Aprendizaje semi-supervisado.

Alejandro Cholaquidis

CMAT-Facultad de Ciencias, UdelaR
Montevideo Uruguay

En conjunto con: R. Fraiman and M. Sued

Seminario de Probabilidad y Estadística

Clasificación

Objetivo

A partir de $\mathcal{D}_n = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$ i.i.d. de $(X, Y) \in \mathcal{F} \times \{0, 1\}$ construir un predictor $g : \mathcal{F} \to \{0, 1\}$, que *minimice* $\mathbb{P}\big(g(X) \neq Y\big)$.

Denotamos

1) $\eta(x) = \mathbb{E}(Y|X = x) = \mathbb{P}(Y = 1|X = x)$ la función de regresión.
2) $g^*(x) = \mathbb{I}_{\{\eta(x)>1/2\}}$ la regla de Bayes.
3) $L^* = \mathbb{P}\big(g^*(X) \neq Y\big)$ riesgo de Bayes.

Si $\eta_n(x) : \mathcal{F} \to [0, 1]$ y $g_n(x) = \mathbb{I}_{\{\eta_n(x)>1/2\}}$, entonces

$$0 \leq \mathbb{P}(g_n(X) \neq Y) - L^* \leq \mathbb{E}|\eta(X) - \eta_n(X)|^2.$$

Notación y objetivos

- $\mathcal{D}^n = (\mathcal{X}^n, \mathcal{Y}^n) = \{(X^1, Y^1), \dots, (X^n, Y^n)\}$ una muestra de $(X, Y) \in S \times \{0, 1\}$, donde $S \subset \mathbb{R}^d$ es compacto. Supondremos que son identicamente distribuidos pero no necesariamente independientes.
- $\mathcal{D}_l = (\mathcal{X}_l, \mathcal{Y}_l) = \{(X_1, Y_1) \dots, (X_l, Y_l)\}$ donde $n \ll l$ es iid de (X, Y). $\mathcal{X}_l$ es conocida pero las etiquetas $\mathcal{Y}_l$ son desconocidas.

Notación y objetivos

- $\mathcal{D}^n = (\mathcal{X}^n, \mathcal{Y}^n) = \{(X^1, Y^1), \ldots, (X^n, Y^n)\}$ una muestra de $(X, Y) \in S \times \{0, 1\}$, donde $S \subset \mathbb{R}^d$ es compacto. Supondremos que son identicamente distribuidos pero no necesariamente independientes.
- $\mathcal{D}_l = (\mathcal{X}_l, \mathcal{Y}_l) = \{(X_1, Y_1) \ldots, (X_l, Y_l)\}$ donde $n \ll l$ es iid de (X, Y). $\mathcal{X}_l$ es conocida pero las etiquetas $\mathcal{Y}_l$ son desconocidas.

Objetivo

Minimizar

$$L(\mathbf{g}_l) := E\left(\frac{1}{l}\sum_{i=1}^{l} \mathbb{I}_{\{i: g_i(\mathcal{X}_l) \neq Y_i\}}\right).$$

Sea $\mathbb{G}_l$ el conjunto de los clasificadores $\mathbf{g}_l : S^l \to \{0, 1\}^l$. Definimos $\mathbf{g}_l^*$ que minimiza $L(\mathbf{g}_l)$. El mínimo de $L(\mathbf{g}_l)$ se obtiene con $(g^*, \ldots, g^*)$, l copias de g^*: sea $(g_1, \ldots, g_l) \in \mathbb{G}_l$, $i = 1, \ldots, l$,

$$P(g_i(\mathcal{X}_l) \neq Y_i) = E\Big(P\big(g_i(\mathcal{X}_l) \neq Y_i | \mathcal{X}_l \setminus X_i\big)\Big) \leq P(g^*(X_i) \neq Y_i) = L^*$$

Versión muestral

Queremos encontrar una sucesión $\mathbf{g}_{n,l} = (g_{n,l,1}, \ldots, g_{n,l,l})$ dependiente de $\mathcal{D}^n$ y $\mathcal{X}_l$, tal que

$$\lim_{l\to\infty} E\left(\frac{1}{l}\#\{i : g_{n,l,i}(\mathcal{X}_l) \neq Y_i, (X_i, Y_i) \in \mathcal{D}_l\}\Big|\mathcal{D}^n\right) - L(\mathbf{g}_l^*) = 0 \quad c.s., \tag{1}$$

cuando n es fijo, y $l \to \infty$.

Denotamos

$$L_n(\mathbf{g}_{n,l}) = E\Big(\#\{i : g_{n,l,i}(\mathcal{X}_l) \neq Y_i, (X_i, Y_i) \in \mathcal{D}_l\}|\mathcal{D}^n\Big). \tag{2}$$

Versión muestral

En cada paso vamos a incorporar $(X_i, \tilde{Y}_i)$ a $\mathcal{D}^n$. En el paso i tenemos

- $\mathcal{T}_{i-1} = \mathcal{D}^n \cup \{(X_{j_1}, \tilde{Y}_{j_1}), \ldots, (X_{j_{i-1}}, \tilde{Y}_{j_{i-1}})\}$, $\mathcal{T}_0 = \mathcal{D}^n$.
- $\mathcal{Z}_i = \{X^1, \ldots, X^n\} \cup \{X_{j_1}, \ldots, X_{j_i}\}$.

Sea $h_l \to 0$, $\mathbf{g}_{n+i-1} \in \mathbb{G}_1$ son clasificadores por núcleos (uniformes) y ventana h_l, construidos con $\mathcal{T}_{i-1}$.

Para cada $X_j \in \mathcal{X}_l \setminus \{X_{j_1}, \ldots, X_{j_{i-1}}\}$, la versión empírica $\hat{\eta}_{i-1}$ de η basada en $\mathcal{T}_{i-1}$, es

$$\hat{\eta}_{i-1}(X_j) = \frac{\sum_{r:(X_r,Y_r)\in\mathcal{D}^n} Y_r \mathbb{I}_{B(X_j,h_l)}(X_r) + \sum_{r:(X_r,\tilde{Y}_r)\in\mathcal{T}_{i-1}\setminus\mathcal{D}^n} \tilde{Y}_r \mathbb{I}_{B(X_j,h_l)}(X_r)}{\sum_{r:(X_r,Y_r)\in\mathcal{T}_{i-1}} \mathbb{I}_{B(X_j,h_l)}(X_r)}.$$

Algoritmo

Inicio: $(\hat{\eta}_0(X_1), \ldots, \hat{\eta}_0(X_l)), \mathcal{Z}_0 = \mathcal{X}^n.$

Algoritmo

Inicio: $(\hat{\eta}_0(X_1), \ldots, \hat{\eta}_0(X_l))$, $\mathcal{Z}_0 = \mathcal{X}^n$.

$1 \leq i < l$: Dado $\hat{\eta}_{i-1}(X_r)$ para todo $X_r \in \mathcal{X}_l \setminus \mathcal{Z}_{i-1}$.

Sea $X_{j_i} \in \mathcal{X}_l \setminus \mathcal{Z}_{i-1}$ tal que $\#\{\mathcal{Z}_{i-1} \cap B(X_{j_i}, h_l)\} > 0$,

$$j_i = \operatorname*{arg\,max}_{j : X_j \in \mathcal{X}_l \setminus \mathcal{Z}_{i-1}} \max \left\{ \hat{\eta}_{i-1}(X_j), 1 - \hat{\eta}_{i-1}(X_j) \right\}. \tag{3}$$

Si hay mas de un j_i que cumple (3), elegimos el que maximiza $\#\{\mathcal{X}_l \cap B(X_{j_i}, h_l)\}$.

$\tilde{Y}_{j_i} := \mathbf{g}_{n+i-1}(X_{j_i})$
$\mathbf{g}_{n+i-1}$ usando $\mathcal{T}_{i-1} = \mathcal{D}^n \cup \{(X_{j_1}, \tilde{Y}_{j_1}), \ldots, (X_{j_{i-1}}, \tilde{Y}_{j_{i-1}})\}$ y h_l.

Algoritmo

Inicio: $(\hat{\eta}_0(X_1), \ldots, \hat{\eta}_0(X_l))$, $\mathcal{Z}_0 = \mathcal{X}^n$.

$1 \leq i < l$: Dado $\hat{\eta}_{i-1}(X_r)$ para todo $X_r \in \mathcal{X}_l \setminus \mathcal{Z}_{i-1}$.

Sea $X_{j_i} \in \mathcal{X}_l \setminus \mathcal{Z}_{i-1}$ tal que $\#\{\mathcal{Z}_{i-1} \cap B(X_{j_i}, h_l)\} > 0$,

$$j_i = \underset{j: X_j \in \mathcal{X}_l \setminus \mathcal{Z}_{i-1}}{\arg\max} \max\left\{\hat{\eta}_{i-1}(X_j), 1 - \hat{\eta}_{i-1}(X_j)\right\}. \tag{3}$$

Si hay mas de un j_i que cumple (3), elegimos el que maximiza $\#\{\mathcal{X}_l \cap B(X_{j_i}, h_l)\}$.

$\tilde{Y}_{j_i} := \mathbf{g}_{n+i-1}(X_{j_i})$
$\mathbf{g}_{n+i-1}$ usando $\mathcal{T}_{i-1} = \mathcal{D}^n \cup \{(X_{j_1}, \tilde{Y}_{j_1}), \ldots, (X_{j_{i-1}}, \tilde{Y}_{j_{i-1}})\}$ y h_l.
Actualizar:
$\mathcal{Z}_i := \mathcal{Z}_{i-1} \cup X_{j_i}$.
$\mathcal{T}_i := \mathcal{D}^n \cup \{(X_{j_1}, \tilde{Y}_{j_1}), \ldots, (X_{j_i}, \tilde{Y}_{j_i})\}$ y calcular $\hat{\eta}_i(X_r)$ con $X_r \in \mathcal{X}_l \setminus \mathcal{Z}_i$.

Algoritmo

Inicio: $(\hat{\eta}_0(X_1), \ldots, \hat{\eta}_0(X_l))$, $\mathcal{Z}_0 = \mathcal{X}^n$.

$1 \leq i < l$: Dado $\hat{\eta}_{i-1}(X_r)$ para todo $X_r \in \mathcal{X}_l \setminus \mathcal{Z}_{i-1}$.

Sea $X_{j_i} \in \mathcal{X}_l \setminus \mathcal{Z}_{i-1}$ tal que $\#\{\mathcal{Z}_{i-1} \cap B(X_{j_i}, h_l)\} > 0$,

$$j_i = \underset{j: X_j \in \mathcal{X}_l \setminus \mathcal{Z}_{i-1}}{\arg\max} \max \left\{ \hat{\eta}_{i-1}(X_j), 1 - \hat{\eta}_{i-1}(X_j) \right\}. \tag{3}$$

Si hay mas de un j_i que cumple (3), elegimos el que maximiza $\#\{\mathcal{X}_l \cap B(X_{j_i}, h_l)\}$.

$\tilde{Y}_{j_i} := \mathbf{g}_{n+i-1}(X_{j_i})$
$\mathbf{g}_{n+i-1}$ usando $\mathcal{T}_{i-1} = \mathcal{D}^n \cup \{(X_{j_1}, \tilde{Y}_{j_1}), \ldots, (X_{j_{i-1}}, \tilde{Y}_{j_{i-1}})\}$ y h_l.
Actualizar:
$\mathcal{Z}_i := \mathcal{Z}_{i-1} \cup X_{j_i}$.
$\mathcal{T}_i := \mathcal{D}^n \cup \{(X_{j_1}, \tilde{Y}_{j_1}), \ldots, (X_{j_i}, \tilde{Y}_{j_i})\}$ y calcular $\hat{\eta}_i(X_r)$ con $X_r \in \mathcal{X}_l \setminus \mathcal{Z}_i$.

Salida $\{(X_{j_1}, \tilde{Y}_{j_1}), \ldots, (X_{j_l}, \tilde{Y}_{j_l})\}$.

Hipótesis

$$
\begin{aligned}
I_1 &= \eta^{-1}((1/2, 1]) \quad I_0 = \eta^{-1}([0, 1/2)) \\
A_1^{\delta} &= I_1 \ominus B(0, \delta) \quad A_0^{\delta} = I_0 \ominus B(0, \delta) \\
B_1^{\delta} &= I_1 \cap B(I_0, \delta) \quad B_0^{\delta} = I_0 \cap B(I_1, \delta)
\end{aligned}
$$

Hipótesis

$$
\begin{aligned}
I_1 &= \eta^{-1}((1/2,1]) \quad I_0 = \eta^{-1}([0,1/2)) \\
A_1^\delta &= I_1 \ominus B(0,\delta) \quad A_0^\delta = I_0 \ominus B(0,\delta) \\
B_1^\delta &= I_1 \cap B(I_0,\delta) \quad B_0^\delta = I_0 \cap B(I_1,\delta)
\end{aligned}
$$

H0) S es estandard

H1) $P_X(I_1) > 0$, $P_X(I_0) > 0$, conexos con borde una variedad $(d-1)$-dimensional C^2.

Hipótesis

$$
\begin{aligned}
I_1 &= \eta^{-1}((1/2, 1]) \quad I_0 = \eta^{-1}([0, 1/2)) \\
A_1^\delta &= I_1 \ominus B(0, \delta) \quad A_0^\delta = I_0 \ominus B(0, \delta) \\
B_1^\delta &= I_1 \cap B(I_0, \delta) \quad B_0^\delta = I_0 \cap B(I_1, \delta)
\end{aligned}
$$

H0) S es estandard

H1) $P_X(I_1) > 0$, $P_X(I_0) > 0$, conexos con borde una variedad $(d-1)$-dimensional C^2.

H2) $P_X(\eta^{-1}(1/2)) = 0$.

Hipótesis

$$
\begin{aligned}
I_1 &= \eta^{-1}((1/2, 1]) \quad I_0 = \eta^{-1}([0, 1/2)) \\
A_1^\delta &= I_1 \ominus B(0, \delta) \quad A_0^\delta = I_0 \ominus B(0, \delta) \\
B_1^\delta &= I_1 \cap B(I_0, \delta) \quad B_0^\delta = I_0 \cap B(I_1, \delta)
\end{aligned}
$$

H0) S es estandard

H1) $P_X(I_1) > 0$, $P_X(I_0) > 0$, conexos con borde una variedad $(d-1)$-dimensional C^2.

H2) $P_X(\eta^{-1}(1/2)) = 0$.

H3) $H1$ y bordes C^3

Hipótesis

$$
\begin{aligned}
I_1 =& \eta^{-1}((1/2,1]) \quad I_0 = \eta^{-1}([0,1/2)) \\
A_1^\delta =& I_1 \ominus B(0,\delta) \quad A_0^\delta = I_0 \ominus B(0,\delta) \\
B_1^\delta =& I_1 \cap B(I_0,\delta) \quad B_0^\delta = I_0 \cap B(I_1,\delta)
\end{aligned}
$$

H0) S es estandard

H1) $P_X(I_1) > 0$, $P_X(I_0) > 0$, conexos con borde una variedad $(d-1)$-dimensional C^2.

H2) $P_X(\eta^{-1}(1/2)) = 0$.

H3) $H1$ y bordes C^3

H4) Sea $h_l \to 0$ tal que $lh_l^{2d}/\log(l) \to \infty$. (X,Y) satisface $H4$ si P_X tiene densidad f, continua tal que para todo $\delta > 0$ existe $\gamma = \gamma(\delta)$, tal que

$$
f(x) - f(y) > \gamma > 0 \text{ for all } x \in (B_1^\delta \cup B_0^\delta)^c \text{ and all } y \in (B_1^{h_l} \cup B_0^{h_l}), \tag{4}
$$

para l suficientemente grande tal que $2h_l < \delta$

Hipótesis

$$
\begin{aligned}
I_1 &= \eta^{-1}((1/2, 1]) \quad I_0 = \eta^{-1}([0, 1/2)) \\
A_1^\delta &= I_1 \ominus B(0, \delta) \quad A_0^\delta = I_0 \ominus B(0, \delta) \\
B_1^\delta &= I_1 \cap B(I_0, \delta) \quad B_0^\delta = I_0 \cap B(I_1, \delta)
\end{aligned}
$$

H0) S es estandard

H1) $P_X(I_1) > 0$, $P_X(I_0) > 0$, conexos con borde una variedad $(d-1)$-dimensional C^2.

H2) $P_X(\eta^{-1}(1/2)) = 0$.

H3) $H1$ y bordes C^3

H4) Sea $h_l \to 0$ tal que $lh_l^{2d}/\log(l) \to \infty$. (X, Y) satisface $H4$ si P_X tiene densidad f, continua tal que para todo $\delta > 0$ existe $\gamma = \gamma(\delta)$, tal que

$$
f(x) - f(y) > \gamma > 0 \text{ for all } x \in (B_1^\delta \cup B_0^\delta)^c \text{ and all } y \in (B_1^{h_l} \cup B_0^{h_l}), \tag{4}
$$

para l suficientemente grande tal que $2h_l < \delta$

H5) $Y^i = g^*(X^i)$ for all $(X^i, Y^i) \in \mathcal{D}^n$, y existen $(X^r, 0) \in \mathcal{D}^n$ y $(X^s, 1) \in \mathcal{D}^n$ with $X^r, X^s \in (B_1^{\delta_0} \cup B_0^{\delta_0})^c$, para algún $\delta_0 > 0$.

Sobre las hipótesis

H5, La muestra inicial tiene que estar bien ubicada: Sean $X|Y = 0 \sim N(0, 1)$, $X|Y = 1 \sim N(1, 1)$ y empezar en $\{(0.4, 1), (0.6, 0)\}$.
H4: $X|Y = 0 \sim U([0, r] \times [0, 1])$, $X|Y = 1 \sim U([r, 0] \times [0, 1])$ es indistinguible de $X|Y = 0 \sim U([0, r'] \times [0, 1])$ y $X|Y = 1 \sim U([r', 0] \times [0, 1])$.
H1, H3, Conexidad:

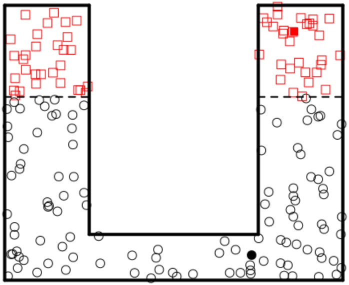

Figure: Los 0 se representan como cuadrados y los 1 como círculos. $\mathcal{D}^n$ puntos con relleno.

Ejemplos

Denotamos $C = \{(x, 1/2\sin(kx)) : -1 \leq x \leq 1\}$. Generamos $l/3$ iid $U[-1,1]^2$, nos quedamos con los que están en $B_{d_H}(C, .15) \cap [-1,1]^2$. Generamos l en $U[-1,1]^2$ y nos quedamos con los que están en $B_{d_H}(C, .15)^c \cap [-1,1]^2$

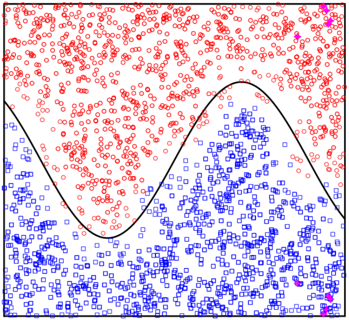

Figure: $k = 4$, $h_l = 0.148$, $l = 2400$. Rojos son clasificados como 1, azules como 0. La muestra inicial en magenta

Ejemplos

Denotamos $C = \{(x, 1/2\sin(kx)) : -1 \leq x \leq 1\}$. Generamos $l/3$ iid $U[-1,1]^2$, nos quedamos con los que están en $B_{d_H}(C, .15) \cap [-1,1]^2$. Generamos l en $U[-1,1]^2$ y nos quedamos con los que están en $B_{d_H}(C, .15)^c \cap [-1,1]^2$

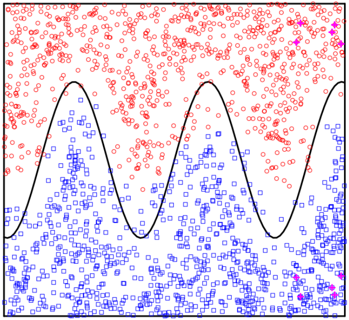

Figure: $k = 8$, $h_l = 0.148$, $l = 2400$. Rojos son clasificados como 1, azules como 0. La muestra inicial en magenta

Ejemplos

Denotamos $C = \{(x, 1/2\sin(kx)) : -1 \leq x \leq 1\}$. Generamos $l/3$ iid $U[-1,1]^2$, nos quedamos con los que están en $B_{d_H}(C, .15) \cap [-1,1]^2$. Generamos l en $U[-1,1]^2$ y nos quedamos con los que están en $B_{d_H}(C, .15)^c \cap [-1,1]^2$

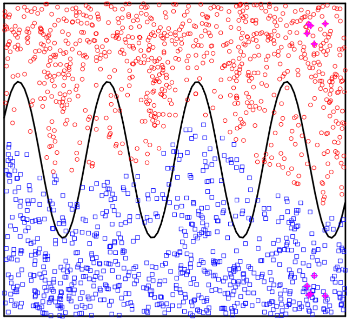

Figure: $k = 12$, $h_l = 0.148$, $l = 2400$. Rojos son clasificados como 1, azules como 0. La muestra inicial en magenta

Some theoretical results

Proposición

Supongamos $H0$ **y** $H1$, si $S \subset \mathbb{R}^d$ es compacto, $d \geq 2$. P_X con densidad continua f. Si $h_l \to 0$ tal que $lh_l^d / \log(l) \to \infty$, entonces, con probabilidad uno, para l suficientemente grande todos los puntos en $\mathcal{X}_l$ son clasificados por el algoritmo.

Some theoretical results

Proposición

Supongamos $H0$ **y** $H1$, si $S \subset \mathbb{R}^d$ es compacto, $d \geq 2$. P_X con densidad continua f. Si $h_l \to 0$ tal que $lh_l^d / \log(l) \to \infty$, entonces, con probabilidad uno, para l suficientemente grande todos los puntos en $\mathcal{X}_l$ son clasificados por el algoritmo.

Lema

Supongamos $H0$, $H1$ **y** $H2$, $\mathcal{D}^n = (\mathcal{X}^n, \mathcal{Y}^n)$ tal que $Y^i = 1$ si y solo si $g^*(X^i) = 1$. Sea $\{(X_{j_1}, \tilde{Y}_{j_1}), \ldots, (X_{j_l}, \tilde{Y}_{j_l})\}$. Sea $h_l \to 0$ tal que $lh_l^d / \log(l) \to \infty$. Sea i el primer índice tal que $g^*(X_{j_i}) \neq \mathbf{g}_{n+i-1}(X_{j_i})$.

1) si $\eta(X_{j_i}) > 1/2$ entonces, $B\big(X_{j_i}, h_l\big) \cap \eta^{-1}\big([0, 1/2)\big) \neq \emptyset$ para todo n, c.s.
2) si $\eta(X_{j_i}) < 1/2$ entonces, $B\big(X_{j_i}, h_l\big) \cap \eta^{-1}\big((1/2, 1]\big) \neq \emptyset$ para todo n, c.s.

Some theoretical results

Proposición

Supongamos $H0$ **y** $H1$, si $S \subset \mathbb{R}^d$ es compacto, $d \geq 2$. P_X con densidad continua f. Si $h_l \to 0$ tal que $lh_l^d / \log(l) \to \infty$, entonces, con probabilidad uno, para l suficientemente grande todos los puntos en $\mathcal{X}_l$ son clasificados por el algoritmo.

Lema

Supongamos $H0$, $H1$ **y** $H2$, $\mathcal{D}^n = (\mathcal{X}^n, \mathcal{Y}^n)$ tal que $Y^i = 1$ si y solo si $g^*(X^i) = 1$. Sea $\{(X_{j_1}, \tilde{Y}_{j_1}), \dots, (X_{j_l}, \tilde{Y}_{j_l})\}$. Sea $h_l \to 0$ tal que $lh_l^d / \log(l) \to \infty$. Sea i el primer índice tal que $g^*(X_{j_i}) \neq \mathbf{g}_{n+i-1}(X_{j_i})$.

1) si $\eta(X_{j_i}) > 1/2$ entonces, $B\big(X_{j_i}, h_l\big) \cap \eta^{-1}\big([0, 1/2)\big) \neq \emptyset$ para todo n, c.s.
2) si $\eta(X_{j_i}) < 1/2$ entonces, $B\big(X_{j_i}, h_l\big) \cap \eta^{-1}\big((1/2, 1]\big) \neq \emptyset$ para todo n, c.s.

Proposición

$H2$, $H3$, $H4$ **y** $H5$. Sea i el primer índice tal que $g^*(X_{j_i}) \neq \mathbf{g}_{n+i-1}(X_{j_i})$. Entonces, con probabilidad 1, para todo $\delta > 0$ existe l_0 tal que si $l > l_0$, $i > \#\{\mathcal{X}_l \cap B(\eta^{-1}(1/2), \delta)^c\}$.

Consistencia

Theorem

Bajo $H0$**,** $H2$**,** $H3$**,** $H4$ **y** $H5$**.**, para todo $n > 2$,

$$\lim_{l\to\infty} E\left(\frac{1}{l}\#\{i : g_{n,l,i}(\mathcal{X}_l) \neq Y_i, (X_i, Y_i) \in \mathcal{D}_l\} \middle| \mathcal{D}^n\right) - L(\mathbf{g}_l^*) = 0 \quad c.s., \tag{5}$$

Devroye, L., Györfi, L. and Lugosi, G. (1996).
A Probabilistic Theory of Pattern Recognition. Springer-Verlag, New York.

Agrawala, A.K. (1970). Learning with a probabilistic teacher.
IEEE Transactions on Automatic Control, (19), 716–723

Belkin, M. and Niyogi, P. (2004). Semi-supervised learning on Riemannian manifolds.
Machine Learning **56** pp. 209–239.

Ben-David, S., Lu, T. and Pal, D.(2008). Does unlabelled data provably help?. Worst-case analysis of the sample complexity of semi-supervised learning.
In 21*st Annual Conference on Learning Theory (COLT)*. Available at `http://www.informatik.uni-trier.de/~ley/db/conf/colt/colt2008.html`.

Chapelle, O., Schölkopf, B. and Zien, A., eds. (2006) Semi-supervised learning.
Adaptative computation and machine learning series. MIT

Fralick, S.C. (1967) Learning recognize patterns without teacher.
IEEE Transactions on Information Theory (13) 57–64.

Haffari, G. and Sarkar, A. (2007) Analysis of Semi-Supervised Learning with the Yarkowsky algoritm.
In Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence, UAI 2007. Vancouver, BC. July 19-22, 2007.

Zhu, X. (2008) Semi-supervised learning literature survey.
`http://pages.cs.wisc.edu/~jerryzhu/research/ssl/semireview.html`