Teoría de Aprendizaje Estadístico I

M.Sc. William Caicedo Torres

Universidad Tecnológica de Bolívar

caicedo77@gmail.com

30 de septiembre de 2016

Qué tenemos hasta ahora?

- Modelos de regresión lineal.
- Modelos de regresión logística.
- La habilidad de obtener un bajo error de entrenamiento (E_{in}) a través de features polinomiales.
- Pero, cómo podemos saber si nuestros modelos van a generalizar bien?
- Generalización $= E_{\text{out}}$ bajo!

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

Es el aprendizaje factible matemáticamente?

- De nada nos sirve que nuestros modelos se comporten bien en el entrenamiento, si no se comportan bien ante entradas nuevas.
- La habilidad de mantener un buen comportamiento ante entradas ausentes del conjunto de entrenamiento se denomina generalización.
- Pero, un bajo E_{in} implica un bajo E_{out}?

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

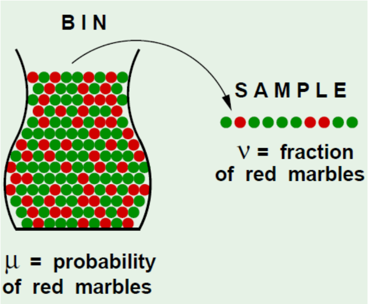
Es el aprendizaje factible matemáticamente?

- De nada nos sirve que nuestros modelos se comporten bien en el entrenamiento, si no se comportan bien ante entradas nuevas.
- La habilidad de mantener un buen comportamiento ante entradas ausentes del conjunto de entrenamiento se denomina generalización.
- Pero, un bajo E_{in} implica un bajo E_{out}?
- R/ No necesariamente. No podemos garantizar que sea posible obtener un E_{out} bajo como consecuencia del entrenamiento.
- Entonces?

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Las probabilidades al rescate

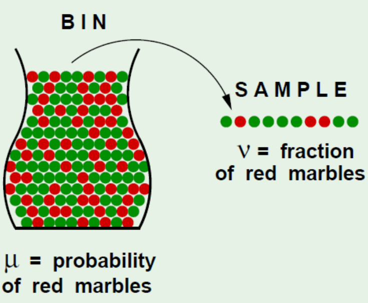
- Consideremos un recipiente lleno de canicas verdes y rojas.
- P[Sacar una canica roja] $= \mu$
- P[Sacar una canica verde] $= 1 - \mu$
- El valor de μ (probabilidad de sacar una canica roja) es desconocido.
- Sacamos N canicas de forma independiente.
- La fracción de canicas rojas en la muestra la denominaremos v



Universidad Tecnológica de Bolívar

ν nos dice algo acerca de μ?

- **NO:** Podríamos obtener una muestra llena de canicas verdes, mientras que la mayoria de canicas en el recipiente podría ser roja.
- **SÍ:** Es probable que la frecuencia muestral ν se aproxime a la frecuencia real dentro del recipiente μ.
- Posible vs Probable.



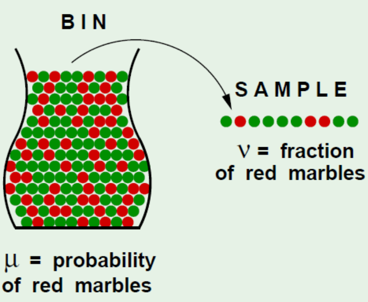
Qué v nos dice algo acerca de μ?

- En una muestra grande (N grande), v probablemente es cercano a μ (dentro de un "radio" ϵ).
- Formalmente,
$$P[|v - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$$
- A esta expresión se le conoce como la **Desigualdad de Hoeffding**.
- En otras palabras, la afirmación "$\mu = v$" es Probablemente Aproximadamente Correcta (P.A.C.).

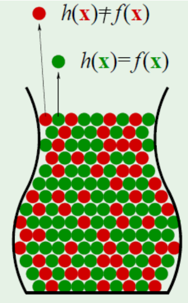
Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

$$P[|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$$

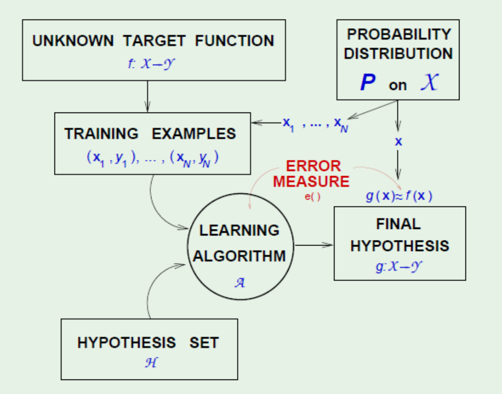
- Válida para todo N y para todo ϵ
- El resultado no depende de μ
- Disyuntiva: N, ϵ y el resultado.



- **Recipiente:** La incógnita es μ.
- **Aprendizaje:** La incógnita es una función $f : x \to y$
- Cada canica es un punto $x \in X$:
  - Nuestra hipótesis acierta en el punto x, ej: $h(x) = f(x)$
  - Nuestra hipótesis falla en el punto x, ej: $h(x) \neq f(x)$



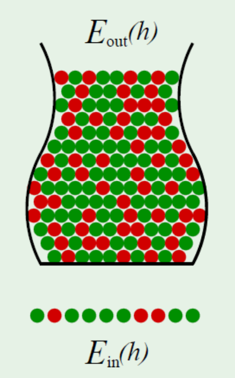
Mecanismo de aprendizaje revisado



Notación para el aprendizaje

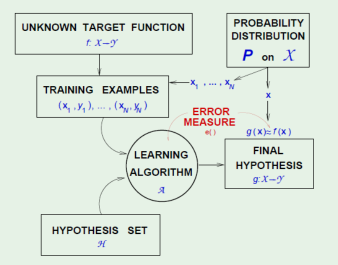
- Tanto μ como v dependen de la hipótesis h escogida.
- v es el error "dentro de la muestra" $\rightarrow E_{\text{in}}(h)$
- μ es el error "fuera de la muestra" $\rightarrow E_{\text{out}}(h)$
- Entonces la desigualdad de Hoeffding se convierte en

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



Ya probamos que el Machine Learning es factible?

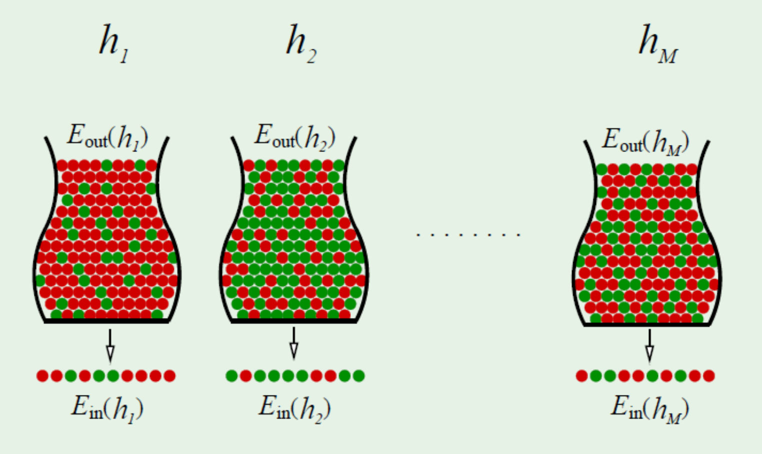
- Respuesta: **NO!** Hoeffding no aplica para múltiples recipientes (hipótesis en nuestro caso).
- Recordemos el mecanismo de aprendizaje:



- El algoritmo de aprendizaje busca entre un conjunto posiblemente infinito, la hipótesis que minimice el error de entrenamiento E_{in}. Cómo adaptamos Hoeffding al caso de múltiples recipientes (hipótesis)?

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Notación para múltiples recipientes/hipótesis



Hoeffding para múltiples hipótesis

- Consideremos la siguiente situación

Hoeffding para múltiples hipótesis

- Consideremos la siguiente situación
- Si tiramos al aire una moneda 10 veces, cual es la probabilidad de que la moneda caiga en cara 10 veces?

Hoeffding para múltiples hipótesis

- Consideremos la siguiente situación
- Si tiramos al aire una moneda 10 veces, cual es la probabilidad de que la moneda caiga en cara 10 veces?
- Respuesta: $\approx 0{,}1\,\%$ $(0{,}5^{10})$

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Hoeffding para múltiples hipótesis

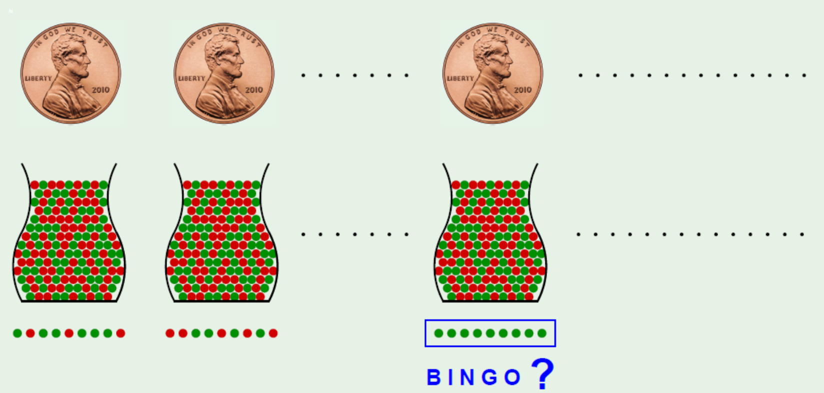
- Consideremos la siguiente situación
- Si tiramos al aire una moneda 10 veces, cual es la probabilidad de que la moneda caiga en cara 10 veces?
- Respuesta: $\approx 0{,}1\,\%$ $(0{,}5^{10})$
- Si tiramos al aire 1000 monedas 10 veces cada una, cuál es la probabilidad de que alguna moneda caiga 10 veces en cara?

Hoeffding para múltiples hipótesis

- Consideremos la siguiente situación
- Si tiramos al aire una moneda 10 veces, cual es la probabilidad de que la moneda caiga en cara 10 veces?
- Respuesta: $\approx 0{,}1\,\%$ $(0{,}5^{10})$
- Si tiramos al aire 1000 monedas 10 veces cada una, cuál es la probabilidad de que alguna moneda caiga 10 veces en cara?
- Respuesta: $\approx 63\,\%$ $(1-(1-0{,}5^{10})^{1000})$

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Hoeffding para múltiples hipótesis



Si tenemos un conjunto lo suficientemente grande de hipótesis, es posible que encontremos una hipótesis con bajo E_{in} por puro azar...

Hoeffding para múltiples hipótesis

- Ahora en vez de considerar hipótesis individuales, vamos a hablar de una hipótesis g sacada de manera aleatoria de un conjunto de hipótesis $\mathcal{H}$
- Entonces considerando cada hipótesis del conjunto como un evento independiente tenemos que

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

$$P[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq 2\sum_{m=1}^{m} e^{-2\epsilon^2 N}$$

$$P[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq 2{\color{red}M}e^{-2\epsilon^2 N}$$

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Union Bound (Cota superior de la unión)

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

- Esta cota se hace inocua, en la medida de que M aumenta. De hecho, si M es INFINITO, la cota pierde todo el sentido (Pregunta, para la Regresión Logística, cual es el tamaño de M?)
- Sin embargo es un primer paso para demostrar la factibilidad del aprendizaje autónomo.
- Punto de partida para cotas más estrictas.
- El veredicto es, el aprendizaje es factible teóricamente.

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Qué sabemos hasta ahora?

- Sabemos que utilizando nuestros algoritmos de optimización, podemos hacer que $E_{\text{in}}(g)$ pequeño.
- Ahora sabemos que el aprendizaje y una correcta generalización son factibles, es decir que

$$E_{\text{out}}(g) \approx E_{\text{in}}(g)$$

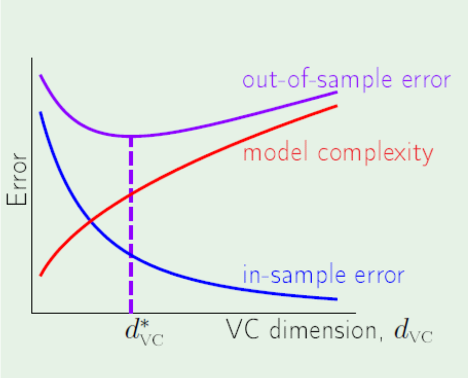
  es probable, dadas las condiciones adecuadas.
- Lo que nos va a permitir la Teoría de Aprendizaje Estadístico es obtener cotas útiles para el error, en el caso de conjuntos de hipótesis potencialmente infinitos.

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

- Con lo que hemos visto, podemos establecer que existe el siguiente *quid pro quo*:

$$\text{Complejidad del modelo} \uparrow \quad E_{\text{in}} \downarrow$$

$$\text{Complejidad del modelo} \uparrow \quad E_{\text{out}} - E_{\text{in}} \uparrow$$



Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Muchas gracias!

Preguntas?

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS