Clase 8: Compromiso entre sesgo y varianza

Responsable: Santiago Ruiz Velasco Fernández

EST-25134, Primavera 2021
Profesor: Dr. Alfredo Garbuno Iñigo
Febrero 9, 2021

1. Introducción

Como hemos estudiado a través de los modelos lineales, para contrarrestar el riesgo de sobreajuste, una alternativa es utilizar una familia de hipótesis $\mathcal{H}$ más restringida, asegurando que se encuentre todo nuestro conocimiento previo. Sin embargo, este conocimiento previo puede causar un sesgo en el aprendizaje. Ahora, para evitar que nuestro aprendizaje esté sesgado, lo que esperaríamos es que nuestra familia de hipótesis no posea sesgo y que sea universalmente aplicable.

Recordemos que una tarea de aprendizaje se define a través de la distribución D definida en $\mathcal{X}$x$\mathcal{Y}$, con $\mathcal{X}$ conjunto de atributos y $\mathcal{Y}$ variables objetivo donde buscamos alguna h: $\mathcal{X} \to \mathcal{Y}$ de tal forma que el error de generalización o el riesgo teórico $L_D(h)$ sea mínimo.

¿Existe algún algoritmo de aprendizaje A y un conjunto de entrenamiento S con $|S| = m$ tal que para toda distribución D, si A recibe las m observaciones independientes e idénticamente distribuidas en D, entonces con alta probabilidad lograremos encontrar esa hipótesis h: $\mathcal{X} \to \mathcal{Y}$ tal que el error de generalización $L_D(h) \ll \varepsilon$? Si se puede contestar afirmativamente esta pregunta, entonces sabremos que existe la familia de hipótesis $\mathcal{H}$ insesgada y universalmente aplicable

2. Teorema: No Free Lunch (NFL)

2.1. Explicación

El nombre de este teorema nos da a entender un poco mejor el contexto de un problema de clasificación con un ejemplo simple: así como no es posible tener un almuerzo que sea completamente gratuito, no existe ninguna distribución donde nuestra mejor hipótesis candidata esté libre de fallos, es decir, para alguna distribución, no importa que tengamos a un modelo candidato que nosotros consideremos el "mejor"teóricamente, este va a fallar. Por este teorema podemos concluir que no existe una familia de hipótesis $\mathcal{H}$ que sea universalmente aplicable.

2.2. Descomposición del error cometido por ERM

Esta descomposición se analizará en dos partes:

1. Calidad de $\mathcal{H}$: Habla del posible sesgo que pueda presentar el aprendizaje.
2. Riesgo de sobreajuste: Habla sobre la varianza que presenta el aprendizaje al observar distintos conjuntos de datos.

La idea principal es que dependiendo del modelo que nosotros utilicemos, el error cometido por el ERM presentará diferentes problemas:

Si nuestro modelo es altamente complejo, este emitirá un aprendizaje con menor sesgo, pero con una mayor varianza. Por otro lado, con un modelo simple, como una regresión lineal, presentará una varianza pequeña, pero tendrá mucho sesgo.

2.3. Enunciación del Teorema y Demostración

Teorema 2.1 (No Free Lunch (NFL)). *Sea A un algoritmo de aprendizaje con dominio finito $\mathcal{X}$ y conjunto objetivo $\mathcal{Y}$={0,1}. Sea m un entero tal que m $<\frac{|\mathcal{X}|}{2}$. Entonces existe una distribución con soporte $\mathcal{X}x\mathcal{Y}$ tal que*

1. *Existe una función de etiquetado f: $\mathcal{X} \to \mathcal{Y}$ con un error de generalización nulo, es decir, $L_D(f) = 0$*
2. *Con probabilidad mayor o igual a $\frac{1}{7}$ sobre posibles realizaciones de nuestros conjuntos de entrenamiento S $\sim D^m$, tenemos que el riesgo empírico de realizar el algoritmo de aprendizaje A a S no puede disminuir más de $\frac{1}{8}$, es decir:*

$$\mathbb{P}_{S\sim D^m}(L_D(A(S)) \geq \tfrac{1}{8}) \geq \tfrac{1}{7}$$

En pocas palabras, aunque encontremos una hipótesis f "perfecta", hay una distribución donde nuestra clasificación resulta no ser tan baja como se esperaba. La intuición de la demostración que veremos a continuación es que cualquier algoritmo que vea la mitad de algún subconjunto de $\mathcal{X}$ no va a tener información suficiente sobre el resto. Esto va a llevar a que exista una función de etiquetado que contradiga las etiquetas ya aprendidas

Demostración. Sea $C \subset \mathcal{X}$ con |C|=2m.

Además, sea T el conjunto que contiene a todas las posibles funciones f: C$\to \mathcal{Y}$. Por cómo está definido, |T|=2^{2m} y T=$\{f_1, ..., f_T\}$.

Sea D_i una distribución que actúa sobre Cx$\mathcal{Y}$ tal que con una observación {x,y}, tenemos que:

$$D_i(\{x,y\}) = \begin{cases} \frac{1}{|C|} & \text{y} = f_i(x) \\ 0 & \text{en otro caso.} \end{cases}$$

Esto nos dice que $L_{D_i}(f_i) = 0$

pd: Para todo A que reciba m observaciones y regrese una proyección de etiquetado A(s): C$\to \mathcal{Y}$ tiene:

$$\max_{i\in\{0,...,T\}} \mathbb{E}_{S\sim D^m} L_{D_i}(A(s)) \geq \tfrac{1}{4} \qquad (1)$$

Esto quiere decir que cualquier algoritmo que usemos tendrá un error mayor a $\frac{1}{4}$.

Sea k =$(2m)^m$ el número de todos los posibles conjuntos de entrenamiento, que etiquetaremos como $S_1, ..., S_k$. Esto se debe porque un conjunto de entrenamiento son realizaciones independientes de la distribución. Por ello para un conjunto de tamaño m estoy asignando 2m puntos posibles.

Si sabemos que tenemos $S_i = (x_1, ..., x_m)$, denotamos por S_j^i el conjunto de muestras de S_j que son etiquetadas por la i-ésima función f_i. Eso quiere decir que $S_j^i = \{(x_r, f_i(x_r))\}_{r=1}^m$.

Si D_i es la distribución que genera los datos, podemos esperar que los conjuntos $S_1, ..., S_k$ bajo esa distribución tienen la misma probabilidad de ser seleccionados. Esto implica que la pérdida esperada usando el algoritmo de aprendizaje con los sobre S es igual al promedio de las pérdidas usando el algoritmo de aprendizaje sobre todos los conjuntos S_j^i con j = 1, ..., k.

$$\mathbb{E}_{S\sim D^m}[L_{D_i}(A(S))] = \frac{1}{k}\sum_{j=1}^{k} L_{D_i}(A(S_j^i)) \qquad (2)$$

A partir de aquí, usaremos como idea principal para continuar la demostración que "máximo $\geq$ *promedio* $\geq$ *mínimo*"

$$\max_{i\in\{0,\dots,T\}} \frac{1}{k}\sum_{j=1}^{k} L_{D_i}(A(S_j^i)) \underset{max\geq prom}{\geq} \frac{1}{T}\sum_{i=1}^{T}\frac{1}{k}\sum_{j=1}^{k} L_{D_i}(A(S_j^i)) = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{T}\sum_{i=1}^{T} L_{D_i}(A(S_j^i))$$
$$\underset{prom\geq min}{\geq} \min_{j\in\{0,\dots,k\}} \frac{1}{T}\sum_{i=1}^{T} L_{D_i}(A(S_j^i)) \qquad (3)$$

Considerando $j \in \{1, ..., k\}$ fija y $S_j = (x_1, ..., x_m)$, sea $V_j = \{v_1, ..., v_p\}$ todos los objetos que no utilizamos para entrenar, es decir, $V_j = C - S_j$, donde p $\geq m.Porlotanto, paratodah : C \to \mathcal{Y}$ y para toda i, se cumple que

$$L_{D_i}(h) = \frac{1}{2m}\sum_{x\in C} \mathbb{I}_{[h(x)\neq f_i(x)]} \geq \frac{1}{2m}\sum_{x\in V_j} \mathbb{I}_{[h(x)\neq f_i(x)]} \geq \frac{1}{2p}\sum_{x\in V_j} \mathbb{I}_{[h(x)\neq f_i(x)]}$$

Entonces, tomando el promedio sobre las diferentes distribuciones, obtenemos que:

$$\frac{1}{T}\sum_{i=1}^{T} L_{D_i}(A(S_j^i)) \geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{2p}\sum_{x\in V_j} \mathbb{I}_{[h(x)\neq f_i(x)]} = \frac{1}{2p}\sum_{x\in V_j}\frac{1}{T}\sum_{i=1}^{T} \mathbb{I}_{[h(x)\neq f_i(x)]}$$
$$\underset{prom\geq min}{\geq} \frac{1}{2}\min_{r\in\{1,\dots,p\}} \frac{1}{T}\sum_{i=1}^{T} \mathbb{I}_{[h(x)\neq f_i(x)]} \qquad (4)$$

Considerando r fija (es decir, v_r fijo, que es un dato con el cual no entrenamos el algoritmo), se pueden partir las funciones $f_1, ..., f_T$ en pares donde cada par $(f_i, f_{i'})$ tiene como característica que para cualquier observación posible $c \in C$, la etiqueta $f_i(c) \neq f_{i'}(c)$ si y solo si $c = v_r$

$\implies$ El conjunto de entrenamiento que $S_i = S_{i'}$

$\implies$ $\mathbb{I}_{[A(S_j^i(v_r)\neq f_i(v_r)]} + \mathbb{I}_{[A(S_j^{i'}(v_r)\neq f_{i'}(v_r)]} = 1$, es decir, uno lo califica mal y el otro lo califica bien.

$$\implies \frac{1}{T}\sum_{i=1}^{T} \mathbb{I}_{[A(S_j^i(v_r)\neq f_i(v_r)]} = \frac{1}{T}\frac{T}{2} = \frac{1}{2} \qquad (5)$$

Esto se debe a que tenemos T términos, y al estar pareadas lo dividimos a la mitad, tomando todas las clasificaciones distintas.

Aplicando (5) en (4), obtenemos que:

$$\frac{1}{T}\sum_{i=1}^{T} L_{D_i}(A(S_j^i)) \geq \frac{1}{2}\min_{r\in\{1,\dots,p\}} \frac{1}{T}\sum_{i=1}^{T} \mathbb{I}_{[h(x)\neq f_i(x)]} = \frac{1}{2}\min_{r\in\{1,\dots,p\}}\frac{1}{2} = \frac{1}{4}$$

Aplicando (4) en (3), obtenemos que:

$$\max_{i\in\{0,\dots,T\}} \frac{1}{k}\sum_{j=1}^{k} L_{D_i}(A(S_j^i)) \geq \min_{j\in\{0,\dots,k\}} \frac{1}{T}\sum_{i=1}^{T} L_{D_i}(A(S_j^i)) \geq \min_{j\in\{0,\dots,k\}}\frac{1}{4} = \frac{1}{4}$$

Aplicando (3) en (2), obtenemos que:

$$\mathbb{E}_{S\sim D^m}[L_{D_i}(A(S))] = \frac{1}{k}\sum_{j=1}^{k} L_{D_i}(A(S_j^i))$$
$$\implies \max_{i\in\{0,\dots,T\}} \mathbb{E}_{S\sim D^m}[L_{D_i}(A(S))] = \max_{i\in\{0,\dots,T\}} \frac{1}{k}\sum_{j=1}^{k} L_{D_i}(A(S_j^i)) \geq \frac{1}{4}.$$

Revisando la aplicación de (3) en (2), podemos observar que llegamos a (1), así demostrando la primera parte.

pd: $\mathbb{E}_{S\sim D^m}[L_{D_i}(A(S))] \geq \frac{1}{4} \implies \mathbb{P}_{S\sim D^m}(L_D(A(S)) \geq 1/8) \geq \frac{1}{7}$

Recordemos que la desigualdad de Markov nos dice que para una variable aleatoria X donde $\mathbb{E}[X] \geq \alpha$, entonces

$$\mathbb{P}(X \geq 1-p) \geq \frac{\alpha-(1-p)}{p}$$

De este resultado, como sabemos que $\mathbb{E}_{S\sim D^m}[L_{D_i}(A(S))] \geq \frac{1}{4}$, concluimos por la desigualdad de Markov que:

$$\mathbb{P}_{S\sim D^m}(L_D(A(S)) \geq 1-\tfrac{7}{8}) \geq \frac{\frac{1}{4}-(1-\frac{7}{8})}{\frac{7}{8}} = \frac{1}{7}$$

□

2.4. ¿Cómo relacionamos NFL con el conocimiento previo

Corollary 2.2. *Sea $\mathcal{X}$ un conjunto infinito y sea $\mathcal{H} = \{f{:}\mathcal{X} \to \{0,1\}\}$. Entonces $\mathcal{H}$ no es aprendible en el sentido PAC (probable y aproximadamente correcto).*

Demostración. Asumamos que $\mathbb{H}$ es aprendible en el sentido PAC. Dada $\varepsilon < \frac{1}{8}$ y $\delta < \frac{1}{7}$, esto implica que existe un algoritmo A y un entero $m = m(\varepsilon, \delta)$ que denota el tamaño de muestra suficiente tal que para toda función de etiquetado f:$\mathcal{X} \to \{0,1\}$, el riesgo de realizabilidad $L_D(f) = 0$, lo que implica que con probabilidad mayor o igual que 1-δ el algoritmo A con muestras m nos dará que:

$$L_d(A(S)) \geq \varepsilon$$

Como $|\mathcal{X}| = \infty > 2m$, por el Teorema NFL tenemos que para todo algoritmo existe una distribución tal que con probabilidad mayor o igual que $\frac{1}{7} > \delta$, obtenermos que:

$$L_D(A(S)) > \tfrac{1}{8} > \varepsilon$$

Esto nos lleva a una contradicción, que es causada por imponer que $\mathcal{H}$ es aprendible PAC. Por lo tanto, $\mathcal{H}$ no es aprendible en el sentido PAC. □

3. Descomposición del Error

Recordemos que denotamos a h_s como el predictor que regresa el principio de minimización de riesgo empírico aplicado a una familia de posibilidades $\mathcal{H}$ ($ERM_{\mathcal{H}}$) El error de generalización h_s se puede descomponer de la siguiente manera:

$$L_D(h_s) = \varepsilon_a + \varepsilon_e$$

$$\varepsilon_a := \text{Error de aproximación}; \qquad \varepsilon_e := \text{Error de estimación}.$$

$$\varepsilon_a = \min_{h\in\mathcal{H}} L_D h; \qquad \varepsilon_e = L_D(h_s) - \varepsilon_a.$$

ε_a representa el riesgo mínimo de la clase (sesgo inductivo), y ε_e es la diferencia entre nuestra aproximación con el método ERM y lo que hubiéramos esperado de la clase de posibilidades que estamos tomando en cuenta. De aquí tomamos en cuenta dos posibilidades:

1. $\mathcal{H}$ es muy flexible, lo cual causa que ε_a disminuya, pero ε_e aumente. A esto lo conocemos como sobreajuste.
2. $\mathcal{H}$ es muy flexible, lo cual causa que ε_a aumente , pero ε_e disminuye. A esto lo conocemos como subajuste.

Esto lleva a que si aumentamos la flexibilidad de nuestra familia de funciones encontremos resultados poco favorables, pero si la familia es muy rígida, esto causaría que los resultados no estén bien generalizados, que hace que no sea lo suficientemente complejo.

Visto de forma más gráfica, si nuestra familia $\mathcal{H}$ es muy flexible, eso ocasionaría que intente conectar con la mayor cantidad de datos posibles, haciendo una gran cantidad de oscilaciones para llegar a esos puntos. En casos donde sea muy difícil llegar a esos resultados, donde los demás modelos indicarían que es un error de observación, una familia más ajustada lo tomaría como un caso particular. Por otro lado, si $\mathcal{H}$ es más rígida, a pesar de su utilidad para la estimación, habría un gran sesgo inductivo, que se puede ver porque no se acerca lo suficiente a los datos. En la siguiente figura se muestra un ejemplo gráfico con esta descripción para facilitar la visualización.

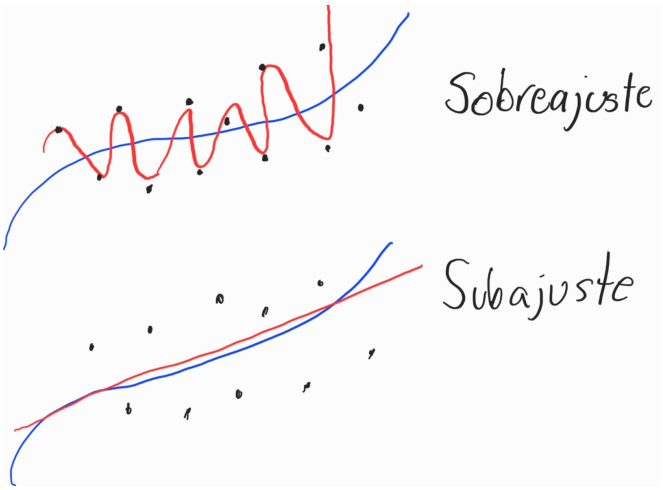

Figura 1 Ejemplo de modelos con familias sobreajustadas y subajustadas.