Teoría de Aprendizaje Estadístico II

M.Sc. William Caicedo Torres

Universidad Tecnológica de Bolívar

caicedo77@gmail.com

7 de octubre de 2016

Material inspirado en el curso de Caltech “Learning from Data”, por Yaser S. Abu Mustafá.

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Qué tenemos hasta ahora?

- Si tenemos una hipótesis, a partir del error E_{in}, podemos acotar la probabilidad de que generalice correctamente:

$$P[|E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon] \leq 2e^{-2\epsilon^2 N} \text{ (Hoeffding)}$$

- Si tenemos un conjunto de hipótesis (Regresión Logística por ejemplo), la probabilidad de generalizar correctamente se puede acotar a partir del error de entrenamiento E_{in} , así:

$$P[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N} \text{ (Unión)}$$

- M Representa el número de hipótesis en el conjunto. La mala noticia es que para casi todos los conjuntos de hipótesis interesantes, M es infinito.

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Una esperanza

- Entonces, de qué nos sirve la cota de la unión?

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

- Entonces, de qué nos sirve la cota de la unión?
- R/ Hasta ahora, en términos prácticos, de nada.
- A menos de que podamos darle a M un valor distinto a infinito.
- Es posible? Si. La cota de la unión sobreestima terriblemente la probabilidad de no generalizar correctamente.
- Para entender lo anterior, hay que recordar de donde viene M.

De donde viene M?

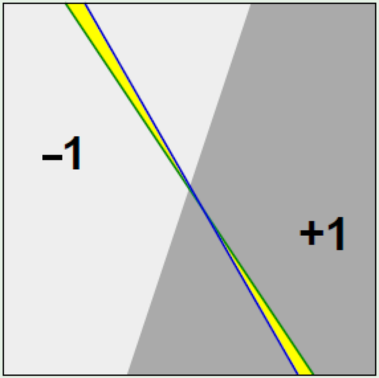
- M representa el número de hipótesis del conjunto en cuestión. Hasta ahora consideramos las hipótesis como eventos disyuntos entre si. De ahí la sumatoria de M términos en la cota de la Unión:

$$P[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq \sum_{m=1}^{M} P[|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon]$$

- Pero la realidad es que, el número **efectivo** de hipótesis es mucho menor.
- Por Qué? Porque 2 hipótesis distintas pueden tener el mismo desempeño, y en términos efectivos contarían como una sola!

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Podemos obtener un mejor estimado de M?



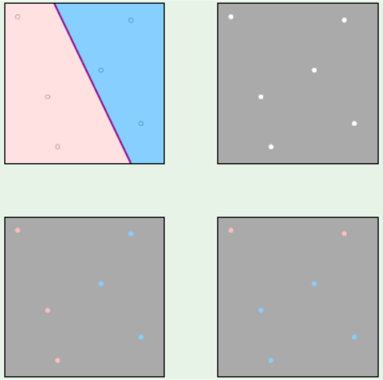
- Por supuesto! Para las 2 hipótesis de la figura, el cambio en las áreas positivas y negativas es mínimo, por lo que:

$$|E_{\text{in}}(h_1) - E_{\text{out}}(h_1)| \approx |E_{\text{in}}(h_2) - E_{\text{out}}(h_2)|$$

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Con qué reemplazamos M?

- En vez de considerar todo el espacio de entradas, consideremos un conjunto finito de puntos, y contemos el número de dicotomías (hipótesis efectivas).



Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

- Una hipótesis: $h : X \to \{0, 1\}$.
- Una dicotomía: $h : \{x_1, x_2, ..., x_N\} \to \{0, 1\}$.
- El número de hipótesis $|\mathcal{H}|$ puede ser infinito.
- El número de dicotomías $|\mathcal{H}(x_1, x_2, ..., x_N)|$ es máximo 2^N.
- Número de dicotomías: Candidato para reemplazar M.

La función de crecimiento

- La función de crecimiento cuenta el máximo de dicotomías posibles (hipótesis efectivas) en N puntos cualquiera:

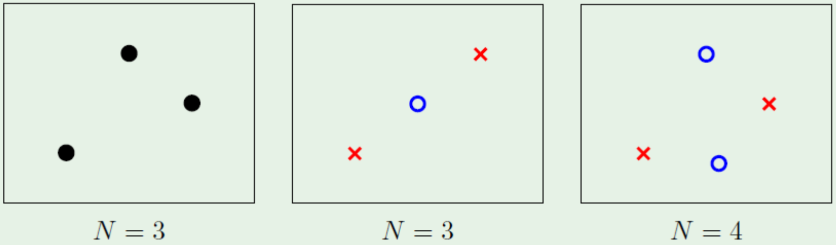
$$m_{\mathcal{H}}(N) = max_{x_1,\ldots,x_N \in X} |\mathcal{H}(x_1, \ldots, x_N)|$$

- La función de crecimiento satisface:

$$m_{\mathcal{H}}(N) \leq 2^N$$

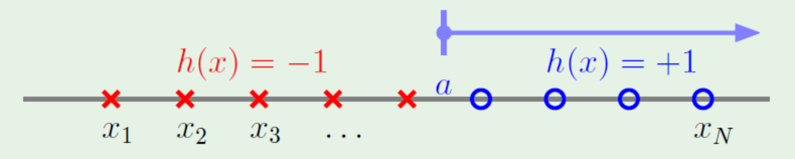
Aplicando la definición de $m_{\mathcal{H}}$: Clasificador Lineal

Cuál es el número máximo de dicotomías posible usando un separador lineal en 3 puntos? En 4 puntos?



$m_{\mathcal{H}}(3) = 8$ $\qquad$ $m_{\mathcal{H}}(4) = 14$

Ejemplos ilustrativos: Rayos positivos

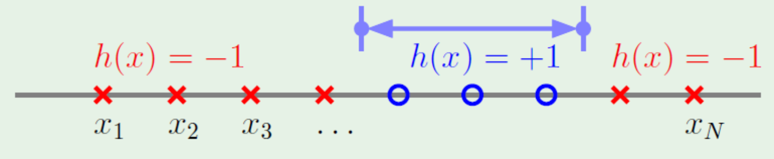


$$\mathcal{H} \text{ es un conjunto de } h : \mathbb{R} \to \{-1, +1\}$$

$$h(x) = \text{sign}(x - a)$$

$$m_{\mathcal{H}}(N) = N + 1$$

Ejemplos ilustrativos: Intervalos Positivos



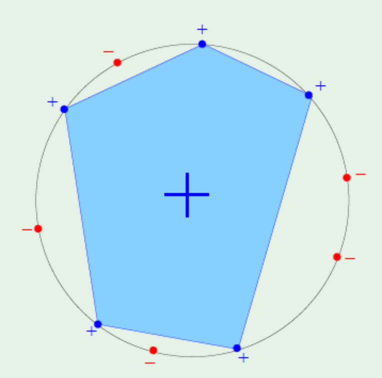
$\mathcal{H}$ es un conjunto de $h : \mathbb{R} \to \{-1, +1\}$

Los extremos del intervalo se ubican en 2 de $N+1$ posibles puntos.

$$m_{\mathcal{H}}(N) = \binom{N+1}{2} + 1 = \frac{1}{2}N^2 + \frac{1}{2}N + 1$$

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

- $\mathcal{H}$ es un conjunto de $h : \mathbb{R}^2 \rightarrow \{-1, +1\}$
- $h(x) = +1$ es una región convexa.
- $m_{\mathcal{H}}(N) = 2^N$
- Los N puntos son "quebrados" por el conjunto convexo.



Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Recapitulando: 3 Funciones de crecimiento

- $\mathcal{H}$ es rayos positivos:

$$m_{\mathcal{H}}(N) = N + 1$$

- $\mathcal{H}$ es intervalos positivos:

$$m_{\mathcal{H}}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1$$

- $\mathcal{H}$ es conjuntos convexos:

$$m_{\mathcal{H}}(N) = 2^N$$

Regresando a la cota de la Unión

- Recordemos que:

$$P[|E_{\text{in}} - E_{\text{out}}| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$

- Qué sucedería si $m_{\mathcal{H}}$ reemplaza a M ?
- Bueno, si $m_{\mathcal{H}}$ es polinomial $\rightarrow$ Excelente!
- Solo nos faltaría probar que $m_{\mathcal{H}}$ es polinomial...

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

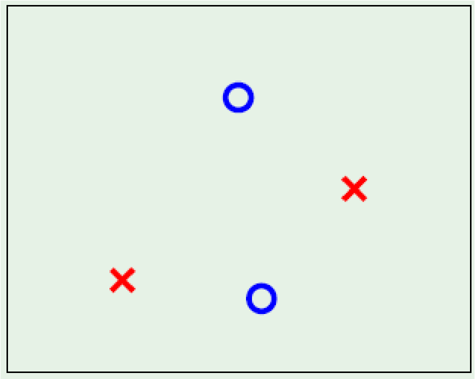
Noción clave: Break point

Definición:

- Si algún conjunto de datos de tamaño k ya no puede ser "quebrado" por $\mathcal{H}$, entonces k es un *break point* para $\mathcal{H}$

$$m_{\mathcal{H}} < 2^N$$

- Para un clasificador lineal en 2D, $k = 4$
- Un conjunto de datos de mayor tamaño tampoco puede ser quebrado.



Break point de los 3 ejemplos tratados

- Rayos positivos $m_{\mathcal{H}}(N) = N + 1$

  Break point $k = 2$ 

- Intervalos positivos $m_{\mathcal{H}}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1$:

  Break point $k = 3$ 

- Conjuntos convexos $m_{\mathcal{H}}(N) = 2^N$:

  Break point $k = \infty$

Resultado principal del día de hoy

- No hay break point $\rightarrow m_{\mathcal{H}} = 2^N$
- Si hay break point $\rightarrow m_{\mathcal{H}}$ es **polinomial** en N.

Muchas gracias!

Preguntas?

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS