Equivalencia entre el método Ada-Boosting y Forward Stageway Additive Model

Andrés Nieto Guadarrama

Lema:

Encontrar la clasificación óptima mediante el método de $Ada\ Boosting$ equivale a implementar el modelo aditivo con la función de costo

$$L\left(y, f(x)\right) \;=\; e^{-yf(x)}.$$

Demostración

Observar que, para cada paso m, el método del modelo aditivo busca al clasificador G_m y al parámetro β_m tales que:

$$(\beta_m, G_m) \;=\; \min_{\beta, G}\left\{\sum_{i=1}^{N} e^{-y_i(f_{m-1}(x_i)+\beta G(x_i))}\right\} \;=\; \min_{\beta, G}\left\{\sum_{i=1}^{N} e^{-y_i f_{m-1}(x_i)} e^{-y_i\beta G(x_i)}\right\} \;=\; \min_{\beta, G}\left\{\sum_{i=1}^{N} w_i^{(m)} e^{-}\right.$$

con $w_i^{(m)} \;=\; e^{-y_i f_{m-1}(x_i)} \;>\; 0$ independiente de G y β. Más aún, partiendo la suma anterior tenemos que

$$\sum_{i=1}^{N} w_i^{(m)} e^{-y_i\beta G(x_i)} \;=\; \sum_{\{i:y_i=G(x_i)\}} w_i^{(m)} e^{-\beta} + \sum_{\{i:y_i\neq G(x_i)\}} w_i^{(m)} e^{\beta} \;=\; \sum_{i=1}^{N} w_i^{(m)} e^{-\beta} - \sum_{\{i:y_i\neq G(x_i)\}} w_i^{(m)} e^{-\beta} +$$

con lo cual,

$$\sum_{i=1}^{N} w_i^{(m)} e^{-y_i\beta G(x_i)} \;=\; \sum_{i=1}^{N} w_i^{(m)} e^{-\beta} + (e^{\beta} - e^{-\beta}) \sum_{\{i:y_i\neq G(x_i)\}} w_i^{(m)} \;=\; \sum_{i=1}^{N} w_i^{(m)} e^{-\beta} + (e^{\beta} - e^{-\beta}) \sum_{i=1}^{N} w_i^{(m}$$

Dado que el único elemento de la suma que se ve afectado por la función clasificadora es $\sum_{i=1}^{N} w_i^{(m)} 1_{\{y_i \neq G(x_i)\}}$, tenemos que para cada paso m

$$G_m \;=\; \min_G \left\{ \sum_{i=1}^{N} w_i^{(m)} 1_{\{y_i \neq G(x_i)\}} \right\}.$$

Notemos que a partir de éstos cálculos ya se consiguió el clasificador G_m y los pesos $w_i^{(m)}$ requeridos como primer paso en el algoritmo $Ada\ Boosting\ M1$.

Por otro lado, sustituyendo a G por G_m en la función objetivo, podemos notar que las sumas $\sum_{i=1}^{N} w_i^{(m)}$ y $\sum_{i=1}^{N} w_i^{(m)} 1_{\{y_i \neq G(x_i)\}}$ no dependen de β, por lo que β_m se encuentra minimizando la función en $\mathbb{R}$:

$$g(\beta) \;=\; Ae^{-\beta} + B(e^{\beta} - e^{-\beta}), = \;(A-B)e^{-\beta} + Be^{\beta}$$

con $A \;=\; \sum_{i=1}^{N} w_i^{(m)}$ y $B \;=\; \sum_{i=1}^{N} w_i^{(m)} 1_{\{y_i \neq G_m(x_i)\}}$

Por condiciones de primer orden, β_m debe cumplir que:

$$0 \;=\; g'(\beta_m) \;=\; (B-A)e^{-\beta_m} + Be_m^{\beta} \;=\; \frac{1}{e^{\beta_m}}((B-A) + Be^{2\beta_m}),$$

esto es,

$$\beta_m \;=\; \frac{1}{2} ln\left(\frac{A-B}{B}\right) \;=\; \frac{1}{2} ln\left(\frac{\sum_{i=1}^{N} w_i^{(m)} 1_{\{y_i = G_m(x_i)\}}}{\sum_{i=1}^{N} w_i^{(m)} 1_{\{y_i \neq G_m(x_i)\}}}\right) \;=\; \frac{1}{2} ln\left(\frac{1-err_m}{err_m}\right)$$

donde $err_m = \frac{B}{A}$ es el error de predicción de la función clasificadora G_m.

Una vez encontradas β_m y G_m, el siguiente paso del modelo aditivo nos lleva a actualizar las funciones f_m de manera que:

$$f_m \;=\; f_{m-1} + G_m\beta_m.$$

Sin embargo, por la forma en que definimos los pesos w_i^m, tenemos que para el siguiente paso:

$$w_i^{m+1} \;=\; e^{-y_i f_m(x_i)} \;=\; e^{-y_i f_{m-1}(x_i)} e^{-y_i G_m(x_i)\beta_m} \;=\; w_i^m e^{-y_i G_m(x_i)\beta_m}$$

Más aún, dado que $-y_i G_m(x_i) \;=\; -1$ si $y_i \;=\; G_m(x_i)$ y $-y_i G_m(x_i) \;=\; 1$ si $y_i \;\neq\; G_m(x_i)$, podemos escribir a $-y_i G_m(x_i)$ como

$$-y_i G_m(x_i) \;=\; 2 1_{\{y_i \neq G_m(x_i)\}} - 1$$

Con lo cual,

$$w_i^{m+1} \;=\; w_i^m e^{2\beta_m 1_{\{y_i \neq Gm(x_i)\}}} e^{-\beta_m}$$

Finalmente, observar que el factor $e^{-\beta_m}$ afecta a todas las w_i^{m+1} para toda $i \;=\; 1, \ldots, N$, por lo que puede omitirse. Por lo tanto, se obtuvieron los mismos valores para β, G y w_i^m que si se hubiera seguido el algoritmo de $AdaBoosting\ M1$ ■