Clase 9: Dimensión VC

Responsable: Pablo Martínez Medina

EST-25134, Primavera 2021
Dr. Alfredo Garbuno Iñigo
Febrero 16, 2021

1. Introducción

Una vez relacionado el Teorema NFL con conocimiento previo es indispensable preguntarse ¿Qué clases de hipótesis son PAC aprendibles? ¿Qué es lo que hace que una clase sea aprendible mientras que otras no lo sean? ¿En verdad todas las clases finitas son aprendibles? ¿Cómo podemos definir de las clases finitas su complejidad muestral? En esta sesión podremos tener una respuesta a las mismas utilizando la clasificación dada por la función de pérdida 0-1.

2. Clases Infinitas $\mathcal{H}$

Anteriomente se vió que las clases finitas son aprendibles y que el grado de complejidad muestral de una clase de hipótesis está acotada superiormente por el logaritmo de su tamaño. Para ejemplificar que el tamaño de las clases de hipótesis no es la manera correcta de caracterizar la complejidad muestral, se presenta el siguiente ejemplo:

Example 2.1. Sea $\mathcal{H}_a = \{h_a : a \in \mathbb{R}\}$, definiendo a la función h_a como:

$$h_a(x) = \begin{cases} 1 & si \quad x \leq a \\ \\ 0 & en \quad c.o.c. \end{cases}$$

Vamos a ver que $\mathcal{H}_a$ es aprendible si usamos ERM y que tiene una complejidad muestral tal que

$$m_{\mathcal{H}}(\varepsilon, \delta) \leq \left\lceil \frac{log(\frac{2}{\delta})}{\varepsilon} \right\rceil.$$

Solución: Sea a^* la constante de $h_{a^*}(x) = \mathbb{1}_{[x \leq a^*]}$ con $\mathcal{L}_{\mathcal{D}} = 0$, con $\mathcal{D}_x$ la distribución marginal de x y sean a_0, a_1 tal que para $a^* \in (a_0, a_1)$ se tiene que $\mathbb{P}(x \in (a_0, a^*)) = \mathbb{P}(x \in (a^*, a_1)) = \varepsilon$. Esto se puede ver reflejado en el lado izquierdo de la Figura 1.

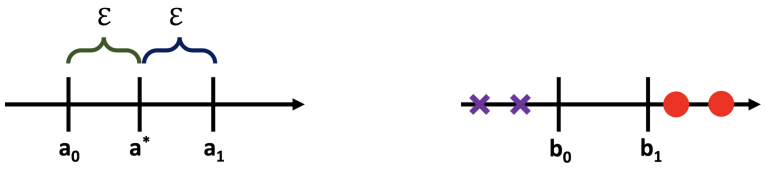

Figura 1 Figuras complementarias al texto. Figura de la recta numérica [izquierda] que muestra que la a^* se encuentra a la mitad del intervalo (a_0, a_1). Figura de la recta numérica [derecha] que muestra el comportamiento del conjunto S

Ahora, se tiene a S, un conjunto de entrenamiento tal que $b_0 = max\{x : (x, 1) \in S\}$ y $b_1 = min\{x : (x, 0) \in S\}$. Esto se puede ver reflejado en el lado derecho de la Figura 1. Gracias al ERM, definamos a $h_S = h_{b_S}$ tal que $b_S \in (b_0, b_1)$. Tenemos que $\mathcal{L}(h_S) \leq \varepsilon$ es suficiente para $b_0 \geq a_0$ y $b_1 \leq a_1$, ahora analizando el complemento, tenemos lo siguiente:

$$\mathbb{P}_{S\sim\mathcal{D}^m}(\mathcal{L}_{\mathcal{D}}(h_S) > \varepsilon) \leq \mathbb{P}_{S\sim\mathcal{D}^m}(b_0 < a_0 \cup b_1 > a_1) \leq \mathbb{P}_S(b_0 < a_0) + \mathbb{P}_S(b_1 > a_1).$$

Cabe mencionar que el evento $b_0 < a_0$ ocurre si todos los ejemplos en S no se encuentran en (a_0, a^*), por lo tanto como $m > \frac{log(\frac{2}{\delta})}{\varepsilon}$, se cumple lo siguiente:

$$\mathbb{P}(b_0 < a_0) = \mathbb{P}(\forall(x, y) \in S, x \notin (a_0, a^*)) = (1 - \varepsilon)^m \leq e^{-\varepsilon m} \leq \frac{\delta}{2}.$$

Es análogo para el caso de $b_1 > a_1$ en donde se obtiene que $\mathbb{P}(b_1 > a_1) \leq \frac{\delta}{2}$.

Nota: Esta solución nos muestra que entonces sí podemos aprender cuando tenemos un espacio infinito de funciones gracias al ERM.

3. Dimensión VC: Vapnik-Chervonenkis

Hasta ahorita, hemos visto que si tenemos clases finitas, entonces tenemos capacidad de aprendizaje; por otro lado, que podemos utilizar el Teorema NFL para las ocasiones en las que restringimos el tamaño de $\mathcal{H}$. Ahora lo que nos interesa conocer es cómo se comporta $\mathcal{H}$ en $\mathbb{C}$ si no lo restringimos, además, es importante mencionar que puede suceder que el adversario escoja, dentro de la familia de distribuciones, un agente que nos perjudique para que no sea PAC aprendible.

Definition 3.1 (Restricción de $\mathcal{H}$ en $\mathbb{C}$)**.** Sea $\mathcal{H}$ una familia de hipótesis donde h: $\chi \to \{0, 1\}$ y sea $\mathbb{C} = \{C_1, ..., C_m\} \subset \chi$. La restricción de $\mathcal{H}$ en $\mathbb{C}$ tal que,

$$\mathcal{H}_{\mathbb{C}} = \{(h(C_1), ..., h(C_m)) : h \in \mathcal{H}, C_j \in \mathbb{C}\}.$$

Definition 3.2 (Fragmentación)**.** Decimos que la clase de funciones $\mathcal{H}$ fragmenta un conjunto finito $\mathbb{C} \subset \chi$ si $\mathcal{H}_{\mathbb{C}}$ considera todas las funciones $\mathbb{C}$ a $\{0, 1\}$ tal que $\mid \mathcal{H}_{\mathbb{C}} \mid = 2^{|\mathbb{C}|}$.

Example 3.3. Tomando la definición de $\mathcal{H}_a$ (la familia de funciones indicadoras) y definiendo a $\mathbb{C} = \{C_1\}$ con $C_1 \in \mathbb{R}$, tenemos que:

1. $a = C_1 + 1 \implies h_a(C_1) = 1$
2. $a = C_1 - 1 \implies h_a(C_1) = 0$

Podemos observar que $(\mathcal{H}_a)_{\mathbb{C}}$ incluye todas las posibles funciones de $\mathbb{C} \to \{0, 1\}$, por lo tanto, $\mathcal{H}$ fragmenta conjuntos de cardinalidad igual a 1.

Ahora, definiendo $\mathbb{C} = \{C_1, C_2\}$ con $C_1 \leq C_2$, vemos que si quisiéramos etiquetar a C_1 y C_2 de tal manera que obtengamos 0 y/o 1, vemos que no se puede debido a la restricción $C_1 \leq C_2$, por lo tanto, $\mathcal{H}_a$ no fragmenta a $\mathbb{C}$ con cardinalidad igual a 2.

Corollary 3.4. *Sea $\mathcal{H}$ una familia de hipótesis $\chi \to \{0,1\}$, S conjunto de entrenamiento con m observaciones. Suponemos que existe $\mathbb{C} \subset \chi, \mid \mathbb{C} \mid = 2m$ y que es fragmentado por $\mathcal{H}$. Entonces $\forall$ A $\exists$ una distribución $\mathcal{D}$ sobre $\chi \times \{0,1\}$ y un predictor $h \in \mathcal{H}$ tal que $\mathcal{L}_{\mathcal{D}}(h) = 0$, pero con probabilidad $\geq \frac{1}{7}$ tenemos que $\mathcal{L}_{\mathcal{D}}(A(S)) \geq \frac{1}{8}$.*

Definition 3.5 (Dimensión Vapnik-Chervonenkis)**.** La dimensión VC de $\mathcal{H}$, $VCdim(\mathcal{H})$, es el tamaño más grande de $\mathbb{C}$ que puede ser fragmentado por $\mathcal{H}$.

Remark 3.6. Si $\mathcal{H}$ puede fragmentar cualquier $\mathbb{C}$ $\implies$ $VCdim(\mathcal{H}) = \infty$.

Teorema 3.7. *Sea $\mathcal{H}$, $VCdim(\mathcal{H}) = \infty \implies \mathcal{H}$ no es PAC.*

4. Ejemplos de dimensión VC

Considerando que $VCdim(\mathcal{H}) = d$, tenemos que:

1. Mostrar que un conjunto $\mathbb{C}$, $\mid \mathbb{C} \mid = d$, es fragmentado por $\mathcal{H}$.
2. Mostrar que cualquier $\mathbb{C}$, $\mid \mathbb{C} \mid = d + 1$, no es fragmentado por $\mathcal{H}$.

4.1. Indicadores

Caso 1: Sea $\mathbb{C} = \{C_1\}$ ——————————————— Sí procede
Caso 2: Sea $\mathbb{C} = \{C_1, C_2\}$ ——————————————— No procede

Por lo tanto, $VCdim(\mathcal{H}) = 1$.

4.2. Intervalos

Sea $\mathcal{H}_{a,b} = \{h_{a,b} : a, b \in \mathbb{R}, a < b\}$ tal que $h_{a,b}(x) = \mathbb{1}_{[x \in (a,b)]}$.

Caso 1: Sea $\mathbb{C} = \{C_1, C_2\}$ ——————————————— Sí procede
Caso 2: Sea $\mathbb{C} = \{C_1, C_2, C_3\}$ ——————————————— No procede

Vemos que sin pérdida de generalidad, considerando $C_1 \leq C_2 \leq C_3$, se presenta un problema ya que no se puede etiquetar con solo 0 y 1, por ejemplo, no se puede $1 \leq 0 \leq 1$. Por lo tanto $VCdim(\mathcal{H}_{a,b}) = 2$.

4.3. Rectángulos

Sea $\mathcal{H} = \{h_\theta : \theta = (a_1, a_2, b_1, b_2), a_1 \leq a_2, b_1 \leq b_2\}$.

$$h_\theta(x_1, x_2) = \begin{cases} 1 & si \quad x_1 \in (a_1, a_2), x_2 \in (b_1, b_2) \\ 0 & en \qquad\qquad\quad c.o.c. \end{cases}$$

En la Figura 2 podemos observar que $VCdim(\mathcal{H}_{a,b}) = 4$. Del lado izquierdo está el caso de 4 puntos (que Sí procede), mientras que del lado derecho es el caso de 5 puntos (que No procede).

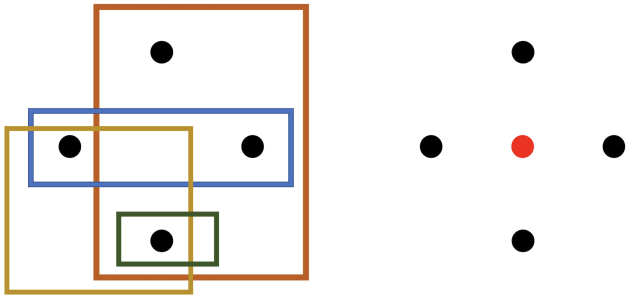

Figura 2 Figuras de casos para rectángulos. Figura de etiquetados de puntos [izquierda] que se muestra que sí pueden ser etiquetados sin ningún problema. Figura de etiquetados [derecha] que muestra un problema porque si suponemos que el punto rojo es negativo, no hay manera de poder etiquetarlos correctamente.

4.4. Clases finitas

Sea $\mathcal{H}$, clase finita, tal que $\forall\ \mathbb{C} \subset |\ \mathcal{H}_\mathbb{C}\ | \leq |\ \mathcal{H}\ |$ y por lo tanto $\mathbb{C}$ no es fragmentado por $\mathcal{H}$ si $|\ \mathcal{H}\ | < 2^{|\mathbb{C}|} \implies VCdim(\mathcal{H}) \leq log_2(|\ \mathcal{H}\ |)$, es decir, está acotado, sin embargo, dicha cota en general se considera como mala, ya que la $VCdim(\mathcal{H})$ de una clase de hipótesis finita puede ser significativamente más pequeña que $log_2(|\ \mathcal{H}\ |)$.

5. El Teorema Fundamental de Aprendizaje Estadístico

Teorema 5.1 (Teorema Fundamental de Aprendizaje Estadístico)**.** *Sea $\mathcal{H}$ tal que $h \in \mathcal{H}$ $h{:}\chi \to \{0, 1\}$ y consideramos la pérdida 0-1. Entonces los siguientes enunciados son equivalentes:*

1. *$\mathcal{H}$ tiene la propiedad de convergencia uniforme.*
2. *ERM es capaz de generar aprendizaje PAC agnóstico con $\mathcal{H}$.*
3. *$\mathcal{H}$ es PAC agnóstico.*
4. *$\mathcal{H}$ es PAC.*
5. *ERM es capaz de generar aprendizaje PAC.*
6. *$\mathcal{H}$ tiene $VCdim(\mathcal{H}) < \infty$.*