DATAMINING SUPERVISE.

III Prédiction non paramétrique.

1°) Prédicteur à noyau.
a) Estimation d'une densité.

Soit $\mathcal{E}_n = \{\mathbf{X_i},\, i = 1...n\}$ un n-échantillon de loi μ absolument continue par rapport à la mesure de Lebesgue sur R^p, de densité continue $f(\mathbf{x})$.

Notons $\widehat{f}_\sigma(\mathbf{x}, \mathcal{E}_n) = \frac{1}{n}\Sigma_{i=1}^{n} \frac{1}{\mathcal{V}_\sigma} 1_{\|\mathbf{X_i} - \mathbf{x}\| < \sigma}$ avec $\mathcal{V}_\sigma = \int_{\|\mathbf{x}\| < \sigma} d\mathbf{x} = C_p.\sigma^p$.

Alors $\int \widehat{f}_\sigma(\mathbf{x}, \mathcal{E}_n)\, d\mathbf{x} = 1$ et $E(\widehat{f}_\sigma(\mathbf{x}, \mathcal{E}_n)) = \frac{1}{\mathcal{V}_\sigma} P(\|\mathbf{X_1} - \mathbf{x}\| < \sigma)$, de l'ordre de $f(\mathbf{x})$ si $f(\mathbf{x})$ varie lentement sur la boule $\{\mathbf{u} |\, \|\mathbf{u} - \mathbf{x}\| < \sigma\}$.

Si $\sigma > Sup_i \|\mathbf{X}_i - \mathbf{x}\|$, $\widehat{f}_\sigma(\mathbf{x}, \mathcal{E}_n) = \frac{1}{\mathcal{V}_\sigma}$;
si $\sigma < inf_{i \neq j} \left\|\mathbf{X_i} - \mathbf{X_j}\right\|$, $\widehat{f}_\sigma(\mathbf{x}, \mathcal{E}_n) = 0$ ou $\frac{1}{n\mathcal{V}_\sigma}$ (donc varie beaucoup).

Plus généralement, un **noyau** $K_\sigma(\mathbf{x}, \mathbf{y})$ est une (suite de) fonction(s) $R^p \times R^p \to R$ d'intégrale 1 (en $\mathbf{x}$) qui converge en loi vers la mesure de Dirac $\delta_\mathbf{x}$ quand $\sigma \to 0$. On lui associe l'estimateur de $f(\mathbf{x})$:

$\widehat{f}_\sigma(\mathbf{x}, \mathcal{E}_n) = \frac{1}{n}\Sigma_{i=1}^{n} K_\sigma(\mathbf{x}, \mathbf{X_i})$.

En pratique, le noyau K_σ est "régulier":
positif (donc $\widehat{f}_\sigma$ est une densité),
symétrique en $\mathbf{y}$ (annule l'erreur du "premier ordre"),
de la forme $K_\sigma(\mathbf{x}, \mathbf{y}) = k_\sigma(\mathbf{x} - \mathbf{y})$, et $k_\sigma(\mathbf{u}) = \frac{1}{\sigma^p} k(\frac{1}{\sigma}\mathbf{u})$, $\int k(\mathbf{u}) d\mathbf{u} = 1$.

Le paramètre σ est la (largeur de la) **fenêtre**; on considère parfois un multi-fenêtrage $\sigma = (\sigma_1, ..., \sigma_p) \in R^p$, $k_\sigma(\mathbf{u}) = \frac{1}{\sigma_1 ... \sigma_p} k(\frac{1}{\sigma_1} u_1, ..., \frac{1}{\sigma_p} u_p)$. $\frac{1}{\sigma^p}$ (ou $\frac{1}{\sigma_1 ... \sigma_p}$) mesure la **complexité** de l'estimateur.

Exemples: pour $p = 1$, le noyau uniforme (défini ci-dessus), le chapeau de gendarme, le noyau gaussien... On en déduit des noyaux sur R^p par produits ou rotation (fonction de $\|\mathbf{u}\|_2$ => noyau **radial**).

b) Erreur quadratique d'estimation $e(\widehat{f}_\sigma(\mathbf{x}))$.

$e(\widehat{f}_\sigma(\mathbf{x}))^2 = E(\widehat{f}_\sigma(\mathbf{x}, \mathcal{E}_n) - f(\mathbf{x}, \mathcal{E}_n))^2 =$
$(E\widehat{f}_\sigma(\mathbf{x}, \mathcal{E}_n) - f(\mathbf{x}))^2 + Var(\widehat{f}_\sigma(\mathbf{x}, \mathcal{E}_n)) = (biais)^2 + variance$.

Dans le cas du noyau uniforme, $E\widehat{f}_\sigma(\mathbf{x}, \mathcal{E}_n) = \frac{1}{\mathcal{V}_\sigma} \int_{\|\mathbf{x}-\mathbf{y}\| < \sigma} f(\mathbf{y}) d\mathbf{y}$ (dans le cas général, $E\widehat{f}_\sigma(\mathbf{x}, \mathcal{E}_n) = (k_\sigma * f)(\mathbf{x})$ (convolution)).

$|E\widehat{f}_\sigma(\mathbf{x}, \mathcal{E}_n) - f(\mathbf{x})| =$
$\frac{1}{\mathcal{V}_\sigma} |\int_{\|\mathbf{y}-\mathbf{x}\| < \sigma} (f(\mathbf{y}) - f(\mathbf{x})) d\mathbf{y} \leq Sup_{\|\mathbf{y}-\mathbf{x}\| < \sigma} |f(\mathbf{y}) - f(\mathbf{x}))|$ tend vers 0 quand $\sigma \to 0$ par continuité de f, et vers $|f(\mathbf{x})|$ quand $\sigma \to +\infty$.
Si de plus f est lipschitzienne, c.à.d. $|f(\mathbf{y}) - f(\mathbf{x})| \leq A \|\mathbf{y} - \mathbf{x}\|$, alors $|E\widehat{f}_\sigma(\mathbf{x}, \mathcal{E}_n) - f(\mathbf{x})| \leq A\sigma$.

$Var(\widehat{f}_\sigma(\mathbf{x}, \mathcal{E}_n)) = \frac{1}{n\mathcal{V}_\sigma^2} \int_{\|\mathbf{x}-\mathbf{y}\| < \sigma} f(\mathbf{y}) d\mathbf{y} - \frac{1}{n\mathcal{V}_\sigma^2} (\int_{\|\mathbf{x}-\mathbf{y}\| < \sigma} f(\mathbf{y}) d\mathbf{y})^2$.

Quand $\sigma \to +\infty$, $Var(\widehat{f}_\sigma(\mathbf{x}, \mathcal{E}_n)) \to 0$;
quand $\sigma \to 0$, le premier terme est de l'ordre de $\frac{1}{n\mathcal{V}_\sigma} f(\mathbf{x})$ et le second $\frac{1}{n} f(\mathbf{x})^2$, négligeable devant le premier; $Var(\widehat{f}_\sigma(\mathbf{x}, \mathcal{E}_n)) \to +\infty$.

Dans un modèle asymptôtique où $n \to +\infty$, on choisira $\sigma_n \to 0$ (pour le biais) et $n\mathcal{V}_{\sigma_n} = nC_p\sigma_n^p \to +\infty$ (pour la variance).

En particulier, dans le cas lipschitzien, $e(\widehat{f}_\sigma(\mathbf{x}))^2 \leq A^2\sigma^2 + \frac{B}{n\sigma^p}$;
à n fixé, le majorant est minimum pour $\sigma_n = \frac{D}{n^{\frac{1}{p+2}}}$,et $e(\widehat{f}_{\sigma_n}(\mathbf{x})) \leq \frac{E}{n^{\frac{1}{p+2}}}$.

c) Prédiction par estimateurs à noyau.

Etant donné l'ensemble d'apprentissage $\mathcal{A}_n$, nous estimons les densités $f_0(\mathbf{x})$ et $f_1(\mathbf{x})$ (supposées continues) des lois μ_0 et μ_1 par
$\widehat{f}_{0,\sigma}(\mathbf{x}, \mathcal{A}_n) = \frac{1}{n_0}\Sigma_{i=1}^n 1_{Y_i=0}.K_\sigma(\mathbf{x}, \mathbf{X_i})$, $\widehat{f}_{1,\sigma}(\mathbf{x}, \mathcal{A}_n) = \frac{1}{n_1}\Sigma_{i=1}^n 1_{Y_i=1}.K_\sigma(\mathbf{x}, \mathbf{X_i})$.

Le prédicteur $\widehat{g}_\sigma(\mathbf{x}, \mathcal{A}_n)$ choisira la classe 1 si
$\Sigma_{i=1}^n 1_{Y_i=1}.K_\sigma(\mathbf{x}, \mathbf{X_i}) - \Sigma_{i=1}^n 1_{Y_i=0}.K_\sigma(\mathbf{x}, \mathbf{X_i}) > 0$
(et la classe 0 si l'expression est négative).

Soit $\varepsilon_{\sigma,\mathcal{A}_n} = \varepsilon_{\widehat{g}_\sigma(\mathbf{x},\mathcal{A}_n)} = P(\widehat{g}_\sigma(\mathbf{X}, \mathcal{A}_n) \neq Y|\mathbf{X} = \mathbf{x}, \mathcal{A}_n)$ l'erreur de prédiction en $\mathbf{x}$ conditionnellement en $\mathcal{A}_n$, le couple "nouveau" $(\mathbf{X}, Y)$ étant supposé **indépendant** de l'ensemble d'apprentissage $\mathcal{A}_n$.
Avec les notations du premier chapitre et $P(Y = 1) = \frac{1}{2}$ pour simplifier, on a $\varepsilon_{\sigma,\mathcal{A}_n}(\mathbf{x}) - \varepsilon^*(\mathbf{x}) = |1 - 2h(\mathbf{x})|.|1_{\widehat{g}_\sigma(\mathbf{x},\mathcal{A}_n)=1} - 1_{g^*(\mathbf{x})=1}|$.
En prenant l'espérance en $\mathbf{X}$:
$\varepsilon_{\sigma,\mathcal{A}_n} - \varepsilon^* = \int |1 - 2h(\mathbf{x})|.|1_{\widehat{g}_\sigma(\mathbf{x},\mathcal{A}_n)=1} - 1_{g^*(\mathbf{x})=1}|.\frac{f_0(\mathbf{x})+f_1(\mathbf{x})}{2}d\mathbf{x}$
$\widehat{g}_\sigma(\mathbf{x}, \mathcal{A}_n) \neq g^*(\mathbf{x})$ entraîne $|1 - 2h(\mathbf{x})| \leq 2|h(\mathbf{x}) - \widehat{h}_\sigma(\mathbf{x}, \mathcal{A}_n)|$ d'où
$\varepsilon_{\sigma,\mathcal{A}_n} - \varepsilon^* \leq 2\int |h(\mathbf{x}) - \widehat{h}_\sigma\mathbf{x}, \mathcal{A}_n)|\frac{f_0(\mathbf{x})+f_1(\mathbf{x})}{2}d\mathbf{x}$
$\varepsilon_{\sigma,\mathcal{A}_n} - \varepsilon^* \leq \int |\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})+f_1(\mathbf{x})} - \frac{\widehat{f}_{1,\sigma}(\mathbf{x},\mathcal{A}_n)}{\widehat{f}_{0,\sigma}(\mathbf{x},\mathcal{A}_n)+\widehat{f}_{1,\sigma}(\mathbf{x},\mathcal{A}_n)}|.(f_0(\mathbf{x}) + f_1(\mathbf{x}))d\mathbf{x}$
$\varepsilon_{\sigma,\mathcal{A}_n} - \varepsilon^* \leq \int \frac{|f_1(\mathbf{x}).\widehat{f}_{0,\sigma}(\mathbf{x},\mathcal{A}_n) - f_0(\mathbf{x}).\widehat{f}_{1,\sigma}(\mathbf{x},\mathcal{A}_n)|}{\widehat{f}_{0,\sigma}(\mathbf{x},\mathcal{A}_n)+\widehat{f}_{1,\sigma}(\mathbf{x},\mathcal{A}_n)}d\mathbf{x}$
$\varepsilon_{\sigma,\mathcal{A}_n} - \varepsilon^* \leq \int |f_0(\mathbf{x}) - \widehat{f}_{0,\sigma}(\mathbf{x}, \mathcal{A}_n)|d\mathbf{x} + \int |f_1(\mathbf{x}) - \widehat{f}_{1,\sigma}(\mathbf{x}, \mathcal{A}_n)|d\mathbf{x}$.

En notant $\varepsilon_{\sigma,n} = E(\varepsilon_{\sigma,\mathcal{A}_n})$, il vient finalement:
$\varepsilon_{\sigma,n} - \varepsilon^* \leq E(\int |f_0(\mathbf{x}) - \widehat{f}_{0,\sigma}(\mathbf{x}, \mathcal{A}_n)|d\mathbf{x} + \int |f_1(\mathbf{x}) - \widehat{f}_{1,\sigma}(\mathbf{x}, \mathcal{A}_n)|d\mathbf{x})$.

$\int e(\widehat{f}_{0,\sigma}(\mathbf{x}))d\mathbf{x} = E(\int (f_0(\mathbf{x}) - \widehat{f}_{0,\sigma}(\mathbf{x}, \mathcal{A}_n))^2 d\mathbf{x}) \to 0$ entraîne $E(\int |f_0(\mathbf{x}) - \widehat{f}_{0,\sigma}(\mathbf{x}, \mathcal{A}_n)|d\mathbf{x} \to 0$ (par exemple, pour f_0 à support borné, appliquer l'inégalité de Cauchy-Schwartz), et de même pour f_1. On en conclut que $\varepsilon_{\sigma_n,n} \to \varepsilon^*$ pour $n \to +\infty$, $\sigma_n \to 0$ et $n\sigma_n^p \to +\infty$.

En réalité, on peut montrer:

Théorème (**consistence universelle**).

Pour toutes lois μ_0 et μ_1, l'erreur de prédiction $\varepsilon_{\sigma_n,n}$ d'un prédicteur associé à un noyau régulier converge vers l'erreur ε^* du prédicteur de Bayes pour $n \to +\infty$, $\sigma_n \to 0$ et $n\sigma_n^p \to +\infty$.

Pour choisir la fenêtre σ qui minimise l'erreur de prédiction (ou plutôt un estimateur de cette erreur), on partitionne l'ensemble d'apprentissage $\mathcal{A}_n$ en $\mathcal{A}_{1,m_1}$ et $\mathcal{A}_{2,m_2}$ indépendants par tirage au hasard($m_1 \gg m_2$). A σ donné, on calcule $\widehat{g}_\sigma$ sur $\mathcal{A}_{1,m_1}$ et on estime son erreur de prédiction sur $\mathcal{A}_{2,m_2}$(principe de la **validation croisée**); on en déduit σ_n qui minimise cette erreur.

Remarques:
- Il est parfois possible d'interpréter l'estimateur à noyau comme un estimateur du développement de la densité sur une base de fonctions dans un certain espace fonctionnel. Pour $K_\sigma(\mathbf{x}, \mathbf{y})$ de type positif, on verra une autre interprétation lors de l'étude de la méthode SVM.
- Dans le cas $\Sigma_{i=1}^n 1_{Y_i=1}.K_\sigma(\mathbf{x}, \mathbf{X_i}) = \Sigma_{i=1}^n 1_{Y_i=0}.K_\sigma(\mathbf{x}, \mathbf{X_i}) = 0$, on ne peut rien prédire (zone "blanche").

2°) Prédicteurs k-NN.

Etant donné l'ensemble d'apprentissage $\mathcal{A}_n$, le prédicteur k-NN choisit au point $\mathbf{x}$ la classe la mieux représentée chez les k voisins $\mathbf{X_i}$ les plus proches (on choisira k impair pour éviter les ex-aequo).
Ce prédicteur peut se déduire d'estimateurs k-NN des densités...

Le paramètre $\frac{n}{k}$ mesure la complexité du prédicteur:
pour $k = n$, la classe la mieux représentée dans $\mathcal{A}_n$ sera choisie en tout point $\mathbf{x}$;
pour $k = 1$, tout point du "polytope de Voronoï" de $\mathbf{X_i}$ appartient à la classe de $\mathbf{X_i}$.

Erreur de prédiction. On peut supposer $\mathbf{x} \in Support(\mu_0) \bigcup Support(\mu_1)$.

$k = 1$.
$\varepsilon_{1,n}(\mathbf{x}) = h(\mathbf{x})(1 - h(\mathbf{X}'_n)) + (1 - h(\mathbf{x}))h(\mathbf{X}'_n)$, où $\mathbf{X}'_n$ désigne le plus proche voisin de $\mathbf{x}$ dans $\mathcal{A}_n$. Quand $n \to +\infty$, $\mathbf{X}'_n \to \mathbf{x}$, donc
$\varepsilon_{1,n}(\mathbf{x}) \to \varepsilon_{1,\infty}(\mathbf{x}) = 2h(\mathbf{x})(1 - h(\mathbf{x})) \leq 2Min(h(\mathbf{x}), 1 - h(\mathbf{x})) = 2\varepsilon^*(\mathbf{x})$.

$k = 3$.
$\varepsilon_{3,n}(\mathbf{x}) = h(\mathbf{x})(1 - h(\mathbf{X}'))(1 - h(\mathbf{X}''))(1 - h(\mathbf{X}''')) + etc...$
quand $n \to +\infty$, $\varepsilon_{3,n}(\mathbf{x}) \to \varepsilon_{3,\infty} =$
$h(\mathbf{x})(1 - h(\mathbf{x}))^3 + 6h(\mathbf{x})^2(1 - h(\mathbf{x}))^2 + (1 - h(\mathbf{x}))h(\mathbf{x})^3 =$
$h(\mathbf{x})(1 - h(\mathbf{x})) + 4h(\mathbf{x})^2(1 - h(\mathbf{x}))^2 \leq \varepsilon_{1,\infty}(\mathbf{x})$.

$k = 2l - 1$.
$\varepsilon_{2l-1,n}(\mathbf{x}) \to \varepsilon_{2l-1,\infty} = \Sigma_{i=0}^{l} C_{2l-1}^{i} h(\mathbf{x})^{i+1}(1 - h(\mathbf{x}))^{2l-1-i}$; on peut montrer que $\varepsilon_{2l-1,\infty}(\mathbf{x}) \leq \varepsilon_{2l-3,\infty}(\mathbf{x}) \leq ... \leq \varepsilon_{1,\infty}(\mathbf{x}) \leq 2\varepsilon^*(\mathbf{x})$.

Théorème (**consistence universelle** du prédicteur k-NN).
Soient g^* le prédicteur de Bayes, ε^* son erreur de prédiction, et une suite d'ensembles d'apprentissage $\mathcal{A}_n$, $n \to +\infty$.
Pour $k_n \to +\infty$, $\frac{k_n}{n} \to 0$, $\varepsilon_{k_n,n} \to \varepsilon^*$.

$Inf(h, 1-h) = \frac{1}{2}(1 - \sqrt{1 - 4h(1-h)})$ pour $h \in [0, \frac{1}{2}[\bigcup]\frac{1}{2}, 1]$; on en déduit le développement en série de $\varepsilon^*(\mathbf{x}) = Inf(h(\mathbf{x}), 1 - h(\mathbf{x}))$:
$\varepsilon^*(\mathbf{x}) = \Sigma_{i=1}^{\infty} \frac{1}{i} C_{2i-2}^{i-1} h(\mathbf{x})^i (1-h(\mathbf{x}))^i$ (développer $(1-u)^\alpha$, $u = h(1-h)$).

$\varepsilon_{2l-1,\infty}(\mathbf{x}) = \Sigma_{i=1}^{l-1} \frac{1}{i} C_{2i-2}^{i-1} h(\mathbf{x})^i (1 - h(\mathbf{x}))^i + 2C_{2l}^{l} h(\mathbf{x})^l (1 - h(\mathbf{x}))^l$
montre que $\varepsilon_{k,\infty} \to \varepsilon^*$ quand $k \to +\infty$ (en prenant l'espérance).

La condition $\frac{k_n}{n} \to 0$ entraînera $|\varepsilon_{k,n} - \varepsilon_{k,\infty}| \to 0$.

Choix du paramètre k_n.
Certains praticiens proposent $k_n \simeq \sqrt{n}$ (règle de "plomb"). On choisira plutôt k_n par validation croisée, comme pour la fenêtre σ du prédicteur à noyau. On remarquera que l'erreur apparente du prédicteur 1-NN est nulle (très forte surparamétrisation) !

Remarques finales:
Les prédicteurs à noyau et k-NN sont des méthodes paresseuses ("lazy"): aucun calcul préliminaire. Mais il faut conserver toute l'information de l'ensemble $\mathcal{A}_n$ pour le calcul de la prédiction en chaque point: coût important du stockage et des calculs. Dans le cas du prédicteur k-NN, un travail préliminaire de **condensation** des données permet d'éliminer les points inutiles et d'accélérer les prédictions futures.
D'autre part, ces algorithmes totalement numériques ne peuvent suggérer aucune interprétation à l'expérimentateur.

3°) Arbres de classification. Méthode CART.

On peut aussi estimer μ_0 et μ_1 dans des parallélotopes d'arêtes parallèles aux axes ("**histogrammes**" p-dimensionnels); le prédicteur associe une classe à chaque parallélotope. Les pas de subdivision sur les axes correspondent à la (multi)fenêtre de l'estimateur à noyau.
Du prédicteur k-NN, on peut retenir l'idée de pas variables; on obtient un pavage de R^p par des parallélotopes de différentes tailles.
Par subdivisions successives de ces pavages, l'ordre naturel d'inclusion conduit à la notion d'**arbre de classification**. Nous allons nous limiter aux arbres **binaires**, et examiner la méthode **CART**.
Dans une première étape, on construit l'arborescence complète: à chaque itération, on partitionne un **noeud** (une partie) de $\{\mathbf{X_i}\}$ en deux sous-noeuds. On élague ensuite l'arbre pour éviter une surparamétrisation.

a) Construction d'un arbre binaire complet.

1-Mesure de l'hétérogénéité **(impureté)** des noeuds .
Soient $p(l|t) = P(classe\,l|node\,t)$, estimée par $\frac{n_{lt}}{n_t}$, et c le nombre de classes (non limité à 2). On peut mesurer l'hétérogénéité d'un noeud par:

l'indice de Gini: $i(t) = \Sigma_{l=1}^{c}\Sigma_{h\neq l}p(l|t)p(h|t) = 1 - \Sigma_{l=1}^{c}p(l|t)^2$

l'entropie de Shannon: h(t)=-$\Sigma_{l=1}^{c}p(l|t)ln(p(l|t))$.

Ces deux fonctions sont nulles si et seulement si $p(l|t) = 1$ pour un l et tous les autres $p(h|t) = 0$, et sont strictement concaves de $p(1|t),...,p(c|t)$ (sommes de fonctions concaves).

2-Réduction de l'impureté par subdivision**.**

Soit s une subdivision du noeud t en deux sous-noeuds s_1 et s_2;
$\Delta i(t,s) = i(t) - i(s_1) - i(s_2) \geq 0$ par concavité de $i(t)$; en effet
$\widehat{p}(l|t) = \frac{n_{lt}}{n_t} = \frac{n_{ls_1}+n_{ls_2}}{n_{s_1}+n_{s_2}} = \frac{\widehat{p}(l|s_1)n_{s_1}+\widehat{p}(l|s_2)n_{s_2}}{n_{s_1}+n_{s_2}} = \widehat{p}(l|s_1)\frac{n_{s_1}}{n_{s_1}+n_{s_2}}+\widehat{p}(l|s_2)\frac{n_{s_2}}{n_{s_1}+n_{s_2}}$.

On a le même résultat pour $h(t)$.

Pour le noeud t, on choisit s qui maximise $\Delta i(t,s)$ (ou $\Delta h(t,s)$) parmi toutes les variables et les partitions possibles de chaque variable: selon son type (binaire, qualitative à k niveaux, ordinale à k niveaux ou quantitative prenant q valeurs), le nombre de subdivisions possibles sera 1, $2^{k-1} - 1$, $k - 1$, $q - 1$.

3-Règle d'arrêt et prédiction dans une feuille**.**
On termine l'algorithme quand $Min(n_{s_1}, n_{s_2}) < v$ ou $\Delta i(t,s) < \epsilon$.
Un noeud terminal est appelé une **feuille**.
A la feuille t, on attribuera une classe par la règle de Bayes, c'est-à-dire on choisira la classe l qui maximise $\widehat{p}(l|t)$.

b) Elagage de l'arbre.

Soient $\mathcal{T}$ un sous-arbre de l'arbre complet $\mathcal{T}_{max}$, et $\varepsilon(\mathcal{T})$ l'erreur apparente associée à $\mathcal{T}$: $\varepsilon(\mathcal{T}) = \frac{1}{n}\Sigma_{f\in\mathcal{T}}\Sigma_{l=1}^{c} m_l(f)$ en notant f une feuille de $\mathcal{T}$ et $m_l(f)$ le nombre d'éléments mal-classés de classe l dans la feuille f .

L'algorithme d'élagage choisit $\mathcal{T}_k$ qui minimise $\varepsilon(\mathcal{T})$ parmi les sous-arbres de $\mathcal{T}_{k+1}$. Il démarre avec l'arbre complet $\mathcal{T}_{max}$.

c) choix de l'arbre final.

On peut minimiser l'erreur apparente $\varepsilon(\mathcal{T})$ sur tous les sous-arbres de $\mathcal{T}_{max}$. C'est en général une mauvaise méthode, surtout pour un ensemble $\mathcal{A}_n$ de petit effectif n (surparamétrisation).
On lui préfèrera la minimisation de l'erreur apparente pénalisée $\varepsilon(\mathcal{T}) + \alpha.f(\mathcal{T})$ où $f(\mathcal{T})$ est le nombre de feuilles du sous-arbre $\mathcal{T}$ (le poids α sera déterminé par validation croisée).

Pour n grand, on pourra effectuer une validation croisée directe: construire l'arbre complet $\mathcal{T}_{max}$ sur $\mathcal{A}_{1,m_1}$; choisir le sous-arbre $\mathcal{T}$ qui minimise l'erreur de prédiction (estimée par) $\varepsilon(\mathcal{T})$ calculée sur $\mathcal{A}_{2,m_2}$.

Dans de nombreux cas, on constate une instabilité de l'arbre final. Parmi les solutions possibles: divisions équiréductrices.

Intérêts de la méthode:
-interprétation aisée des résultats (on travaille directement sur les variables);
-variables quantitatives ou qualitatives;
-nombre de classes non restreint à 2 (arbre de régression pour $Y \in R$).

Connexions avec les modèles logistiques, la régression ascendante descendante (pour la régression).

4°) Estimation de l'erreur de prédiction.

Les méthodes qui précèdent montrent l'importance de l'estimation de l'erreur de prédiction: pour évaluer la largeur de la fenêtre, le paramètre k du prédicteur k-NN, le niveau d'élagage...

Notre méthode de validation croisée (rudimentaire) partage l'ensemble d'apprentissage $\mathcal{A}_n$ en $\mathcal{A}_{1,m_1}$ et $\mathcal{A}_{2,m_2}$ indépendants (par tirage au hasard), avec $m_1 \gg m_2$;
on estime le prédicteur g par $\widehat{g}$ sur $\mathcal{A}_{1,m_1}$, d'autant meilleur que m_1 est grand;
on estime l'erreur de prédiction $\varepsilon_{\widehat{g}}$ de $\widehat{g}$ par $\widehat{\varepsilon}_{\widehat{g}}$ sur $\mathcal{A}_{1,m_2}$, de loi binômiale $\mathcal{B}(m_2, \varepsilon_{\widehat{g}})$ conditionnellement à $\mathcal{A}_{1,m_1}$, estimateur d'autant meilleur que m_2 est grand.

Comme $m_1 + m_2 = n$, plusieurs stratégies sont possibles:

a) $m_1 = n$, $m_2 = 0$: on estime $\varepsilon_{\widehat{g}}$ par l'**erreur apparente** $\widehat{\varepsilon}_{\mathcal{A}_n}$ **pénalisée** par le nombre de paramètres d (ou une mesure de la complexité) pour éviter la surparamétrisation; ainsi, le BIC (critère d'information bayésien) ou le MDL (description de longueur minimum) estiment $\varepsilon_{\widehat{g}}$ par $\widehat{\varepsilon}_{\mathcal{A}_n} + \alpha.\frac{d.ln(n)}{n}$ (où le paramètre α doit être estimé)...

b) validation croisée par **rotation** et "**leave-one out**":
on partage $\mathcal{A}_n$ en h parties $\mathcal{A}_{l,\frac{n}{h}}$ indépendantes (par tirage au hasard);
pour chaque $l = 1...h$, on estime le prédicteur sur $\bigcup_{j \neq l} \mathcal{A}_{j,\frac{n}{h}}$, et son erreur de prédiction par $\widehat{\varepsilon}_l$ sur $\mathcal{A}_{l,\frac{n}{h}}$. L'erreur de prédiction estimée finale est $\widehat{\varepsilon}_g = \frac{1}{h}\Sigma_{l=1}^{h}\widehat{\varepsilon}_l$.

Le cas ultime $h = n$ est le **jackknife**, que l'on peut utiliser en particulier pour le prédicteur à noyau et le prédicteur 1-NN.

Exercice: Ensemble de Highleyman (suite).

1°) Construire (sous R) un estimateur à noyau gaussien. En déduire un prédicteur à noyau pour diverses valeurs de la fenêtre; estimer les erreurs de prédiction et proposer une fenêtre optimale.

2°) Effectuer la prédiction par prédicteur k-NN pour diverses valeurs de k; estimer les erreurs de prédiction et proposer une valeur de k optimale.

3°) Effectuer la prédiction par méthode CART.