Clase 10: Boosting

Responsable: Mariana Pérez-Cong Sánchez

EST-25134, Primavera 2021
Dr. Alfredo Garbuno Iñigo
20 de febrero de 2021

1. Introducción

- Quisieramos ver una generalizacion sobre los predictores lineales
- Que sean capaces de atacar el compromiso de sesgo y varianza.
  - Mayor complejidad $\Rightarrow$ error de aproximación pequeño
- Boosting nos permite controlar este compromiso por medio de un parámetro. Empieza con un modelo sencillo pero gana complejidad en el proceso de aprendizaje.
- Vamos a ver **Ada Boost** (Adaptative Boosting): combina predicciones de manera lineal.

2. Capacidad de aprendizaje Débil

Tenemos:

- PAC con $m_H(\varepsilon, \delta)$
- Teorema fundamental de Aprendizaje Estadístico $\rightarrow$ ERM
- ¿Existe un algoritmo eficiente con un error ligeramente mejor que lanzar una moneda?

Definición:
Un algoritmo A es $\gamma - \textit{débil}$ para H si $\exists m_H : (0,1) \rightarrow \mathbb{N}$ tal que $\forall \delta \in (0,1)$, D y $f : X \rightarrow \{\pm 1\}$ y si se mantiene la hipótesis de realizabiladidad,entonces con $m > m_H(\delta)$ iid de D, el algoritmo regresa un candidato $h \in H$ tal que con probabilidad mayor o igual a $1 - \delta$, $L_D(h) \leq \frac{1}{2} - \gamma$

- Decimos que H es $\gamma - \textit{débil}$ si existe un algoritmo $\gamma - \textit{débil}$ para dicha clase

Ejemplo: Sea $X = \mathbb{R}$ y H los clasificadores en tres partes:

$$H = \{h_{\theta_1,\theta_2,b} : \theta_1, \theta_2 \in \mathbb{R}, \theta_1 < \theta_2, b \in \{\pm 1\}\}$$

$$h_{\theta_1,\theta_2,b}(x) = \begin{cases} -b & \text{si } x < \theta \text{ o } x > \theta_2 \\ -b & \text{si } x \in [\theta_1, \theta_2] \end{cases}$$

Sea $B = \{f : f(x) = signo(x - \theta) \cdot b : \theta \in \mathbb{R}, b \in \{\pm 1\}\}$. Veremos que ERM_b es $\gamma - debil$ para H con $\gamma = 1/2$. $\exists h \in B$ que tendrá un error de clasificacion $Ld(h) < \frac{1}{3}$
$VCdim(B) = 2 \Rightarrow m_H(\gamma) \geq \frac{log(1/\gamma)}{\varepsilon}$ (módulo una constante).

Entonces, Con prob $\geq 1 - \delta$, ERM tiene un error de generalizacion $\leq 1/3 + \varepsilon$.
Si consideramos $\varepsilon = 1/12 \Rightarrow 1/3 + 1/12 = 1/2 - 1/12$
$\therefore ERM_b$ es $\gamma - debil$

2.1. Implementacion

Sea $X = \mathbb{R}^d$; $H_{DS} = \{g : g(x) = signo(\theta - x_i), x \in \mathbb{R}^d, \theta \in \mathbb{R}, i = \{1, ..., d\}\}$ Sea $S = \{(x_i, y_i)\}_{i=1}^m$.
Vamos a ver cómo minimizar $L_s(h)$y cómo se pueden usar estos predictores débiles para Ada Boost.

Sea $\mathcal{D}$ un vector de probabilidades en $\mathbb{R}^m$ ($D_i \geq 0, \sum \mathcal{D}_i = 1$). El modelo recibe $\mathcal{D}, S$ y regresa un predictor $h : X \to \{\pm 1\}$, donde h minimiza el riesgo con respecto a $\mathcal{D}$

$$L_{\mathcal{D}}(h) = \sum_{i=1}^{m} \mathcal{D}\, i \mathbb{1}_{[h(x^{(i)}) \neq y^{(i)}]}$$

$\forall h \in H_{DS}$ es $h = h(\theta, i)$ Entonces queremos minimizar:

$$\min_j \min_\theta \left(\sum_{i:y^{(i)}=1} \mathcal{D}_i \ \ \mathbb{1}_{[x_j^{(i)} > \theta]} + \sum_{i:y^{(i)}=-1} \mathcal{D}_i \ \ \mathbb{1}_{[x_j^{(i)} < \theta]} \right)$$

Si consideramos fija j, podemos ordenar:

$$x_j^{(1)} \leq ... \leq x_j^{(m)}$$

$$\Theta_j = \left\{ \frac{x_j^{(i)} + x_j^{(i+1)}}{2} : i \in 1, .., m-1 \right\} \cup \left\{ x_j^{(1)} - 1, x_j^{(m)} + 1 \right\}$$

Como $\forall \theta \in \mathbb{R} \quad \exists \theta' \in \Theta_j$ que realiza las mismas predicciones, en lugar de optimizar $\mathbb{R}$, optimimizamos Θ_j

3. Ada Boost

$S = (x^{(i)}, y^{(i)})_{i=1}^m$ donde $y^{(i)} = f(x^{(i)})$.
Al tiempo t:

1. Definimos una distribución sobre S $\mathcal{D}^{(t)} \in \mathbb{R}_+^m$
2. Utilizamos un algoritmo débil con $S, \mathcal{D}^{(t)}$. Obtenemos h_t con error

   $$\varepsilon := L_{D^{(t)}}(h_t) \leq 1/2 - \gamma$$

   Con prob $\geq 1 - \delta$
3. Asignamos una contribución a h_t con $w_t = \frac{1}{2} log(1/\varepsilon_t - 1)$
4. Calculamos una nueva distribución:

   $$\tilde{\mathcal{D}}_i^{(t+1)} = \mathcal{D}_i^{(t)} \, exp\Big(-w_t y^{(i)} h_t(x^{(i)})\Big)$$

   $$\mathcal{D}_i^{(t+1)} = \frac{\tilde{\mathcal{D}_i}^{(t+1)}}{\sum_j \tilde{\mathcal{D}_j}^{(t+1)}}$$

Repetimos los pasos 1)-4) T veces de forma que:

$$h_s(x) = signo\left(\sum_{t=1}^{T} w_t h_t(x)\right)$$

Teorema 3.1. *Sea S un conjunto de entrenamiento y asumamos que en cada iteración de AdaBoost tenemos $\varepsilon_t \leq 1/2 - \gamma$, entonces el error de entrenamiento:*

$$L_s(hs) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{[h_s(x^{(i)}) \neq y^{(i)}]} \leq exp^{(-2\gamma^2 T)}$$

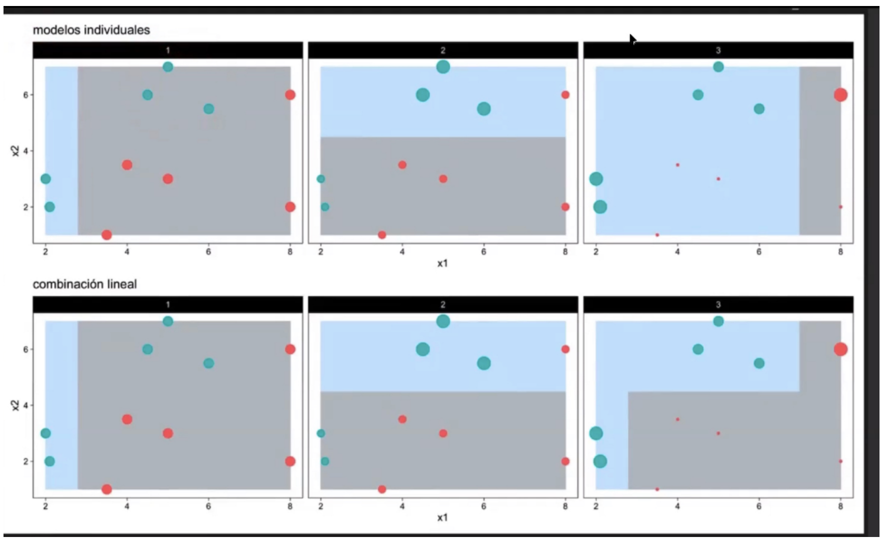

Figura 1 Modelos lineales vs. Combinación lineal

- Los modelos débiles pueden fallar con probabilidad δ. La probabilidad de éxito al tiempo T está acotada por $1 - \delta T$
- ¿Como podemos garantizar que el error de estimación $(Ls(h_s) - \min_{h \in H} L_{\mathcal{D}}(h))$ es pequeño?
  $VCdim(H) \rightarrow$ *pequeña* $\rightarrow$ garantiza buen desempeño
- $L(B,T) = \{h(x) = signo\left(\sum_{t=1}^{T} w_t h - t(x)\right) : x \in \mathbb{R}^d, wt \in \mathbb{R}^T, h_t \in R\}$
  $VCdim(L(B,T)) \leq TVCdim(B)$
  Por lo tanto T nos ayuda a controlar el compromiso entre sesgo y varianza