Teoría de Aprendizaje Estadístico III

M.Sc. William Caicedo Torres

Universidad Tecnológica de Bolívar

caicedo77@gmail.com

14 de octubre de 2016

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

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.
- Para librarnos de infinitas hipótesis, hablaremos mejor de dicotomías (número efectivo de hipótesis en N puntos).

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Qué tenemos hasta ahora?

- Vamos a reemplazar M por la función de crecimiento: Número de dicotomías que el algoritmo puede generar para N puntos:

$$m_{\mathcal{H}}(N)$$

- break point: Valor de N a partir del cual $m_{\mathcal{H}}(N) < 2^N$
- Si un algoritmo no tiene break point $\rightarrow m_{\mathcal{H}} = 2^N$
- Si un algoritmo tiene break point $\rightarrow m_{\mathcal{H}}$ es **polinomial** en N (La generalización es factible).

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

$m_{\mathcal{H}}(N)$ es polinomial si hay un break point

- Si existe un break point k, entonces

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

Recapitulando: 3 Funciones de crecimiento

$$\sum_{i=0}^{k-1} \binom{N}{i}$$

- $\mathcal{H}$ es rayos positivos (break point $\mathbf{k = 2}$):

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

- $\mathcal{H}$ es intervalos positivos (break point $\mathbf{k = 3}$):

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

- $\mathcal{H}$ es Regresión Logística en 2D (break point $\mathbf{k = 4}$):

$$m_{\mathcal{H}}(N) =? \leq \frac{1}{6}N^3 + \frac{5}{6}N + 1$$

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Reemplazando a M

- Recordemos que la idea es sacar a M de aquí

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

- Reemplazando M (y considerando algunos detalles matemáticos) tendríamos que:

$$P[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N} \text{ (Vapnik-Chervonenkis)}$$

La cota VC

$$P[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq 4 m_{\mathcal{H}}(2N) e^{-\frac{1}{8}\epsilon^2 N}$$

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

La cota VC

- La cota VC es uno de los resultados teóricos más importantes del Machine Learning.
- La intuición que captura es la siguiente: Entre más grande sea el poder de un conjunto de hipótesis (mayor cantidad de dicotomías puede generar en N puntos), más datos se necesitan para permitir que el error de prueba se encuentre cerca del error de entrenamiento.
- En otras palabras, la complejidad del clasificador puede afectar negativamente nuestras posibilidades de generalizar.
- La cota VC es el punto de partida para la implementación de un entrenamiento basado en la minimización del riesgo estructural, en vez del riesgo empírico (el que hemos venido minimizando hasta ahora).

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

La dimensión VC

- Dentro del glosario utilizado por Vapnik, encontramos el término dimensión VC.
- La dimensión VC es una medida de la complejidad de un conjunto de hipótesis, similar al break point.
- De hecho la dimensión VC de un conjunto de hipótesis es el correspondiente break point menos uno.
- De acuerdo a lo anterior, podemos definir la dimensión VC de un conjunto de hipótesis $\mathcal{H}$, denotada por $d_{vc}(\mathcal{H})$, como el mayor valor de N para el que la función de crecimiento $m_{\mathcal{H}}(N) = 2^N$.
- En otras palabras, la dimensión VC de un conjunto de hipótesis $\mathcal{H}$ es el mayor número de puntos que $\mathcal{H}$ puede "quebrar".

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Dimensión VC: Ejemplos ilustrativos

- Rayos positivos

$$d_{vc} = 1$$

- Regresión Logística en 2D:

$$d_{vc} = 3 \text{ (Para d dimensiones, } d_{vc} = d + 1)$$

- Conjuntos convexos:

$$d_{vc} = \infty$$

Interpretación de la dimensión VC

- Tomemos por ejemplo la Regresión Logística:

$$d_{vc} = d + 1$$

- Qué es $d + 1$ en la Regresión Logística?

Interpretación de la dimensión VC

- Tomemos por ejemplo la Regresión Logística:

$$d_{vc} = d + 1$$

- Qué es $d + 1$ en la Regresión Logística?
- R/ El número de parámetros θ...

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Interpretación de la dimensión VC

- Tomemos por ejemplo la Regresión Logística:

$$d_{vc} = d + 1$$

- Qué es $d + 1$ en la Regresión Logística?
- R/ El número de parámetros θ...
- d_{vc} el es número de parámetros efectivos asociados a un conjunto de hipótesis $\mathcal{H}$.

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Número de datos de entrenamiento necesarios en función de la dimensión VC

- Recordemos la cota VC:

$$P[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}$$

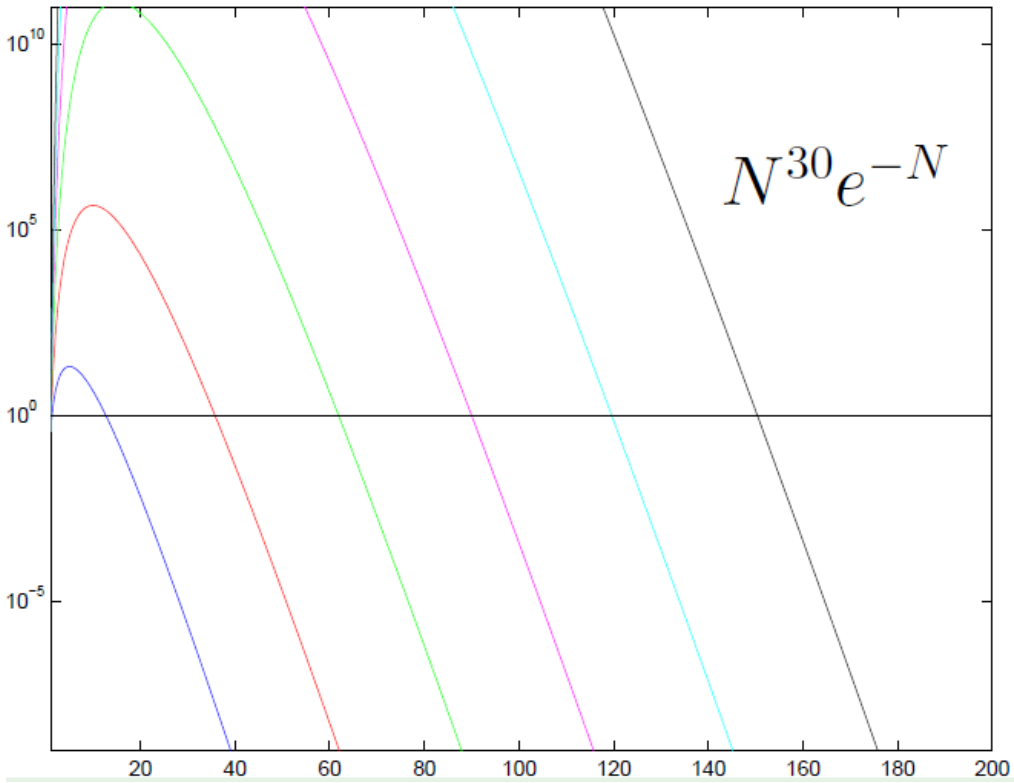
- Hagamos $4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N} = \delta$
- Para un δ y ϵ determinados, cómo depende N de d_{vc}?
- Para responder lo anterior, recordemos que si un conjunto de hipótesis $\mathcal{H}$ tiene dimensión VC no infinita, $m_{\mathcal{H}}(N)$ está acotada por un polinomio de grado $k-1$, o lo que es lo mismo, de grado d_{vc}.
- Teniendo en cuenta esto último, miremos el comportamiento de $N^d e^{-N}$

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Número de datos de entrenamiento necesarios en función de la dimensión VC

- Cómo cambia la cota superior de la probabillidad del error de generalización ($N^d e^{-N}$) en función de la cantidad de datos?
- Regla heurística:

$$N \geq 10 d_{vc}$$



Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Reorganizando la cota de generalización

- Comenzamos por la cota VC:

$$P[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}$$

- Dejamos la diferencia ϵ en función de δ:

$$\delta = 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N} \rightarrow \epsilon = \sqrt{\frac{8}{N}\ln\frac{4m_{\mathcal{H}}(2N)}{\delta}}$$

- Entonces, con probabilidad $\geq 1 - \delta$,

$$|E_{\text{out}} - E_{\text{in}}| \leq \Omega(N, \mathcal{H}, \delta)$$

- Lo que es lo mismo que decir:

$$E_{\text{out}} \leq E_{\text{in}} + \Omega(N, \mathcal{H}, \delta)$$

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Conclusiones Teoría Aprendizaje Estadístico I

- El aprendizaje es factible matemáticamente.
- La incertidumbre nos obliga a expresar esto a través de una afirmación probablemente aproximadamente correcta (PAC).
- El break point y la dimensión VC son maneras de expresar la complejidad de un conjunto de hipótesis. A mayor complejidad, mayor poder del conjunto de hipótesis en cuestión.
- El número efectivo de hipótesis que un conjunto $\mathcal{H}$ puede generar se expresa a través de la función de crecimiento.
- Si $\mathcal{H}$ tiene dimensión VC $< \infty$, la función de crecimiento $m_{\mathcal{H}}$ está acotada por un polinomio grado d_{vc}.

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Conclusiones Teoría Aprendizaje Estadístico II

- La dimensión VC puede interpretarse como el número de parámetros efectivos de un clasificador.
- A mayor complejidad del conjunto de hipótesis, más riesgo hay de que el error de prueba sea demasiado mayor que el error de entrenamiento (overfitting).
- Para mitigar dicho riesgo, la cantidad de ejemplos de entrenamiento N debe aumentar de forma proporcional a la dimensión VC.
- Para mitigar el riesgo de overfitting, hay que tener en cuenta la complejidad del clasificador en el entrenamiento: Regularización y Minimización del riesgo estructural.

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Muchas gracias!

Preguntas?

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS