An optimal aggregation type classifier

Alejandro Cholaquidis

CMAT-Facultad de Ciencias, UdelaR

Seminario de Probabilidad y Estadística
Abril 2014

On the classification problem

Goal:

From a training sample $\mathcal{D}_n = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ i.i.d. of (X, Y) with values in $\mathcal{F} \times \{0, 1\}$ we want a predictor $g : \mathcal{F} \to \{0, 1\}$, which *minimize* $\mathbb{P}\big(g(X) \neq Y\big)$.

Let us denote

1) $\eta(x) = \mathbb{E}(Y|X = x) = \mathbb{P}(Y = 1|X = x)$ the regression function.
2) $g^*(x) = \mathbb{I}_{\{\eta(x) > 1/2\}}$ the Bayes rule.
3) $L^* = \mathbb{P}\big(g^*(X) \neq Y\big)$ the optimal Bayes risk.

The classical estimator of $\eta(x)$ is

$$\eta_n(X) = \sum_{i=1}^{n} W_{ni}(X) Y_i,$$

for some weights $W_{ni}(X) = W_{ni}(X, X_1, \ldots, X_n)$.

On the classification problem: Stone Theorem

Theorem

Stone 1977: *If $\mathcal{F} = \mathbb{R}^d$, and the weights fulfills:*

1) *There exists a constant $C \geq 1$ such that for every nonnegative Borel function*

$$\mathbb{E}\Big(\sum_{i=1}^{n} |W_{ni}| f(X_i)\Big) \leq C\mathbb{E}\big(f(X)\big) \quad \forall n \geq 1$$

2) *there exists a $D \geq 1$ such that* $\mathbb{P}\Big(\sum_{i=1}^{n} |W_{ni}(X)| \leq D\Big) = 1 \quad \forall n \geq 1$

3) $\sum_{i=1}^{n} |W_{ni}(X)| \mathbb{I}_{\{\|X_i - X\| > a\}} \to 0$ *in probability,* $\forall a > 0$

4) $\sum_{i=1}^{n} W_{ni}(X) \to 1$ *in probability, and*

5) $max_i |W_{ni}(X)| \to 0$

then

$$\mathbb{E}\Big(\mathbb{P}\big(\mathbb{I}_{\{\eta_n(X) \geq 1/2\}} \neq Y | D_n\big)\Big) \to \mathbb{P}\Big(\mathbb{I}_{\{\eta(X) \geq 1/2\}} \neq Y\Big).$$

k-NN and the infinite dimensional case

k-NN

For $W_{ni}(X) = \frac{1}{k}\mathbb{I}_{\{X_i \in k_n(X)\}}$, where $X_i \in k_n(X)$ if X_i is one of the k nearest neighbours of X, conditions 1 to 5 holds if $n \to \infty$ and $k_n/n \to 0$.

Theorem

Let $(X, Y) \in \mathcal{F} \times \{0, 1\}$ where $(\mathcal{F}, d)$ is a separable metric space. Suppose that the probability measure μ of X is a Borel measure. If η satisfies the Besicovitch Condition:

$$\lim_{\delta \to 0} \mu\left\{ x \in \mathcal{F} : \frac{1}{\mu\big(B(x,\delta)\big)} \int_{B(x,\delta)} \big|\eta(y) - \eta(x)\big| d\mu(y) > \varepsilon \right\} = 0, \tag{1}$$

then, taking $k_n \to \infty$ and $k_n/n \to 0$, k-NN is consistent:

$$\mathbb{E}\Big(\mathbb{P}\big(\mathbb{I}_{\{\eta_n(X) \geq 1/2\}} \neq Y | \mathcal{D}_n\big)\Big) \to \mathbb{P}\Big(\mathbb{I}_{\{\eta(X) \geq 1/2\}} \neq Y\Big).$$

On the Besicovitch condition

If $(\mathcal{F}, d)$ is complete and separable, Besicovitch condition holds for all measurable function $f : \mathcal{F} \to \mathbb{R}$ if and only if:

$$\lim_{\delta \to 0} \mu\left\{ x \in \mathcal{F} : \left| \frac{1}{\mu\big(B(x,\delta)\big)} \int_{B(x,\delta)} f(y) d\mu(y) - f(x) \right| > \varepsilon \right\} = 0. \tag{2}$$

Condition (2) holds if any of the following conditions holds:

- f is a continuous function.
- μ is doubling: $\exists C \geq 1$ such that $\forall r > 0$, $\mu\big(B(x,2r)\big) \leq C\mu\big(B(x,r)\big)$.
- $(\mathcal{F}, d)$ is a Vitali space: for every $\mathcal{G}$ collection of closed balls such that for all $a \in \mathcal{F}$, $\inf\big\{r > 0 : B(a,r) \in \mathcal{G}\big\} = 0$ then there exists a disjoint covering of $\mathcal{F}$ with balls in $\mathcal{G}$, a.s. with respect to μ.
- $\mathcal{F}$ is finite dimensional: $\exists n \in \mathbb{N}$ and $s \in (0, +\infty)$: $\bigcap_{x \in \mathcal{G}} B\big(x, r(x)\big) = \emptyset$,

  $\forall \mathcal{G} \subset \mathcal{F}$ finite set such that:

$$1)\ \#\mathcal{G} > n \quad 2)\ r(x) \in (0,s) \quad 3)\ x \notin B\big(y, r(y)\big)\ \ \forall x \neq y \in \mathcal{G}$$

Mojirsheibani 1999

1) With $\mathcal{D}_n$ build up $g_{ni} : \mathbb{R}^d \to \{0,1\}, i = 1, \ldots, M$
2) Taking $\mathbf{g_n}(x) \doteq \big(g_{n1}(x), \ldots, g_{nM}(x)\big)$, we define the classifier:

$$g_T(x) = \mathbb{I}_{\{T_n(\mathbf{g_n}(x)) > 1/2\}},$$

where

$$T_n(\mathbf{g_n}(x)) = \sum_{j=1}^{n} W_{n,j}(x) Y_j, \quad x \in \mathcal{F}, \tag{3}$$

and the weights $W_{n,j}(x)$ given by

$$W_{n,j}(x) = \frac{\mathbb{I}_{\{\mathbf{g_n}(x) = \mathbf{g_n}(X_j)\}}}{\sum_{j=1}^{n} \mathbb{I}_{\{\mathbf{g_n}(x) = \mathbf{g_n}(X_j)\}}}. \tag{4}$$

Here, $0/0$ is assumed to be 0.

Mojirsheibani 1999, consistency results

Notation:

- Taking $\mathcal{T}_n \in \left(\mathbb{R}^d \times \{0,1\}\right)^n$, the sets: $\Pi_n(\mathcal{T}_n) = \{A_{n,1}, \ldots, A_{n,2^M}\}$ define a partition of $\mathbb{R}^d$ where

$$A_{n,i} = \left\{x \in \mathbb{R}^d : \left(g_{n,1}(x), \ldots, g_{n,M}(x)\right) = \nu_i \text{ where } \nu_i \in \{0,1\}^M\right\}.$$

  The family of all partitions induced by $g_{n,1}, \ldots, g_{n,M}$ is:

$$\Omega_n = \bigcup_{\mathcal{T}_n \in \left(\mathbb{R}^d \times \{0,1\}\right)^n} \Pi_n(\mathcal{T}_n).$$

- Let $B = \{x_1, \ldots, x_n\}$, define $\Delta_n(\Omega_n)$ the number of distinct partitions:

$$\left\{\Pi_n(\mathcal{T}_n^1) \cap B, \ldots, \Pi_n(\mathcal{T}_n^r) \cap B\right\},$$

  of the finite set B that are induced by partitions $\left\{\Pi_n(\mathcal{T}_n^1), \ldots, \Pi_n(\mathcal{T}_n^r)\right\} \in \Omega_n$.

Mojirsheibani 1999, consistency results

Theorem

If, as $n \to \infty$,

- $M_n / \log(n) \to 0$
- $n^{-1} \log(\Delta_n) \to 0$
- *for every* $\gamma > 0$ *and* $\delta \in (0,1)$

$$\inf_{S \subset \mathbb{R}^d : \mu(S) \geq 1-\delta} \mu\Big\{ x : diam\big(A_n[x] \cap S\big) > \gamma \Big\} \to 0 \quad a.s.$$

if we denote $g_T^*(x) = \mathbb{I}_{\left\{\mathbb{P}(Y=1|g_n(x))>1/2\right\}}$ *the optimal combined classifier, then*

$$\mathbb{P}\big(g_T(X) \neq Y | \mathcal{D}_n\big) - \mathbb{P}\big(g_T^*(X) \neq Y | \mathcal{D}_n\big) \to 0 \quad a.s.$$

The classifier

1) Split $\mathcal{D}_n$ into
   - $\mathcal{D}_k = \{(X_1, Y_1), \ldots, (X_k, Y_k)\}$
   - $\mathcal{E}_l = \{(X_{k+1}, Y_{k+1}), \ldots, (X_n, Y_n)\}$ with $l = n - k \geq 1$.
2) With $\mathcal{D}_k$ build up $g_{ki} : \mathcal{F} \to \{0, 1\}$, $i = 1, \ldots, M$
3) Taking $\mathbf{g_k}(x) \doteq \big(g_{k1}(x), \ldots, g_{kM}(x)\big)$, we define the classifier:

$$g_T(x) = \mathbb{I}_{\{T_n(\mathbf{g_k}(x)) > 1/2\}},$$

where

$$T_n(\mathbf{g_k}(x)) = \sum_{j=k+1}^{n} W_{n,j}(x) Y_j, \quad x \in \mathcal{F}, \tag{5}$$

and the weights $W_{n,j}(x)$ given by

$$W_{n,j}(x) = \frac{\mathbb{I}_{\{\mathbf{g_k}(x) = \mathbf{g_k}(X_j)\}}}{\sum_{j=k+1}^{n} \mathbb{I}_{\{\mathbf{g_k}(x) = \mathbf{g_k}(X_j)\}}}. \tag{6}$$

Here, $0/0$ is assumed to be 0.

The classifier, example

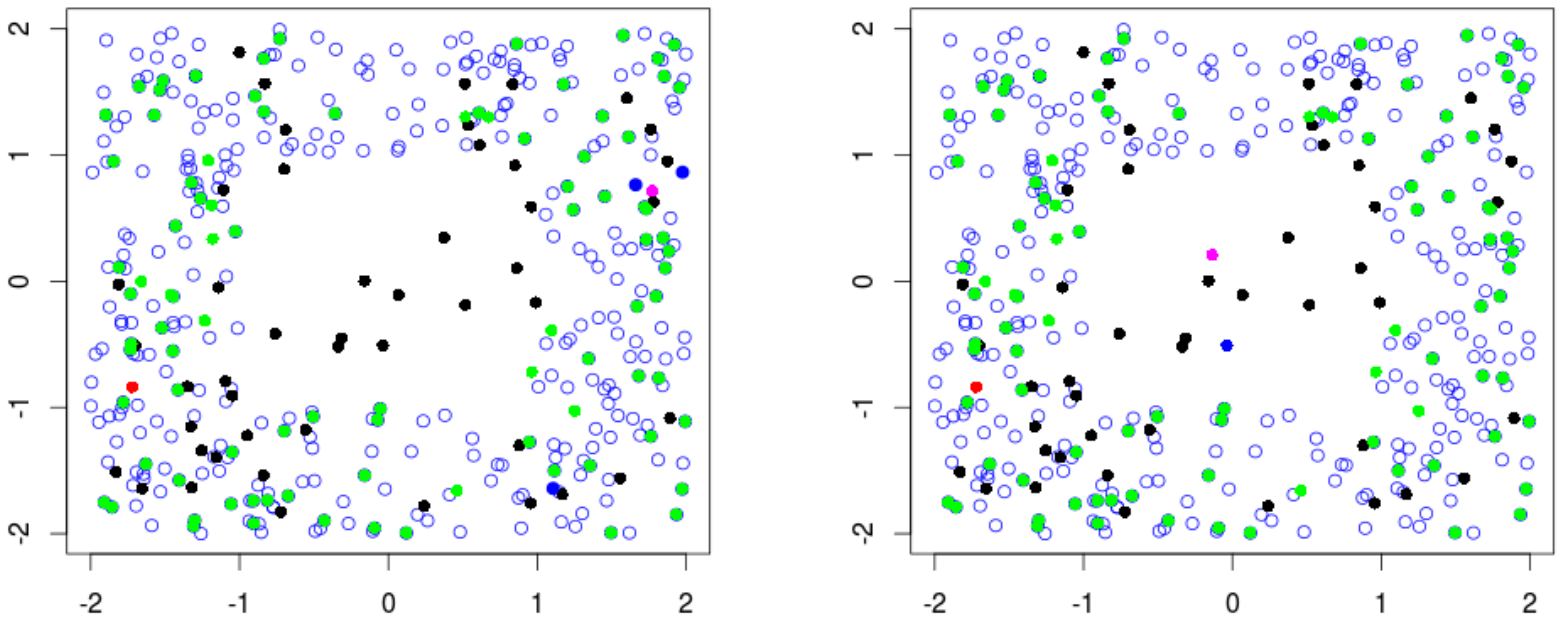

Figure : An example in $\mathbb{R}^2$ where the 0-class points are shown in black dots and the 1-class points are shown in blue circles. In both pictures the voters of the point in red are shown as green dots while the voters of he point in magenta are shown with blue dots.

A more flexible version

For $0 \leq \alpha < 1$ a more flexible version of the classifier, called $g_T(x, \alpha)$, can be defined replacing the weights in (6) by

$$W_{n,j}(x) = \frac{\mathbb{I}_{\{\frac{1}{M}\sum_{i=1}^{M}\mathbb{I}_{\{g_{ki}(x)=g_{ki}(X_j)\}} \geq 1-\alpha\}}}{\sum_{j=k+1}^{n}\mathbb{I}_{\{\frac{1}{M}\sum_{i=1}^{M}\mathbb{I}_{\{g_{ki}(x)=g_{ki}(X_j)\}} \geq 1-\alpha\}}}. \tag{7}$$

Observe that if we choose $\alpha = 0$ in (7) we obtain original weights given in (6).

Consistency

Theorem

Consistency: *Suppose that for every $i = 1, \ldots, M$ the classifier g_{ki} converges in probability to g^* as $k \to \infty$. If $\mathbb{P}(Y = 1 | g^*(X) = 1) > 1/2$ and $\mathbb{P}(Y = 0 | g^*(X) = 0) > 1/2$, then*

$$\lim_{k \to +\infty} \lim_{l \to +\infty} \mathbb{P}_{\mathcal{D}_k}(g_T(X) \neq Y) - L^* = 0.$$

Theorem

Optimality: *Let $\mathbb{C} \doteq \{0, 1\}^M$. For $\nu \in \mathbb{C}$ we define:*

$$A^0_\nu \doteq \bigcap_{i=1}^{M} g_{ki}^{-1}(\nu(i)) \cap \{Y = 0\}, \quad A^1_\nu \doteq \bigcap_{i=1}^{M} g_{ki}^{-1}(\nu(i)) \cap \{Y = 1\}$$

$\forall \nu \in \mathbb{C}$, we assume that $\mathbb{P}_{\mathcal{D}_k}\big((X, Y) \in A^1_\nu\big) \neq \mathbb{P}_{\mathcal{D}_k}\big((X, Y) \in A^0_\nu\big)$ *a.s., then*

$$\lim_{l \to \infty} \mathbb{P}_{\mathcal{D}_k}\big(g_T(X) \neq Y\big) \leq \mathbb{P}_{\mathcal{D}_k}\big(g_{ki}(X) \neq Y\big),$$

for each $i = 1, \ldots, M$. Therefore, $\lim_{l \to \infty} \mathbb{P}_{\mathcal{D}_k}\big(g_T(X) \neq Y\big) \leq \min_{1 \leq i \leq M} \mathbb{P}_{\mathcal{D}_k}\big(g_{ki}(X) \neq Y\big)$.

Generating the sample $\mathcal{D}_n$

1) Generate $Z_1 \ldots, Z_N$ iid uniform in $[0, 1]$,
   - if $Z_i > p$ we generate $X_i \sim \mu_1$ and $Y_i = 1$,
   - if $Z_i \leq p$ we generate $X_i \sim \mu_0$ and set $Y_i = 0$.
2) The first $n = N/2$ pairs (X_i, Y_i) for the training sample, the remaining for the testing sample.
3) We consider M nearest neighbour classifiers with the numbers of neighbours taken as follows:
   - We fix M consecutive odd numbers.
   - We choose at random M different odd integers between 1 and $\min\{\sum_{i=1}^{k} Y_i, k - \sum_{i=1}^{k} Y_i\}$.

In $\mathbb{R}^{150}$

- $M = 8$ neighbour rules with $k = 5, 7, 9, 11, 13, 15, 17, 19$
- $p = 1/6$
- $\mu_1 \sim U([-2,2]^{150})$
- $\mu_0 \sim U\big(\tau_v([-2,2]^{150})\big)$ where τ_v is the translation along $(v, \ldots, v) \in \mathbb{R}^{150}$, for $v = 1/4$.

n (k)	$g_T(\cdot)$	$g_T(\cdot, 1/4)$	g_{k1}	g_{k2}	g_{k3}	g_{k4}	g_{k5}	g_{k6}	g_{k7}	g_{k8}
400	.046	.056	.071	.067	.066	.067	.068	.069	.071	.073
(300)			(.074)	(.072)	(.072)	(.073)	(.074)	(.077)	(.080)	(.082)
600	.043	.052	.067	.062	.061	.061	.061	.062	.063	.065
(400)			(.072)	(.069)	(.068)	(.068)	(.069)	(.071)	(.073)	(.076)
800	.037	.045	.062	.057	.055	.055	.055	.056	.056	.057
(600)			(.066)	(.061)	(.060)	(.060)	(.060)	(.061)	(.062)	(.064)
1000	.035	.043	.061	.055	.053	.052	.052	.052	.053	.054
(700)			(.065)	(.060)	(.058)	(.057)	(.057)	(.058)	(.059)	(.060)

Table : $\mathbb{R}^{150}$ with fixed number of neighbours

In $\mathbb{R}^{150}$

We take at random $M = 10$ values for k, and compare with cross validated nearest neighbour classifier.

n	k	$g_T(\cdot)$	$g_T(\cdot, 1/8)$	$g_T(\cdot, 1/4)$	gcv_n	gcv_k
400	300	.052	.065	.077	.068	.073
600	400	.049	.063	.074	.061	.068
800	500	.048	.062	.073	.056	.061
1000	700	.047	.061	.072	.053	.058

Table : $\mathbb{R}^{150}$ with the number of neighbours chosen at random, compared with cross validation

FDA

- $M = 6$ neighbour rules with $k = 3, 9, 15, 21, 27, 33$
- $p = 1/8$
- if $Z_i < p$ then $X_i = B(t)$ where $B(t)$ is the Brownian bridge on a grid of length 100, and $Y_i = 0$
- if $Z_i \geq p$, $X_i = B(t) + 4t(1-t)$ and $Y_i = 1$.

n (k)	$g_T(\cdot)$	$g_T(\cdot, 1/4)$	g_{k1}	g_{k2}	g_{k3}	g_{k4}	g_{k5}	g_{k6}
400	.074	.072	.079	.071	.070	.069	.070	.070
(300)			(.079)	(.072)	(.070)	(.070)	(.071)	(.072)
600	.072	.070	.080	.071	.070	.069	.069	.069
(400)			(.080)	(.071)	(.070)	(.070)	(.070)	(.070)
800	.071	.069	.080	.071	.070	.069	.069	.068
(500)			(.079)	(.071)	(.070)	(.069)	(.069)	(.069)
1000	.071	.068	.080	.071	.069	.069	.068	.068
700			(.079)	(.071)	(.069)	(.069)	(.068)	(.068)

Table : Functional Data with fixed number of neighbours and the values of g_{ki}

Baíllo, A.; Cuevas, A. y Fraiman, R. (2010) - *Classification methods for functional data*. The Oxford Handbook of Functional Data Analysis.

Billingsley, P. (1999) *Convergence of probability measures, s.e.*. Wiley.

Cérou, F. y Guyader, A. (2005) *Nearest neighbor classification in infinite dimension*. Rapport de Recherche 5536, INRIA.

Devroye, L.; Györfi, L. y Lugosi, G. (1996) *A probabilistic theory of pattern recognition.*. Springer–Verlag, New York.

Forzani, L.; Fraiman, R. y Llop, P. (2012) *Consistent nonparametric regression for functional data under the Stone-Besicovitch conditions*

Mojirsheibani, M. (1999) Combining classifiers via discretization. *JASA*, **94**(446), 600–609.

Mojirsheibani, M. (2002) An almost surely optimal combined classification rule. *J. Multivariate Anal.* **81**(1), 28–46.

Stone, C. (1977) *Consistent Nonparametric Regression*. The Annals of Statistics.