Google expliqué par la Belle au Bois Dormant

Manon Bulla, Axel Blampain,

Thomas Duquesne, Antoine Sulon

6A

1. Le paradoxe de la Belle au Bois Dormant

Il s'agit d'un paradoxe probabiliste.

Le dimanche soir, des expérimentateurs endorment la Belle puis font un tirage au sort avec une pièce non truquée:

- Face : ils réveillent (on réveille) la Belle le lundi matin et les expérimentateurs ont un entretien avec elle ;
- Pile : ils réveillent (on réveille) la Belle le lundi matin et les expérimentateurs ont un entretien avec elle, puis ils l'endorment à nouveau avec un traitement qui lui fait oublier la journée du lundi ; les expérimentateurs la réveillent une seconde fois le mardi matin et ont un second entretien avec elle.

La Belle est bien au courant de cette procédure. Lors de chacun des entretiens, on demande à la Belle si, selon elle, on a tiré pile ou face le dimanche soir. Deux points de vue divergent:

- Pour les expérimentateurs, il y a bien une chance sur deux pour tirer pile ou face le dimanche soir;
- Mais les jours suivants la situation est plus subtile : face occasionne deux fois moins de réveils que pile (et aussi les lundis sont bien sûr plus nombreux que les mardis).

À chaque réveil, la Belle fait le raisonnement suivant:

- il existe pour elle trois cas possibles : face-lundi, pile-lundi et pile-mardi ;
- face-lundi et pile-lundi ont la même probabilité (on dit que ces deux cas sont équiprobables) car on suppose la pièce bien équilibrée ;
- pile-lundi et pile-mardi ont la même probabilité aussi car le protocole les rend aussi fréquents l'un que l'autre ;
- ces trois cas (qu'elle ne peut pas dissocier) sont donc équiprobables.

Parmi ces trois cas, un est pour face et deux pour pile. On applique la définition des probabilités : le rapport entre le nombre de cas favorables (pour pile ou face) divisé par le nombre de cas totaux (ici 3).

La Belle répond :

- face avec une probabilité de 1/3
- et pile avec une probabilité de 2/3.

Selon le point de vue, la probabilité diffère ce qui justifie le terme de paradoxe.

2. Les chaines de Markov

Andreï Markov est un mathématicien russe né en 1856 et mort en 1922. Une chaine de Markov est un moyen visuel d'analyser un problème de probabilité. Cette technique est utilisée dans le cas des systèmes sans mémoire.

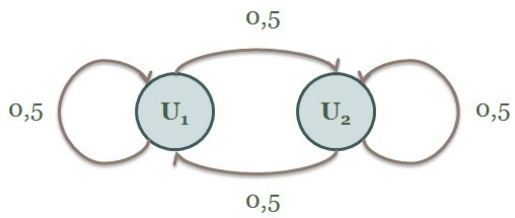
a. Pièce de monnaie non truquée

Lançons une pièce de monnaie plusieurs fois, notons ces instants t, t+1, t+2, t+3, … Et les 2 états possibles de la pièce U1 pour pile et U2 face.

Pour passer de t à t+1, il y a 4 transitions possibles:

- de U1 vers U1, on reste dans le même état, pile-pile
- de U1 vers U2
- de U2 vers U2
- de U1 vers U2

Dans ce cas, on suppose que la pièce est non truquée et que chaque transition a des probabilités d'apparaitre (probabilités de transition) égales. Regroupons en un dessin les 2 états et 4 transitions. Cette représentation est ce qu'on appelle une chaine de Markov.



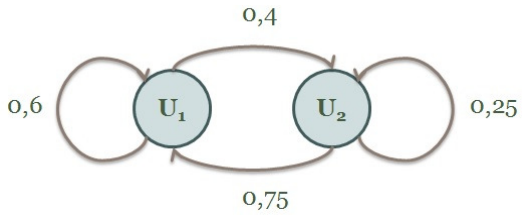
Sur cette chaîne de Markov, les états sont notés par des cercles et les transitions par des flèches partant de l'état initial vers l'état final et ayant une valeur qui est la probabilité de transition. La somme des probabilités sortant d'un état est toujours égale à 1.

A chaque instant t, le processus emprunte l'une des flèches partant de l'état dans lequel il était.

b. Pièce de monnaie truquée

Prenons maintenant une pièce truquée. La pièce étant truquée, on peut voir que les probabilités de transitions ne sont plus égales et la probabilité dépend alors de l'état actuel.

Considérons la chaine de Markov suivante :



Notons la probabilité d'apparition d'un état sous la forme d'un vecteur :

$$\overrightarrow{V_t} = \begin{pmatrix} p_t(U_1) \\ p_t(U_2) \end{pmatrix}$$

Par exemple, pour l'état pile

$$\overrightarrow{V_t} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Et, d'une manière générale :

$$\overrightarrow{V_t} = \begin{pmatrix} p \\ 1-p \end{pmatrix} \text{ où } 0 \leq p \leq 1$$

Dans le cadre de la pièce truquée, on peut calculer :

$$\begin{cases} p_{t+1}(U_1) = 0{,}6.p_t(U_1) + 0{,}75.p_t(U_2) \\ p_{t+1}(U_2) = 0{,}4.p_t(U_1) + 0{,}25.p_t(U_2) \end{cases}$$

Où, sous forme matricielle :

$$\begin{pmatrix} p_{t+1}(U_1) \\ p_{t+1}(U_2) \end{pmatrix} = \begin{pmatrix} 0{,}6 & 0{,}75 \\ 0{,}4 & 0{,}25 \end{pmatrix} \begin{pmatrix} p_t(U_1) \\ p_t(U_2) \end{pmatrix}$$

La matrice

$$T = \begin{pmatrix} 0{,}6 & 0{,}75 \\ 0{,}4 & 0{,}25 \end{pmatrix}$$

est appelée matrice de transition entre l'état t et l'état t+1.

On a donc, pour calculer les probabilités d'apparition des états U1 et U2 aux instants t+1, t+2, … t+n, les relations :

$$V_{t+1} = T.V_t$$

$$\begin{aligned} V_{t+2} &= T.V_{t+1} \\ &= T^2.V_t \end{aligned}$$

$$\vdots$$

$$V_{t+n} = T^n.V_t$$

Dans le cas de la pièce truquée, si l'on donne à n différentes valeurs croissantes, on remarque que le processus converge vers un vecteur de probabilité de transition : $\begin{pmatrix} \frac{15}{23} \\ \frac{8}{23} \end{pmatrix}$

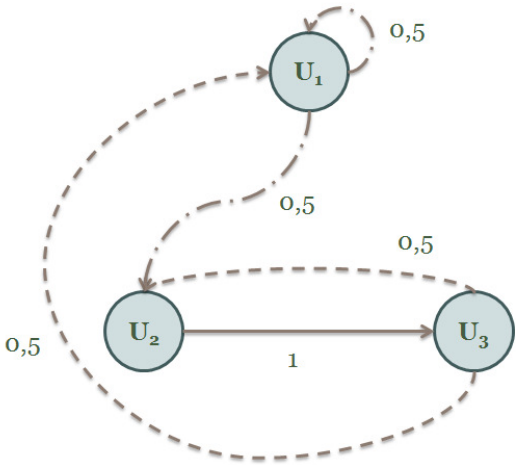
Ce sont les probabilités d'obtenir pile et face de cette pièce truquée. On peut montrer que ces probabilités sont indépendantes de l'état initial (pile ou face).

c. La Belle au bois dormant

Vérifions que la chaine de Markov illustrant le paradoxe reste cohérente:

- U1: lundi-face
- U2: lundi pile
- U3: mardi pile

La chaine et la matrice de transition correspondantes sont :



$$T = \begin{pmatrix} 0{,}5 & 0 & 0{,}5 \\ 0{,}5 & 0 & 0{,}5 \\ 0 & 1 & 0 \end{pmatrix}$$

Après un état U1, lundi face, le test de la semaine est terminé étant donné que les expérimentateurs ne réveillent la Belle une seconde fois, le mardi, seulement s'ils tombent sur pile le dimanche soir. On passe donc au dimanche suivant et on recommence le protocole. Ainsi, après U1 succède un autre état U1, lundi face, avec une probabilité $\frac{1}{2}$ ou un état U2, lundi pile, également avec une probabilité de $\frac{1}{2}$.

Obtenir l'état U2, lundi pile, conduit directement à l'état U3, mardi pile, avec une probabilité de 1 étant donné qu'il n'y a aucune autre manière d'atteindre cet état U3.

Inévitablement, après un état U3, on obtient là aussi la fin du test hebdomadaire. On revient donc encore au dimanche soir. A l'état U3 succède l'état U1, lundi face, avec une probabilité de $\frac{1}{2}$ et l'état U2, lundi pile, avec également une probabilité de $\frac{1}{2}$.

La matrice a été créée en tenant compte des probabilités de tomber sur U1 pour la première ligne, celles de U2 pour la seconde et celles de U3 pour la dernière. De l'état U1, il y a une probabilité de 1/2, de tomber sur U1, de l'état U2, tomber sur U1 est impossible, et de l'état U3, $\frac{1}{2}$ chance de tomber sur U1. Le même raisonnement est appliqué pour U2 et U3.

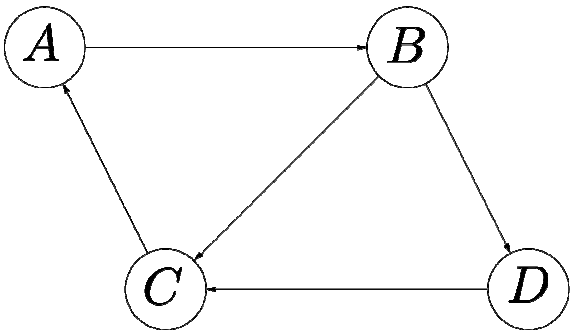
De même que précédemment, l'application de cette matrice de transition à n'importe quel état initial converge vers le vecteur $\begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix}$

3. L'algorithme du PageRank de Google

Cet algorithme permet à Google de trier les pages par une mesure de la probabilité de passer par cette page à partir d'une autre. Il est basé sur le modèle de « *l'internaute aléatoire* » qui, une fois sur une page, va, de manière aléatoire en choisir une autre.

La modélisation de cette marche aléatoire est une chaine de Markov.

Considérons le web (simplissime) suivant (les flèches représentent les liens entre les pages A, B, C et D):



La matrice de transition (traduisant la probabilité de passer d'une page à l'autre) est la suivante :

$$T = \left(\begin{array}{cccc|c} A & B & C & D & \\ \hline 0 & 0 & 1 & 0 & A \\ 1 & 0 & 0 & 0 & B \\ 0 & \frac{1}{2} & 0 & 1 & C \\ 0 & \frac{1}{2} & 0 & 0 & D \end{array}\right)$$

Si l'on part de la page A, l'état initial est:

$$\vec{V}_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

En appliquant la matrice de transition à cet état initial, on converge vers l'état :

$$\vec{V}_n = T^n.\vec{V}_O = \begin{pmatrix} 0.2857 \\ 0.2857 \\ 0.2857 \\ 0.1429 \end{pmatrix}$$

Les composantes de ce vecteur sont une mesure du PageRank de chacune des pages de web envisagé ci-dessus.

4. Conclusions

Nous avons pu, grâce aux chaines de Markov, effleurer la complexité des algorithmes de classement des pages par Google. Bien sûr il reste encore beaucoup de pistes à explorer (notamment pour envisager le cas de pages « terminus » qui ne peuvent aboutir à aucune autre).

5. Sources

- Google PageRank et Chaîne de Markov, PHAN Tran Thanh Du, École Nationale Supérieure de Cognitique, Institut Polytechnique de Bordeaux (23 décembre 2015)
- https://images.math.cnrs.fr/Comment-Google-classe-les-pages-web
- https://images.math.cnrs.fr/Markov-et-la-Belle-au-bois-dormant

Remarque : ce beau projet aurait dû être défendu lors de la 11ème édition du congrès Dédra-math-isons à Louvain-la-Neuve le 21 avril 2020.