Méthode des k plus proches voisins

Clément Erignoux
sous la direction de Sylvain Arlot

21 fevrier 2011

Table des matières

1 Definition des notions

On cherche dans cette problématique de classification à donner une étiquette à une variable aléatoire X, en observant un échantillon étiqueté $D_n = ((X_1, Y_1), ..., (X_n, Y_n))$

Définition 1 (Observation)**.** *On appelle* ***observation*** *un ensemble de données notée dans la suite X_i à valeurs dans $\mathbb{R}^d$.*

Définition 2 (Classe)**.** *On appelle* ***classe*** *d'une observation, notée y, la nature inconnue de l'observation à valeur dans un ensemble fini $\{1, ..., M\}$.*

Il faut maintenant préciser ce qu'on entend par prédiction : on cherche une fonction qui associe à une observation sa classe .

Définition 3 (Classifieur)**.** *On appelle* ***classifieur*** *une application g telle que*

$$\begin{aligned} g : \mathbb{R}^d &\mapsto \{1, ..., M\} \\ X &\mapsto g(X) \end{aligned}$$

$g(X)$ est la prédiction de la classe de X par le classifieur g.

Dans la pratique, un classifieur se construit grâce aux observations : la prédiction faite dépend de ce qu'on a observé précédemment. Remarquons qu'il est nécessaire de posséder un échantillon du type $(X_1, Y_1, ..., X_n, Y_n)$ où X_i est l'observation et Y_i sa classe. En effet, posséder l'échantillon $(X_1, ..., X_n)$ n'est pas suffisant pour pouvoir extrapoler la classe Y_{n+1} d'une nouvelle observation X_{n+1}.

Définition 4 (Classificateur empirique)**.** *Dans la pratique, un* ***classifieur*** *g_n se construit mesurablement par rapport à $(X_1, Y_1..., X_n, Y_n)$:*

$$\begin{aligned} g_n : \mathbb{R}^d \times (\mathbb{R}^d \times \{1, ..., M\})^n &\mapsto \{1, ..., M\} \\ (X, (X_1, Y_1, ..., X_n, Y_n)) &\mapsto g(X; X_1, Y_1, ..., X_n, Y_n) \end{aligned}$$

On considère, dans la suite, que les couples (X_i, Y_i) sont des variables aléatoires indépendantes identiquement distribuées (v.a. i.i.d.).

Cependant, une prédiction n'est pas parfaite et il convient de mesurer l'erreur entre la prédiction et la réalité.

Définition 5 (Erreur)**.** *Pour une distribution (X, Y) donnée et pour un classifieur g, on note $L(g)$ l'erreur d'un classifieur définie par :*

$$L(g) = \mathbb{P}\{g(X) \neq Y\} .$$

Proposition 1 (Classificateur de Bayes, erreur de Bayes)**.** *Pour une distribution (X, Y) donnée, il existe un meilleur classifieur appelé* ***classifieur de Bayes*** *noté g^* vérifiant :*

$$g^* = \underset{g:\mathbb{R}^d \mapsto \{1,...,M\}}{\arg\min} \mathbb{P}\{g(X) \neq Y\}$$

On appelle ***erreur de Bayes*** *l'erreur du classifieur de Bayes notée $L^* = L(g^*)$.*

Définition 6. *L'erreur d'un classifieur empirique est définie de la même manière par :*

$$L_n = L(g_n) = \mathbb{P}\{g_n(X; X_1, Y_1, ..., X_n, Y_n) \neq Y | X_1, Y_1, ..., X_n, Y_n\}$$

Le classifieur de Bayes étant le meilleur, on cherchera à construire des classifieurs s'approchant le plus possible de ce dernier.

Définition 7. *On appelle* ***règle*** *une suite de classifieur notée $(g_n)_{n\in\mathbb{N}}$. On dira qu'une règle (g_n) est consistante si :*

$$\lim_{n \leftrightarrow +\infty} \mathbb{E}(L_n) = L^* .$$

Il sera dit universellement consistant, si cette consistance est vérifiée indépendamment de la loi μ de l'échantillon.

A partir de maintenant, nous nous intéressons à une règle de classification particulière, la règle des k-plus proches voisins. Comme son nom l'indique, on étiquette chaque échantillon par l'étiquette majoritaire dans ses k-plus proches voisins. On s'intéressera principalement à deux problèmes : d'une part prouver la consistance faible universelle de cette règle dans le cas où k/n tend vers l'infini. De l'autre, énoncer quelques inégalités sur l'erreur asymptotique de la règle des k plus proches voisins, pour k impair fixé.

2 Cas $k/n \to \infty$, consistance faible universelle

2.1 Enoncé du résultat

Le but de cette section est de prouver la consistance faible universelle de la methode des k plus proches voisins. Un tel résultat est donné par le théorème suivant, dû à Stone (1977) :

Théorème 1. *Si $k \to \infty$ et $k/n \to 0$, alors indépendamment de la loi de l'échantillon, on a consistance faible, c'est à dire*

$$\mathbb{E}(L_n) \underset{n\to\infty}{\to} L^* .$$

Afin de prouver le théorème précédent, il suffit en réalité de montrer que la méthode des k plus proches voisins vérifie les conditions du théorème suivant, plus général, dû également à Stone.

Théorème 2. *Supposons que pour toute loi de X, les poids $W_{n,i}$ satisfont les conditions suivantes :*

1. *Il existe une constante c telle que, pour toute fonction f mesurable et positive, satisfaisant $\mathbb{E}(f(X)) < \infty$,*
$$\mathbb{E}\left(\sum_{i=1}^{n} W_{n,i}(X) f(X_i)\right) \leq c\,\mathbb{E}(f(X)) .$$
2. *Pour tout $a > 0$,*
$$\lim_{n\to\infty} \mathbb{E}\left(\sum_{i=1}^{n} W_{n,i}(X) \mathbb{1}_{\{\|X_i - X\| > a\}}\right) = 0 .$$
3.
$$\lim_{n\to\infty} \mathbb{E}\left(\max_{1\leq i\leq n} W_{n,i}(X)\right) = 0 .$$

Alors, g_n est universellement consistant.

Preuve du théorème 1 : Notons à partir de maintenant $G_{k,X} = \{k \text{ plus proches voisins de } X\}$. Comme annoncé précédemment, nous allons montrer que le théorème précédent est applicable pour $W_{n,i} = \mathbb{1}_{X_i \in G_{k,X}}/k$.

Pour ce qui est du 1), il faut montrer que

$$\mathbb{E}\left(\sum_{i=1}^{n} \frac{1}{k} \mathbb{1}_{X_i \in G_{k,X}} f(X_i)\right) \leq \mathbb{E}(cf(X)) ,$$

pour une certaine constante c. Posons $\gamma_d = E\left(\left(1 + 2/\sqrt{2-\sqrt{3}}\right)^d - 1\right)$ (E désigne la partie entière), nous allons démontrer le lemme suivant :

Lemme 1. *Soit f une fonction intégrable, n fixé, $k \leq n$. Alors,*

$$\mathbb{E}\left(\sum_{i=1}^{n} \frac{1}{k} \mathbb{1}_{X_i \in G_{k,X}} f(X_i)\right) \leq \mathbb{E}(\gamma_d f(X)) ,$$

2.2 Preuve du lemme 1

Pour $\theta \in]0, \pi/3[$, on note $C(x, \theta)$ le cône centré en x, issu de 0 et d'angle θ, à savoir l'ensemble des points tels que l'angle entre x et y soit inférieur à θ, ou encore l'ensemble des $y \in \mathbb{R}^d$ tels que

$$\frac{(x, y)}{\|x\|\|y\|} \geq \cos(\theta) \ .$$

Remarquons tout d'abord que pour $\theta = \pi/6$, soient $y, z \in C(x, \pi/6)$, supposons $\|y\| \leq \|z\|$, alors $\|z - y\| \leq \|z\|$. En effet,

$$\begin{aligned} \|y - z\|^2 &= \|y\|^2 + \|z\|^2 - 2\|y\|\|z\| \frac{(y, z)}{\|y\|\|z\|} \\ &\leq \|y\|^2 + \|z\|^2 - 2\|y\|\|z\| cos(\pi/3) \\ &= \|z\|^2 \left(1 + \frac{\|y\|^2}{\|z\|^2} - \frac{\|z\|^2}{\|y\|^2}\right) \\ &< \|z\|^2 \ . \end{aligned}$$

Montrons maintenant le résultat suivant :

Lemme 2. *Soit $\theta \in]0, \pi/2[$, alors, en posant*

$$\gamma_d(\theta) = E\left(\left(1 + \frac{1}{\sin(\theta)}\right)^d - 1\right) \ ,$$

où E désigne la partie entière, il existe un ensemble $\{x_1, ..., x_{\gamma_d(\theta)}\}$ tel que

$$\mathbb{R}^d = \bigcup_{i=1}^{\gamma_d(\theta)} C(x_i, \theta) \ .$$

Autrement dit, on peut recouvrir $\mathbb{R}^d$ par $\gamma_d(\theta)$ cônes issus de 0 et d'angle θ. Remarquons que pour $\theta = \pi/6$, on retrouve le γ_d précédent.

Preuve du lemme 2 Considérons un tel recouvrement fini, de cardinal N et montrons que son cardinal peut être ramené à γ_d. Sans perdre de généralité, on peut supposer que les x_i de norme 1 et que $\|x_i - x_j\| \geq r$ pour tout $j \neq i$. Notons S_i la sphère de centre x_i et de rayon $r = 2sin(\theta/2)$. Remarquons, en notant S la sphère unité dans $\mathbb{R}^d$, que

$$S \cap S_i = S \cap C(x_i, \theta) \ .$$

Notons S_i' la sphère centrée en x_i et de rayon $r/2$. Les S_i' sont disjointes et $\bigcup S_i' \subset S(0, 1 + r/2) - S(0, r/2)$, car $\theta < \pi/3$ donc $r < 1$. Alors, en posant $v_d = vol(S)$, on obtient que

$$N v_d \left(\frac{r}{2}\right)^d \leq v_d \left(1 + \frac{r}{2}\right)^d - v_d \left(\frac{r}{2}\right)^d \ ,$$

soit

$$N \leq \left(1 + \frac{r}{2}\right)^d - 1 = \left(1 + \frac{1}{sin(\theta/2)}\right)^d - 1 = \gamma_d(\theta) \ . \quad \blacksquare$$

Retour à la preuve du lemme 1 :

On recouvre donc désormais $\mathbb{R}^d$ grâce à γ_d cônes $X + C(x_j, \pi/6)$, et on marque dans chaque cône le X_i le plus proche de X, s'il existe. Si X_i est dans le cône $X + C(x_j, \pi/6)$, et n'est pas marqué, alors il ne peut pas être le plus proche voisin de X_i dans $\{X_1, ..., X_{i-1}, X, X_{i+1}, ..., X_n\}$. On procède de même

pour les k plus proches voisins de X dans chaque cône, et on les marque tous s'il y en a moins de k Par un argument similaire, si $X_i \in X + C(x_j, \pi/6)$ et n'est pas marqué, alors il ne peut être dans les k plus proches voisins de X_i dans$\{X_1, ..., X_{i-1}, X, X_{i+1}, ..., X_n\}$. (On utilise ici l'argument donné en début de démonstration du lemme) Par conséquent, si f est une fonction positive ou nulle, par égalité en loi de X et de X_i, on a

$$\begin{aligned}\sum_{i=1}^{n} \mathbb{E}\left(\mathbb{1}_{X_i \in G_{k,X}} f(X_i)\right) &= \sum_{i=1}^{n} \mathbb{E}\left(\mathbb{1}_{X \in G_{k,X_i}(\{X_1,...,X_{i-1},X,X_{i+1},...,X_n\})} f(X)\right) \\ &\leq \mathbb{E}\left(f(X) \sum_{i=1}^{n} \mathbb{1}_{\{X_i \text{est marqué}\}}\right) \\ &\leq k\gamma_d \mathbb{E}(f(X)) \ ,\end{aligned}$$

par majoration du nombre de X_i marqués. Ceci conclut la preuve du lemme 1. ■

2.3 Fin de la preuve du théorème de Stone

Le lemme 1 nous donne donc directement la condition 1) du théorème 1. Pour ce qui est de la condition 2), commençons par remarquer que

$$\mathbb{E}\left(\frac{1}{k}\sum_{i=1}^{n} \mathbb{1}_{X_i \in G_{k,X}} \mathbb{1}_{\|X_i - X\| > \varepsilon}\right) \underset{n\to\infty}{\longrightarrow} 0 \text{ dès lors que}$$

$$\mathbb{P}(\|X_{(k)}(X) - X\| > \varepsilon) \underset{n\to\infty}{\longrightarrow} 0 \ ,$$

où $X_{(k)}(X)$ est le k-ième plus proche voisin de X. Nous allons désormais montrer le résultat suivant :

Lemme 3. *Soit x dans le support de μ, supposons $\lim_{n\to\infty} k/n = 0$, alors $\|X_{(k)}(x) - x\| \underset{n\to\infty}{\longrightarrow} 0$ p.s. Si X est indépendant de l'échantillon et est également de loi μ, alors*

$$\|X_{(k)}(X) - X\| \underset{n\to\infty}{\longrightarrow} 0 \ p.s.$$

Preuve lemme 3 : Soit $\varepsilon > 0$, par définition, on a $\mu(S(x,\varepsilon)) > 0$. Remarquons que

$$\|X_{(k)}(x) - x\| > \varepsilon \Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \in S(x,\varepsilon)} < k/n \ .$$

Par loi forte des grands nombres, la somme tend ves $\mu(x,\varepsilon) > 0$ p.s, alors que k/n tend vers 0 par hypothèse. On en déduit que cet évènement asymptotique se produit avec probabilité nulle, i.e

$$\|X_{(k)}(x) - x\| \underset{n\to\infty}{\longrightarrow} 0 \ p.s.$$

Pour la seconde égalité, notons que $\mathbb{P}(X \in supp(\mu)) = 1$, et donc pour tout $\varepsilon > 0$,

$$\mathbb{P}(\|X_{(k)}(X) - X\| > \varepsilon) = \mathbb{P}(X \in supp(\mu))\mathbb{P}(\|X_{(k)}(X) - X\| > \varepsilon \ \mid \ X \in supp(\mu)) \ .$$

La seconde quantité tend vers 0 par convergence dominée, ce qui donne la convergence en probabilités. Si k ne dépend pas de n, alors $\|X_{(k)}(X) - X\|$ est décroissante pour $n \geq k$, et converge donc également $p.s.$. Si $k = k_n$ vérifie $k_n/n \to 0$, alors, par un argument similaire, on obtient que la suite décroissante de variables aléatoires $\sup_{m\geq n}\|X_{(k)}(X) - X\|$ converge vers 0 en probabilités, et donc $p.s$ également. Le lemme 3 est donc démontré. ■

En réalité, la seule convergence en probabilités suffisait pour le problème qui nous intéresse ici. Le point 2) est donc également démontré.

Le point 3) est évident compte tenu du fait que $k \to \infty$. Nous pouvons donc appliquer le théorème de Stone, ce qui démontre directement le théorème 1. ■

3 Résultats à k fixé

3.1 Comportement asymptotique de L_n

Pour k impair fixé, nous allons désormais étudier le comportement asymptotique de L_n. Notons

$$L_{kNN} = \mathbb{E}\left(\sum_{j=0}^{k}\binom{k}{j}\eta^j(X)(1-\eta(X))^{k-j}(\eta(X)\mathbb{1}_{j<k/2} + (1-\eta(X))\mathbb{1}_{j>k/2})\right) .$$

Dans ces conditions, on a le résultat suivant :

Théorème 3. *Soit k impair fixé. Alors, pour la règle des k plus proches voisins, on a*

$$\lim_{n\to\infty}\mathbb{E}(L_n) = L_{kNN}$$

Afin de démontrer ce résultat, on se ramène à montrer que $\mathbb{E}(L'_n) \longrightarrow L_{kNN}$, pour L'_n une modification de L_n que nous allons définir maintenant. Comme η est une fonction mesurable, donc suffisamment régulière, et que $\|x - X_{(k)}(x)\|$ est faible, les valeurs de $\|\eta(X_{(i)}(x)\|$ devraient être proches de $\eta(x)$ pour tout i assez petit. On définit alors une règle auxiliaire d'étiquetage g'_n dans laquelle les $Y_{(i)}(x)$ sont remplacés par des bernouillis i.i.d. de paramètre $\eta(x)$. On peut montrer facilement que la probabilité d'erreur entre les deux règles est faible. Plus précisément, on suppose données des paires $(X_1, U_1), ..., (X_n, U_n)$, où les X_i sont définis comme précédemment, et les U_i sont des variables aléatoires i.i.d. $U([0,1])$. En posant alors $Y_i = \mathbb{1}_{U_i \geq \eta(X_i)}$, on obtient une famille (X_i, U_i) de même distribution que l'échantillon initial (X, Y). On définit alors $Y'_i = \mathbb{1}_{U_i \geq \eta(x)}$. On admettra que l'on peut alors se restreindre à l'étude de la règle approchée $g'_n(x)$ déterminée par le signe de $\psi(x, Y'_{(1)}(x), ..., Y'_{(k)}(x))$. On définit finalement

$$L'_n = \mathbb{P}(signe(\psi(x, Y'_{(1)}(x), ..., Y'_{(k)}(x))) \neq signe(2Y-1) \mid D'_n) .$$

On peut alors montrer, grâce à la corrélation forte entre g_n et g'_n, que $\mathbb{E}(L_n - L'_n) \underset{n\to\infty}{\longrightarrow} 0$. Il suffit donc de montrer le résultat voulu pour L'_n afin de l'avoir directement pour L_n.

Or on a pour tout n

$$\begin{aligned}\mathbb{E}(L'_n) &= \mathbb{P}\left(Z_1 + ... + Z_k > \frac{k}{2}, Y = 0\right) + \mathbb{P}\left(Z_1 + ... + Z_k < \frac{k}{2}, Y = 1\right) \\ &= \mathbb{P}\left(Z_1 + ... + Z_k > \frac{k}{2}, Z_0 = 0\right) + \mathbb{P}\left(Z_1 + ... + Z_k < \frac{k}{2}, Z_0 = 1\right) ,\end{aligned}$$

où les Z_i sont des bernouillis de paramètre $\eta(X)$, ce qui donne directement le résultat souhaité. ■

3.2 Diverses inégalités de la probabilité d'erreur L_{kNN}

Nous allons dans cette partie énoncer quelques inégalités sur l'erreur L_{kNN}. On va là encore se limiter au cas où k est impair.

Théorème 4. *Pour toute distribution, on a*

$$L^* \leq ... \leq L_{(2k+1)NN} \leq L_{(2k-1)NN} \leq ... \leq L_{3NN} \leq L_{NN} \leq 2L^*$$

$$L_{kNN} = \mathbb{E}(\alpha_k(\eta(X))) \text{ , où l'on a posé}$$

$$\alpha_k(p) = \min(p, 1-p) + |2p-1| \, \mathbb{P}\left(Binomial(k, \min(p, 1-p)) > \frac{k}{2} \right) .$$

Théorème 5. *Pour tout k et pour toute distribution, on a*

$$L_{kNN} \leq L^* + \frac{1}{\sqrt{ke}} .$$

Preuve : Par la représentation précédente de L_{kNN}, on a

$$\begin{aligned}
L_{kNN} - L^* &\leq \sup_{p \in [0,1/2]} (1-2p) \mathbb{P}\left(\mathcal{B}(k,p) > \frac{k}{2} \right) \\
&= \sup_{p \in [0,1/2]} (1-2p) \mathbb{P}\left(\frac{(\mathcal{B}(k,p) - kp)}{k} > \frac{1}{2} - p \right) \\
&\leq \sup_{p \in [0,1/2]} (1-2p) e^{-2k(1/2-p)^2} \text{ (Inégalité de Hoeffding)} \\
&= \sup_{u \in [0,1]} u e^{-ku^2/2} \\
&= \frac{1}{\sqrt{ke}} . \quad \blacksquare
\end{aligned}$$

Théorème 6 (Györfi, 1978)**.** *Pour toute distribution et pour tout k,*

$$L_{kNN} \leq L^* + \sqrt{\frac{2L_{NN}}{k}} .$$

Preuve : Pour $p \leq 1/2$, on a

$$\begin{aligned}
\mathbb{P}\left(\mathcal{B}(k,p) > \frac{k}{2} \right) &= \mathbb{P}\left(\mathcal{B}(k,p) - kp > k\left(\frac{1}{2} - p \right) \right) \\
&\leq \frac{\mathbb{E}(|B - kp|)}{k(1/2-p)} \text{ (Inégalité de Markov)} \\
&\leq \frac{Var(B)}{k(1/2-p)} \text{ (Inégalité de Cauchy-Schwarz)} \\
&= \frac{2\sqrt{p(1-p)}}{\sqrt{k}(1-2p)}
\end{aligned}$$

On en déduit que

$$\begin{aligned}
L_{kNN} - L^* &\leq \mathbb{E}\left(\frac{2}{\sqrt{k}} \sqrt{\eta(X)(1-\eta(X))} \right) \\
&\leq \frac{2}{\sqrt{k}} \sqrt{\mathbb{E}(\eta(X)(1-\eta(X)))} \text{ (Inégalité de Jensen)} \\
&= \frac{2}{\sqrt{k}} \sqrt{\frac{L_{NN}}{2}} \\
&= \sqrt{\frac{2L_{NN}}{k}} . \quad \blacksquare
\end{aligned}$$

Théorème 7 (Devroye, 1981). *For all distribution, et $k \geq 3$ impair,*

$$L_{kNN} \leq L^* \left(1 + \frac{\gamma}{\sqrt{k}} \left(1 + O_{k\to\infty}\left(k^{-1/6}\right)\right)\right) ,$$

où $\gamma = \sup_{r>0} 2r\mathbb{P}(N > r) \simeq 0,4$, où N suit une loi normale $(0,1)$.

remarque : La règle des k plsu proches voisins n'est pas "intelligente", dans le sens où $\mathbb{E}(L_n)$ n'est pas décroissante.

Références

[1] Luc Devroye, Laszlo Györfi, and Gábor Lugosi. A probabilistic theory of pattern recognition. *Applications of Mathematics*, 31, 1996.