Éléments de processus stochastiques

Projet sur les chaînes de Markov

3ème BAC IC, Majeures en biomédical, électricité, informatique, physique

Profs. Pierre Geurts et Louis Wehenkel
Assistant : Arnaud Joly

Année académique 2013-2014

Ce document décrit le projet du cours d'Eléments de Processus Stochastiques pour les étudiants de troisième année en Bac ingénieur civil qui ont choisi la majeure en biomédical, électricité, informatique, ou physique.

Le travail sera réalisé par groupe de trois étudiants. Les consignes générales pour réaliser le travail, et celles relatives la manière de rédiger les rapports et préparer la présentation orale, sont les mêmes pour toutes les sections et sont communiquées par le Professeur V. Denoël, coordinateur du cours.

Des éléments complémentaires de théorie nécessaires pour réaliser ce projet seront expliqués lors des séances d'encadrement du travail. Nous encourageons vivement les étudiants à se documenter eux-mêmes sur les sujets abordés dans ce projet. Des pointeurs vers des sources de théorie/pratique seront fournis au fur et à mesure de la réalisation du travail.

Le travail sera encadré au moyen de séances de questions/réponses qui seront programmées au fur et à mesure de l'avancement du projet. Ces séances auront lieu les mardis de 10h45 à 12h45 au local R53 (Europe B4). Seules les questions posées lors de ces séances donneront lieu à des réponses des encadrants.

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 Première partie du projet

Cette première partie du projet vise à introduire progressivement le modèle PageRank utilisé par Google et à en faire une analyse détaillée.

1.1 Etude du modèle de base

Analyse du graphe A_1 :

Soit la matrice A_1 suivante :

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

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

1. Représentez par une figure le graphe correspondant à la matrice A_1.
2. 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 Q (avec $[Q]_{i,j} = prob(X_{t+1} = j | X_t = i)$) d'une chaîne de Markov modélisant ce surfeur. Représentez le diagramme d'états de la chaîne de Markov correspondante.
3. En vous servant de MATLAB, calculez les quantités suivantes pour des valeurs de t croissantes :
   — $prob(X_t = i)$ en supposant que le surfeur est parti d'une page choisie au hasard parmi toutes les pages,
   — $prob(X_t = i)$ en supposant que le surfeur démarre toujours de la page 1,
   — Q^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.
4. En déduire la distribution stationnaire π_∞ de la chaîne de Markov définie par $[\pi_\infty]_j = \lim_{t\to\infty} prob(X_t = j)$.
5. En supposant que $[\pi_\infty]_j, \forall j = 1, \ldots, 4$, définit le score PageRank des 4 pages web du graphe A_1, discutez le classement obtenu sur base de la structure des liens entre les 4 pages. Est-ce que ce classement vous semble intuitif ?

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.

6. Générez une réalisation aléatoire de longueur T de la chaîne de Markov en démarrant d'une page choisie aléatoirement selon la distribution stationnaire. Calculez pour chaque page le nombre de fois que le surfeur passe par cette page rapporté à la longueur de la réalisation. Observez l'évolution de ces valeurs pour chaque page lorsque T croît.
7. Que concluez-vous de cette expérience ? Reliez ce résultat à la théorie.

Analyse des graphes A_2 et A_3 :

Soit les deux matrices d'adjacence A_2 et A_3 ci-dessous :

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

1. Refaites les expériences du point 1 et 3 ci-dessus, avec les graphes définis par A_2 et A_3, et sur base de ces expériences, discutez des éventuelles limitations du modèle de surfeur présenté à la section 1.
2. Discutez le calcul de la distribution stationnaire π_∞ via la résolution du système d'équations $\pi_\infty Q = \pi_\infty$ et $\sum_i [\pi_\infty]_i = 1$ d'abord dans le cas du graphe A_1, ensuite dans le cas des graphes A_2 et A_3.

1.2 Téléportation

Le modèle du surfeur présenté à la section précédente a deux limitations :

— Il ne précise pas comment traiter les pages web qui ne contiennent aucun lien (on les appelle des "dangling nodes"),

— La distribution stationnaire peut ne pas être unique.

Le modèle de surfeur utilisé pour calculer le score PageRank est 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 et code : Le fichier `graphes.mat` contient la matrice d'adjacence G et la liste U des URLs pour 500 pages obtenues par un parcours du web en largeur d'abord à partir de la page `http://www.montefiore.ulg.ac.be`. La fonction `surfer.m` [2] qui a été utilisée pour calculer ce graphe vous est également fournie. Notez qu'étant donné la manière dont le graphe G a été généré, le graphe obtenu est nécessairement connexe.

Questions :

2. Source : `http://www.mathworks.nl/moler/index_ncm.html`.

1. Décrivez comment calculer la matrice de transition correspondant au modèle de surfeur avec téléportation.
2. Montrez 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. Expliquez (en notations probabilistes d'abord et en notations matricielles ensuite) le principe du calcul de la probabilité de passer au temps t_2 par une page p_2 sachant qu'on était sur la page p_1 en t_1 (avec $t_1 < t_2$) et qu'on sera sur la page p_3 en t_3 (avec $t_3 > t_2 > t_1$). Implémentez le calcul avec MATLAB et utilisez votre code pour calculer cette probabilité pour $t_1 = 1$, $t_2 = 10$, $t_3 = 20$, $p_1 =$`http://www.montefiore.ulg.ac.be` et p_2 et p_3 respectivement la deuxième et la première page du classement PageRank (avec $\alpha = 0.15$ et la chaîne initialisée avec sa distribution stationnaire).

1.3 Effet du α

α est un paramètre qui influence le score PageRank d'une page. On aimerait analyser de manière fouillée l'impact de ce paramètre sur le résultat du classement PageRank.

Questions :

1. Montrez que le score PageRank de toute page est au moins α/N où N est le nombre total de pages. Qu'est-ce que cela a comme implication sur la différence en terme de score PageRank entre pages lorsque α tend vers 1 ?
2. Étudiez la stabilité du classement obtenu lorsque la valeur de α est modifiée. Vous pourriez quantifier cette stabilité par exemple en traçant l'évolution du rang de certaines pages du graphe lorsque α croît mais nous vous encourageons à explorer d'autres métriques si nécessaire.
3. Quand on veut calculer les scores PageRank pour un graphe de grande taille, la méthode souvent utilisée consiste à calculer πQ^t pour une distribution initiale π par exemple uniforme et des valeurs de t croissantes en s'arrêtant dès que la distance euclidienne entre πQ^{t-1} et πQ^t devient inférieure à un seuil ϵ. Etudiez empiriquement l'impact de la valeur de α sur la vitesse de convergence de cette méthode. Expliquez intuitivement les résultats obtenus.

2 Seconde partie du projet

Dans la seconde partie du projet, on vous demande d'étudier le problème de l'estimation des paramètres d'une chaîne de Markov à partir d'une réalisation de cette chaîne. On abordera dans un premier temps le problème général de l'estimation de la matrice de transition et dans un second temps le problème de l'estimation du paramètre de téléportation α du modèle de surfeur aléatoire.

Dans le cadre de cette partie, vous êtes libres d'aborder toutes questions annexes qui vous sembleraient intéressantes et vous pouvez utiliser toutes les sources que vous souhaitez, pour autant que vous les citiez correctement.

2.1 Estimation de la matrice de transition

Pour motiver le problème de l'estimation de la matrice de transition, on se place dans le scénario suivant :

— On dispose de deux traces de navigation Tr_1 et Tr_2 obtenues respectivement de l'observation de deux surfeurs S_1 et S_2 se déplaçant sur le même graphe. Ces deux surfeurs ne correspondent pas nécessairement à des surfeurs aléatoires tels que définis plus haut.

— On dispose de 20 traces de navigation supplémentaires. L'origine exacte de ces traces est inconnue mais on sait cependant que 10 traces proviennent du surfeur S_1 et 10 traces proviennent du surfeur S_2.

— Sur base uniquement de ces données, on vous demande d'identifier l'origine de chacune des 20 traces (S_1 ou S_2).

Les deux traces Tr_1 et Tr_2 ainsi que les 20 traces à identifier vous sont fournies dans le fichier `traces.mat` (variables Tr1, Tr2 et Trmat). Les traces Tr_1 et Tr_2 sont de longueur 5000, alors que les traces à identifier sont de longueur 1000. Le réseau sous-jacent, supposé inconnu, est composé de 50 nœuds (numérotés de 1 à 50 dans les traces qui vous sont fournies).

Nous vous suggérons d'adopter la démarche suivante :

1. Discutez de la manière d'estimer la matrice de transition d'une chaîne de Markov à partir d'une réalisation de cette chaîne en utilisant la méthode du maximum de vraisemblance.
2. Expliquez comment utiliser les deux modèles estimés à partir des traces Tr_1 et Tr_2 pour décider de l'origine d'une nouvelle trace.
3. En vous servant de MATLAB, appliquez cette approche sur les traces fournies dans le fichier `traces.mat` et fournissez dans votre rapport une table reprenant pour chacune des 20 traces l'origine prédite ainsi que la probabilité associée.
4. Quelles sont les hypothèses sous-jacentes à l'approche proposées ? Que pouvez-vous dire de la fiabilité de vos prédictions et des limitations de la démarche adoptée en général ?
5. Discutez d'autres applications potentielles pratiques de la démarche illustrée ici.

2.2 Estimation de la valeur de α

Google souhaiterait pouvoir adapter le score PageRank à chaque utilisateur. Une adaptation simple consisterait à utiliser un paramètre de téléportation α différant d'un utilisateur à l'autre et reflétant les habitudes de surf de chacun. Sur base du graphe du réseau et à partir d'une trace de navigation, on souhaite donc pouvoir estimer la valeur du paramètre α de téléportation, en faisant l'hypothèse que le surfeur suit le modèle de surfeur aléatoire développé dans la première partie du projet.

On vous propose d'aborder ce problème en plusieurs étapes :

1. Expliquez d'abord comment vous pourriez estimer la valeur de α à partir d'une trace et en supposant le graphe connu.
2. Appliquez ensuite le ou les approches proposées au point précedent sur des traces obtenues à partir de graphes et de valeurs de α de votre choix. Comparez vos estimations à la valeur réelle du paramètre α et analysez les résultats obtenus.

3. En supposant que dans le problème de la section 2.1, les deux surfeurs S_1 et S_2 sont en fait des surfeurs aléatoires qui ne diffèrent que de leur valeur de α, comment résolveriez-vous le problème de l'identification de l'origine des 20 traces ?
4. Appliquez cette idée aux données fournies dans le fichier `traces.mat` (le graphe vous est founi dans la variable G) et comparez vos prédictions à celles obtenues précédemment. Si ces prédictions diffèrent, expliquez lesquels vous semblent les plus fiables et pourquoi.
5. Discutez de manière générale le problème de l'estimation d'un modèle de surfeur à partir d'une trace en supposant cette fois qu'à la fois le graphe et la valeur de α sont inconnus.

Remarques/suggestions :

— La méthode au maximum de vraisemblance n'est pas la seule méthode possible pour construire une estimateur de α. Vous êtes libres d'étudier tout estimateur supplémentaire qui vous semblerait pertinent pour cette application.

— Vous êtes libres d'utiliser les critères qui vous semblent les plus pertinents pour comparer les estimations aux paramètres réels. Le critère le plus classique pour estimer la qualité de l'estimation d'un paramètre à valeur réelle est l'erreur quadratique. Dans le contexte de notre application, il pourrait être intéressant également de comparer les scores PageRank obtenus à partir du modèle estimé aux score PageRank obtenus à partir du modèle réel.

— La qualité des estimations va dépendre de plusieurs paramètres du problème tels que la valeur réelle de α, la densité des liens dans le réseau, la longueur de la trace utilisée, la présence de dangling nodes, etc. Il peut être intéressant d'analyser, théoriquement et/ou sur base de simulation, l'impact de ces différents paramètres sur la qualité des estimations.

3 Références

Toutes les références, les données, et les codes relatifs au projet seront collectés sur la page web des projets (`http://www.ajoly.org/students`).