Apprentissage Statistique en Génomique

Dell'Aiera Clément, Prévosteau Clément

Résumé

Dans ce document, nous présentons une application des méthodes apprises au cours donné par Jean-Philippe Vert intitulé *Machine Learning for Computational Statistics*. Il s'agit d'extraire un profil génétique à partir des niveaux d'expression de 4654 gènes récoltés sur 184 individus qui ont ou n'ont pas rechuté après une chimiothérapie, profil dont le but est de savoir, étant donné les niveaux d'expression de ces gènes chez un patient, s'il risque de réchuter ou non, en vue d'adapter le traitement *a priori*.

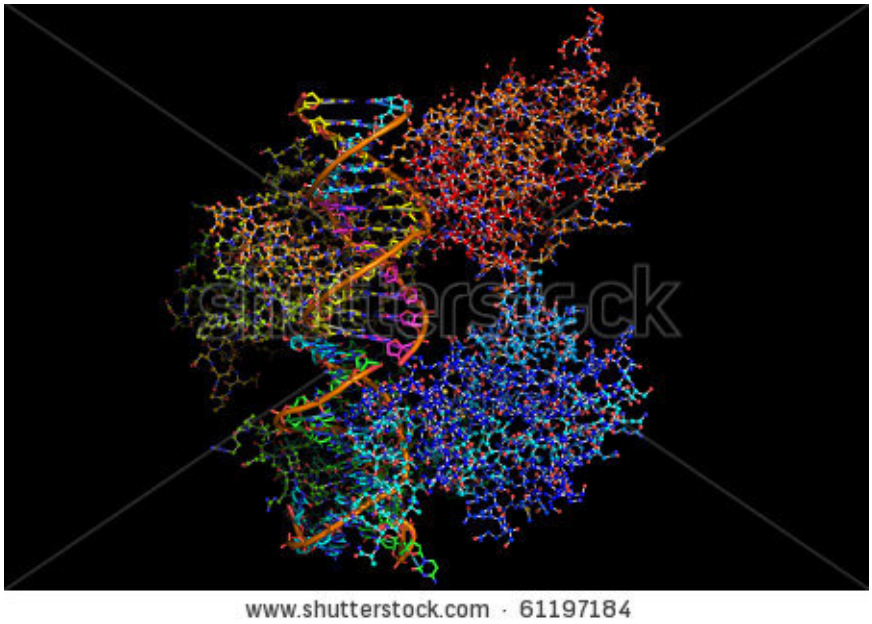

FIGURE 1 – A ball and stick model of molecules of protein p53 binding to a strand of DNA. The majority of human cancers involve mutations which make this protein inactive, from `http://fr.123rf.com/photo_7823791_.html`

Table des matières

1 Support Vector Machine

Nous avons entraîné des SVM sur la base d'apprentissage des 184 patients initiaux, en faisant varier différents paramètres qu'offrait le package *kernlab* sous *R*. Par exemple, nous avons testé les noyaux : linéaires, gaussiens, Anova, TanH, Bessel et Laplace. A chaque fois, la marge a été déterminée par *cross-validation* sur les 184 patients, en divisant l'échantillon en 5 classes.

Voici les marges optimales obtenues sur les 3 noyaux les plus efficaces, à savoir linéaires, gaussiens et Anova :

Noyau	**Marge**
Linéaire	0.0009765625
Gaussien	0.125
Anova	

TABLE 1 – Marges optimales des SVM obtenues par *cross-validation*

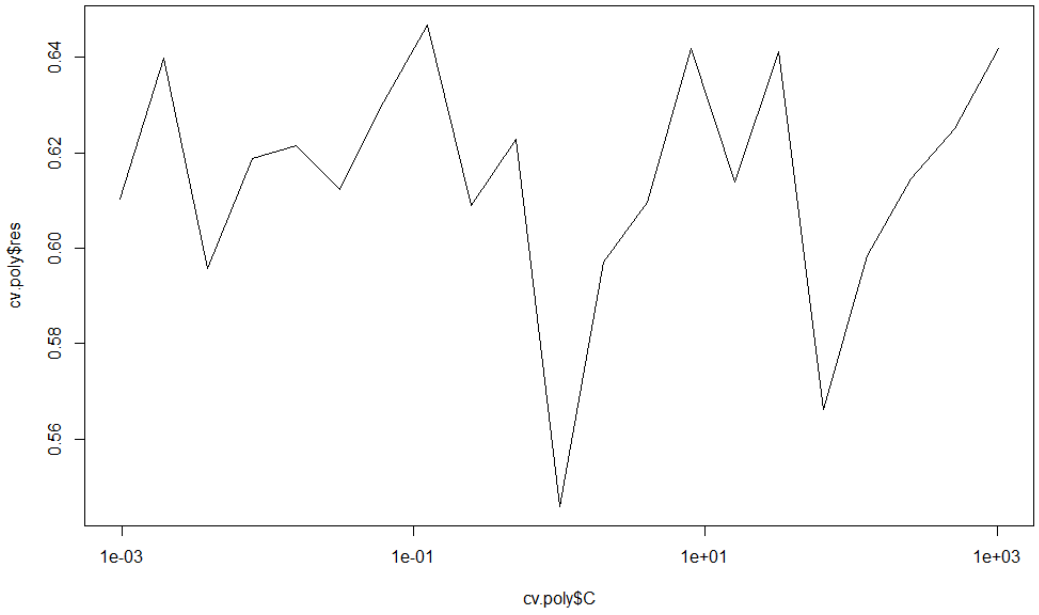

FIGURE 2 – Courbe de l'AUC pour un noyau polynomial de degré 2

L'échantillonage en 5 sous-classes étant aléatoire, la marge$C$ obtenue varie à chaque itération. Afin d'obtenir un résultat plus robuste, nous avons itéré la

cross-validation 100 fois, puis fais la moyenne des marges optimales obtenues.

1.1 Commentaires sur le code

Nous avons principalement utilisé le package *kernlab*. Le code du projet est entièrement disponible à la page :

`https://github.com/cdellaie/MachineLearningforComputationalStat`

1.2 Une mesure théorique de la connexité d'un graphe

Voici la définition de la constante d'isopérimétrie d'un graphe. Provenant de la théorie des graphes expanseurs, elle mesure en quelque sorte, lorsque le graphe est fini, si le graphe est "très connexe", où l'on dit qu'un graphe est très connexe si, après lui avoir enlevé beaucoup d'arêtes, il le reste.

Si X désigne un graphe fini, si $A \subset X$ est un sous-graphe de X, on définit la frontière de A comme :

$$\partial A := \{(x, y) \text{ arêtes t.q.} x \in A, y \notin A\}$$

Alors la constante d'isopérimétrie associée à X est :

$$h(X) = \inf \left\{ \frac{|\partial A|}{|A|} : A \subset X \text{ t.q. } 1 \leq |A| \leq \frac{|X|}{2} \right\}$$

où $|.|$ dénote le cardinal d'un ensemble.

Un sous-graphe linéaire est constitué d'une suite de sommets qui forment un chemin, *i.e.* une suite d'arêtes v_j telles que $(v_j, v_{j+1}) \in E$, où E est l'ensemble des arêtes du graphe.