

부스팅 방법론 (Boosting)

① 아이디어 부스트

- 이미 존재한 모형 집합을 사용하는 것임. ⇒ 관측치를 모형 집합에 포함될 개별 모형 선택.
- 약 분류기 (weak classifier): 부스팅 방법에서 성능이 떨어지는 개별 모형

- 개별 모형은 2중치형 1과 -1을 예측하는 이진 분류 모형이다. (Binary Classification)

$$K_m(x) = \begin{cases} 1 \\ -1 \end{cases}$$

- 최종 분류를 위한 "개별 모형의 집합"은 "위원회" (Committee) "C"라고 한다.
- 위원회 C는 다수결 방법으로 사용하지 않고, 각자의 개별 모형의 결과를 다음처럼 "가중 선형 결합"한 값을 판별함수로 사용한다.

$$C_{(m-1)}(x_i) = \text{Sign}(\alpha_1 k_1(x_i) + \dots + \alpha_{m-1} k_{m-1}(x_i))$$

- 부스팅 방법에서는 하나의 개별 모형에서 시작하여, 개별 모형을 하나씩 추가하여 위원회의 멤버 수를 증가시킨다. 이러한 방법을 stagewise Additive Modeling.
- 위원회가 될 개별 모형을 선택하는 방법으로는 지수 손실 함수 (Exponential Loss Function) L_i 를 사용한다.

$$E = \sum_{Y_i = K_m(x_i)} L_i^{(m)} = \sum_{Y_i = K_m(x_i)} e^{-Y_i C_{m-1}(x_i)}$$

- 위 식에서 $-Y_i C_{m-1}(x_i)$ 는 기존의 위원회의 예측값이 올바른 경우 -1, 틀린 경우 +1

$$Y_i = C_{m-1} \Rightarrow -Y_i C_{m-1} = -1$$

$$Y_i \neq C_{m-1} \Rightarrow -Y_i C_{m-1} = +1$$

- m번째 멤버의 모든 후보에 대해 위 손실 함수를 적용하여 가장 값이 작은 후보를 m번째 멤버로 선정한다.
- 멤버가 될 개별 모형이 선택된 후에는, 그 개별 모형에 대한 최선의 가중치 α_m 를 결정해야 한다.

$$E_m = \frac{\sum_{Y_i \neq K_m(x_i)} L_i^{(m)}}{\sum_{i=1}^N L_i^{(m)}}$$

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