Aprendizaje estadístico

Fernando Lozano

Universidad de los Andes

15 de septiembre de 2017

Supuestos (modelo PAC)

- Espacio de entrada $\mathbf{x} \in \mathcal{X}$.

Supuestos (modelo PAC)

- Espacio de entrada $\mathbf{x} \in \mathcal{X}$.
- Distribución $\mathcal{D}$:

Supuestos (modelo PAC)

- Espacio de entrada $\mathbf{x} \in \mathcal{X}$.
- Distribución $\mathcal{D}$: $\mathbf{x} \sim \mathcal{D}$.

Supuestos (modelo PAC)

- Espacio de entrada $\mathbf{x} \in \mathcal{X}$.
- Distribución $\mathcal{D}$: $\mathbf{x} \sim \mathcal{D}$.
- Concepto (clasificador): $c \subseteq \mathcal{X}$

Supuestos (modelo PAC)

- Espacio de entrada $\mathbf{x} \in \mathcal{X}$.
- Distribución $\mathcal{D}$: $\mathbf{x} \sim \mathcal{D}$.
- Concepto (clasificador): $c \subseteq \mathcal{X}$
- $\mathcal{C}$: clase de conceptos (clasificadores o hipótesis) ($c \in \mathcal{C}$)

Supuestos (modelo PAC)

- Espacio de entrada $\mathbf{x} \in \mathcal{X}$.
- Distribución $\mathcal{D}$: $\mathbf{x} \sim \mathcal{D}$.
- Concepto (clasificador): $c \subseteq \mathcal{X}$
- $\mathcal{C}$: clase de conceptos (clasificadores o hipótesis) ($c \in \mathcal{C}$)
- Datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$:

Supuestos (modelo PAC)

- Espacio de entrada $\mathbf{x} \in \mathcal{X}$.
- Distribución $\mathcal{D}$: $\mathbf{x} \sim \mathcal{D}$.
- Concepto (clasificador): $c \subseteq \mathcal{X}$
- $\mathcal{C}$: clase de conceptos (clasificadores o hipótesis) ($c \in \mathcal{C}$)
- Datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$:
  - $\mathbf{x}_i \sim \mathcal{D}$, independientes.

Supuestos (modelo PAC)

- Espacio de entrada $\mathbf{x} \in \mathcal{X}$.
- Distribución $\mathcal{D}$: $\mathbf{x} \sim \mathcal{D}$.
- Concepto (clasificador): $c \subseteq \mathcal{X}$
- $\mathcal{C}$: clase de conceptos (clasificadores o hipótesis) ($c \in \mathcal{C}$)
- Datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$:
  - $\mathbf{x}_i \sim \mathcal{D}$, independientes.
  - $y_i = c(\mathbf{x}_i) \equiv I_c(\mathbf{x}_i)$

Algoritmo de aprendizaje

Algoritmo de aprendizaje

- Objetivo es aprender c (concepto objetivo) a partir de los datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$

Algoritmo de aprendizaje

- Objetivo es aprender c (concepto objetivo) a partir de los datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$
- Identificar hipótesis $h \in \mathcal{H}$ que dada una instancia $\mathbf{x}$ se capaz de identificar si $\mathbf{x} \in c$ con precisión.

Algoritmo de aprendizaje

- Objetivo es aprender c (concepto objetivo) a partir de los datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$
- Identificar hipótesis $h \in \mathcal{H}$ que dada una instancia $\mathbf{x}$ se capaz de identificar si $\mathbf{x} \in c$ con precisión.
- Conoce que $c \in \mathcal{C}$

Algoritmo de aprendizaje

- Objetivo es aprender c (concepto objetivo) a partir de los datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$
- Identificar hipótesis $h \in \mathcal{H}$ que dada una instancia $\mathbf{x}$ se capaz de identificar si $\mathbf{x} \in c$ con precisión.
- Conoce que $c \in \mathcal{C}$
- Criterio de error:

Algoritmo de aprendizaje

- Objetivo es aprender c (concepto objetivo) a partir de los datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$
- Identificar hipótesis $h \in \mathcal{H}$ que dada una instancia $\mathbf{x}$ se capaz de identificar si $\mathbf{x} \in c$ con precisión.
- Conoce que $c \in \mathcal{C}$
- Criterio de error:

$$e(h) = \mathbf{P}_{\mathcal{D}}\left[c(\mathbf{x}) \neq h(\mathbf{x})\right]$$

Algoritmo de aprendizaje

- Objetivo es aprender c (concepto objetivo) a partir de los datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$
- Identificar hipótesis $h \in \mathcal{H}$ que dada una instancia $\mathbf{x}$ se capaz de identificar si $\mathbf{x} \in c$ con precisión.
- Conoce que $c \in \mathcal{C}$
- Criterio de error:

$$e(h) = \mathbf{P}_{\mathcal{D}} \left[c(\mathbf{x}) \neq h(\mathbf{x}) \right]$$

- Eficiencia:

Algoritmo de aprendizaje

- Objetivo es aprender c (concepto objetivo) a partir de los datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$
- Identificar hipótesis $h \in \mathcal{H}$ que dada una instancia $\mathbf{x}$ se capaz de identificar si $\mathbf{x} \in c$ con precisión.
- Conoce que $c \in \mathcal{C}$
- Criterio de error:

$$e(h) = \mathbf{P}_{\mathcal{D}}\left[c(\mathbf{x}) \neq h(\mathbf{x})\right]$$

- Eficiencia:
  - Número de datos m es pequeño.

Algoritmo de aprendizaje

- Objetivo es aprender c (concepto objetivo) a partir de los datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$
- Identificar hipótesis $h \in \mathcal{H}$ que dada una instancia $\mathbf{x}$ se capaz de identificar si $\mathbf{x} \in c$ con precisión.
- Conoce que $c \in \mathcal{C}$
- Criterio de error:

$$e(h) = \mathbf{P}_{\mathcal{D}}\left[c(\mathbf{x}) \neq h(\mathbf{x})\right]$$

- Eficiencia:
  - Número de datos m es pequeño.
  - Tiempo de computación es pequeño.

Precisión y confianza

Precisión y confianza

- Nos gustaría un algoritmo que obtuviera $h \in \mathcal{H}$ con $e(h) < \varepsilon$ para $\varepsilon \lll$

Precisión y confianza

- Nos gustaría un algoritmo que obtuviera $h \in \mathcal{H}$ con $e(h) < \varepsilon$ para $\varepsilon \lll$
- Note que en general, h depende de los datos $\mathbf{x}_i \sim \mathcal{D}$.

Precisión y confianza

- Nos gustaría un algoritmo que obtuviera $h \in \mathcal{H}$ con $e(h) < \varepsilon$ para $\varepsilon \lll$
- Note que en general, h depende de los datos $\mathbf{x}_i \sim \mathcal{D}$.
- $e(h)$ es una variable aleatoria.

Precisión y confianza

- Nos gustaría un algoritmo que obtuviera $h \in \mathcal{H}$ con $e(h) < \varepsilon$ para $\varepsilon \lll$
- Note que en general, h depende de los datos $\mathbf{x}_i \sim \mathcal{D}$.
- $e(h)$ es una variable aleatoria.
- Es probable que el algoritmo falle:

Precisión y confianza

- Nos gustaría un algoritmo que obtuviera $h \in \mathcal{H}$ con $e(h) < \varepsilon$ para $\varepsilon \lll$
- Note que en general, h depende de los datos $\mathbf{x}_i \sim \mathcal{D}$.
- $e(h)$ es una variable aleatoria.
- Es probable que el algoritmo falle: no pueda encontrar una hipótesis ε-buena.

Precisión y confianza

- Nos gustaría un algoritmo que obtuviera $h \in \mathcal{H}$ con $e(h) < \varepsilon$ para $\varepsilon \lll$
- Note que en general, h depende de los datos $\mathbf{x}_i \sim \mathcal{D}$.
- $e(h)$ es una variable aleatoria.
- Es probable que el algoritmo falle: no pueda encontrar una hipótesis ε-buena.
- Queremos que con alta probabilidad h sea ε-buena:

Precisión y confianza

- Nos gustaría un algoritmo que obtuviera $h \in \mathcal{H}$ con $e(h) < \varepsilon$ para $\varepsilon \lll$
- Note que en general, h depende de los datos $\mathbf{x}_i \sim \mathcal{D}$.
- $e(h)$ es una variable aleatoria.
- Es probable que el algoritmo falle: no pueda encontrar una hipótesis ε-buena.
- Queremos que con alta probabilidad h sea ε-buena:

$$\mathbf{P}_{\mathcal{D}}\left[e(h) \geq \varepsilon\right] \leq \delta$$

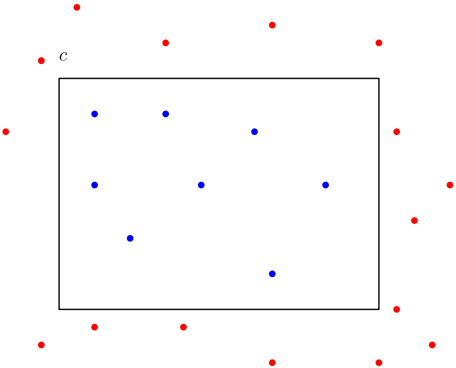
Precisión y confianza

- Nos gustaría un algoritmo que obtuviera $h \in \mathcal{H}$ con $e(h) < \varepsilon$ para $\varepsilon \lll$
- Note que en general, h depende de los datos $\mathbf{x}_i \sim \mathcal{D}$.
- $e(h)$ es una variable aleatoria.
- Es probable que el algoritmo falle: no pueda encontrar una hipótesis ε-buena.
- Queremos que con alta probabilidad h sea ε-buena:

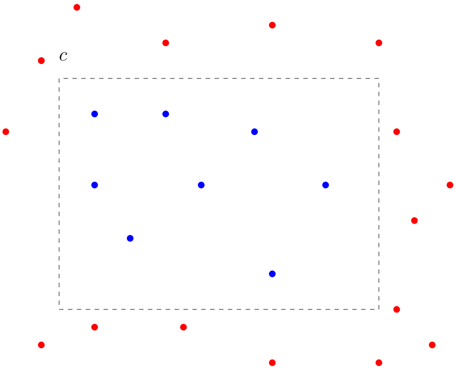
$$\mathbf{P}_{\mathcal{D}}\left[e(h) \geq \varepsilon\right] \leq \delta$$

Ejemplo: Aprendizaje de rectángulos en $\mathbb{R}^2$

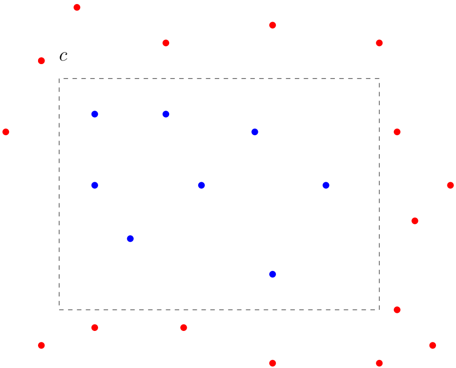
Ejemplo: Aprendizaje de rectángulos en $\mathbb{R}^2$



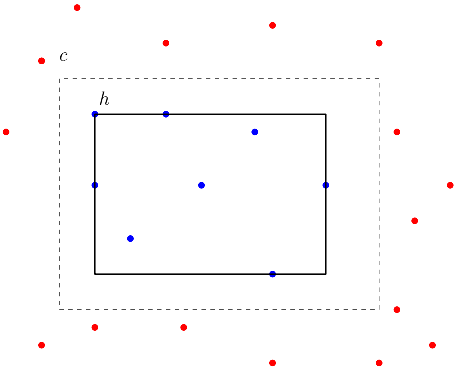
Ejemplo: Aprendizaje de rectángulos en $\mathbb{R}^2$



Algoritmo (consistente)



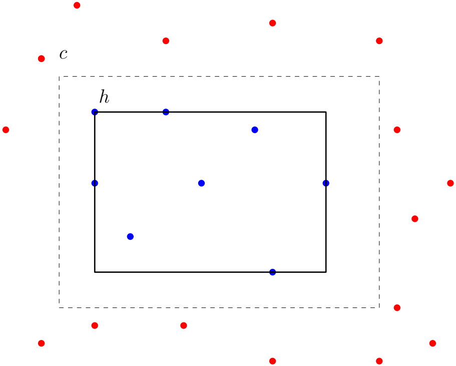
Algoritmo (consistente)



Análisis

- Dados $\varepsilon, \delta > 0$,

$$\mathbf{P}_{\mathcal{D}}\left[e(h) \geq \varepsilon\right] \leq \delta?$$

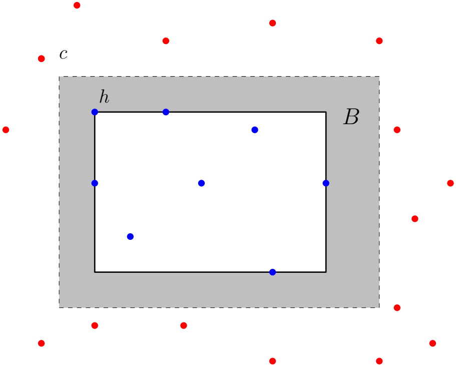


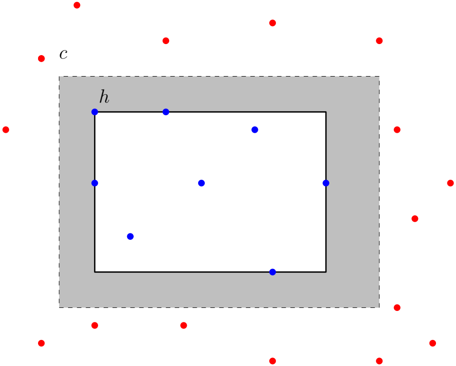
Análisis

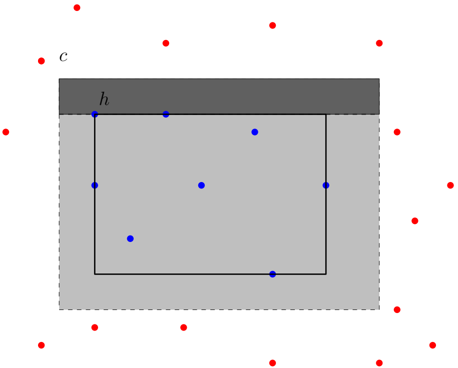
- Dados $\varepsilon, \delta > 0$,

$$\mathbf{P}_{\mathcal{D}}\left[e(h) \geq \varepsilon\right] \leq \delta?$$

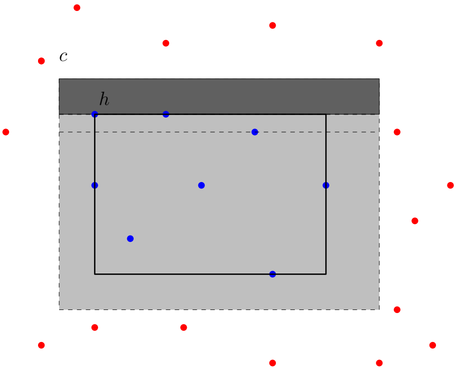
- Queremos $\mathbf{P}_{\mathcal{D}}\left[B\right] \leq \varepsilon$



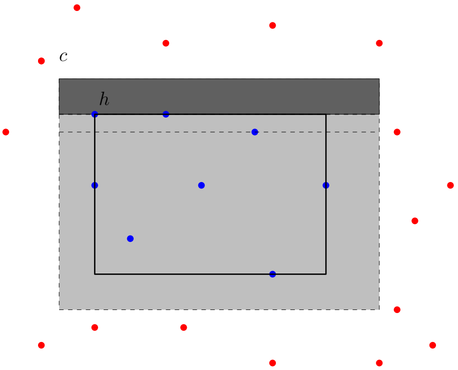

c
h



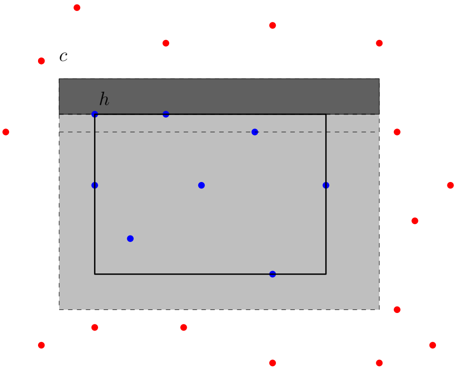
- Franja T'



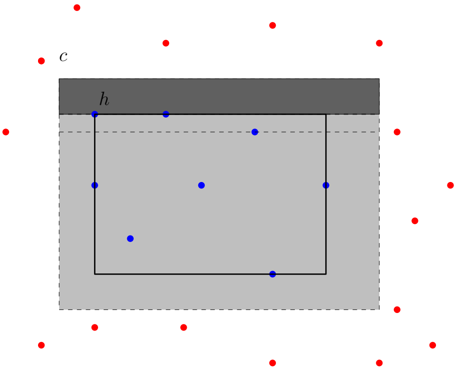
- Franja T'
- Franja T con $\mathbf{P}_{\mathcal{D}}\left[T\right] = \frac{\varepsilon}{4}$



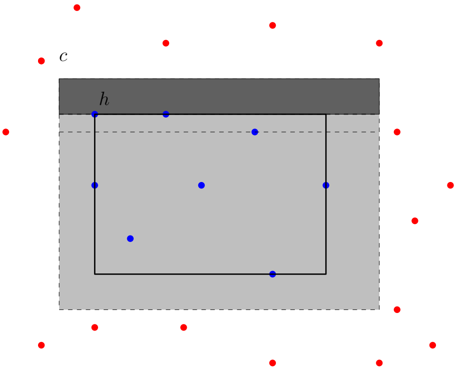
- Franja T'
- Franja T con $\mathbf{P}_{\mathcal{D}}\left[T\right] = \frac{\varepsilon}{4}$
- $\mathbf{P}_{\mathcal{D}}\left[T'\right] > \frac{\varepsilon}{4} \Leftrightarrow$ no hay puntos en T.



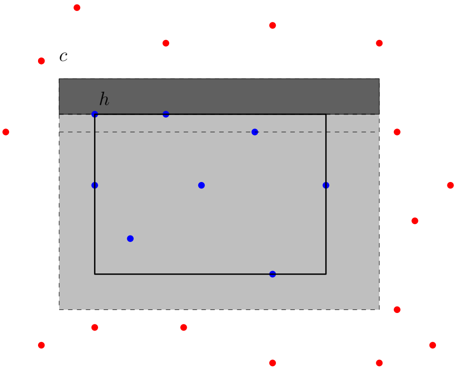
- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x} \notin T\right] = 1 - \frac{\varepsilon}{4}$



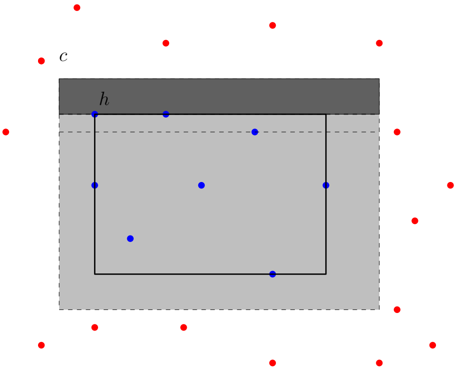
- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x} \notin T\right] = 1 - \frac{\varepsilon}{4}$
- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x}_1, \ldots, \mathbf{x}_m \notin T\right] =$



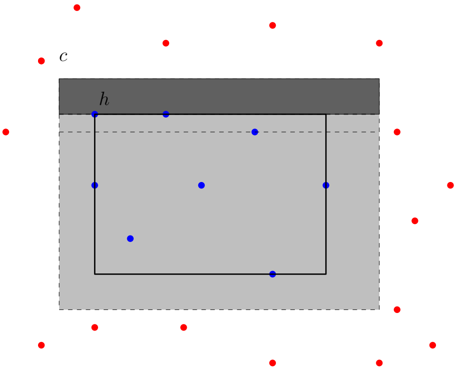
- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x} \notin T\right] = 1 - \frac{\varepsilon}{4}$
- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x}_1, \ldots, \mathbf{x}_m \notin T\right] = \left(1 - \frac{\varepsilon}{4}\right)^m$



- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x} \notin T\right] = 1 - \frac{\varepsilon}{4}$
- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x}_1, \ldots, \mathbf{x}_m \notin T\right] = \left(1 - \frac{\varepsilon}{4}\right)^m$
- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x}_1, \ldots, \mathbf{x}_m \notin B\right]$



- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x} \notin T\right] = 1 - \frac{\varepsilon}{4}$
- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x}_1, \ldots, \mathbf{x}_m \notin T\right] = \left(1 - \frac{\varepsilon}{4}\right)^m$
- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x}_1, \ldots, \mathbf{x}_m \notin B\right] \leq 4\left(1 - \frac{\varepsilon}{4}\right)^m$



- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x} \notin T\right] = 1 - \frac{\varepsilon}{4}$
- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x}_1, \ldots, \mathbf{x}_m \notin T\right] = \left(1 - \frac{\varepsilon}{4}\right)^m$
- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x}_1, \ldots, \mathbf{x}_m \notin B\right] \leq 4\left(1 - \frac{\varepsilon}{4}\right)^m$

- Queremos escoger m que satisfaga

$$4\left(1-\frac{\varepsilon}{4}\right)^{m} \leq \delta$$

- Queremos escoger m que satisfaga

$$4\left(1-\frac{\varepsilon}{4}\right)^m \leq \delta$$

usando $1-x \leq e^{-x}$ tenemos

- Queremos escoger m que satisfaga

$$4\left(1-\frac{\varepsilon}{4}\right)^m \leq \delta$$

usando $1-x \leq e^{-x}$ tenemos

$$4e^{-\frac{\varepsilon m}{4}} \leq \delta$$

- Queremos escoger m que satisfaga

$$4\left(1-\frac{\varepsilon}{4}\right)^m \leq \delta$$

usando $1-x \leq e^{-x}$ tenemos

$$4e^{-\frac{\varepsilon m}{4}} \leq \delta$$

o

$$m \geq \frac{4}{\varepsilon}\ln\frac{4}{\delta}$$

- Queremos escoger m que satisfaga

$$4\left(1-\frac{\varepsilon}{4}\right)^m \leq \delta$$

usando $1-x \leq e^{-x}$ tenemos

$$4e^{-\frac{\varepsilon m}{4}} \leq \delta$$

o

$$m \geq \frac{4}{\varepsilon}\ln\frac{4}{\delta}$$

- El algoritmo consistente con por lo menos $\frac{4}{\varepsilon}\ln\frac{4}{\delta}$ datos produce con probabilidad por lo menos $1-\delta$ una hipótesis que clasifica mal un nuevo dato con probabilidad máxima de ε.

Probablemente (δ) Aproximadamente (ε) Correcto

Probablemente (δ) Aproximadamente (ε) Correcto

- $e(h) \leq \varepsilon$ pequeño: h es aproximadamente correcta.

Probablemente (δ) Aproximadamente (ε) Correcto

- $e(h) \leq \varepsilon$ pequeño: h es aproximadamente correcta.
- Datos son aleatorios

Probablemente (δ) Aproximadamente (ε) Correcto

- $e(h) \leq \varepsilon$ pequeño: h es aproximadamente correcta.
- Datos son aleatorios $\Rightarrow$ Algoritmo puede fallar con cierta probabilidad δ.

Probablemente (δ) Aproximadamente (ε) Correcto

- $e(h) \leq \varepsilon$ pequeño: h es aproximadamente correcta.
- Datos son aleatorios $\Rightarrow$ Algoritmo puede fallar con cierta probabilidad δ.

$$m \text{ datos} \longrightarrow \begin{cases} e(h) < \varepsilon & h \quad \text{“buena”} \\ e(h) \geq \varepsilon & h \quad \text{“mala”} \end{cases}$$

Probablemente (δ) Aproximadamente (ε) Correcto

- $e(h) \leq \varepsilon$ pequeño: h es aproximadamente correcta.
- Datos son aleatorios $\Rightarrow$ Algoritmo puede fallar con cierta probabilidad δ.

$$m \text{ datos} \longrightarrow \begin{cases} e(h) < \varepsilon & h \quad \text{“buena”} \\ e(h) \geq \varepsilon & h \quad \text{“mala”} \end{cases}$$

$$\mathbf{P}_{\mathcal{D}}\left[h \quad \text{sea ‘mala”}\right] \leq \delta$$

Probablemente (δ) Aproximadamente (ε) Correcto

- $e(h) \leq \varepsilon$ pequeño: h es aproximadamente correcta.
- Datos son aleatorios $\Rightarrow$ Algoritmo puede fallar con cierta probabilidad δ.

$$m \text{ datos} \longrightarrow \begin{cases} e(h) < \varepsilon & h \quad \text{“buena”} \\ e(h) \geq \varepsilon & h \quad \text{“mala”} \end{cases}$$

$$\mathbf{P}_{\mathcal{D}}\left[h \quad \text{sea ‘mala”}\right] \leq \delta$$

- $\delta \ll$

Probablemente (δ) Aproximadamente (ε) Correcto

- $e(h) \leq \varepsilon$ pequeño: h es aproximadamente correcta.
- Datos son aleatorios $\Rightarrow$ Algoritmo puede fallar con cierta probabilidad δ.

$$m \text{ datos} \longrightarrow \begin{cases} e(h) < \varepsilon & h \quad \text{``buena''} \\ e(h) \geq \varepsilon & h \quad \text{``mala''} \end{cases}$$

$$\mathbf{P}_{\mathcal{D}}\left[h \quad \text{sea `mala''}\right] \leq \delta$$

- $\delta \ll \; \Rightarrow h$ es probablemente aproximadamente correcta.

Probablemente (δ) Aproximadamente (ε) Correcto

- $e(h) \leq \varepsilon$ pequeño: h es aproximadamente correcta.
- Datos son aleatorios $\Rightarrow$ Algoritmo puede fallar con cierta probabilidad δ.

$$m \text{ datos} \longrightarrow \begin{cases} e(h) < \varepsilon & h \quad \text{“buena”} \\ e(h) \geq \varepsilon & h \quad \text{“mala”} \end{cases}$$

$$\mathbf{P}_{\mathcal{D}}\left[h \quad \text{sea ‘mala”}\right] \leq \delta$$

- $\delta \ll \ \Rightarrow h$ es probablemente aproximadamente correcta.
- Cualquier $\mathcal{D}$.

Probablemente (δ) Aproximadamente (ε) Correcto

- $e(h) \leq \varepsilon$ pequeño: h es aproximadamente correcta.
- Datos son aleatorios $\Rightarrow$ Algoritmo puede fallar con cierta probabilidad δ.

$$m \text{ datos} \longrightarrow \begin{cases} e(h) < \varepsilon & h \quad \text{“buena”} \\ e(h) \geq \varepsilon & h \quad \text{“mala”} \end{cases}$$

$$\mathbf{P}_{\mathcal{D}}\left[h \quad \text{sea ‘mala”}\right] \leq \delta$$

- $\delta \ll \ \Rightarrow h$ es probablemente aproximadamente correcta.
- Cualquier $\mathcal{D}$.
- Cualquier $c \in \mathcal{C}$

Probablemente (δ) Aproximadamente (ε) Correcto

- $e(h) \leq \varepsilon$ pequeño: h es aproximadamente correcta.
- Datos son aleatorios $\Rightarrow$ Algoritmo puede fallar con cierta probabilidad δ.

$$m \text{ datos} \longrightarrow \begin{cases} e(h) < \varepsilon & h \quad \text{“buena”} \\ e(h) \geq \varepsilon & h \quad \text{“mala”} \end{cases}$$

$$\mathbf{P}_{\mathcal{D}}\left[h \quad \text{sea ‘mala”}\right] \leq \delta$$

- $\delta \ll \; \Rightarrow h$ es probablemente aproximadamente correcta.
- Cualquier $\mathcal{D}$.
- Cualquier $c \in \mathcal{C}$
- Tiempo de corrida polinomial en $\frac{1}{\delta}$ y $\frac{1}{\varepsilon}$

Clases de hipótesis finitas

- $|\mathcal{H}| < \infty$

Clases de hipótesis finitas

- $|\mathcal{H}| < \infty$
- Algoritmo A: observa m datos y retorna h_A consistente.

Clases de hipótesis finitas

- $|\mathcal{H}| < \infty$
- Algoritmo A: observa m datos y retorna h_A consistente.
- Sea $B = \{h \in \mathcal{H} \,:\, e(h) > \varepsilon\}$,

Clases de hipótesis finitas

- $|\mathcal{H}| < \infty$
- Algoritmo A: observa m datos y retorna h_A consistente.
- Sea $B = \{h \in \mathcal{H} \,:\, e(h) > \varepsilon\}$,

$$\mathbf{P}_{\mathcal{D}}\left[h_A \in B\right] \leq \mathbf{P}_{\mathcal{D}}\left[\exists h \in B \,:\, h \text{ es consistente con los datos}\right]$$

Clases de hipótesis finitas

- $|\mathcal{H}| < \infty$
- Algoritmo A: observa m datos y retorna h_A consistente.
- Sea $B = \{h \in \mathcal{H} \, : \, e(h) > \varepsilon\}$,

$$\mathbf{P}_{\mathcal{D}}\left[h_A \in B\right] \leq \mathbf{P}_{\mathcal{D}}\left[\exists h \in B \, : \, h \text{ es consistente con los datos}\right]$$

- Para $h \in B$ fija:

$$\mathbf{P}_{\mathcal{D}}\left[h \text{ es consistente}\right] = \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_1) = c(\mathbf{x}_1) \wedge \cdots \wedge h(\mathbf{x}_m) = c(\mathbf{x}_m)\right]$$

Clases de hipótesis finitas

- $|\mathcal{H}| < \infty$
- Algoritmo A: observa m datos y retorna h_A consistente.
- Sea $B = \{h \in \mathcal{H} \,:\, e(h) > \varepsilon\}$,

$$\mathbf{P}_{\mathcal{D}}\left[h_A \in B\right] \leq \mathbf{P}_{\mathcal{D}}\left[\exists h \in B \,:\, h \text{ es consistente con los datos}\right]$$

- Para $h \in B$ fija:

$$\mathbf{P}_{\mathcal{D}}\left[h \text{ es consistente}\right] = \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_1) = c(\mathbf{x}_1) \wedge \cdots \wedge h(\mathbf{x}_m) = c(\mathbf{x}_m)\right]$$

Clases de hipótesis finitas

- $|\mathcal{H}| < \infty$
- Algoritmo A: observa m datos y retorna h_A consistente.
- Sea $B = \{h \in \mathcal{H} \,:\, e(h) > \varepsilon\}$,

$$\mathbf{P}_{\mathcal{D}}\left[h_A \in B\right] \leq \mathbf{P}_{\mathcal{D}}\left[\exists h \in B \,:\, h \text{ es consistente con los datos}\right]$$

- Para $h \in B$ fija:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[h \text{ es consistente}\right] &= \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_1) = c(\mathbf{x}_1) \wedge \cdots \wedge h(\mathbf{x}_m) = c(\mathbf{x}_m)\right] \\ &= \prod_{i=1}^{m} \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_i) = c(\mathbf{x}_i)\right]\end{aligned}$$

Clases de hipótesis finitas

- $|\mathcal{H}| < \infty$
- Algoritmo A: observa m datos y retorna h_A consistente.
- Sea $B = \{h \in \mathcal{H} \; : \; e(h) > \varepsilon\}$,

$$\mathbf{P}_{\mathcal{D}}\left[h_A \in B\right] \leq \mathbf{P}_{\mathcal{D}}\left[\exists h \in B \; : \; h \text{ es consistente con los datos}\right]$$

- Para $h \in B$ fija:

$$\begin{aligned}
\mathbf{P}_{\mathcal{D}}\left[h \text{ es consistente}\right] &= \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_1) = c(\mathbf{x}_1) \wedge \cdots \wedge h(\mathbf{x}_m) = c(\mathbf{x}_m)\right] \\
&= \prod_{i=1}^{m} \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_i) = c(\mathbf{x}_i)\right] \\
&\leq (1 - \varepsilon)^m
\end{aligned}$$

- Sumando sobre todas las posiblidades:

- Sumando sobre todas las posiblidades:

$$\mathbf{P}_{\mathcal{D}}\left[\exists h \in B\ :\ h \,\text{es consistente con los datos}\right] \leq |B|\,(1-\varepsilon)^m$$

- Sumando sobre todas las posiblidades:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[\exists h \in B\ :\ h \,\text{es consistente con los datos}\right] &\leq |B|\,(1-\varepsilon)^m \\ &\leq |\mathcal{H}|\,(1-\varepsilon)^m\end{aligned}$$

- Sumando sobre todas las posiblidades:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[\exists h \in B \;:\; h \,\text{es consistente con los datos}\right] &\leq |B|\,(1-\varepsilon)^m \\ &\leq |\mathcal{H}|\,(1-\varepsilon)^m \\ &\leq |\mathcal{H}|\,e^{-\varepsilon m}\end{aligned}$$

- Sumando sobre todas las posiblidades:

$$\begin{aligned}
\mathbf{P}_{\mathcal{D}}\left[\exists h \in B \;:\; h \,\text{es consistente con los datos}\right] &\leq |B|\,(1-\varepsilon)^m \\
&\leq |\mathcal{H}|\,(1-\varepsilon)^m \\
&\leq |\mathcal{H}|\,e^{-\varepsilon m}
\end{aligned}$$

- Para ε, δ dados, podemos calcular:

- Sumando sobre todas las posiblidades:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[\exists h \in B \;:\; h \text{ es consistente con los datos}\right] &\leq |B|\,(1-\varepsilon)^m \\ &\leq |\mathcal{H}|\,(1-\varepsilon)^m \\ &\leq |\mathcal{H}|\, e^{-\varepsilon m}\end{aligned}$$

- Para ε, δ dados, podemos calcular:

$$|\mathcal{H}|\, e^{-\varepsilon m} \leq \delta$$

- Sumando sobre todas las posiblidades:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[\exists h \in B\ :\ h \text{ es consistente con los datos}\right] &\leq |B|\,(1-\varepsilon)^m \\ &\leq |\mathcal{H}|\,(1-\varepsilon)^m \\ &\leq |\mathcal{H}|\,e^{-\varepsilon m}\end{aligned}$$

- Para ε, δ dados, podemos calcular:

$$\begin{aligned}|\mathcal{H}|\,e^{-\varepsilon m} &\leq \delta \\ m &\geq \frac{1}{\varepsilon}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)\end{aligned}$$

- Sumando sobre todas las posiblidades:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[\exists h \in B \;:\; h \,\text{es consistente con los datos}\right] &\leq |B|\,(1-\varepsilon)^m \\ &\leq |\mathcal{H}|\,(1-\varepsilon)^m \\ &\leq |\mathcal{H}|\,e^{-\varepsilon m}\end{aligned}$$

- Para ε, δ dados, podemos calcular:

$$\begin{aligned}|\mathcal{H}|\,e^{-\varepsilon m} &\leq \delta \\ m &\geq \frac{1}{\varepsilon}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)\end{aligned}$$

- O para m, δ dados, podemos decir que con probabilidad por lo menos $1-\delta$

$$e(h) \leq \frac{1}{m}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)$$

Complejidad para clases de hipótesis infinitas.

Complejidad para clases de hipótesis infinitas.

- Espacio de entrada $\mathcal{X}$, clase de conceptos $\mathcal{C}$.

Complejidad para clases de hipótesis infinitas.

- Espacio de entrada $\mathcal{X}$, clase de conceptos $\mathcal{C}$.
- $S = \{x_1, x_2, \ldots, x_m\} \subset \mathcal{X}$

Complejidad para clases de hipótesis infinitas.

- Espacio de entrada $\mathcal{X}$, clase de conceptos $\mathcal{C}$.
- $S = \{x_1, x_2, \ldots, x_m\} \subset \mathcal{X}$
- $\Pi_{\mathcal{C}}(S) = \{c \cap S \ : \ c \in \mathcal{C}\}$

Complejidad para clases de hipótesis infinitas.

- Espacio de entrada $\mathcal{X}$, clase de conceptos $\mathcal{C}$.
- $S = \{x_1, x_2, \ldots, x_m\} \subset \mathcal{X}$
- $\Pi_{\mathcal{C}}(S) = \{c \cap S \ : \ c \in \mathcal{C}\}$
- $\Pi_{\mathcal{C}}(S) = \left\{ \begin{bmatrix} c(x_1) & c(x_2) & \ldots & c(x_m) \end{bmatrix} \ : \ c \in \mathcal{C} \right\}$

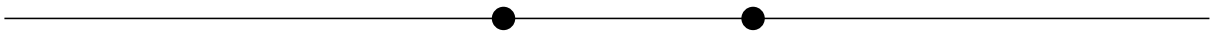
Complejidad para clases de hipótesis infinitas.

- Espacio de entrada $\mathcal{X}$, clase de conceptos $\mathcal{C}$.
- $S = \{x_1, x_2, \ldots, x_m\} \subset \mathcal{X}$
- $\Pi_{\mathcal{C}}(S) = \{c \cap S \ : \ c \in \mathcal{C}\}$
- $\Pi_{\mathcal{C}}(S) = \left\{ \begin{bmatrix} c(x_1) & c(x_2) & \ldots & c(x_m) \end{bmatrix} \ : \ c \in \mathcal{C} \right\}$

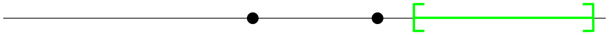
Definición

S es pulverizado (shattered) por $\mathcal{C}$ si $|\Pi_{\mathcal{C}}(S)| = 2^m$ (es decir $\Pi_{\mathcal{C}}(S) = \{0,1\}^{|S|}$)

Ejemplo: $\mathcal{C}$ es la clase de intervalos en $\mathbb{R}$

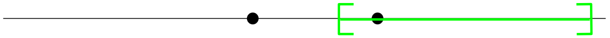


Ejemplo: $\mathcal{C}$ es la clase de intervalos en $\mathbb{R}$

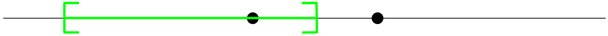


Ejemplo: $\mathcal{C}$ es la clase de intervalos en $\mathbb{R}$

Ejemplo: $\mathcal{C}$ es la clase de intervalos en $\mathbb{R}$



Ejemplo: $\mathcal{C}$ es la clase de intervalos en $\mathbb{R}$

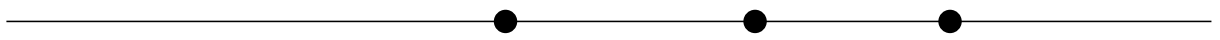


- Dos puntos son pulverizados por $\mathcal{C}$.

Ejemplo: $\mathcal{C}$ es la clase de intervalos en $\mathbb{R}$

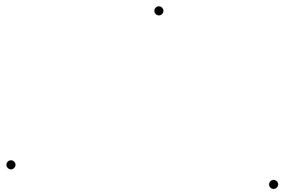
- Dos puntos son pulverizados por $\mathcal{C}$.

Ejemplo: $\mathcal{C}$ es la clase de intervalos en $\mathbb{R}$

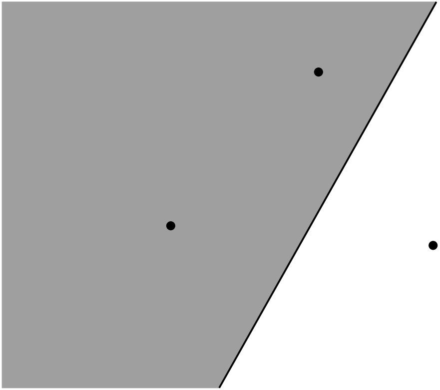


- Dos puntos son pulverizados por $\mathcal{C}$.
- Ningún conjunto de tres puntos es pulverizado por $\mathcal{C}$.

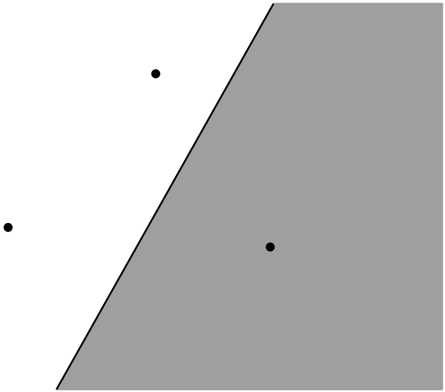
Ejemplo: $\mathcal{C}$ es la clase de semiplanos en $\mathbb{R}^2$



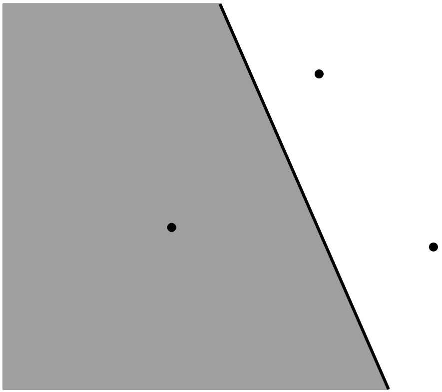
Ejemplo: $\mathcal{C}$ es la clase de semiplanos en $\mathbb{R}^2$



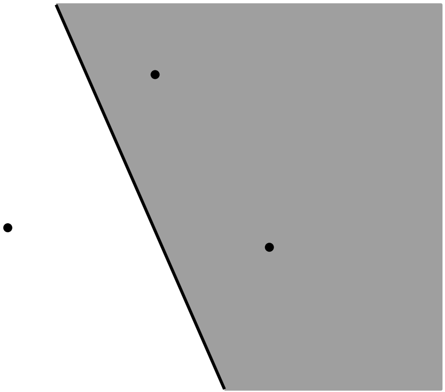
Ejemplo: $\mathcal{C}$ es la clase de semiplanos en $\mathbb{R}^2$



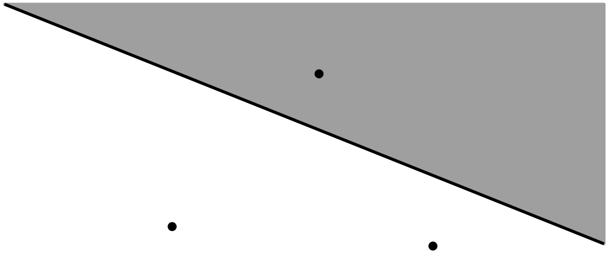
Ejemplo: $\mathcal{C}$ es la clase de semiplanos en $\mathbb{R}^2$



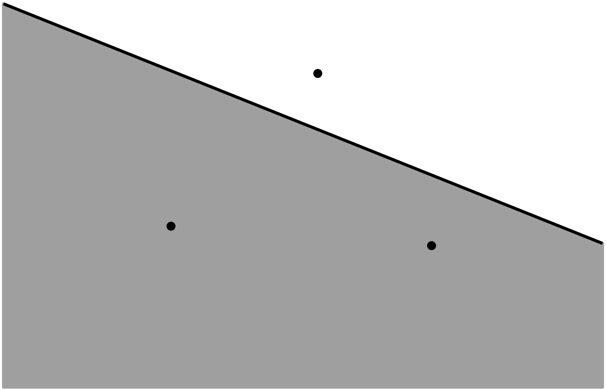
Ejemplo: $\mathcal{C}$ es la clase de semiplanos en $\mathbb{R}^2$



Ejemplo: $\mathcal{C}$ es la clase de semiplanos en $\mathbb{R}^2$

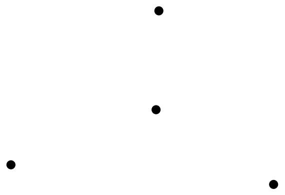


Ejemplo: $\mathcal{C}$ es la clase de semiplanos en $\mathbb{R}^2$



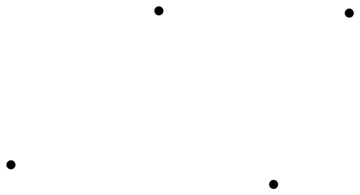
- Tres puntos son pulverizados por $\mathcal{C}$.

Ejemplo: $\mathcal{C}$ es la clase de semiplanos en $\mathbb{R}^2$



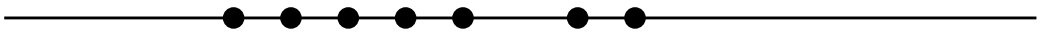
- Tres puntos son pulverizados por $\mathcal{C}$.
- 4 puntos?

Ejemplo: $\mathcal{C}$ es la clase de semiplanos en $\mathbb{R}^2$

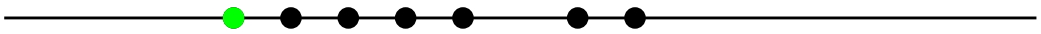


- Tres puntos son pulverizados por $\mathcal{C}$.
- 4 puntos?

Ejemplo: $\mathcal{C}$ es la clase de subconjuntos finitos en $\mathbb{R}$

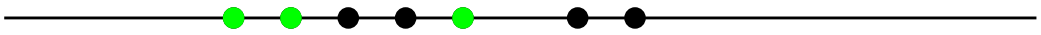


Ejemplo: $\mathcal{C}$ es la clase de subconjuntos finitos en $\mathbb{R}$

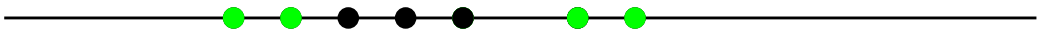


Ejemplo: $\mathcal{C}$ es la clase de subconjuntos finitos en $\mathbb{R}$

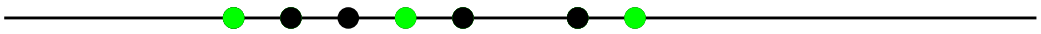
Ejemplo: $\mathcal{C}$ es la clase de subconjuntos finitos en $\mathbb{R}$



Ejemplo: $\mathcal{C}$ es la clase de subconjuntos finitos en $\mathbb{R}$



Ejemplo: $\mathcal{C}$ es la clase de subconjuntos finitos en $\mathbb{R}$



- Cualquier conjunto finito de puntos es pulverizado por $\mathcal{C}$.

La función de crecimiento

La función de crecimiento

Definición

Para un número natural m se define la función de crecimiento de $\mathcal{C}$ como:

$$\Pi_{\mathcal{C}}(m) = \max \{|\Pi_{\mathcal{C}}(S)| \; : \; |S| = m\}$$

La función de crecimiento

Definición

Para un número natural m se define la función de crecimiento de $\mathcal{C}$ como:

$$\Pi_{\mathcal{C}}(m) = \max\{|\Pi_{\mathcal{C}}(S)| \ : \ |S| = m\}$$

- Medida de complejidad de $\mathcal{C}$.

La función de crecimiento

Definición

Para un número natural m *se define la* *función de crecimiento de* $\mathcal{C}$ *como:*

$$\Pi_{\mathcal{C}}(m) = \max\left\{|\Pi_{\mathcal{C}}(S)| \;:\; |S| = m\right\}$$

- Medida de complejidad de $\mathcal{C}$.
- Si $\Pi_{\mathcal{C}}(m)$ crece rápidamente, conjuntos de puntos grandes pueden ser clasificados de más formas diferentes.

La función de crecimiento

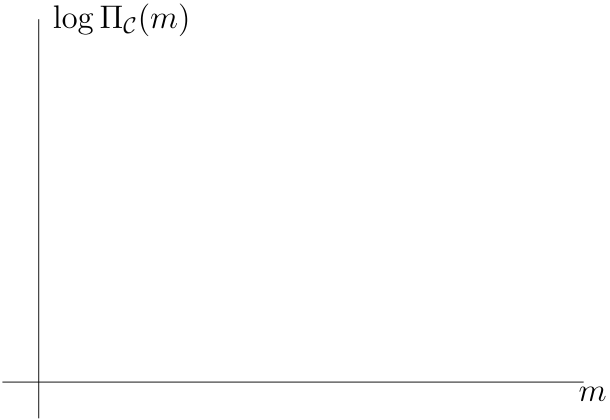
Definición

Para un número natural m se define la *función de crecimiento de $\mathcal{C}$* *como:*

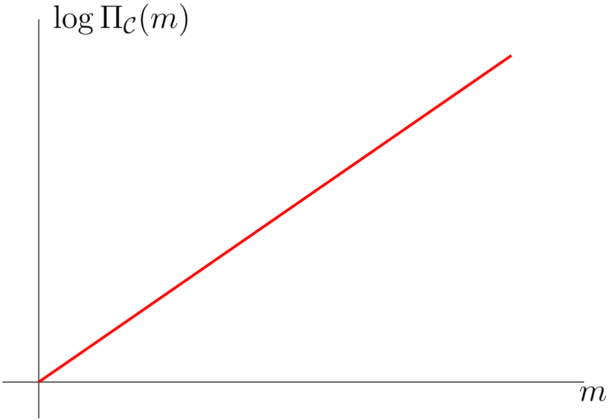
$$\Pi_{\mathcal{C}}(m) = \max\left\{|\Pi_{\mathcal{C}}(S)| \ : \ |S| = m\right\}$$

- Medida de complejidad de $\mathcal{C}$.
- Si $\Pi_{\mathcal{C}}(m)$ crece rápidamente, conjuntos de puntos grandes pueden ser clasificados de más formas diferentes.
- Cómo se comporta $\Pi_{\mathcal{C}}(m)$?

Comportamiento de $\Pi_{\mathcal{C}}(m)$

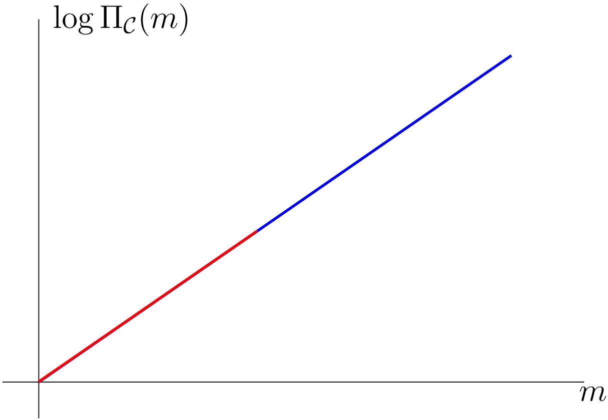


Comportamiento de $\Pi_{\mathcal{C}}(m)$



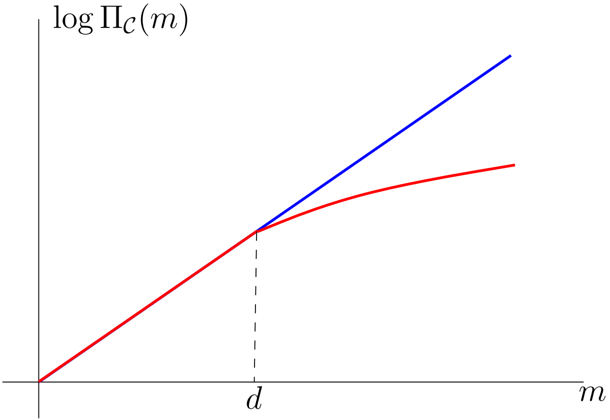
- $\Pi_{\mathcal{C}}(m) = 2^m$ (crecimiento exponencial).

Comportamiento de $\Pi_{\mathcal{C}}(m)$



- $\Pi_{\mathcal{C}}(m) = 2^m$ (crecimiento exponencial).

Comportamiento de $\Pi_{\mathcal{C}}(m)$



- $\Pi_{\mathcal{C}}(m) = 2^m$ (crecimiento exponencial).
- $\Pi_{\mathcal{C}}(m) = O(m^d)$ (crecimiento polinomial).

La dimensión de Vapnik Chervonenkis

Definición

La dimensión de Vapnik-Chervonenkis (o dimensión VC) de una clase de conceptos $\mathcal{C}$ se define como:

$$VC(\mathcal{C}) = \text{máx}\left\{m \;:\; \Pi_{\mathcal{C}}(m) = 2^m\right\}$$

(es decir, la cardinalidad máxima de un conjunto que es pulverizado por $\mathcal{C}$)

La dimensión de Vapnik Chervonenkis

Definición

La *dimensión de Vapnik-Chervonenkis (o dimensión VC)* *de una clase de conceptos* $\mathcal{C}$ *se define como:*

$$VC(\mathcal{C}) = \max\{m \,:\, \Pi_{\mathcal{C}}(m) = 2^m\}$$

(es decir, la cardinalidad máxima de un conjunto que es pulverizado por $\mathcal{C}$*)*

En los ejemplos:

- Intervalos en $\mathbb{R}$

La dimensión de Vapnik Chervonenkis

Definición

La dimensión de Vapnik-Chervonenkis (o dimensión VC) de una clase de conceptos $\mathcal{C}$ se define como:

$$VC(\mathcal{C}) = \text{máx}\,\{m \;:\; \Pi_{\mathcal{C}}(m) = 2^m\}$$

(es decir, la cardinalidad máxima de un conjunto que es pulverizado por $\mathcal{C}$)

En los ejemplos:

- Intervalos en $\mathbb{R} \longrightarrow VC(\mathcal{C}) =$

La dimensión de Vapnik Chervonenkis

Definición

La dimensión de Vapnik-Chervonenkis (o dimensión VC) de una clase de conceptos $\mathcal{C}$ se define como:

$$VC(\mathcal{C}) = \max \{m \; : \; \Pi_{\mathcal{C}}(m) = 2^m\}$$

(es decir, la cardinalidad máxima de un conjunto que es pulverizado por $\mathcal{C}$)

En los ejemplos:

- Intervalos en $\mathbb{R} \longrightarrow VC(\mathcal{C}) = 2$

La dimensión de Vapnik Chervonenkis

Definición

La dimensión de Vapnik-Chervonenkis (o dimensión VC) de una clase de conceptos $\mathcal{C}$ se define como:

$$VC(\mathcal{C}) = \max \{m \,:\, \Pi_{\mathcal{C}}(m) = 2^m\}$$

(es decir, la cardinalidad máxima de un conjunto que es pulverizado por $\mathcal{C}$)

En los ejemplos:

- Intervalos en $\mathbb{R} \longrightarrow VC(\mathcal{C}) = 2$
- Semiplanos en $\mathbb{R}^2$

La dimensión de Vapnik Chervonenkis

Definición

La *dimensión de Vapnik-Chervonenkis (o dimensión VC)* *de una clase de conceptos* $\mathcal{C}$ *se define como:*

$$VC(\mathcal{C}) = \text{máx}\{m \, : \, \Pi_{\mathcal{C}}(m) = 2^m\}$$

(es decir, la cardinalidad máxima de un conjunto que es pulverizado por $\mathcal{C}$*)*

En los ejemplos:

- Intervalos en $\mathbb{R} \longrightarrow VC(\mathcal{C}) = 2$
- Semiplanos en $\mathbb{R}^2 \longrightarrow VC(\mathcal{C}) =$

La dimensión de Vapnik Chervonenkis

Definición

La *dimensión de Vapnik-Chervonenkis (o dimensión VC)* *de una clase de conceptos* $\mathcal{C}$ *se define como:*

$$VC(\mathcal{C}) = \max\{m \,:\, \Pi_{\mathcal{C}}(m) = 2^m\}$$

(es decir, la cardinalidad máxima de un conjunto que es pulverizado por $\mathcal{C}$*)*

En los ejemplos:

- Intervalos en $\mathbb{R} \longrightarrow VC(\mathcal{C}) = 2$
- Semiplanos en $\mathbb{R}^2 \longrightarrow VC(\mathcal{C}) = 3$

La dimensión de Vapnik Chervonenkis

Definición

La dimensión de Vapnik-Chervonenkis (o dimensión VC) de una clase de conceptos $\mathcal{C}$ se define como:

$$VC(\mathcal{C}) = \text{máx}\,\{m \,:\, \Pi_{\mathcal{C}}(m) = 2^m\}$$

(es decir, la cardinalidad máxima de un conjunto que es pulverizado por $\mathcal{C}$)

En los ejemplos:

- Intervalos en $\mathbb{R} \longrightarrow VC(\mathcal{C}) = 2$
- Semiplanos en $\mathbb{R}^2 \longrightarrow VC(\mathcal{C}) = 3$
- Conjuntos finitos en $\mathbb{R}$

La dimensión de Vapnik Chervonenkis

Definición

La *dimensión de Vapnik-Chervonenkis (o dimensión VC)* *de una clase de conceptos* $\mathcal{C}$ *se define como:*

$$VC(\mathcal{C}) = \text{máx}\,\{m\ :\ \Pi_{\mathcal{C}}(m) = 2^m\}$$

(es decir, la cardinalidad máxima de un conjunto que es pulverizado por $\mathcal{C}$*)*

En los ejemplos:

- Intervalos en $\mathbb{R} \longrightarrow VC(\mathcal{C}) = 2$
- Semiplanos en $\mathbb{R}^2 \longrightarrow VC(\mathcal{C}) = 3$
- Conjuntos finitos en $\mathbb{R} \longrightarrow VC(\mathcal{C}) =$

La dimensión de Vapnik Chervonenkis

Definición

La dimensión de Vapnik-Chervonenkis (o dimensión VC) de una clase de conceptos $\mathcal{C}$ se define como:

$$VC(\mathcal{C}) = \text{máx}\,\{m \,:\, \Pi_{\mathcal{C}}(m) = 2^m\}$$

(es decir, la cardinalidad máxima de un conjunto que es pulverizado por $\mathcal{C}$)

En los ejemplos:

- Intervalos en $\mathbb{R} \longrightarrow VC(\mathcal{C}) = 2$
- Semiplanos en $\mathbb{R}^2 \longrightarrow VC(\mathcal{C}) = 3$
- Conjuntos finitos en $\mathbb{R} \longrightarrow VC(\mathcal{C}) = \infty$

Lema

Lema de Sauer

Si $VC(\mathcal{C}) = d < \infty$, *entonces*

$$\Pi_{\mathcal{C}}(m) \leq \left(\frac{em}{d}\right)^d = O(m^d)$$

Aprendibilidad y la dimensión VC

Teorema

Si $h \in \mathcal{H}$ *es consistente, entonces con probabilidad por lo menos* $1 - \delta$

$$e(h) \leq O\left(\frac{\ln \Pi_{\mathcal{C}}(2m) + \ln \frac{1}{\delta}}{m}\right)$$

Aprendibilidad y la dimensión VC

Teorema

Si $h \in \mathcal{H}$ *es consistente, entonces con probabilidad por lo menos* $1 - \delta$

$$e(h) \leq O\left(\frac{\ln \Pi_{\mathcal{C}}(2m) + \ln \frac{1}{\delta}}{m}\right)$$

- Equivalentemente, igualando el lado derecho a ε podemos decir que si h es consistente,

$$\mathbf{P}_{\mathcal{D}}\left[e(h) \geq \varepsilon\right] \leq O\left(\Pi_{\mathcal{C}}(2m)e^{-m\varepsilon/2}\right)$$

Aprendibilidad y la dimensión VC

Teorema

Si $h \in \mathcal{H}$ *es consistente, entonces con probabilidad por lo menos* $1 - \delta$

$$e(h) \leq O\left(\frac{\ln \Pi_{\mathcal{C}}(2m) + \ln \frac{1}{\delta}}{m}\right)$$

- Equivalentemente, igualando el lado derecho a ε podemos decir que si h es consistente,

$$\mathbf{P}_{\mathcal{D}}\left[e(h) \geq \varepsilon\right] \leq O\left(\Pi_{\mathcal{C}}(2m)e^{-m\varepsilon/2}\right)$$

- Qué sucede si $VC(\mathcal{H}) = d < \infty$?

- Si $VC(\mathcal{H}) = d < \infty \Rightarrow \Pi_{\mathcal{C}}(2m) = O\left(m^d\right)$

- Si $VC(\mathcal{H}) = d < \infty \Rightarrow \Pi_{\mathcal{C}}(2m) = O\left(m^d\right)$

$$e(h) \leq O\left(d\frac{\ln m}{m} + \ln\frac{1}{\delta}\right)$$

- Si $VC(\mathcal{H}) = d < \infty \Rightarrow \Pi_{\mathcal{C}}(2m) = O\left(m^d\right)$

$$e(h) \leq O\left(d\frac{\ln m}{m} + \ln\frac{1}{\delta}\right)$$

o

$$\mathbf{P}_{\mathcal{D}}\left[e(h) \geq \varepsilon\right] \leq O\left(m^d e^{-m\varepsilon/2}\right)$$

- Si $VC(\mathcal{H}) = d < \infty \Rightarrow \Pi_{\mathcal{C}}(2m) = O\left(m^d\right)$

$$e(h) \leq O\left(d\frac{\ln m}{m} + \ln\frac{1}{\delta}\right)$$

o

$$\mathbf{P}_{\mathcal{D}}\left[e(h) \geq \varepsilon\right] \leq O\left(m^d e^{-m\varepsilon/2}\right)$$

- Es decir,
dimensión VC finita+Algoritmo consistente=aprendibilidad PAC

Modelo PAC-agnóstico

Modelo PAC-agnóstico

- Espacio de entrada $z = (\mathbf{x}, y) \in \mathcal{Z}$

Modelo PAC-agnóstico

- Espacio de entrada $z = (\mathbf{x}, y) \in \mathcal{Z}$
- Distribución $\mathcal{D}$:

Modelo PAC-agnóstico

- Espacio de entrada $z = (\mathbf{x}, y) \in \mathcal{Z}$
- Distribución $\mathcal{D}$: $(\mathbf{x}, y) \sim \mathcal{D}$

Modelo PAC-agnóstico

- Espacio de entrada $z = (\mathbf{x}, y) \in \mathcal{Z}$
- Distribución $\mathcal{D}$: $(\mathbf{x}, y) \sim \mathcal{D}$
- Datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$, $(\mathbf{x}_i, y_i) \sim \mathcal{D}$, independientes.

Modelo PAC-agnóstico

- Espacio de entrada $z = (\mathbf{x}, y) \in \mathcal{Z}$
- Distribución $\mathcal{D}$: $(\mathbf{x}, y) \sim \mathcal{D}$
- Datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$, $(\mathbf{x}_i, y_i) \sim \mathcal{D}$, independientes.
- No se asume que existe un clasificador a aprender.

Modelo PAC-agnóstico

- Espacio de entrada $z = (\mathbf{x}, y) \in \mathcal{Z}$
- Distribución $\mathcal{D}$: $(\mathbf{x}, y) \sim \mathcal{D}$
- Datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$, $(\mathbf{x}_i, y_i) \sim \mathcal{D}$, independientes.
- No se asume que existe un clasificador a aprender.
- Clase de hipótesis $\mathcal{H}$.

Modelo PAC-agnóstico

- Espacio de entrada $z = (\mathbf{x}, y) \in \mathcal{Z}$
- Distribución $\mathcal{D}$: $(\mathbf{x}, y) \sim \mathcal{D}$
- Datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$, $(\mathbf{x}_i, y_i) \sim \mathcal{D}$, independientes.
- No se asume que existe un clasificador a aprender.
- Clase de hipótesis $\mathcal{H}$.
- Algoritmo retorna hipótesis $h \in \mathcal{H}$ a partir de los datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$

Modelo PAC-agnóstico

- Espacio de entrada $z = (\mathbf{x}, y) \in \mathcal{Z}$
- Distribución $\mathcal{D}$: $(\mathbf{x}, y) \sim \mathcal{D}$
- Datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$, $(\mathbf{x}_i, y_i) \sim \mathcal{D}$, independientes.
- No se asume que existe un clasificador a aprender.
- Clase de hipótesis $\mathcal{H}$.
- Algoritmo retorna hipótesis $h \in \mathcal{H}$ a partir de los datos $\{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$
- Criterio de error:

$$e(h) = \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}) \neq y\right]$$

Ejemplos

Ejemplos

- $\mathbf{x} \sim \mathcal{D}$ y $y = c(\mathbf{x})$ para algún concepto c desconocido.

Ejemplos

- $\mathbf{x} \sim \mathcal{D}$ y $y = c(\mathbf{x})$ para algún concepto c desconocido.
- $\mathbf{x} \sim \mathcal{D}$ y $y = c(\mathbf{x}) \oplus \eta$ donde $\eta \in \{1, 0\}$ es ruido de clasificación con $\mathbb{P}(\eta = 1) = p$.

Ejemplos

- $\mathbf{x} \sim \mathcal{D}$ y $y = c(\mathbf{x})$ para algún concepto c desconocido.
- $\mathbf{x} \sim \mathcal{D}$ y $y = c(\mathbf{x}) \oplus \eta$ donde $\eta \in \{1, 0\}$ es ruido de clasificación con $\mathbb{P}(\eta = 1) = p$.
- $\mathbf{x} \sim \mathcal{D}$ y $\mathbb{P}(y = 1 \mid \mathbf{x}) = \alpha(\mathbf{x})$.

Ejemplos

- $\mathbf{x} \sim \mathcal{D}$ y $y = c(\mathbf{x})$ para algún concepto c desconocido.
- $\mathbf{x} \sim \mathcal{D}$ y $y = c(\mathbf{x}) \oplus \eta$ donde $\eta \in \{1, 0\}$ es ruido de clasificación con $\mathbb{P}(\eta = 1) = p$.
- $\mathbf{x} \sim \mathcal{D}$ y $\mathbb{P}(y = 1 \mid \mathbf{x}) = \alpha(\mathbf{x})$.

Aprendibilidad Agnóstica

- En general no es posible lograr $e(h) = 0$.

Aprendibilidad Agnóstica

- En general no es posible lograr $e(h) = 0$.
- Comparamos con el error de la mejor hipótesis:

$$e^* = \inf_{h \in \mathcal{H}} e(h)$$

Aprendibilidad Agnóstica

- En general no es posible lograr $e(h) = 0$.
- Comparamos con el error de la mejor hipótesis:

$$e^* = \inf_{h \in \mathcal{H}} e(h)$$

- Aprendibilidad: $\forall \delta, \varepsilon$, algoritmo retorna h con

$$\mathbf{P}_{\mathcal{D}} \left[e(h) < e^* + \varepsilon \right] \geq 1 - \delta$$

Minimización de riesgo empírico (ERM)

Minimización de riesgo empírico (ERM)

- Algoritmo: minimizar error en los datos:

$$\hat{e}(h) = \frac{1}{m}\sum_{i=1}^{m} I_{\{h(\mathbf{x}_i)\neq y_i\}}$$

Minimización de riesgo empírico (ERM)

- Algoritmo: minimizar error en los datos:

$$\hat{e}(h) = \frac{1}{m}\sum_{i=1}^{m} I_{\{h(\mathbf{x}_i)\neq y_i\}}$$

- En la práctica ERM puede ser difícil computacionalmente (e.g. NP completo).

Minimización de riesgo empírico (ERM)

- Algoritmo: minimizar error en los datos:

$$\hat{e}(h) = \frac{1}{m}\sum_{i=1}^{m} I_{\{h(\mathbf{x}_i)\neq y_i\}}$$

- En la práctica ERM puede ser difícil computacionalmente (e.g. NP completo).
- Función de error no derivable,

Minimización de riesgo empírico (ERM)

- Algoritmo: minimizar error en los datos:

$$\hat{e}(h) = \frac{1}{m} \sum_{i=1}^{m} I_{\{h(\mathbf{x}_i) \neq y_i\}}$$

- En la práctica ERM puede ser difícil computacionalmente (e.g. NP completo).
- Función de error no derivable, en la práctica se usa función de error sustituta.

Minimización de riesgo empírico (ERM)

- Algoritmo: minimizar error en los datos:

$$\hat{e}(h) = \frac{1}{m} \sum_{i=1}^{m} I_{\{h(\mathbf{x}_i) \neq y_i\}}$$

- En la práctica ERM puede ser difícil computacionalmente (e.g. NP completo).
- Función de error no derivable, en la práctica se usa función de error sustituta.
- Se debe controlar complejidad de la clase de hipótesis $\mathcal{H}$.

Clases de hipótesis finitas

Clases de hipótesis finitas

- En este caso, existe $h^* \in \mathcal{H}$ con

$$e(h^*) = e^* = \min_{h \in \mathcal{H}} e(h)$$

Clases de hipótesis finitas

- En este caso, existe $h^* \in \mathcal{H}$ con

$$e(h^*) = e^* = \min_{h \in \mathcal{H}} e(h)$$

Teorema

Si $|\mathcal{H}| < \infty$, y existe un algoritmo A de minimización de riesgo empírico, $\mathcal{H}$ es PAC-agnóstico aprendible

Clases de hipótesis finitas

- En este caso, existe $h^* \in \mathcal{H}$ con

$$e(h^*) = e^* = \min_{h \in \mathcal{H}} e(h)$$

Teorema

Si $|\mathcal{H}| < \infty$, y existe un algoritmo A de minimización de riesgo empírico, $\mathcal{H}$ es PAC-agnóstico aprendible

- Clave: con alta probabilidad (sobre los datos de entrenamiento) $e(h)$ debe estar cercano a $\hat{e}(h)$:

Clases de hipótesis finitas

- En este caso, existe $h^* \in \mathcal{H}$ con

$$e(h^*) = e^* = \min_{h \in \mathcal{H}} e(h)$$

Teorema

Si $|\mathcal{H}| < \infty$, y existe un algoritmo A de minimización de riesgo empírico, $\mathcal{H}$ es PAC-agnóstico aprendible

- Clave: con alta probabilidad (sobre los datos de entrenamiento) $e(h)$ debe estar cercano a $\hat{e}(h)$:

$$|\hat{e}(h) - e(h)| \leq \varepsilon$$

uniformemente sobre $\mathcal{H}$

Clases de hipótesis finitas

- En este caso, existe $h^* \in \mathcal{H}$ con

$$e(h^*) = e^* = \min_{h \in \mathcal{H}} e(h)$$

Teorema

Si $|\mathcal{H}| < \infty$, y existe un algoritmo A de minimización de riesgo empírico, $\mathcal{H}$ es PAC-agnóstico aprendible

- Clave: con alta probabilidad (sobre los datos de entrenamiento) $e(h)$ debe estar cercano a $\hat{e}(h)$:

$$|\hat{e}(h) - e(h)| \leq \varepsilon$$

uniformemente sobre $\mathcal{H}$

- Si A selecciona h con error empírico pequeño, con alta probabilidad $e(h)$ será pequeño.

Lema

Para $h \in \mathcal{H}$, $\mathbf{P}_{\mathcal{D}}\left[|\hat{e}(h) - e(h)| \geq \varepsilon\right] \leq 2e^{-2\varepsilon^2 m}$

Lema

Para $h \in \mathcal{H}$, $\mathbf{P}_{\mathcal{D}}\left[|\hat{e}(h) - e(h)| \geq \varepsilon\right] \leq 2e^{-2\varepsilon^2 m}$

Demostración.

Lema

Para $h \in \mathcal{H}$, $\mathbf{P}_{\mathcal{D}}\left[|\hat{e}(h) - e(h)| \geq \varepsilon\right] \leq 2e^{-2\varepsilon^2 m}$

Demostración.

Chernoff □

Lema

Para $h \in \mathcal{H}$, $\mathbf{P}_{\mathcal{D}}\left[|\hat{e}(h) - e(h)| \geq \varepsilon\right] \leq 2e^{-2\varepsilon^2 m}$

Demostración.

Chernoff □

Lema

$$\mathbf{P}_{\mathcal{D}}\left[\max_{h\in\mathcal{H}} |\hat{e}(h) - e(h)| \geq \varepsilon\right] \leq 2\,|\mathcal{H}|\,e^{-2\varepsilon^2 m}$$

Lema

Para $h \in \mathcal{H}$, $\mathbf{P}_{\mathcal{D}}\left[|\hat{e}(h) - e(h)| \geq \varepsilon\right] \leq 2e^{-2\varepsilon^2 m}$

Demostración.

Chernoff □

Lema

$$\mathbf{P}_{\mathcal{D}}\left[\max_{h\in\mathcal{H}} |\hat{e}(h) - e(h)| \geq \varepsilon\right] \leq 2\,|\mathcal{H}|\, e^{-2\varepsilon^2 m}$$

Demostración.

$$\mathbf{P}_{\mathcal{D}}\left[\max_{h\in\mathcal{H}} |\hat{e}(h) - e(h)| \geq \varepsilon\right] = \mathbf{P}_{\mathcal{D}}\left[\bigcup_{h\in\mathcal{H}} \{z \in \mathcal{Z} \,:\, |\hat{e}(h) - e(h)| \geq \varepsilon\}\right]$$

Lema

Para $h \in \mathcal{H}$, $\mathbf{P}_{\mathcal{D}}\left[|\hat{e}(h) - e(h)| \geq \varepsilon\right] \leq 2e^{-2\varepsilon^2 m}$

Demostración.

Chernoff □

Lema

$$\mathbf{P}_{\mathcal{D}}\left[\max_{h\in\mathcal{H}} |\hat{e}(h) - e(h)| \geq \varepsilon\right] \leq 2\,|\mathcal{H}|\,e^{-2\varepsilon^2 m}$$

Demostración.

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[\max_{h\in\mathcal{H}} |\hat{e}(h) - e(h)| \geq \varepsilon\right] &= \mathbf{P}_{\mathcal{D}}\left[\bigcup_{h\in\mathcal{H}} \{z \in \mathcal{Z} \,:\, |\hat{e}(h) - e(h)| \geq \varepsilon\}\right] \\ &\leq \sum_{h\in\mathcal{H}} \mathbf{P}_{\mathcal{D}}\left[|\hat{e}(h) - e(h)| \geq \varepsilon\right]\end{aligned}$$

Lema

Para $h \in \mathcal{H}$, $\mathbf{P}_{\mathcal{D}}\left[|\hat{e}(h) - e(h)| \geq \varepsilon\right] \leq 2e^{-2\varepsilon^2 m}$

Demostración.

Chernoff □

Lema

$$\mathbf{P}_{\mathcal{D}}\left[\max_{h\in\mathcal{H}} |\hat{e}(h) - e(h)| \geq \varepsilon\right] \leq 2\,|\mathcal{H}|\,e^{-2\varepsilon^2 m}$$

Demostración.

$$\begin{aligned}
\mathbf{P}_{\mathcal{D}}\left[\max_{h\in\mathcal{H}} |\hat{e}(h) - e(h)| \geq \varepsilon\right] &= \mathbf{P}_{\mathcal{D}}\left[\bigcup_{h\in\mathcal{H}} \{z \in \mathcal{Z} \,:\, |\hat{e}(h) - e(h)| \geq \varepsilon\}\right] \\
&\leq \sum_{h\in\mathcal{H}} \mathbf{P}_{\mathcal{D}}\left[|\hat{e}(h) - e(h)| \geq \varepsilon\right] \\
&\leq 2\,|\mathcal{H}|\,e^{-2\varepsilon^2 m}
\end{aligned}$$

Demostración (teorema)

Demostración (teorema)

- $$\mathbf{P}_{\mathcal{D}}\left[\max_{h\in\mathcal{H}}\left|\hat{e}(h)-e(h)\right|\geq\varepsilon\right]\leq 2\left|\mathcal{H}\right|e^{-2\varepsilon^2 m}\leq\delta$$

Demostración (teorema)

- $$\mathbf{P}_{\mathcal{D}}\left[\max_{h\in\mathcal{H}}|\hat{e}(h)-e(h)|\geq\varepsilon\right]\leq 2\,|\mathcal{H}|\,e^{-2\varepsilon^2 m}\leq\delta$$

  si $\varepsilon\geq\left(\frac{1}{2m}\ln\frac{2|\mathcal{H}|}{\delta}\right)^{1/2}$

Demostración (teorema)

- $$\mathbf{P}_{\mathcal{D}}\left[\max_{h\in\mathcal{H}}|\hat{e}(h)-e(h)|\geq\varepsilon\right]\leq 2\,|\mathcal{H}|\,e^{-2\varepsilon^2 m}\leq\delta$$

  si $\varepsilon\geq\left(\frac{1}{2m}\ln\frac{2|\mathcal{H}|}{\delta}\right)^{1/2}$
- Es decir, con probabilidad por lo menos $1-\delta$, para la $h\in\mathcal{H}$ que retorna A:

Demostración (teorema)

- $$\mathbf{P}_{\mathcal{D}}\left[\max_{h\in\mathcal{H}}|\hat{e}(h)-e(h)|\geq\varepsilon\right]\leq 2\,|\mathcal{H}|\,e^{-2\varepsilon^2 m}\leq\delta$$

  si $\varepsilon \geq \left(\frac{1}{2m}\ln\frac{2|\mathcal{H}|}{\delta}\right)^{1/2}$
- Es decir, con probabilidad por lo menos $1-\delta$, para la $h\in\mathcal{H}$ que retorna A:

$$e(h)\leq\hat{e}(h)+\varepsilon$$

Demostración (teorema)

- $$\mathbf{P}_{\mathcal{D}}\left[\max_{h\in\mathcal{H}}|\hat{e}(h)-e(h)|\geq\varepsilon\right]\leq 2\,|\mathcal{H}|\,e^{-2\varepsilon^2 m}\leq\delta$$

  si $\varepsilon \geq \left(\frac{1}{2m}\ln\frac{2|\mathcal{H}|}{\delta}\right)^{1/2}$
- Es decir, con probabilidad por lo menos $1-\delta$, para la $h\in\mathcal{H}$ que retorna A:

$$\begin{aligned} e(h) &\leq \hat{e}(h)+\varepsilon \\ &\leq \hat{e}(h^*)+\varepsilon \end{aligned}$$

Demostración (teorema)

- $$\mathbf{P}_{\mathcal{D}}\left[\max_{h\in\mathcal{H}}|\hat{e}(h)-e(h)|\geq\varepsilon\right]\leq 2\,|\mathcal{H}|\,e^{-2\varepsilon^2 m}\leq\delta$$

  si $\varepsilon \geq \left(\frac{1}{2m}\ln\frac{2|\mathcal{H}|}{\delta}\right)^{1/2}$
- Es decir, con probabilidad por lo menos $1-\delta$, para la $h\in\mathcal{H}$ que retorna A:

$$\begin{aligned} e(h) &\leq \hat{e}(h)+\varepsilon \\ &\leq \hat{e}(h^*)+\varepsilon \\ &\leq (e(h^*)+\varepsilon)+\varepsilon \end{aligned}$$

Demostración (teorema)

- $$\mathbf{P}_{\mathcal{D}}\left[\max_{h\in\mathcal{H}}|\hat{e}(h)-e(h)|\geq\varepsilon\right]\leq 2\,|\mathcal{H}|\,e^{-2\varepsilon^2 m}\leq\delta$$

  si $\varepsilon \geq \left(\frac{1}{2m}\ln\frac{2|\mathcal{H}|}{\delta}\right)^{1/2}$
- Es decir, con probabilidad por lo menos $1-\delta$, para la $h\in\mathcal{H}$ que retorna A:

$$\begin{aligned} e(h) &\leq \hat{e}(h)+\varepsilon \\ &\leq \hat{e}(h^*)+\varepsilon \\ &\leq (e(h^*)+\varepsilon)+\varepsilon = e(h^*)+2\varepsilon \end{aligned}$$

Demostración (teorema)

- $$\mathbf{P}_{\mathcal{D}}\left[\max_{h\in\mathcal{H}}|\hat{e}(h)-e(h)|\geq\varepsilon\right]\leq 2\,|\mathcal{H}|\,e^{-2\varepsilon^2 m}\leq\delta$$

  si $\varepsilon\geq\left(\frac{1}{2m}\ln\frac{2|\mathcal{H}|}{\delta}\right)^{1/2}$
- Es decir, con probabilidad por lo menos $1-\delta$, para la $h\in\mathcal{H}$ que retorna A:

$$\begin{aligned} e(h) &\leq \hat{e}(h)+\varepsilon \\ &\leq \hat{e}(h^*)+\varepsilon \\ &\leq (e(h^*)+\varepsilon)+\varepsilon = e(h^*)+2\varepsilon \\ &\leq e^*+\left(\frac{2}{m}\ln\frac{2\,|\mathcal{H}|}{\delta}\right)^{1/2} \end{aligned}$$

Demostración (teorema)

- $$\mathbf{P}_{\mathcal{D}}\left[\max_{h\in\mathcal{H}} |\hat{e}(h) - e(h)| \geq \varepsilon\right] \leq 2\,|\mathcal{H}|\,e^{-2\varepsilon^2 m} \leq \delta$$

  si $\varepsilon \geq \left(\frac{1}{2m}\ln\frac{2|\mathcal{H}|}{\delta}\right)^{1/2}$
- Es decir, con probabilidad por lo menos $1-\delta$, para la $h \in \mathcal{H}$ que retorna A:

$$\begin{aligned}
e(h) &\leq \hat{e}(h) + \varepsilon \\
&\leq \hat{e}(h^*) + \varepsilon \\
&\leq (e(h^*) + \varepsilon) + \varepsilon = e(h^*) + 2\varepsilon \\
&\leq e^* + \left(\frac{2}{m}\ln\frac{2\,|\mathcal{H}|}{\delta}\right)^{1/2}
\end{aligned}$$

- o, dados ε, δ,

$$m \geq \frac{2}{\varepsilon^2}\ln\frac{2\,|\mathcal{H}|}{\delta}$$

Clases de hipótesis infinitas

Clases de hipótesis infinitas

- Medida de complejidad es la dimensión VC.

Clases de hipótesis infinitas

- Medida de complejidad es la dimensión VC.
- Dimensión VC finita + Algoritmo ERM = Aprendibilidad PAC-Agnóstica.

Clases de hipótesis infinitas

- Medida de complejidad es la dimensión VC.
- Dimensión VC finita + Algoritmo ERM = Aprendibilidad PAC-Agnóstica.
- Clave: Convergencia uniforme de riesgo empírico a probabilidad de error.

Clases de hipótesis infinitas

- Medida de complejidad es la dimensión VC.
- Dimensión VC finita + Algoritmo ERM = Aprendibilidad PAC-Agnóstica.
- Clave: Convergencia uniforme de riesgo empírico a probabilidad de error.

Teorema

$$\mathbf{P}_{\mathcal{D}}\left[\sup_{h\in\mathcal{H}} |\hat{e}(h) - e(h)| \geq \varepsilon\right] \leq 4\Pi_{\mathcal{C}}(2m)e^{-\varepsilon^2 m/8}$$

Convergencia uniforme ⇒ Aprendibilidad (bosquejo)

Convergencia uniforme ⇒ Aprendibilidad (bosquejo)

- Si $|\hat{e}(h) - e(h)| \leq \varepsilon$ con probabilidad por lo menos $1 - \delta$, por un razonamiento similar al del caso finito, podemos decir (con probabilidad $\geq 1 - \delta$
$$e(h) \leq e^* + 2\varepsilon$$

Convergencia uniforme ⇒ Aprendibilidad (bosquejo)

- Si $|\hat{e}(h) - e(h)| \leq \varepsilon$ con probabilidad por lo menos $1-\delta$, por un razonamiento similar al del caso finito, podemos decir (con probabilidad $\geq 1-\delta$

$$e(h) \leq e^* + 2\varepsilon$$

- Por el teorema de convergencia uniforme, esto es cierto (con probabilidad $\geq 1-\delta$) si

$$4\Pi_{\mathcal{C}}(2m)e^{-\varepsilon^2 m/8} \leq \delta$$

Convergencia uniforme ⇒ Aprendibilidad (bosquejo)

- Si $|\hat{e}(h) - e(h)| \leq \varepsilon$ con probabilidad por lo menos $1 - \delta$, por un razonamiento similar al del caso finito, podemos decir (con probabilidad $\geq 1 - \delta$

$$e(h) \leq e^* + 2\varepsilon$$

- Por el teorema de convergencia uniforme, esto es cierto (con probabilidad $\geq 1 - \delta$) si

$$4\Pi_{\mathcal{C}}(2m)e^{-\varepsilon^2 m/8} \leq \delta$$

- Si $VC(\mathcal{H}) = d < \infty$, sabemos que $\Pi_{\mathcal{C}}(m) = O\left(m^d\right)$.

Convergencia uniforme $\Rightarrow$ Aprendibilidad (bosquejo)

- Si $|\hat{e}(h) - e(h)| \leq \varepsilon$ con probabilidad por lo menos $1-\delta$, por un razonamiento similar al del caso finito, podemos decir (con probabilidad $\geq 1-\delta$

$$e(h) \leq e^* + 2\varepsilon$$

- Por el teorema de convergencia uniforme, esto es cierto (con probabilidad $\geq 1-\delta$) si

$$4\Pi_{\mathcal{C}}(2m)e^{-\varepsilon^2 m/8} \leq \delta$$

- Si $VC(\mathcal{H}) = d < \infty$, sabemos que $\Pi_{\mathcal{C}}(m) = O\left(m^d\right)$.
- Esto quiere decir que (con probabilidad $\geq 1-\delta$)

$$e(h) \leq e^* + \left(\frac{32}{m}\left(d\ln\frac{2em}{d} + \ln\frac{4}{\delta}\right)\right)^{1/2}$$

Despejando para m, es suficiente tener:

$$m \geq \frac{64}{\varepsilon^2}\left(2d\ln\frac{12}{\varepsilon} + \ln\frac{4}{\delta}\right)$$