$

o f est mesurable de $\mathbb{R}^2$ dans $\mathbb{R}$ et les $(x_n)_{n\in\mathbb{N}}$ sont les ralisations d'une suite de variables alatoires $(X_n)_{n\in\mathbb{N}}$ toutes de mme loi et telle que toutes les runions possibles $(X_n)_{n\in\mathbb{I}\subset\mathbb{N}}$ soient indpendantes de Z_0.

3.3 Chaînes de Markov caches

Voil, normalement on ne devrait pas avoir besoin de plus de notions que a pour aborder les HMM. Pour ceux qui veulent approfondir les outils et ustensiles des chaînes de Markov, je conseillerai le poly du GT de Franois Bienvenu sur les modles matriciels de populations [**?**] (ou bien, videment, un vrai bouquin de votre BU, mais e sera probablement super srieux et ennuyeux alors que nos notes nous sont vraiment cool). Nous pouvons maintenant nous poser la question qui nous dmangeait tant : *Mais o se cachent donc les chanes de Markov caches* ?

Bon, posons-nous une question un peu plus fine : Qu'est-ce qu'une chane de Markov *cache* ? C'est avant tout un modle. Et la meilleur faon d'introduire un modle, c'est de prsenter un phnomne trs simple qui se modlise trs bien avec. Je ne ferais pas la mme erreur que celle de mon professeur de premire anne de master qui tait all chercher la probabilit d'apparition des lots CpG proximit des promoteurs dans l'ADN. Ce phnomne d'apparition de ces lots non-mthyls est trs intressant modliser avec des chanes de Markov caches mais bon... Il y a vraiment plus simple comme intro...

Prenons deux mecs bourrs la sortie d'un bar. Nous allons faire la supposition suivante : l'un (Pierrot) est vraiment foutu et entame donc une marche alatoire. L'autre (Robert) a un peu moins bu : il essaye comme il le peut de suivre Pierrot pour ne pas le perdre de vue. Mais bon, il en a quand mme un bon coup dans le nez et les choses sont moins simples qu'elles en ont l'air : il n'y voit plus rien. Pierrot-alatoire se sent rassur de la prsence bienveillante de son ami Robert qui essaye de le suivre. Mais il ne peut pas s'arrter de marcher et a beaucoup de mal contrler ses jambes. Lorsqu'il s'carte de son ami, il pousse un cri de dtresse "HAAAAOOOO". Quand il s'en rapproche, il pousse un cri rassur "HOOOOOOO". Robert utilise ces cris pour le reprer et le suivre.

Le problme est le suivant : Pierrot pousse des cris approximatifs et se trompe parfois.

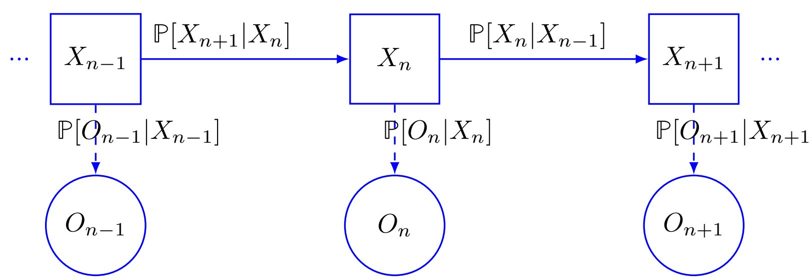
<u>Formalisons cette situation :</u> Il y a deux "tats" possible pour Pierrot : soit il se rapproche (tat x_1), soit il s'carte (tat x_2). Notons X_n la variable alatoire qui reprsente l'tat dans lequel se trouve Pierrot l'instant n. Lorsque Pierrot prend une direction, il a du mal s'en dfaire tout de suite.

Je veux dire par l que lorsqu'il commence s'carter, il essaye de se sortir de cette situation mais c'est difficile. Disons qu'il a, chaque instant n o il s'carte, une probabilit de revenir vers Robert $\mathbb{P}[X_{n+1} = x_1 | X_n = x2] = 0.4$; Bien entendu cela implique la probabilit de continuer s'carter $\mathbb{P}[X_{n+1} = x_2 | X_n = x2] = 0.6$. D'autre part, si Pierrot sent qu'il se rapproche, il fait attention ne pas perdre ce bon plie. Sa probabilit de rester dans cet tat de rapprochement est donc leve $\mathbb{P}[X_{n+1} = x_1 | X_n = x_1] = 0.7$; ce qui implique $\mathbb{P}[X_{n+1} = x_2 | X_n = x_1] = 0.3$. Comme Pierrot est sacrment bourr, chaque instant il oublie le pass. Bref, $\mathbb{P}[X_{n+1} | X_n, X_{n-1} ... X_0] = \mathbb{P}[X_1 | X_0]$. On est clairement face une chane de Markov deux tats. Plus haut, j'ai parl de marche alatoire : il s'agissait de chanes de Markov nombre infini d'tats (la variable alatoire reprsentant l'tat l'instant n de la chane que je notais Z_n pouvait prendre n'importe quelle valeur entire entre $-\infty$ et $+\infty$ selon la grandeur de n alors qu'ici notre variable X_n de la chane-Pierrot prend comme valeur x_1 ou bien x_2 quelque soit n).

Dans une HMM, les probablilits lies la chane dcrivent ce qu'on appelle les **lois de transition** : elles rgissent le passge d'un tat l'autre.

Pour ce qui est des cris que pousse Pierrot, crivons qu'il y a deux cris possibles : o_1 (HOOOOO) et o_2 (HAAAOO). Notons aussi O_n la variable alatoire qui reprsente le cri pouss l'instant n. Nous utilisons des $\mathcal{O}$ parce qu'il s'agit d'$\mathcal{O}$bservations. Disons donc que la probabilit que Pierrot pousse le bon cri quand il est dans l'tat de rapprochement x_1 est $\mathbb{P}[O_{n+1} = o_1 | X_n = x_1] = 0.8$ et que la probabilit qu'il pousse le bon cri quand il s'carte (x_2) est $\mathbb{P}[O_{n+1} = o_2 | X_n = x_2] = 0.7$. Ces probabilit sont appeles **lois d'mission**. Comme on suppose ici (et c'est une trs grosse supposition !) qu'il n'y a que deux cris possible, $\mathbb{P}[O_{n+1} = o_2 | X_n = x_1] = 0.2$ et $\mathbb{P}[O_{n+1} = o_1 | X_n = x_2] = 0.3$.

La manire la plus simple de reprsenter graphiquement ce processus est de dessiner un "peigne" : le temps s'coule de gauche droite et de l'instant n l'instant $n+1$ on passe de l'tat X_n l'tat X_{n+1}. chaque instant, un signal observable (cri) est mis (verticalement). Sur le graphe suivant, les tats sont represents par des carrs, et les signaux observables par des ronds :



Intressons-nous maintenant quelques petites proprits qui dcoulent d'une telle structure...

3.3.1 Quelques proprits qui dcoulent de cette structure

3.4 L'algorithme EM

3.5 L'agorithme de Viterbi

References

[1] Jacq, A., Bienvenu, F. (2013) *Introduction l'apprentissage*. sance GT n.8

[2] Jacod, J. (2003) *Chaînes de Markov, Processus de Poisson et Applications*. Poly. de cours

[3] Jacod, J. (2004) *Processus de Markov, application la dynamique des populations*. Poly. de cours

[4] Metalshine (2047) *Les distances n'existent pas.*

[5] Bienvenu, F. (2013) *Introduction aux modles matriciels de populations*. sance GT n.6,7