Derivación del problema de optimización en Boosting

Percepción - ETSInf

1. Planteamiento

Se dispone de un conjunto de L clasificadores débiles: $\mathcal{G} = \{G_1, \ldots, G_L\}$ binarios ($G_l(x) \in \{-1, 1\}$) y un conjunto de entrenamiento $\mathcal{X} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ con $y_n \in \{-1, 1\}$.

En Boosting, en la i-ésima iteración se toma el clasificador $C_i \in \mathcal{G}$ de menor error sobre los datos de $\mathcal{X}$, ponderando el error de cada muestra por un peso asociado a dicha iteración ($w^{(i)}$). El clasificador final $G(x)$ obtenido tras m iteraciones se define por:

$$G(x) = G^{(m)}(x) = \sum_{i=1}^{m} \alpha_i \, C_i(x)$$

donde $C_i \in \mathcal{G}$ y α_i es su peso asociado para la construcción del clasificador final.

Esto puede verse como:

$$G^{(m)}(x) = \sum_{i=1}^{m} \alpha_i \, C_i(x) = G^{(m-1)}(x) + \alpha_m C_m(x)$$

Se define el criterio de error E como la pérdida exponencial en cada dato:

$$E = \sum_{n=1}^{N} \exp(-y_n \, G^{(m)}(x_n))$$

Como puede verse, si signo de clasificador $G^{(m)}(x_n)$ y la etiqueta y_i coinciden, la exponencial queda elevada a un valor final negativo, lo cual da un valor reducido. En cambio, si los signos difieren, la exponencial queda elevada a un valor positivo, lo cual le da un valor más alto. Por tanto, guarda lógica con el hecho de que las muestras mal clasificadas aporten más pérdida exponencial.

Aplicando la definición previa para $G^{(m)}$, la pérdida exponencial queda como:

$$E = \sum_{n=1}^{N} \exp(-y_n\, G^{(m-1)}(x_n) - y_n\, \alpha_m\, C_m(x_n)) =$$

$$\sum_{n=1}^{N} \exp(-y_n\, G^{(m-1)}(x_n)) \exp(-y_n\, \alpha_m\, C_m(x_n))$$

Ahora se define el peso $w_n^{(m)}$ para cada muestra x_n en función de la pérdida exponencial hasta la iteración previa, es decir:

$$w_n^{(m)} = \exp(-y_n\, G^{(m-1)}(x_n))$$

De esta forma, la pérdida exponencial para la iteración m queda definida por:

$$E = \sum_{n=1}^{N} w_n^{(m)}\, \exp(-y_n\, \alpha_m\, C_m(x_n))$$

Este sumatorio se puede separar entre las muestras bien clasificadas ($y_n \cdot C_m(x_n) = 1$) y mal clasificadas ($y_n \cdot C_m(x_n) = -1$), de forma que:

$$E = \sum_{y_n = C_m(x_n)} w_n^{(m)}\, \exp(-\alpha_m) + \sum_{y_n \neq C_m(x_n)} w_n^{(m)}\, \exp(\alpha_m)$$

Sumando y restando $\sum_{y_n \neq C_m(x_n)} w_n^{(m)}\, \exp(-\alpha_m)$ quedaría:

$$E = \sum_{y_n = C_m(x_n)} w_n^{(m)}\, \exp(-\alpha_m) + \sum_{y_n \neq C_m(x_n)} w_n^{(m)}\, \exp(-\alpha_m) +$$

$$\sum_{y_n \neq C_m(x_n)} w_n^{(m)}\, \exp(\alpha_m) - \sum_{y_n \neq C_m(x_n)} w_n^{(m)}\, \exp(-\alpha_m)$$

Los dos primeros términos recorren todas las muestras (las bien y las mal clasificadas), mientras que los dos últimos términos recorren sobre las muestras mal clasificadas y se puede sacar factor común $w_n^{(m)}$; de esta manera, nos queda:

$$E = \sum_{n=1}^{N} w_n^{(m)}\, \exp(-\alpha_m) + \sum_{y_n \neq C_m(x_n)} w_n^{(m)}\, (\exp(\alpha_m) - \exp(-\alpha_m))$$

Sobre esta definición ya estamos en condiciones de resolver el problema de optimización para C_m y para α_m.

2. Obtención del clasificador C_m

Partiendo de la expresión de pérdida exponencial obtenida:

$$E = \sum_{n=1}^{N} w_n^{(m)} \exp(-\alpha_m) + \sum_{y_n \neq C_m(x_n)} w_n^{(m)} \left(\exp(\alpha_m) - \exp(-\alpha_m)\right)$$

Se busca la minimización de E respecto a C_m. Se puede ver que el primer sumatorio es independiente de C_m, con lo que se puede ignorar. Respecto al segundo sumatorio, se puede asumir que $\exp(\alpha_m) - \exp(-\alpha_m)$ es constante. De esta forma, la pérdida exponencial puede aproximarse como:

$$E \approx \sum_{y_n \neq C_m(x_n)} w_n^{(m)}$$

Recurriendo a la definición de $w_n^{(m)}$, nos queda:

$$E \approx \sum_{y_n \neq C_m(x_n)} \exp(-y_n \, G^{(m-1)}(x_n))$$

Es decir, para minimizar E se debe escoger el clasificador C_m que minimice el error de clasificación sobre los datos ponderados según la iteración previa.

3. Obtención del peso α_m

Partiendo de la expresión de pérdida exponencial obtenida:

$$E = \sum_{n=1}^{N} w_n^{(m)} \exp(-\alpha_m) + \sum_{y_n \neq C_m(x_n)} w_n^{(m)} \left(\exp(\alpha_m) - \exp(-\alpha_m)\right)$$

En este caso, la minimización para α_m se hace derivando respecto a la misma e igualando a 0. De esta forma:

$$\frac{dE}{d\alpha_m} = -\sum_{n=1}^{N} w_n^{(m)} \exp(-\alpha_m) + \sum_{y_n \neq C_m(x_n)} w_n^{(m)} \exp(\alpha_m) + \sum_{y_n \neq C_m(x_n)} w_n^{(m)} \exp(-\alpha_m)$$

Descomponiendo el primer sumatorio en un sumatorio para muestras bien clasificadas y otro para muestras mal clasificadas, quedará:

$$\frac{dE}{d\alpha_m} = -\sum_{y_n=C_m(x_n)}^{N} w_n^{(m)}\ \exp(-\alpha_m) - \sum_{y_n\neq C_m(x_n)}^{N} w_n^{(m)}\ \exp(-\alpha_m)$$
$$+ \sum_{y_n\neq C_m(x_n)} w_n^{(m)}\ \exp(\alpha_m) \quad + \sum_{y_n\neq C_m(x_n)} w_n^{(m)}\ \exp(-\alpha_m)$$

Cancelando los dos sumatorios iguales (segundo y cuarto), queda finalmente:

$$\frac{dE}{d\alpha_m} = -\sum_{y_n=C_m(x_n)} w_n^{(m)}\ \exp(-\alpha_m) + \sum_{y_n\neq C_m(x_n)} w_n^{(m)}\ \exp(\alpha_m)$$

Igualando a cero:

$$-\sum_{y_n=C_m(x_n)} w_n^{(m)}\ \exp(-\alpha_m) + \sum_{y_n\neq C_m(x_n)} w_n^{(m)}\ \exp(\alpha_m) = 0$$
$$\sum_{y_n=C_m(x_n)} w_n^{(m)}\ \exp(-\alpha_m) = \sum_{y_n\neq C_m(x_n)} w_n^{(m)}\ \exp(\alpha_m)$$

Sacando la parte común de los sumatorios (las exponenciales, pues los pesos dependen de la y_n):

$$\exp(-\alpha_m) \sum_{y_n=C_m(x_n)} w_n^{(m)} = \exp(\alpha_m) \sum_{y_n\neq C_m(x_n)} w_n^{(m)}$$

Por tanto:

$$\frac{\sum_{y_n=C_m(x_n)} w_n^{(m)}}{\sum_{y_n\neq C_m(x_n)} w_n^{(m)}} = \frac{\exp(\alpha_m)}{\exp(-\alpha_m)} = \exp(2\alpha_m) \rightarrow 2\alpha_m = \ln\left(\frac{\sum_{y_n=C_m(x_n)} w_n^{(m)}}{\sum_{y_n\neq C_m(x_n)} w_n^{(m)}}\right)$$

$$\alpha_m = \frac{1}{2}\ \ln\left(\frac{\sum_{y_n=C_m(x_n)} w_n^{(m)}}{\sum_{y_n\neq C_m(x_n)} w_n^{(m)}}\right)$$

Se puede redefinir α_m en términos del error en la iteración m. Para ello, definimos:

$$\epsilon_m = \sum_{y_n\neq C_m(x_n)} w_n^{(m)}$$

entonces:

$$\alpha_m = \frac{1}{2}\ \ln\left(\frac{1-\epsilon_m}{\epsilon_m}\right)$$