Aprendizaje Estadístico
Alfredo Garbuño

Notas clase 11 de marzo

Escribano:
Alejandro Chávez

Marzo 2021

Índice general

Capítulo 1

Regularización y estabilidad

Hoy veremos que CLA y CSA son realmente logrables/aprendibles. Hay instancias de esos dos que son uniformes y por lo tanto aprendibles, con la regla más sencilla: ERM. El único problema es que, en general, no es cierto que la mayoría de las CLA o CSA seana aprendibles, entonces necesitamos un nuevo paradigma que nos garantice ello: minimización de pérdida regularizada.

Idea: tomamos una función de pérdida y sumamos algo que regularice.

Nota: regularización nos sirve para medir la complejidad de la hipótesis y actúa como un estabilizador).

1.1. Minimización de pérdida regularizada (RLM)

RLM = pérdida empírica + función de regularización.

El objetivo: $\underset{w\in H}{\operatorname{argmin}} L_s(w) + R(w)$. Queremos, pues, encontrar un balance entre modelos simples y soluciones con error empírico pequeño. $R()$ normalmente se escoje con conocimiento de dominio, y las opciones clásicas son $R(w) = \lambda * \|w\|_2^2$ ó $R(w) = \lambda * \|w\|_1^2$

Regularización de Tikhonov (se usa para problemas inversos): supongamos que tenemos un problema de regresión de la forma $\frac{1}{2m}\|x_w - y\|^2 + \frac{\lambda}{2}\|w\|^2$, el cual buscamos minimizar. Ya hemos visto que $\frac{1}{2m}\|x_w - y\|^2$ tiene un mínimo. La ventaja de $\frac{\lambda}{2}\|w\|^2$ es que no afecta mucho ese mínimo y, además, la solución de $\Delta_w(L_s(w) + R(w) = 0$ es $(\lambda_m * \mathbb{1}_{pxp})w = x^T y * 2w = (x^T x + \lambda_m * \mathbb{1}_{pxp})^{-1} x^T y$. Por ejemplo, si $P(w) = w_0 + w_1 x + \cdots + w_p x^p$ y $\lambda >> 0$, esperaríamos que el modelo fuera parsimonioso y que los términos mayores se fueran eliminando.

Verificaremos que $R(w)$ estabiliza y permite el sobreajuste. En particular, veremos un resultados en que

$$E_S(L_D(A(S))) \leq \min_{w\in H} L_D(w) + \varepsilon$$

Otra forma de pensar el problema es: $\min_{w\in H} L_D(w)$ sujeto a $\|w\|_2^2 \leq \theta$

1.2. Noción de estabilidad

Sea A un algoritmo de aprendizaje y sea $S(\sim D^m) = (Z_1, \ldots, Z_m)$. Hablamos de sobreajuste cuando $|L_D(A(S)) - L_S(A(S))|$ es grande. Sea $S^{(i)} = (Z_1, \ldots, Z_i, Z', Z_{i+1}, \ldots, Z_m)$ con Z' independiente de los anteriores y $Z' \sim D^m$. Lo que esperaríamos de un buen argumento es:

$$l(A(S^{(i)})), Z_i) * l(A(S)), Z_i) \geq 0$$

.

Teorema 1. *Supongamos que D es una distribución. Sea $S(\sim D^m) = (Z_1, \dots, Z_m)$, y sea $Z' \sim D^m$ una observación independiente. Denotamos como $U(m)$ a la distribución uniforme en el conjunto de índices $\{1, \dots, m\}$. Entonces*

$$E_S(L_D(A(S)) - L_S(A(S))) = E_{\substack{S\sim D^m \\ i\sim U(m)}}(l(A(S^{(i)})), Z_i) - l(A(S)), Z_i))$$

Demostración. $E_S(L_D(A(S)) = E_{S,Z'}(l(A(S), Z')) = E_{\substack{S\sim D^m \\ i\sim U(m)}}(l(A(S^{(i)})), Z_i)$ Por otro lado, $E_S(L_S(A(S)) = E_{S,i}(l(A(S), Z_i)) = E_S(\frac{1}{n}\sum_{i=1}^m l(A(S), Z_i)$ □

Definición 1. *Sea $\varepsilon : \mathbb{N} \to \mathbb{R}$ monótonamente decreciente. Decimos que el algoritmo A es estable en promedio, bajo reemplazos individuales con tasa $\varepsilon(m)$, si $\forall D$ se tiene que*

$$E_{\substack{S\sim D^m \\ i\sim U(m)}}(l(A(S^{(i)})), Z_i) - l(A(S)), Z_i)) \leq \varepsilon(m)$$

1.3. Regularización como estabilizador

Definición 2. *Una función f es λ-fuertemente convexa si $\forall u, w$ y $\alpha \in (0, 1)$ tenemos que*

$$f(\alpha w + (1-\alpha)u) \leq \alpha f(w) + (1-\alpha)f(u) - \frac{\lambda}{2}\alpha(1-\alpha)\|w-u\|^2$$

Lema 1.

1. *$f(w) = \lambda\|w\|^2$ es 2λ-fuertemente convexa.*
2. *Si f es λ-fuertemente convexa y g es convexa $\Rightarrow f+g$ es fuertemente convexa.*
3. *Si f es λ-fuertemente convexa y u es el minimizador de $f \Rightarrow \forall w, f(w) - f(u) \geq \frac{\lambda}{2}\|w-u\|^2$*

Demostración. Del inciso 3 del Lema 1.
$f(u + \alpha(w-u)) - f(u) \leq \alpha f(w) - \alpha f(u) - \frac{\lambda}{2}\alpha(1-\alpha)\|w-u\|^2$
$\Rightarrow \frac{f(u+\alpha(w-u))-f(u)}{\alpha} \leq f(w) - f(u) - \frac{\lambda}{2}(1-\alpha)\|w-u\|^2$ Si $\alpha \to 0$, el término de la derecha equivale a la derivada evaluada en el minimizados □

Falta ver que RLM es estable. Consideremos $S, Z', S^{(i)}$ como arriba, y A RLM. Entonces

$$A(S) = \underset{w\in H}{\operatorname{argmin}} L_s(w) + \lambda\|w\|^2 \text{ y } f_S(w) = L_S(w) + \lambda\|w\|^2 \text{ (2λ-fuertemente convexa).}$$

Además, $f_S(w) - f_S(A(s)) \geq \lambda\|v - A(S)\|^2$

$$\Rightarrow f_S(w) - f_S(u) = L_S(v) + \lambda\|v\|^2 - L_S(u) - \lambda\|u\|^2 L_S(u) = L_{S^{(i)}}(v) + \frac{l(v, Z_i) - l(v, Z')}{m}$$

$$\therefore \lambda\|A(S^{(i)}) - A(S)\| \leq f_S(A(S^{(i)})) - f_S(A(S)) \leq \frac{l(A(S^{(i)}), Z_i) - l(A(S), Z_i)}{m} + \frac{l(A(S), Z') - l(A(S^{(i)}), Z')}{m}$$

1.3.1. Bajo Lipschitz

$$l(A(S^{(i)}), Z_i) - l(A(S), Z_i) \leq \rho\|A(S^{(i)}) - A(S)\| \Rightarrow \|A(S^{(i)}) - A(S)\| \leq \frac{2\rho}{\lambda m}$$

$$\Rightarrow E_S(L_D(A(S)) - L_S(A(S))) \leq \frac{2\rho^2}{\lambda m}$$

1.4. Control de sobreajuste y estabilidad

$E_S(L_D(A(S))) = E_S(L_S(A(S))) + E_S(L_D(A(S)) - L_S(A(S))$ Si λ crece, el error empírico también.

Dado que $A(S) = \text{argmin} L_S(w) + \lambda\|w\|^2$,

$$L_S(A(S)) \leq L_S(A(S)) + \lambda\|A(S)\|^2 \leq L_S(w) + \lambda\|w\|^2$$

$$\Rightarrow E_S(L_S(A(S))) \leq L_D(w) + \lambda\|w\|^2, E_S(L_D(A(S))) \leq L_D(w) + \lambda\|w\|^2 + estabilidad$$

$$\therefore E_S(L_D(A(S))) \leq L_D(w) + \lambda\|w\|^2 + \frac{2\rho}{\lambda m}$$