Apprentissage statistique: TD8
Data-Dependant Partitioning

Emile Contal
http://econtal.perso.math.cnrs.fr/teaching

14 mars 2016

Exercice 1. On considère ici les classifieurs par partitionnement de $\mathbb{R}$ où les partitions dépendent des données D_n de façon déterministes suivant la règle $\mathcal{P}_n = \pi_n(D_n)$. Soit la séquence de règles (π_n), le classifieur associé fonctionne par vote à l'intérieur des cellules $A_n(x)$. On notera $\mathcal{F}_n$ la famille de toutes les partitions possibles pour la règle π_n : $\mathcal{F}_n = \big\{\pi_n(D) : D \in (\mathbb{R} \times \{0,1\})^n\big\}$. En utilisant le théorème et les lemmes suivants, montrer qu'un classifieur par partitionnement de $\mathbb{R}$ tel que les intervalles contiennent au moins a_n et au plus b_n points, est fortement consistant lorsque $a_n \to \infty$ et $\frac{b_n}{n} \to 0$.

Théorème 1 (Consistance forte)**.** *On note* $\operatorname{diam}(A) = \sup_{x,y\in A} \|x-y\|$*,* S_M *la boule fermée de rayon* M *centrée à l'origine, et* $\mathcal{P}^{(M)}$ *la restriction d'une partition* $\mathcal{P}$ *à la boule* S_M*. De même si* $\mathcal{F}$ *est une collection de partitions, on note* $\mathcal{F}^{(M)} = \big\{\mathcal{P}^{(M)} : \mathcal{P} \in \mathcal{F}\big\}$*. On introduit* $\Delta_n(\mathcal{F}^{(M)})$ *une mesure de complexité* [1] *d'une famille de partition. Soient* $\mathcal{F}_1, \mathcal{F}_2, \dots$ *comme précédemment, si on a :*

$$\forall M < \infty \,, \frac{\log \Delta_n(\mathcal{F}_n^{(M)})}{n} \to 0 \,,$$
$$\forall S_M \text{ et } \gamma > 0 \,, \mu\Big(\Big\{x : \operatorname{diam}\big(A_n(x) \cap S_M\big) > \gamma\Big\}\Big) \to 0 \text{ p.s,}$$

alors le classifieur associé est fortement consistant.

Lemme 1. *Soit* $\mathcal{F}^{(M)}$ *une famille de partitions telle qu'il existe une constante* N *vérifiant pour tout* $P^{(M)} \in \mathcal{F}^{(M)}$ *que* $\big|P^{(M)}\big| \leq N$*, alors :*

$$\Delta_n(\mathcal{F}^{(M)}) \leq 2^N \Delta_n^*(\mathcal{F}^{(M)}) \,,$$

où $\Delta_n^*(\mathcal{F}^{(M)})$ *est le nombre maximal de façons de partitionner* n *points avec des éléments de* $\mathcal{F}^{(M)}$*.*

Lemme 2. *Soit* $\mathcal{I}$ *l'ensemble des intervalles de* $\mathbb{R}$*,* $\sup_{I\in\mathcal{I}} |\mu(I) - \mu_n(I)| \to 0$ *p.s.*

Lemme 3. *Soit* H *l'entropie binaire, alors* $\log \binom{s}{t} \leq sH(\frac{t}{s})$*.*

[1] On ne donne pas sa définition ici mais on pourra la borner en utilisant le Lemme 1. Les plus curieux iront voir le *shatter coefficient* et la *VC dimension*.