

Aprendizaje semi-supervisado.

Alejandro Cholaquidis

CMAT-Facultad de Ciencias, UdelaR
Montevideo Uruguay

En conjunto con: R. Fraiman and M. Sued

IV Jornadas de Estadística Aplicada

1 Aprendizaje supervisado

2 Semi-supervisado

- Algoritmo
- Ejemplos

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 $g : \mathcal{F} \rightarrow \{0, 1\}$, que *minimice* $\mathbb{P}(g(X) \neq Y)$.

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) $L^* = \mathbb{P}(g^*(X) \neq Y)$ riesgo de Bayes.

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 idénticamente 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

- $\mathcal{D}^n = (\mathcal{X}^n, \mathcal{Y}^n) = \{(X^1, Y^1), \dots, (X^n, Y^n)\}$ una muestra de $(X, Y) \in S \times \{0, 1\}$, donde $S \subset \mathbb{R}^d$ es compacto. **Supondremos que son idénticamente 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.**

Objetivo

Minimizar

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

Definimos \mathbf{g}_l^* que minimiza $L(\mathbf{g}_l)$.

El mínimo de $L(\mathbf{g}_l)$ se obtiene con (g^*, \dots, g^*) , l copias de g^* .

Versión muestral

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

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

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

Denotamos

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

IDEA:

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

IDEA:

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

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

IDEA:

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

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

IDEA:

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

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

Sea $h_l \rightarrow 0$, para cada $X_j \in \mathcal{X}_l \setminus \{X_{j_1}, \dots, X_{j_{i-1}}\}$, definimos $\hat{\eta}_{i-1}$ (basado en \mathcal{T}_{i-1}), como

IDEA:

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

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

Sea $h_l \rightarrow 0$, para cada $X_j \in \mathcal{X}_i \setminus \{X_{j_1}, \dots, X_{j_{i-1}}\}$, definimos $\hat{\eta}_{i-1}$ (basado en \mathcal{T}_{i-1}), como

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

Ejemplo

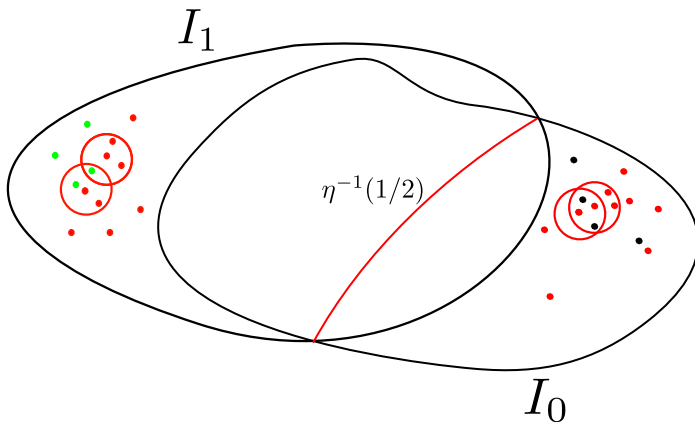


Figure: En verde los puntos de \mathcal{D}^n con etiqueta 1, en negro los de \mathcal{D}^n con etiqueta 0. En rojo la muestra \mathcal{X}_l

Ejemplo

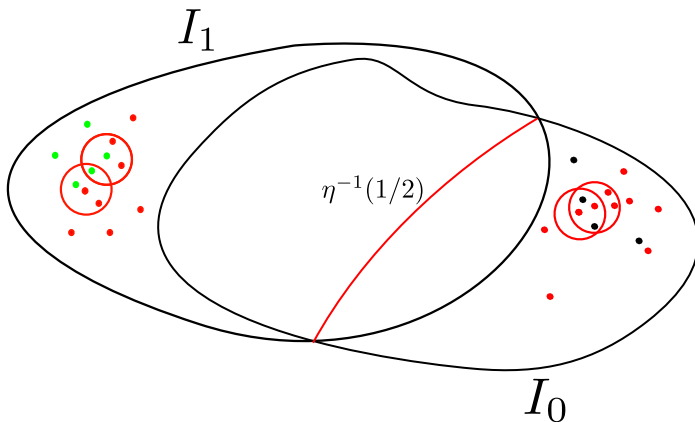


Figure: En verde los puntos de \mathcal{D}^n con etiqueta 1, en negro los de \mathcal{D}^n con etiqueta 0. En rojo la muestra \mathcal{X}_l

Ejemplo

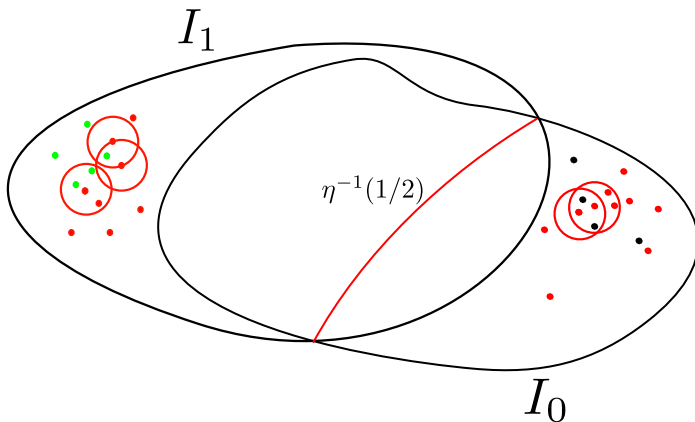


Figure: En verde los puntos de \mathcal{D}^n con etiqueta 1, en negro los de \mathcal{D}^n con etiqueta 0. En rojo la muestra \mathcal{X}_l

Ejemplo

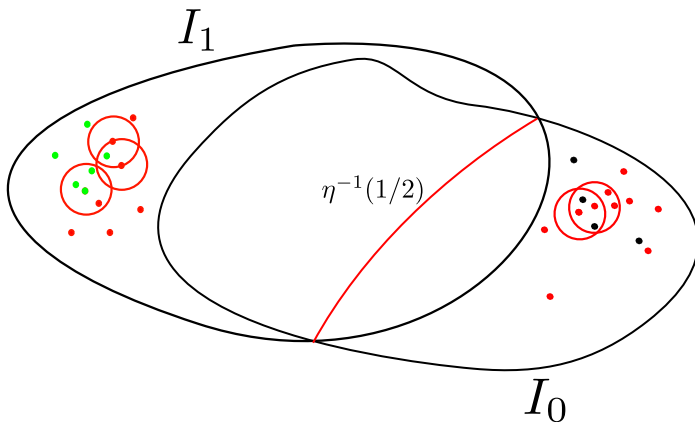


Figure: En verde los puntos de \mathcal{D}^n con etiqueta 1, en negro los de \mathcal{D}^n con etiqueta 0. En rojo la muestra \mathcal{X}_l

Algoritmo

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

Algoritmo

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

$1 \leq i < l$: 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 $\#\{\mathcal{Z}_{i-1} \cap B(X_{j_i}, h_l)\} > 0$,

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

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

$\tilde{Y}_{j_i} = 1$ si y solo si $\hat{\eta}_{i-1}(X_{j_i}) > 1/2$

Algoritmo

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

$1 \leq i < l$: 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 $\#\{\mathcal{Z}_{i-1} \cap B(X_{j_i}, h_l)\} > 0$,

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

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

$\tilde{Y}_{j_i} = 1$ si y solo si $\hat{\eta}_{i-1}(X_{j_i}) > 1/2$

Actualizar:

$\mathcal{Z}_i := \mathcal{Z}_{i-1} \cup X_{j_i}$.

$\mathcal{T}_i := \mathcal{D}^n \cup \{(X_{j_i}, \tilde{Y}_{j_i}), \dots, (X_{j_i}, \tilde{Y}_{j_i})\}$ y calcular $\hat{\eta}_i(X_r)$ con $X_r \in \mathcal{X}_l \setminus \mathcal{Z}_i$.

Algoritmo

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

$1 \leq i < l$: 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 $\#\{\mathcal{Z}_{i-1} \cap B(X_{j_i}, h_l)\} > 0$,

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

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

$\tilde{Y}_{j_i} = 1$ si y solo si $\hat{\eta}_{i-1}(X_{j_i}) > 1/2$

Actualizar:

$\mathcal{Z}_i := \mathcal{Z}_{i-1} \cup X_{j_i}$.

$\mathcal{T}_i := \mathcal{D}^n \cup \{(X_{j_1}, \tilde{Y}_{j_1}), \dots, (X_{j_i}, \tilde{Y}_{j_i})\}$ y calcular $\hat{\eta}_i(X_r)$ con $X_r \in \mathcal{X}_l \setminus \mathcal{Z}_i$.

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

Hipótesis

$$I_1 = \eta^{-1}((1/2, 1]) \quad I_0 = \eta^{-1}([0, 1/2))$$

$$B_1^\delta = I_1 \cap B(I_0, \delta) \quad B_0^\delta = I_0 \cap B(I_1, \delta)$$

H3) Sea $h_l \rightarrow 0$ tal que $lh_l^{2d} / \log(l) \rightarrow \infty$. (X, Y) satisface H4 si P_X tiene densidad f , continua tal que para todo $\delta > 0$ existe $\gamma = \gamma(\delta)$, tal que

$$f(x) - f(y) > \gamma > 0 \text{ para todo } x \in (B_1^\delta \cup B_0^\delta)^c \text{ para todo } y \in (B_1^{h_l} \cup B_0^{h_l}), \quad (4)$$

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

Hipótesis

$$I_1 = \eta^{-1}((1/2, 1]) \quad I_0 = \eta^{-1}([0, 1/2))$$

$$B_1^\delta = I_1 \cap B(I_0, \delta) \quad B_0^\delta = I_0 \cap B(I_1, \delta)$$

- H3)** Sea $h_l \rightarrow 0$ tal que $lh_l^{2d} / \log(l) \rightarrow \infty$. (X, Y) satisface H4 si P_X tiene densidad f , continua tal que para todo $\delta > 0$ existe $\gamma = \gamma(\delta)$, tal que

$$f(x) - f(y) > \gamma > 0 \text{ para todo } x \in (B_1^\delta \cup B_0^\delta)^c \text{ para todo } y \in (B_1^{h_l} \cup B_0^{h_l}), \quad (4)$$

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

- H4)** $Y^i = g^*(X^i)$ para todo $(X^i, Y^i) \in \mathcal{D}^n$, y existen $(X^r, 0) \in \mathcal{D}^n$ y $(X^s, 1) \in \mathcal{D}^n$ donde $X^r, X^s \in (B_1^{\delta_0} \cup B_0^{\delta_0})^c$, para algún $\delta_0 > 0$.

1 Aprendizaje supervisado

2 Semi-supervisado

- Algoritmo
- Ejemplos

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 $B_{d_H}(C, .15)^c \cap [-1, 1]^2$

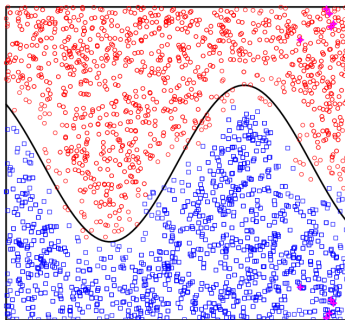


Figure: $k = 4$, $h_l = 0.148$, $l = 2400$. Rojos son clasificados como 1, azules como 0. La muestra inicial en magenta

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 $B_{d_H}(C, .15)^c \cap [-1, 1]^2$

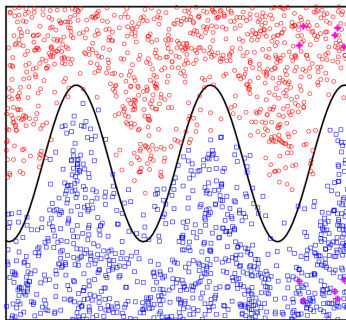










Figure: $k = 8$, $h_l = 0.148$, $l = 2400$. Rojos son clasificados como 1, azules como 0. La muestra inicial es magenta

-  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 *21st 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 algorithm.
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>