Aprendizaje Estadístico
Alfredo Garbuño

Primer examen parcial

Equipo:
Carlos Lezama
Jorge Rizo
Alejandro Chávez

Marzo 2021

Índice general

Capítulo 1

Parte teórica

1.1. Problema 1

Sabemos que una función convexa f es β-suave si satisface que $f(v) \leq f(w) + \langle \Delta f(w), v - w \rangle + \frac{\beta}{2}\|v - w\|^2$, $\forall v, w \in D_f$.

Sean:

- A un algoritmo de aprendizaje.
- $S(\sim D^m) = (z_1, \ldots, z_m)$.
- $S^{(i)} = (z_1, \ldots, z_i, z', z_{i+1}, \ldots, z_m)$ con z' independiente de los anteriores y $z' \sim D^m$.
- $\ell(\cdot, z_i)$ una función de pérdida.

Tenemos, pues, lo siguiente:

$$\begin{aligned}
\ell(A(S^{(i)}), z_i) &\leq \ell(A(S), z_i) + \langle \Delta\ell(A(S), z_i), A(S^{(i)}) - A(S) \rangle + \frac{\beta}{2}\|A(S^{(i)}) - A(S)\|^2 \\
\Leftrightarrow \ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) &\leq \langle \Delta\ell(A(S), z_i), A(S^{(i)}) - A(S) \rangle + \frac{\beta}{2}\|A(S^{(i)}) - A(S)\|^2 \\
&\leq \|\Delta\ell(A(S), z_i), A(S^{(i)}) - A(S)\| \|A(S^{(i)}) - A(S)\| + \frac{\beta}{2}\|A(S^{(i)}) - A(S)\|^2 \\
&\leq \|A(S^{(i)}) - A(S)\| \sqrt{2\beta\ell(A(S^{(i)}), z_i)} + \frac{\beta}{2}\|A(S^{(i)}) - A(S)\|^2
\end{aligned}$$

Dada la simetría de cada $z_i \in (S \cup S^{(i)})$ iid, se cumple para $\ell(\cdot, z_i)$ que:

$$\ell(A(S^{(i)}), z_i) - \ell(A(S^i), z_i) \leq \|A(S) - A(S^{(i)})\| \sqrt{2\beta\ell(A(S^{(i)}), z_i)} + \frac{\beta}{2}\|A(S) - A(S^{(i)})\|^2$$

De la demostración del lema sobre convexidad fuerte (vista en clase), sabemos que:

$$\lambda\|A(S^{(i)}) - A(S)\|^2 \leq \frac{\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)}{m} + \frac{\ell(A(S), z') - \ell(A(S^{(i)}), z')}{m}$$

$$\Rightarrow \lambda\|A(S^{(i)}) - A(S)\| \leq \frac{1}{m}\left(\|A(S^{(i)}) - A(S)\|\sqrt{2\beta\ell(A(S), z_i)} + \frac{\beta}{2}\|A(S^{(i)}) - A(S)\|^2 \right.$$
$$\left. + \|A(S^{(i)}) - A(S)\|\sqrt{2\beta\ell(A(S), z')} + \frac{\beta}{2}\|A(S^{(i)}) - A(S)\|^2\right)$$

$$\Rightarrow \|A(S^{(i)}) - A(S)\| \leq \frac{1}{\lambda m}\left(\sqrt{2\beta}\left(\sqrt{\ell(A(S), z_i)} + \sqrt{\ell(A(S^{(i)}), z')}\right) + \beta\|A(S^{(i)}) - A(S)\|^2\right)$$

$$\Rightarrow \|A(S^{(i)}) - A(S)\|\left(1 - \frac{1}{\lambda m}\right) \leq \frac{\sqrt{2\beta}}{\lambda m}\left(\sqrt{\ell(A(S), z_i)} + \sqrt{\ell(A(S^{(i)}), z')}\right)$$

$$\Rightarrow \|A(S^{(i)}) - A(S)\| \leq \frac{\lambda m\sqrt{2\beta}}{\lambda m(\lambda m - \beta)}\left(\sqrt{\ell(A(S), z_i)} + \sqrt{\ell(A(S^{(i)}), z')}\right)$$

Si asumimos que $\lambda \geq \frac{2\beta}{m}$, tenemos que:

$$\|A(S^{(i)}) - A(S)\| \leq \frac{\sqrt{8\beta}}{\lambda m}\left(\sqrt{\ell(A(S), z_i)} + \sqrt{\ell(A(S^{(i)}), z')}\right)$$

.

Así pues,

$$\begin{aligned}
\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) &\leq \|A(S^{(i)}) - A(S)\|\sqrt{2\beta\ell(A(S^{(i)}), z_i)} + \frac{\beta}{2}\|A(S^{(i)}) - A(S)\|^2 \\
&\leq \left(\frac{4\beta}{\lambda m} + \frac{8\beta^2}{\lambda^2 m^2}\right)\left(\sqrt{\ell(A(S), z_i)} + \sqrt{\ell(A(S^{(i)}), z')}\right)^2 \\
&\leq \frac{8\beta}{\lambda m}\left(\sqrt{\ell(A(S), z_i)} + \sqrt{\ell(A(S^{(i)}), z')}\right)^2 \\
&\leq \frac{24\beta}{\lambda m}\left(\ell(A(S), z_i) + \ell(A(S^{(i)}), z')\right)
\end{aligned}$$

Podemos entonces acotar el error de generalización para $\lambda \geq \frac{2\beta}{m}$ como sigue:

$$\begin{aligned}
\mathbb{E}(\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)) &\leq \frac{24\beta}{\lambda m}\mathbb{E}\left((\ell(A(S), z_i) + \ell(A(S^{(i)}), z'))\right) \\
&= \frac{48\beta}{\lambda m}\mathbb{E}\left(L_S(A(S))\right.
\end{aligned}$$

Tenemos el siguiente teorema de clase:

Teorema 1. *Supongamos que D es una distribución. Sea $S(\sim D^m) = (z_1, \ldots, z_m)$, y sea $z' \sim D^m$ una observación independiente. Denotamos como $U(m)$ a la distribución uniforme en el conjunto de índices $\{1, \ldots, m\}$. Entonces*

$$\mathbb{E}_S(L_D(A(S)) - L_S(A(S))) = \mathbb{E}_{\substack{S\sim D^m \\ i\sim U(m)}}(\ell(A(S^{(i)})), z_i) - \ell(A(S)), z_i))$$

Por lo tanto, juntando el teorema con la última ecuación, tenemos que

$$\mathbb{E}_{\substack{S\sim D^m \\ i\sim U(m)}}(L_D(A(S)) - L_S(A(S))) \leq \frac{48\beta}{\lambda m}\mathbb{E}\left(L_S(A(S))\right.$$

1.2. Problema 2

1.2.1. a

Al asumir $\ell(0, z) \leq C$ con $C > 0$, $\forall z \sim D$, tenemos que:

$$\begin{aligned} L_S(A(S)) &\leq L_S(A(S)) + \lambda\|A(S)\|^2 \\ &\leq L_S(0)\lambda\|0\|^2 \\ &= L_S(0) \\ &\leq C \end{aligned}$$

Concluimos que

$$\mathbb{E}_{\substack{S\sim D^m \\ i\sim U(m)}}(L_D(A(S))L_S(A(S))) \leq \frac{48\beta}{\lambda m}C$$

1.2.2. b

Dada la propiedad de linealidad del valor esperado, podemos reescribir el riesgo esperado de un algoritmo de aprendizaje como sigue: $\mathbb{E}_S(L_D(A(S))) = \mathbb{E}_S(L_S(A(S))) + \mathbb{E}_S(L_D(A(S)) - L_S(A(S)))$. De eso, sabemos que $\mathbb{E}_S(L_D(A(S)) - L_S(A(S)))$ denota nuestro error de generalización o estabilidad del algoritmo; y $\mathbb{E}_S(L_D(A(S)))$, nuestro error de ajuste. Nótese que todo eso concuerda con que "a mayor regularización, mejor estabilidad, pero mayor sesgo" [**1**]

Antes acotamos superiormente la estabilidad de un algoritmo en aprendizaje bajo el principio de ERM con regularizador $\lambda\|w\|^2$.

Con el fin de acotar el riesgo esperado, podemos utilizar el enfoque de aprendizaje no-uniforme tal que, siguiendo el principio inductivo de minimización de riesgo estructural (SRM), busquemos el predictor que minimice $L_D(A(S)) \leq L_S(h) + \epsilon$.

Sabemos que $L_D(A(S)) \leq L_S(h) + \lambda\|A(S)\|^2 \leq L_S(u) + \lambda\|u\|^2$, con u un vector arbitrario. Nota: por el principio de RLM, $A(S) = \arg\min(L_S(u)\lambda\|u\|^2)$.

Es fácil ver que $\mathbb{E}_S(L_D(A(S))) \leq L_D(u) + \lambda\|u\|^2$. Para $\mathbb{E}_S(L_S(A(w)) = L_D(w)$, sustituyendo en x, tenemos $\mathbb{E}_S(L_D(A(S))) \leq L_D(u) + \lambda\|u\|^2 + \mathbb{E}_S(L_D(A(S))) - L_S(h)$. Para nuestro caso de estudio con una función convexa β-suave, tenemos que para $\lambda \geq \frac{2\beta}{m}$:

$$\mathbb{E}_S(L_D(A(S))) \leq \left(1 + \frac{48\beta}{\lambda m}\right)\mathbb{E}_S(L_S(A(S)) \leq \left(1 + \frac{48\beta}{\lambda m}\right)(L_D(w) + \lambda\|w\|^2$$

Es fácil ver que podemos llegar a un resultado similar, acotando el riesgo esperado a través del principio de SRM.

Bibliografía

[1] Goodfellow, Ian and Bengio, Yoshua and Courville, Aaron. *Deep Learning*. MIT Press, Reading, Cambridge, 2017, p. 107-117