Résumé long-RJS2015

Maxime Brunin

5 novembre 2015

Résumé long

Dans cette présentation, nous étudions le problème de la détection de ruptures à l'aide de méthodes à noyau sous l'angle du compromis temps de calcul–précision, particulièrement pertinent dans le cadre de grands jeux de données (big data).

Plus précisément, notre approche de détection de ruptures (basée sur les noyaux positifs semi-définis) a pour but de détecter des changements dans la distribution d'observations recueillies au cours du temps entre les instants $1, \dots, n$ dans un contexte *offline*. Ces changements interviennent à des instants appelés instants de ruptures, notés $\{\tau_t^*\}_{0 \leq t \leq D^*}$ avec la convention $\tau_0^* = 1$ et $\tau_{D^*}^* = n+1$. L'idée d'introduire des noyaux permet de recoder des changements dans la distribution des observations initiales par des changements dans la moyenne de nouvelles "observations" appartenant à l'espace auto-reproduisant $\mathcal{H}$ associé au noyau k choisi et appelé $RKHS$. Une application de ce genre de stratégie consiste à étudier les variations du nombre de copies d'ADN le long du génome dans l'étude des différents types de cancers.

Pour répondre à cette question, nous adoptons le formalisme de la sélection de modèle dans lequel chaque segmentation candidate (liste d'instants de ruptures) est naturellement reliée à un modèle qu'il s'agit de choisir. Nous présenterons deux approches concurrentes que nous comparerons en termes de performance statistique et de temps de calcul afin de mettre en avant les forces et faiblesses de chacune : (i) l'approche exacte basée sur la programmation dynamique et (ii) l'approche approchée basée sur l'heurisitique de segmentation binaire.

Pour la première, on procède en deux étapes. D'abord, il s'agit de trouver la meilleure segmentation en D segments qui minimise le risque empirique. Puis, on réalise la sélection de la meilleure segmentation (de la meilleure dimension) minimisant un critère pénalisé. Appliquée aux noyaux, cette stratégie bien comprise sur le plan théorique a une complexité de $O\big(D_{\max} n^4\big)$ en temps et $O\big(D_{\max} n^2\big)$ en espace, ce qui devient vite prohibitif quand n augmente. Par ailleurs, cette approche repose également sur la calibration de la pénalité à minimiser qui impose de plus le calcul de toutes les meilleures segmentations jusqu'à une grande valeur de $D_{\max}$. Cette étape indispensable pour une bonne performance statistique est très coûteuse algorithmiquement.

Pour la deuxième approche, la segmentation binaire est une heuristique itérative permettant de ne récupérer qu'un minimiseur local du risque empirique sur l'espace des segmentations en D segments pour chaque $1 \leq D \leq D_{\max}$. L'implémentation de l'heuristique de segmentation binaire à noyau a une compexité $O\big(n^2\big)$ en temps et $O(n)$ en espace ce qui est un avantage par rapport à la programmation dynamique (naïve). L'inconvénient de la seg-

mentation binaire à noyau est que la détermination d'un temps d'arrêt $\hat{D}$ correspondant à la meilleure segmentation demeure une question ouverte et correspond à l'un des objectifs de mon travail de thèse. À cet égard, l'article de Fryzlewicz (2014) constitue une piste intéressante.