Introduction aux processus stochastiques

Projet 1 - Chaînes de Markov en temps discret

Algorithme PageRank

Profs. Céline Esser et Pierre Geurts

Année académique 2018-2019

Ce travail est à réaliser par groupe de 2 étudiants et il doit être remis via la plateforme de soumission de Montefiore pour le dimanche 21/04/2019 à 23h59. L'archive soumise, au format zip, devra contenir les fichiers suivants :

— Votre rapport avec les réponses aux questions au format pdf.

— Le code éventuel permettant de répondre aux questions.

Le travail peut être réalisé au choix en utilisant MATLAB, Mathematica, R, Python, Java, ou C.

Contexte général et objectifs

Lors d'une requête sur le moteur de recherche Google, les pages Web renvoyées à l'utilisateur sont triées en prenant en compte, entre autres, le score *PageRank* de la page Web. Ce dernier dépend uniquement des liens qui existent entre les pages Web et peut être interprété, pour une page donnée, comme la probabilité qu'un surfeur se déplaçant sur le Web consulte à un moment donné cette page.

On peut en effet se représenter le Web comme un graphe où chaque sommet représente une page Web et dans lequel il existe un arc dirigé entre deux pages n_1 et n_2 si la page n_1 contient un lien vers la page n_2. Les déplacements d'un surfeur peuvent alors être vus comme une marche aléatoire sur ce graphe, en supposant qu'à chaque changement de page le surfeur choisit au hasard une des pages référencées par la page sur laquelle il se trouve.

Ce comportement peut se modéliser par une chaîne de Markov dont chaque état possible correspond à une des pages du Web.

Sous certaines conditions, le score *PageRank* d'une page Web est alors la probabilité qu'un surfeur quelconque se trouve sur cette page Web à un moment donné, et peut se modéliser par la distribution stationnaire de la chaîne de Markov représentant la marche aléatoire de surfeurs du Web.

Dans ce projet, on se propose d'utiliser ce problème pour illustrer différents concepts importants relatifs aux processus de Markov en temps discrets et à valeurs discrètes (chaînes de Markov).

1 Etude du modèle de base

Soit la matrice A_1 suivante :

$$A_1 = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \end{pmatrix}$$

représentant la matrice d'adjacence [1] d'un graphe dirigé représentant des liens entre cinq pages Web hypothétiques.

Questions :

1. Soit un surfeur se déplaçant sur ce graphe en choisissant sur chaque page rencontrée un lien au hasard parmi tous les liens présents sur la page et effectuant ce choix indépendamment des pages précédemment visitées. Construisez la matrice de transition $\mathbf{P}$ (avec $p_{i,j} = \mathbb{P}(X_{t+1} = j | X_t = i)$) d'une chaîne de Markov modélisant ce surfeur. Représentez le graphe de transition de la chaîne de Markov correspondante.
2. Calculez (sur ordinateur) les quantités suivantes pour des valeurs de t croissantes :
   — $\mathbb{P}(X_t = i)$ en supposant que le surfeur est parti d'une page choisie au hasard parmi toutes les pages, i.e. la page initiale est tirée dans une loi uniforme discrète.
   — $\mathbb{P}(X_t = i)$ en supposant que le surfeur démarre toujours de la page 1,
   — $\mathbf{P}^t$, c'est-à-dire la t-ième puissance de la matrice de transition.
   Représentez l'évolution des deux premières grandeurs sur un graphe. Discutez et expliquez les résultats obtenus sur base de la théorie.
3. En déduire la distribution stationnaire π_∞ de la chaîne de Markov définie par $[\pi_\infty]_j = \lim_{t\to\infty} \mathbb{P}(X_t = j)$.
4. En supposant que $[\pi_\infty]_j, \forall j = 1, \ldots, 5$, définit le score *PageRank* des 5 pages web du graphe A_1, discutez le classement obtenu sur base de la structure des liens entre les 5 pages.
5. Générez une réalisation aléatoire de longueur T de la chaîne de Markov en démarrant d'une page de votre choix. Calculez pour chaque page le nombre de fois que le surfeur passe par cette page rapporté à la longueur de la réalisation. Observez graphiquement l'évolution de ces valeurs pour chaque page lorsque T croît. Reliez ce résultat à la théorie.
6. Soit les deux nouvelles matrices d'adjacence A_2 et A_3 ci-dessous :

$$A_2 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \qquad A_3 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \end{pmatrix}$$

1. Dans ce projet, l'élément $[A]_{i,j}$ d'une matrice d'adjacence A sera à 1 s'il existe un arc du sommet i vers le sommet j dans le graphe, à 0 sinon. *PageRank* ne prenant pas en compte les auto-référencements, dans tous nos graphes, on supposera qu'il n'y a jamais d'arc d'un sommet vers lui-même (i.e., $[A]_{i,i} = 0\ \forall i$). Nous verrons néanmoins plus loin que le modèle de téléportation ajoutera des liens artificiels de ce type.

A partir des graphes définis par A_2 et A_3, représentez le diagramme d'états des chaînes de Markov correspondant au modèle du surfeur du point 1 et refaites les expériences des point 2 et 5 ci-dessus pour ces deux chaînes. Sur base de ces expériences, discutez des éventuelles limitations du modèle de surfeur présenté au point 1 pour definir le score pagerank d'un page web.

2 Téléportation

En plus des limitations mises en évidence au point 6 de la section précédente, le modèle du surfeur présenté jusqu'ici a une autre limitation importante : Il ne précise pas comment traiter les pages web qui ne contiennent aucun lien (on les appelle des "dangling nodes"),

Le modèle de surfeur utilisé pour calculer le score *PageRank* est dès lors modifié de la manière suivante :

— Un lien artificiel est ajouté de chaque dangling node vers lui-même et vers toutes les autres pages.

— Avec une probabilité α le surfeur a la possibilité de se téléporter vers une page choisie aléatoirement parmi toutes les pages (en ce compris la page où il se trouve). Avec une probabilité $1 - \alpha$, il choisit un lien aléatoire sur la page courante. La téléportation permet de modéliser un utilisateur qui déciderait au bout d'un moment d'entrer une nouvelle URL dans la barre du navigateur plutôt que de suivre un des liens sur la page où il se trouve.

Données : Le fichier `graph_adjmat.csv` contient la matrice d'adjacence G et le fichier `liste_urls.txt` la liste des URLs pour 1000 pages obtenues par un parcours du web en largeur d'abord à partir de la page `https://en.wikipedia.org/wiki/Markov_chain`. Notez qu'étant donné la manière dont le graphe G a été généré, le graphe obtenu est nécessairement connexe.

Questions :

1. Donnez l'expression mathématique correspondant à la matrice de transition du modèle de surfeur avec téléportation.
2. Démontrez sur base de la théorie que la matrice de transition est bien telle que la distribution stationnaire est unique dès que $\alpha > 0$.
3. Quelles sont les 10 pages web de score *PageRank* le plus élévé sur base du graphe G en utilisant la valeur $\alpha = 0.15$?
4. Calculez (sur ordinateur) les deux pages les plus proches et les deux pages les plus éloignées, où la distance entre deux pages est déterminée par le nombre de clicks moyens nécessaires pour passer de l'une à l'autre (toujours lorsque $\alpha = 0.15$).

3 Références

Toutes les références, les données, et les codes relatifs au projet seront collectés sur la page web suivante : `http://www.montefiore.ulg.ac.be/~geurts/math1222.html`.