Detrás del Método Plug-In

Fundamentos Matemáticos de la Clasificación

Daniela Rodriguez

UTDT - CONICET

Qué es la clasificación?

Qué es la clasificación?

¿Es usted un enfermo coronario?

queremos hacer un diagnostico a partir de variables clinicas

Las variables registradas son:

1. $X_1 =$ Presión Sanguínea
2. $X_2 =$Genero: F - M
3. $X_3 =$ Fumador: Si - No
4. $X_4 =$Colesterol
5. $X_5 =$Actividad Física: Horas semanales de ejercicio
6. $X_6 =$ TV: Horas semanales de TV
7. $X_7 =$Altura
8. $X_8 =$Peso

$$Y = \begin{cases} 1 & \text{presencia de enfermedad coronaria} \\ 0 & \text{caso contrario} . \end{cases}$$

Qué es la clasificación?

¿Le damos un crédito?

Podemos predecir si un cliente va a pagar?

Las variables registradas son:

1. Veraz
2. Demográficas
3. Salario

$$Y = \begin{cases} 1 & \text{pagador} \\ 0 & \text{caso contrario} \end{cases} .$$

$$Y = \begin{cases} B & \text{Bajo riesgo crediticio} \\ M & \text{Medio riesgo crediticio} \\ A & \text{Alto riesgo crediticio} \end{cases} .$$

Qué es la clasificación?

Dos Objetivos

- Cuales son y como interactuan las caracteristicas de los objetos para explicar las etiquetas.
- Dado un nuevo objeto que etiqueta le asigno.

Preguntas naturales:

Que son los objectos a clasificar?

En que espacio(s) viven?

Cómo medir una buena/mala tarea de clasificación?

Cómo optimizar la tarea de clasificación?

Cómo automatizar la tarea de clasificación?

Clasificación - Pajaritos

Las aves parásitas de cría ponen huevos en nidos de otras especies (hospedador), las cuales incuban los huevos y crían al pichón parásito.

En un bosque de talas de la provincia de Buenos Aires hay dos especies hospederas que son indistinguibles a simple vista.

Pero una de las principales diferencias entre estas especies radica en el grado de discriminación y remoción de huevos parásitos de sus nidos.

Clasificación - Pajaritos

La especie “aceptadora“ de huevos parásitos, remueve del nido sólo el 30% de los huevos parásitos. Y la especie “rechazadora“ remueve el 80% de los huevos parásitos. Sea

$$Y = \begin{cases} 1 & \text{rechazador} \\ 0 & \text{aceptador}. \end{cases}$$

El 90% de los nidos son de la especie “aceptadora“ y el 10% a la especie “rechazadora“.

$P(Y = 0) = 0.9$

$P(Y = 1) = 0.1$

Objetivo del problema:

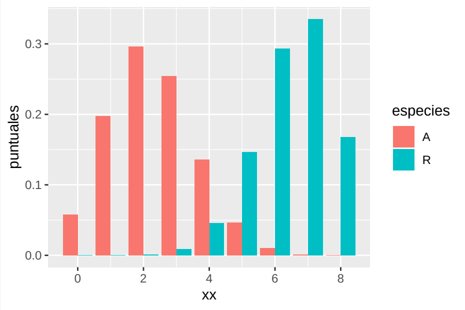
conociendo el número de huevos removidos predecir si la especie es rechazadora o aceptadora.

Clasificación - Pajaritos

Supongamos que en un nido se colocan $n = 8$ huevos parasitarios.
Sea X = número de huevos removidos. Sabemos,

$$X|_{Y=0} \sim Bi(8, 0.3), \text{ es decir } p_{X|_{Y=0}}(x) = \binom{8}{x} 0.3^x 0.7^{8-x}$$

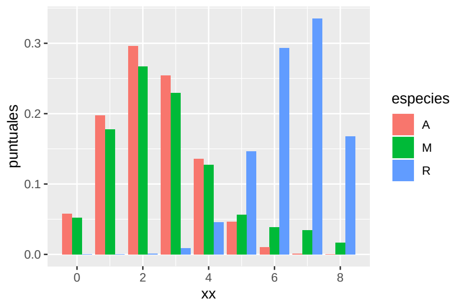
$$X|_{Y=1} \sim Bi(8, 0.8), \text{ es decir } p_{X|_{Y=1}}(x) = \binom{8}{x} 0.8^x 0.2^{8-x}$$



Cuál es la distribución de X?? $p_X(x)$?

$$\begin{aligned} p_X(x) &= p_{X|Y=1}(x)P(Y=1) + p_{X|Y=0}(x)P(Y=0) \\ &= 0.1\binom{8}{x}0.8^x 0.2^{8-x} + 0.9\binom{8}{x}0.3^x 0.7^{8-x} \end{aligned}$$

0	1	2	3	4	5	6	7	8
0.052	0.178	0.267	0.230	0.127	0.057	0.038	0.035	0.017



Cómo hacemos para decidir mirando un nido con 8 huevos parasitarios y dependiendo cuantos huevos sean removidos si la especie es rechazadora o aceptadora.

- Si se remueven 5 huevos; es decir si $X = 5$ ¿de qué clase de nido diría que se trata?
- Si se remueven 3 huevos; ($X = 3$) ¿de qué clase de nido diría que se trata?

Etiquetando Pajaritos

Una Regla (de clasificación) asigna a $x \in \{0, 1, \ldots, 8\}$ un tipo de hospedador: $\{0, 1\}$ ($0 = A, 1 = R$) es decir, buscamos

$$h : \{0, 1, \ldots, 8\} \to \{0, 1\}$$

Por ejemplo, si remueve 5 huevos o más es rechazadora

$$h_1(x) = \begin{cases} 0 & x \in \{0, 1, 2, 3, 4\} \\ 1 & x \in \{5, 6, 7, 8\} \end{cases}$$

Otro clasificador, si remueve un número par de huevos es rechazadora

$$h_2(x) = \begin{cases} 0 & x \in \{1, 3, 5, 7\} \\ 1 & x \in \{0, 2, 4, 6, 8\} \end{cases}$$

Cuál es más razonable??

Clasificación: Marco Teórico

Información que tengo::

$X \in \mathcal{X}$, $Y \in \mathcal{Y}$.

(en nuestro caso $\mathcal{X} = \{0, 1, \ldots, 8\}$ y $\mathcal{Y} = \{0, 1\}$)

Clasificador:

Regla de clasificación es una función que asigna a cada $x \in \mathcal{X}$ un elemento $y \in \mathcal{Y}$

$$\text{Clasificador} \quad h : \mathcal{X} \to \mathcal{Y}$$

Toda regla de clasificación induce una partición de $\mathcal{X}$.

Buscamos la mejor *h* en que sentido?

Minimizar el Error de Clasificación

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

Etiquetando Pajaritos

Que error de Clasificación $L(h) = \mathbb{P}(h(X) \neq Y)$ cometo con el clasificador h_1?

$$h_1(x) = \begin{cases} 0 & x \in \{0,1,2,3,4\} \\ 1 & x \in \{5,6,7,8\} \end{cases}$$

$$\begin{aligned}
&P(h_1(X) \neq Y) \\
&= P(h_1(X) \neq Y | Y = 1) P(Y = 1) + P(h_1(X) \neq Y | Y = 0) P(Y = 0) \\
&= P(X \in \{0,1,2,3,4\} | Y = 1) P(Y = 1) + P(X \in \{5,6,7,8\} | Y = 0) P(Y = 0) \\
&= \sum_{x=0}^{4} \binom{8}{x} 0.8^x 0.2^{8-x} 0.1 + \sum_{x=5}^{8} \binom{8}{x} 0.3^x 0.7^{8-x} 0.9 \\
&= 0.05779905
\end{aligned}$$

Etiquetando Pajaritos

Y con el clasificador h_2?

$$h_2(x) = \begin{cases} 0 & x \in \{1,3,5,7\} \\ 1 & x \in \{0,2,4,6,8\} \end{cases}$$

$$\begin{aligned}
&P(h_2(X) \neq Y) \\
=&P(h_2(X) \neq Y | Y=1)P(Y=1) + P(h_2(X) \neq Y | Y=0)P(Y=0) \\
=&P(X \in \{1,3,5,7\} | Y=1)P(Y=1) + P(X \in \{0,2,4,6,8\} | Y=0)P(Y=0) \\
=&\sum_{x=1,3,5,7} \binom{8}{x} 0.8^x 0.2^{8-x} 0.1 + \sum_{x=0,2,4,6,8} \binom{8}{x} 0.3^x 0.7^{8-x} 0.9 \\
=&0.4994551
\end{aligned}$$

Etiquetando Pajaritos

$$h_1(x) = \begin{cases} 0 & x \in \{0, 1, 2, 3, 4\} \\ 1 & x \in \{5, 6, 7, 8\} \end{cases}$$

$$P(h_1(X) \neq Y) = .05779905$$

$$h_2(x) = \begin{cases} 0 & x \in \{1, 3, 5, 7\} \\ 1 & x \in \{0, 2, 4, 6, 8\} \end{cases}$$

$$P(h_2(X) \neq Y) = 0.4994551$$

Hay un etiquetado de Pajaritos mejor?

Por ejemplo con la familias de clasificadores h_t,

$$h_t(x) = \begin{cases} 0 & x \in \{0, \ldots t-1\} \\ 1 & x \in \{t, \ldots, 8\} \end{cases}$$

$$P(h_t(X) \neq Y) = \sum_{x=0}^{t-1} \binom{8}{x} 0.8^x 0.2^{8-x} 0.1 + \sum_{x=t}^{8} \binom{8}{x} 0.3^x 0.7^{8-x} 0.9$$

Hay un etiquetado de Pajaritos mejor?

Por ejemplo con la familias de clasificadores h_t,

$$h_t(x) = \begin{cases} 0 & x \in \{0, \dots t-1\} \\ 1 & x \in \{t, \dots, 8\} \end{cases}$$

$$P(h_t(X) \neq Y) = \sum_{x=0}^{t-1} \binom{8}{x} 0.8^x 0.2^{8-x} 0.1 + \sum_{x=t}^{8} \binom{8}{x} 0.3^x 0.7^{8-x} 0.9$$

X	0	1	2	3	4
$P(h_t(X) \neq Y)$	0.9000	0.8481	0.6702	0.4035	0.1757

X	5	6	7	8	todos A
$P(h_t(X) \neq Y)$	0.0578	0.0305	0.0508	0.0833	0.100

Clasificación Binaria

- Variable de respuesta: Y toma valores en $\{0, 1\}$.

- Una regla de clasificación $h : \mathcal{X} \to \{0, 1\}$ que asigne a cada $x \in \mathcal{X}$ una etiqueta 0 o 1.

- La mejor regla es la que minimiza la probabilidad de error $P(h(X) \neq Y)$.

- El problema fundamental en la clasificación es etiquetar un nuevo punto $X = x$ minimizando la probabilidad de error.

Función de regresión

$$m^*(x) = P(Y = 1|X = x).$$

Notemos que $P(Y = 1|X = x) = E(I_{\{1\}}(Y)|X = x)$

Regla de clasificación óptima - Regla de Bayes

La regla óptima de clasificación, conocida como la **Regla de Bayes**, es:

$$h^*(x) = \begin{cases} 1 & \text{si } m^*(x) > 1/2 \\ 0 & \text{si } m^*(x) \leq 1/2 \end{cases}$$

Esta regla minimiza la probabilidad de error de clasificación $P(h(X) \neq Y)$.

$$h^*(x) = \begin{cases} 1 & \text{si } P(Y = 1|X = x) > 1/2 \\ 0 & \text{si } P(Y = 1|X = x) \leq 1/2 \end{cases}$$

Teorema

Sea

$$h^*(x) = I_{\{m^*(x)>1/2\}} = I_{\{P(Y=1|X=x)>1/2\}}$$

luego $P(h^*(X) \neq Y) = \operatorname{argmin}_h P(h(X) \neq Y)$.

donde el mínimo se toma sobre todas las funciones medibles de $\mathcal{X} \to \{0, 1\}$

Demostración Regla de Bayes

Queremos encontrar una función $h : \mathcal{X} \to \{0, 1\}$ que minimice el error de clasificación:

$$P(h(X) \neq Y) = E[I_{h(X)\neq Y}]$$

Podemos reescribir la probabilidad de error utilizando propiedades de la esperanza condicional:

$$P(h(X) \neq Y) = E[P(h(X) \neq Y|X)]$$

Fijado x minimizamos $P(h(x) \neq Y|X = x)$.

- Si $h(x) = 0$, entonces

$$P(h(x) \neq Y|X = x) = P(Y = 1|X = x) = m^*(x).$$

- Si $h(x) = 1$, entonces

$$P(h(x) \neq Y|X = x) = P(Y = 0|X = x) = 1 - P(Y = 1|X = x) = 1 - m^*(x).$$

Demostración Regla de Bayes

Es decir,

$$P(h(x) \neq Y|X = x) = I_{\{h(x)=0\}} m^*(x) + I_{\{h(x)=1\}} (1 - m^*(x)).$$

Analogamente para la propuesta de regla óptima $h^*(x) = I_{\{m^*(x)>1/2\}}$, tenemos

$$P(h^*(x) \neq Y|X = x) = I_{\{h^*(x)=0\}} m^*(x) + I_{\{h(^*x)=1\}} (1 - m^*(x)).$$

Y restando

$$\begin{aligned} P(h(x) \neq Y|X = x) &- P(h^*(x) \neq Y|X = x) = \\ &= \left[I_{\{h(x)=0\}} - I_{\{h^*(x)=0\}} \right] m^*(x) \\ &+ \left[I_{\{h(x)=1\}} - I_{\{h^*(x)=1\}} \right] (1 - m^*(x)) \end{aligned}$$

Demostración Regla de Bayes

$$
\begin{aligned}
P(h(x) \neq Y|X=x) &- P(h^*(x) \neq Y|X=x) = \\
&= \left[I_{\{h(x)=0\}} - I_{\{h^*(x)=0\}}\right] m^*(x) \\
&+ \left[I_{\{h(x)=1\}} - I_{\{h^*(x)=1\}}\right] (1 - m^*(x))
\end{aligned}
$$

Notemos $\left[I_{\{h(x)=0\}} - I_{\{h^*(x)=0\}}\right] = -\left[I_{\{h(x)=1\}} - I_{\{h^*(x)=1\}}\right]$

$$P(h(x) \neq Y|X=x) - P(h^*(x) \neq Y|X=x) = \left[I_{\{h(x)=1\}} - I_{\{h^*(x)=1\}}\right] (1 - 2m^*(x))$$

$$P(h(x) \neq Y|X=x) - P(h^*(x) \neq Y|X=x) = \left[I_{\{h^*(x)=1\}} - I_{\{h(x)=1\}}\right] (2m^*(x) - 1)$$

Demostración Regla de Bayes

$$P(h(x) \neq Y|X = x) - P(h^*(x) \neq Y|X = x) = \left[I_{\{h^*(x)=1\}} - I_{\{h(x)=1\}}\right](2m^*(x) - 1)$$

Por definición la regla óptima es $h^*(x) = I_{\{m^*(x)>1/2\}}$. Luego,

$h^*(x) = 1$ si y solo si $2m^*(x) - 1 > 0$.
$I_{\{h^*(x)=1\}} - I_{\{h(x)=1\}} \geq 0$ si y solo si $2m^*(x) - 1 > 0$.

$h^*(x) = 0$ si y solo si $2m^*(x) - 1 < 0$.
$I_{\{h^*(x)=1\}} - I_{\{h(x)=1\}} \leq 0$ si y solo si $2m^*(x) - 1 < 0$.

En resumen,

$$P(h(x) \neq Y|X = x) - P(h^*(x) \neq Y|X = x) \geq 0.$$

Finalmente se tiene que $P(h(x) \neq Y) \geq P(h^*(x) \neq Y)$.

Resumiendo: Regla de Bayes

La regla óptima de clasificación, **Regla de Bayes**, es:

$$h^*(x) = I_{\{m^*(x)>1/2\}} = \begin{cases} 1 & \text{si } P(Y=1|X=x) > 1/2 \\ 0 & \text{si } P(Y=1|X=x) \leq 1/2 \end{cases}$$

equivalentemente $I_{\{m^*(x)>1/2\}} = I_{\{m^*(x)>1-m^*(x)\}}$

$$h^*(x) = \begin{cases} 1 & \text{si } P(Y=1|X=x) > P(Y=0|X=x) \\ 0 & \text{si } P(Y=1|X=x) \leq P(Y=0|X=x) \end{cases}$$

Resumiendo: Regla de Bayes

En la practica, llamemos $p = P(Y = 1)$

- Si ambos son discretos

$$h^*(x) = \begin{cases} 1 & \text{si } pP(X = x|Y = 1) > (1 - p)P(X = x|Y = 0) \\ 0 & \text{en caso contrario} \end{cases}$$

- Si X f_0, f_1 son las densidad condicionales de X dado $Y = 0$ o 1 respectivamente

$$h^*(x) = \begin{cases} 1 & \text{si } f_1(x)p > (1 - p)f_0(x) \\ 0 & \text{en caso contrario} \end{cases}$$

Resumiendo: Regla de Bayes

El error de Clasificación de una regla h es:

$$L(h) = P(h(X) \neq Y) = 1 - E\left(I_{\{h(x)=1\}} m^*(x) + I_{\{h(x)=0\}}(1 - m^*(x)\right)$$

Error de Bayes: El error de Clasificación de la regla óptima h^*

$$L(h^*) = P(h(X) \neq Y) = E\left(min\{m^*(x), 1 - m^*(x)\}\right)$$

Clasificador Óptimo: Pátito feo

La probabilidad de ser rechazador $p = P(Y = 1) = 0.1$.
Clasificamos como 1 a los x que satisfagan

$$0.1 p_{X|_{Y=1}}(x) = 0.1 \binom{8}{x} 0.8^x 0.2^{8-x} > 0.9 p_{X|_{Y=0}}(x) 0.9 \binom{8}{x} 0.3^x 0.7^{8-x}$$

$$h_{opt}(x) = \begin{cases} 1 & x \in \{6, 7, 8\} \\ 0 & x \in \{0, 1, 2, 3, 4, 5\} \end{cases}$$

$L(h_{opt}) = \mathbb{P}(h_{opt}(X) \neq Y) = 0.0305$

Que hacemos si no conocemos las condicionales?

En la práctica, $m^*(x) = P(Y = 1|X = x)$ es desconocida y la estimaremos a partir de los datos de entrenamiento $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Estos datos provienen de la misma distribución de (X, Y) y se asumen independientes.

Estimar la regla de clasificación abre una puerta enorme a diferentes opciones

- Bayes Naive
- Perceptrón (clasificador lineal Rosenblat 1962)
- LDA (Fisher 1932) y QDA (supuesto de normalidad)
- Regresión logística. (modelo para $m(x) = (1 + e^{-\beta^T x})^{-1}$)
- K-vecinos más cercanos (k-NN). : (estimación no paramétrica basada en la proporción de vecinos de clase 1)
- Máquinas de vectores de soporte (SVM), Random Forest, Arboles, Redes Neuronales,....

Método Plug-in

El método **plug-in** es una forma de estimar que se obtiene sustituyendo lo teórico desconocido por una estimación basada en la muestra.

En este caso reemplazaremos la función de regresión teórica $m^*(x)$ por una estimación $m_n(x)$ basada en la muestra.

La regla plug-in se define como:

$$h_n(x) = I_{\{m_n(x) > 1/2\}}$$

donde $m_n(x)$ es un estimador de $m^*(x)$.

Método Plug-in

La regla plug-in se define como:

$$h_n(x) = I_{\{m_n(x) > 1/2\}}$$

donde $m_n(x)$ es un estimador de $m^*(x)$.

El rendimiento de una regla plug-in depende de cuán bien $m_n(x)$ aproxime a $m^*(x)$.

La consistencia de $m_n(x)$ (es decir, si $m_n(x) \to m^*(x)$ en algún sentido a medida que $n \to \infty$) es crucial para la consistencia de la regla de clasificación $h_n(x)$.

Método Plug-in

Sea $h^*(x)$ la reglas de clasificación óptima, $m^*(x)$ la función de regresión teórica desconocida y $L^* = L(h^*)$ su error de clasificación.

Sea m una aproximación o estimación de m^* y $h(x)$ la regla de clasificación inducida por $m(x)$ con $L(h)$ su error de clasificación.

Buscamos vincular la diferencia de los errores de clasificación $L(h) - L^*$ en terminos de $m(x)$ y $m^*(x)$.

Sabemos que $L(h) - L^* \geq$ o buscaremos una cota superior para esta diferencia.

Método Plug-in

Teorema

Sean $h(x)$ y $h^*(x)$ las reglas de clasificación inducidas por $m(x)$ y $m^*(x)$, con m una aproximacion de m^*.

$$L(h) - L^* = 2\int_{\mathcal{X}} |m^*(x) - \frac{1}{2}| I_{\{h(x)\neq h^*(x)\}} dP_X(x)$$

$$L(h) - L^* \leq 2\int_{\mathcal{X}} |m^*(x) - m(x)| dP_X(x)$$

Demostración:

Habiamos mostrado que

$$P(h(x) \neq Y|X = x) - P(h^*(x) \neq Y|X = x) = \left[I_{\{h^*(x)=1\}} - I_{\{h(x)=1\}}\right](2m^*(x) - 1)$$

Podemos ver fácilmente (analizando los casos)

$$\left[I_{\{h^*(x)=1\}} - I_{\{h(x)=1\}}\right](2m^*(x) - 1) = |2m^*(x) - 1| \left[I_{\{h^*(x)\neq h(x)\}}\right]$$

Método Plug-in

Es decir,

$$P(h(x) \neq Y|X = x) - P(h^*(x) \neq Y|X = x) = |2m^*(x) - 1| \left[I_{\{h^*(x) \neq h(x)\}} \right]$$

Por lo tanto

$$L(h) - L^* = \int_{\mathcal{X}} |2m^*(x) - 1| I_{\{h^*(x) \neq h(x)\}} dP_X(x)$$

Por otra parte si $h^*(x) \neq h(x)$ tenemos que

$$|m^*(x) - 1/2| \leq |m^*(x) - m(x)|$$

Esto sale mirando cada caso, por ejemplo si $h^*(x) = 1$ y $h(x) = 0$, $m^*(x) \geq 1/2$ y $m(x) < 1/2$ luego $m^*(x) - 1/2 < m^*(x) - m(x)$. Y análoga la otra.

Método Plug-in

Como consecuencia del Teorema

$$0 \leq L(h) - L^* \leq 2E(|m^*(X) - m(X)|) \leq 2\sqrt{E((m^*(X) - m(X))^2)}.$$

Por lo tanto aproximar bien a m^* por m implica que el error de la regla h inducida por m estará cerca del error óptimo L^*.

Otro Método

Otra opción posible para construir un clasificador es restringir la clase de donde buscar el óptimo.

La regla de Bayes busca el mejor entre todas las funciones medibles. Pero por ejemplo podemos restringirnos solamente a clasificadores lineales.

Sea $\mathcal{C} = \{\phi : \mathcal{X} \to \{0, 1\}\}$ un subconjunto de las posibles reglas de clasificación y busquemos aquella(/s) que mimimizan $L(\phi) = P(\phi(X) \neq Y)$,

$$\inf_{\phi \in \mathcal{C}} L(\phi) \geq L^*$$

La diferencia $\inf_{\phi \in \mathcal{C}} L(\phi) - L^*$ no es aleatoria, depende de la clase.

Como estimo en este enfoque

Como $L(\phi) = P(\phi(X) \neq Y)$ en general no podemos calcularla.

Dada una muestra $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ podemos estimar el error de clasificación a partir de su versión empirica

$$\widehat{L}_n(\phi) = \frac{1}{n} \sum_{i=1}^{n} I_{\{\phi(X_i) \neq Y_i\}}$$

Y definimos $\widehat{\phi} = \mathrm{argmin}_{\phi \in \mathcal{C}} \widehat{L}_n(\phi)$.

Como estimo en este enfoque

Tenemos

$$\inf_{\phi\in\mathcal{C}} L(\phi) = \inf_{\phi\in\mathcal{C}} P(\phi(X)\neq Y)$$

$$\widehat{\phi} = \mathrm{argmin}_{\phi\in\mathcal{C}} \widehat{L}_n(\phi) = \mathrm{argmin}_{\phi\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} I_{\{\phi(X_i)\neq Y_i\}}$$

Por la ley de los grandes números para cada ϕ, tenemos que $\widehat{L}_n(\phi)\to L(\phi)$. Pero eso no alcanza el comportamiento de $\widehat{L}_n(\widehat{\phi})$.

Notemos que hay dos errores

- $\inf_{\phi\in\mathcal{C}} L(\phi) - L^*$
  Si $\mathcal{C}$ es chica estaremos lejos del error óptimo de Bayes.

- $|\widehat{L}_n(\widehat{\phi}) - L(\widehat{\phi})| \leq \sup_{\phi\in\mathcal{C}} |\widehat{L}_n(\phi) - L(\phi)|$
  Si $\mathcal{C}$ es grande podríamos estar cometiendo sobre ajuste y esto $L_n(\widehat{\phi})$ podría darnos muy chico pero lejos de $L(\widehat{\phi})$.

Como estimo en este enfoque

Veamos que

$$L(\widehat{\phi}) - \inf_{\phi\in\mathcal{C}} L(\phi) \leq 2\sup_{\phi\in\mathcal{C}} |\widehat{L}_n(\phi) - L(\phi)|$$

$$L(\widehat{\phi}) - \inf_{\phi\in\mathcal{C}} L(\phi) = L(\widehat{\phi}) - \widehat{L}_n(\widehat{\phi}) + \widehat{L}_n(\widehat{\phi}) - \inf_{\phi\in\mathcal{C}} L(\phi) \tag{1}$$

Sea $\widetilde{\phi}$ una regla cualquier en $\mathcal{C}$ luego

$$\begin{aligned}
\widehat{L}_n(\widehat{\phi}) - \inf_{\phi\in\mathcal{C}} L(\phi) &\leq \widehat{L}_n(\widetilde{\phi}) - \inf_{\phi\in\mathcal{C}} L(\phi) = \inf_{\phi\in\mathcal{C}} (\widehat{L}_n(\widetilde{\phi}) - L(\phi)) \\
&\leq \inf_{\phi\in\mathcal{C}} |\widehat{L}_n(\widetilde{\phi}) - L(\phi)| \leq |\widehat{L}_n(\widetilde{\phi}) - L(\widetilde{\phi})| \\
&\leq \sup_{\phi\in\mathcal{C}} |\widehat{L}_n(\phi) - L(\phi)|
\end{aligned}$$

Finalmente de (1)

$$L(\widehat{\phi}) - \inf_{\phi\in\mathcal{C}} L(\phi) \leq |\widehat{L}_n(\widehat{\phi}) - L(\widehat{\phi})| + \sup_{\phi\in\mathcal{C}} |\widehat{L}_n(\phi) - L(\phi)| \leq 2\sup_{\phi\in\mathcal{C}} |\widehat{L}_n(\phi) - L(\phi)|$$

Como estimo en este enfoque

$L(\phi) = P(\phi(X) \neq Y)$
$\widehat{L}_n(\phi) = \frac{1}{n}\sum_{i=1}^{n} I_{\{\phi(X_i)\neq Y_i\}}$

$$\begin{aligned} L(\widehat{\phi}) - \inf_{\phi\in\mathcal{C}} L(\phi) &\leq 2\sup_{\phi\in\mathcal{C}} |\widehat{L}_n(\phi) - L(\phi)| \\ |\widehat{L}_n(\widehat{\phi}) - L(\widehat{\phi})| &\leq \sup_{\phi\in\mathcal{C}} |\widehat{L}_n(\phi) - L(\phi)| \end{aligned}$$

Convergencia Uniforme Sobre la Clase

$$\sup_{\phi\in\mathcal{C}} |\widehat{L}_n(\phi) - L(\phi)| \to 0$$

Es decir,

$$P(\sup_{\phi\in\mathcal{C}} |\widehat{L}_n(\phi) - L(\phi)| > \epsilon) \to 0?$$

Convergencia Uniforme de la medida empírica

- **Desigualdad de Hoeffding:** Si $X_1, \ldots, X_n$ son independientes, $E(X_i) = 0$ y $a_i \leq X_i \leq b_i$ para cualquier $i = 1, \ldots, n$ entonces, dado un valor de $\epsilon > 0$, vale que:

$$P\left(\sum_{i=1}^{n} X_i \geq \epsilon\right) \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$$

- En particular, si $X_i \sim_{iid} Be(p)$, usando la desigualdad de Hoeffding para las variables $Z_i = X_i - p$, vale que

$$P\left(|\bar{Z}_n - p| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

Convergencia Uniforme de la medida empírica

$$L(\phi) = P(\phi(X) \neq Y)$$
$$\widehat{L}_n(\phi) = \frac{1}{n}\sum_{i=1}^{n} I_{\{\phi(X_i)\neq Y_i\}}$$

Por la Desigualdad de Hoeffing para cada ϕ,
$Z_i = I_{\{\phi(X_i)\neq Y_i\}} - P(\phi(X) \neq Y)$

$$P\left(|L_n(\phi) - L(\phi)| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

Por lo tanto,

$$\sup_{\phi\in\mathcal{C}} P\left(|L_n(\phi) - L(\phi)| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

Convergencia Uniforme Sobre la Clase

$$P(\sup_{\phi\in\mathcal{C}} |\widehat{L}_n(\phi) - L(\phi)| > \epsilon) \to 0?$$

Convergencia Uniforme de la medida empírica

Sea P una medida de probabilidad y $P\,g = \int g\, dP = E_P(g)$. Dada

$(X_1, Y_1), \ldots, (X_n, Y_n)$ definimos la medida empírica P_n que asigna $\frac{1}{n}$ a cada (X_i, Y_i). Entonces, $P_n\, g = \int g\, dP_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i, Y_i)$

Luego

$$L(\phi) = P(\phi(X) \neq Y) = P\, I_{\{\phi(X)\neq Y\}}$$

$$\widehat{L}_n(\phi) = \frac{1}{n}\sum_{i=1}^{n} I_{\{\phi(X_i)\neq Y_i\}} = P_n\, I_{\{\phi(X)\neq Y\}}$$

Definamos la clase de funciones $\mathcal{F} = \{g(x,y) = I_{\{\phi(x)\neq y\}} : \phi \in \mathcal{C}\}$

Convergencia Uniforme Sobre la Clase

$$\sup_{\phi\in\mathcal{C}} |\widehat{L}_n(\phi) - L(\phi)| = \sup_{g\in\mathcal{F}} |\widehat{P}_n(g) - P(g)| \to 0$$

Convergencia Uniforme de la medida empírica

Convergencia Uniforme Sobre la Clase

$$\sup_{g \in \mathcal{F}} |\widehat{P}_n(g) - P(g)| \to 0$$

Si la clase de funciones $\mathcal{F} = \{g(x, y) = I_{\{\phi(x) \neq y\}} : \phi \in \mathcal{C}\}$ es finita ($N = |\mathcal{F}|$). Entonces,

$$\sup_{g \in \mathcal{F}} |\widehat{P}_n(g) - P(g)| \leq 2N\, e^{-2n\epsilon^2}$$

Y si no??
Tenemos que saber mediar cuán compleja es la Clases de Funciones.
Una opción es usando de la Teoría de Vapnik-Chervonenkis.

Resumiendo

- Importancia de la Caracterización poblacional (Regla óptima de Bayes)
- Tener buenos estimadores para enchufar (m y m^*)
- Versiónes empiricas de las función de pérdida y control de las clases sobre las que se busca el óptimo.
- Solo para clasificación?

El método plug-in como fábrica de estimadores

El método plug-in como fábrica de estimadores

Minimizar o Maximizar funciones de Riesgos.

El método plug-in como fábrica de estimadores

Minimizar o Maximizar funciones de Riesgos.

Clasificación

$$\widehat{g} = \operatorname*{argmin}_{g \in \mathcal{G}} \frac{1}{n} \sum I(Y_i \neq g(X_i))$$

El método plug-in como fábrica de estimadores

Minimizar o Maximizar funciones de Riesgos.

Clasificación

$$\widehat{g} = \operatorname*{argmin}_{g \in \mathcal{G}} \frac{1}{n} \sum I(Y_i \neq g(X_i))$$

M-estimadores

$$\widehat{\theta} = \operatorname*{argmin}_{\theta} \frac{1}{n} \sum m(X_i, \theta)$$

El método plug-in como fábrica de estimadores

Minimizar o Maximizar funciones de Riesgos.

Clasificación

$$\widehat{g} = \operatorname*{argmin}_{g \in \mathcal{G}} \frac{1}{n} \sum I(Y_i \neq g(X_i))$$

M-estimadores

$$\widehat{\theta} = \operatorname*{argmin}_{\theta} \frac{1}{n} \sum m(X_i, \theta)$$

K-medias

$$\widehat{\mathbf{c}} = \operatorname*{argmin}_{\mathbf{c} \in \mathbb{R}^{\mathbf{k}}} \frac{1}{n} \sum \min_{\mathbf{c}} \|X_i - c_j\|^2$$

El método plug-in como fábrica de estimadores

Minimizar o Maximizar funciones de Riesgos.

Clasificación

$$\widehat{g} = \operatorname*{argmin}_{g \in \mathcal{G}} \frac{1}{n} \sum I(Y_i \neq g(X_i))$$

M-estimadores

$$\widehat{\theta} = \operatorname*{argmin}_{\theta} \frac{1}{n} \sum m(X_i, \theta)$$

K-medias

$$\widehat{\mathbf{c}} = \operatorname*{argmin}_{\mathbf{c} \in \mathbb{R}^{\mathbf{k}}} \frac{1}{n} \sum \min_{\mathbf{c}} \|X_i - c_j\|^2$$

Componentes principales

$$\widehat{\mathbf{a}} = \operatorname*{argmax}_{\mathbf{a}} \frac{1}{n} \sum (\mathbf{a}^t (X_i - \bar{X}))^2$$

El método plug-in como fábrica de estimadores

Minimizar o Maximizar funciones de Riesgos.

Clasificación

$$\widehat{g} = \operatorname*{argmin}_{g \in \mathcal{G}} \frac{1}{n} \sum I(Y_i \neq g(X_i))$$

M-estimadores

$$\widehat{\theta} = \operatorname*{argmin}_{\theta} \frac{1}{n} \sum m(X_i, \theta)$$

K-medias

$$\widehat{\mathbf{c}} = \operatorname*{argmin}_{\mathbf{c} \in \mathbb{R}^{\mathbf{k}}} \frac{1}{n} \sum \min_{\mathbf{c}} \|X_i - c_j\|^2$$

Componentes principales

$$\widehat{\mathbf{a}} = \operatorname*{argmax}_{\mathbf{a}} \frac{1}{n} \sum (\mathbf{a}^t (X_i - \bar{X}))^2$$

Regresión lineal

$$\widehat{\beta} = \operatorname*{argmin}_{\beta} \frac{1}{n} \sum (Y_i - \beta^t X_i)^2$$

El método plug-in como fábrica de estimadores

Minimizar o Maximizar funciones de Riesgos.

Clasificación

$$\widehat{g} = \operatorname*{argmin}_{g \in \mathcal{G}} \frac{1}{n} \sum I(Y_i \neq g(X_i)) \qquad P(I_{\{Y \neq g(X)\}}) \; (Y, X) \sim P$$

M-estimadores

$$\widehat{\theta} = \operatorname*{argmin}_{\theta} \frac{1}{n} \sum m(X_i, \theta) \qquad P(m(X, \theta))$$

K-medias

$$\widehat{\mathbf{c}} = \operatorname*{argmin}_{\mathbf{c} \in \mathbb{R}^{\mathbf{k}}} \frac{1}{n} \sum \min_{\mathbf{c}} \|X_i - c_j\|^2 \qquad P(\min_{\mathbf{c}} \|X - c_j\|^2) \; \mathbf{c} = (c_1, \dots, c_k)$$

Componentes principales

$$\widehat{\mathbf{a}} = \operatorname*{argmax}_{\mathbf{a}} \frac{1}{n} \sum (\mathbf{a}^t (X_i - \bar{X}))^2 \qquad P((\mathbf{a}^t X)^2 - (P\mathbf{a}^t X)^2)$$

Regresión lineal

$$\widehat{\beta} = \operatorname*{argmin}_{\beta} \frac{1}{n} \sum (Y_i - \beta^t X_i)^2 \qquad P((Y - \beta^t X)^2) \; (Y, X) \sim P$$

El método plug-in como fábrica de estimadores

En todos estos casos tenemos un estimador del parámetro o función de interés definido como

$$\widehat{\gamma} = \operatorname*{argmin}_{\gamma} \frac{1}{n} \sum_{i=1}^{n} R(X_i, \gamma) = \operatorname*{argmin}_{\gamma} P_n R(X, \gamma)$$

y podemos pensar en su versión poblacional

$$\gamma(P) = \operatorname*{argmin}_{\gamma} P\, R(X, \gamma).$$

El método plug-in como fábrica de estimadores

En todos estos casos tenemos un estimador del parámetro o función de interés definido como

$$\widehat{\gamma} = \underset{\gamma}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} R(X_i, \gamma) = \underset{\gamma}{\operatorname{argmin}} P_n R(X, \gamma)$$

y podemos pensar en su versión poblacional

$$\gamma(P) = \underset{\gamma}{\operatorname{argmin}} P\, R(X, \gamma).$$

Es decir, $\widehat{\gamma} = \gamma(P_n)$.
Quisieramos probar que $\widehat{\gamma} \to \gamma(P)$.
Para ello no interesa estudiar

$$\sup_{\gamma} |P_n\, R(X, \gamma) - P\, R(X, \gamma)| \to 0$$

El método plug-in como fábrica de estimadores

La principal dificultad de este tipo de argumentos para probar consistencia es verificar los supuestos de convergencia uniforme de las funciones de riesgo,

$$\sup_{\gamma} |P_n\, R(X,\gamma) - P\, R(X,\gamma)| \to 0$$

es decir,

$$\sup_{\gamma} |\frac{1}{n}\sum_{i=1}^{n} R(X_i,\gamma) - P\, R(X,\gamma)| \to 0 \qquad \text{c.s.}$$

Si se cumple esa convergencia la clase $\mathcal{F} = \{R(x, \gamma : \gamma \in \Gamma\}$ de todas las funciones que indexa el parámetro γ se llaman clases Glivenko-Cantelli.