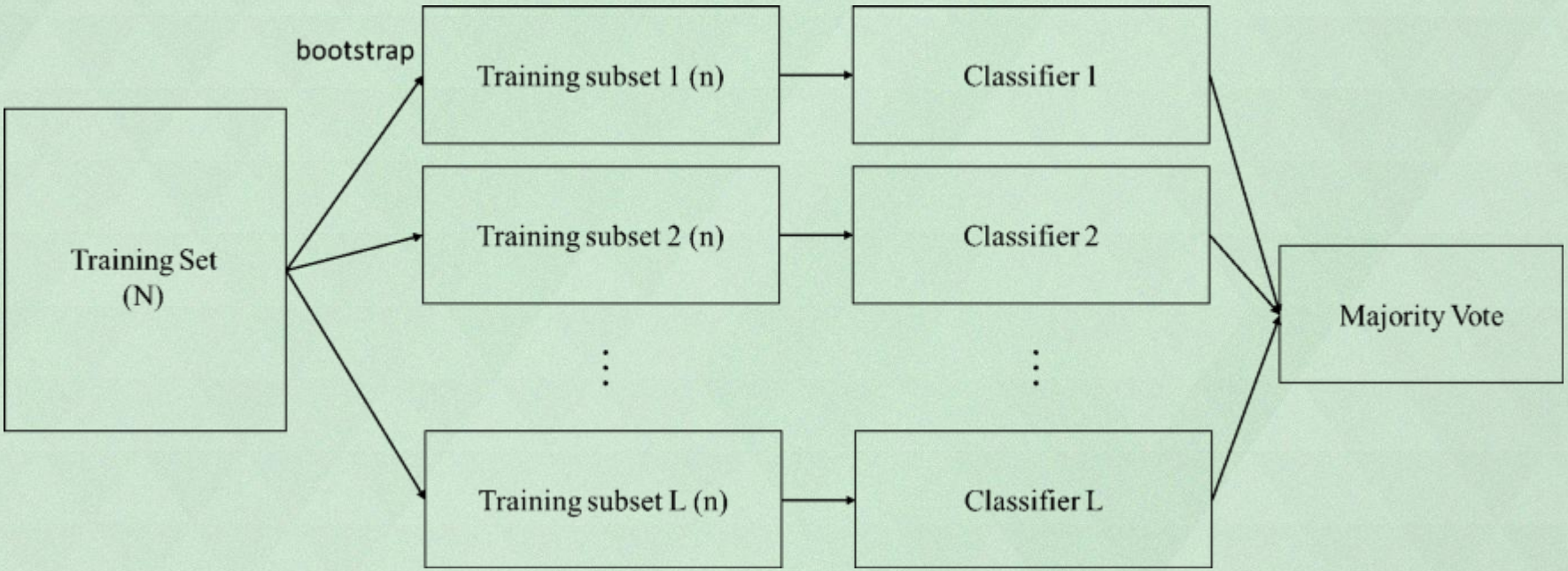
自適應提升(增強)樹
AdaBoost (adaptive boosting)

雲科財金系 張子溥 2018.11.21

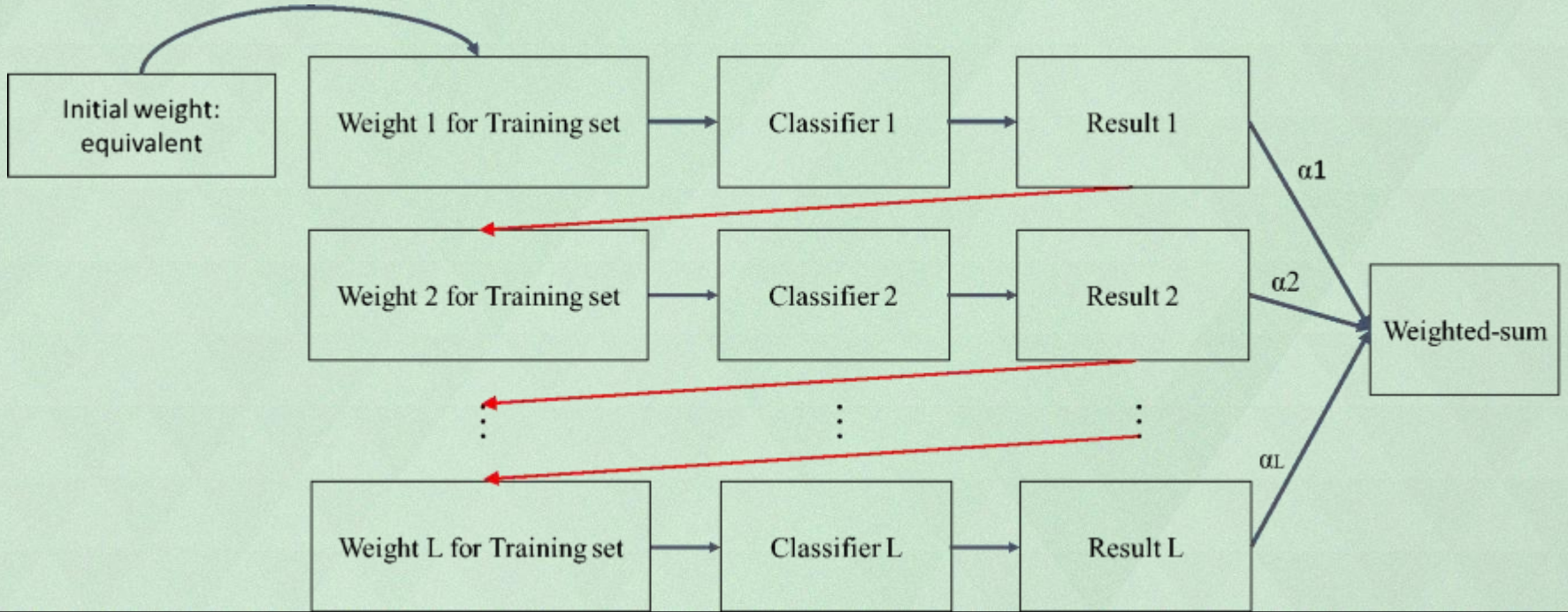
Bootstrap aggregating (Bagging)

- Bagging概念很簡單，從訓練資料中隨機抽取(取出後放回，n<N)樣本訓練多個分類器(要多少個分類器自己設定)，每個分類器的權重一致最後用投票方式(Majority vote)得到最終結果



Boosting

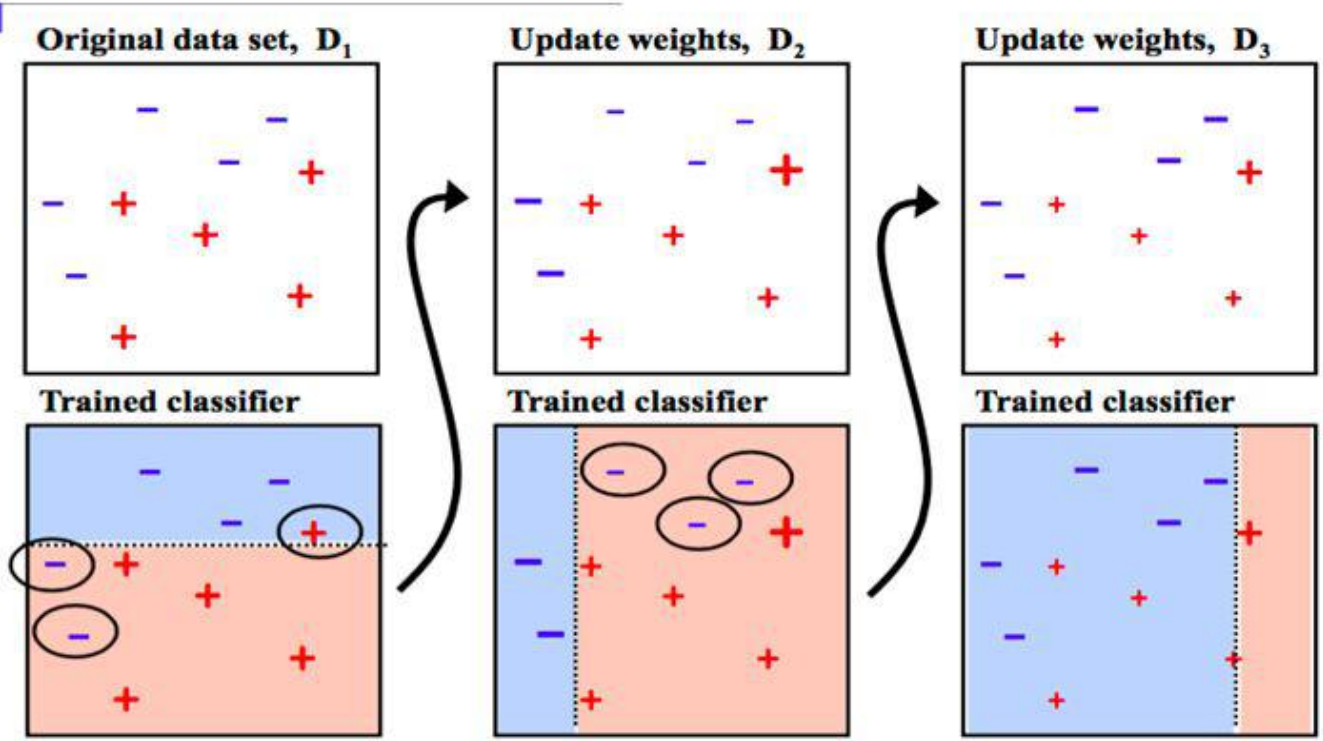
- Boosting算法是將很多個弱的分類器(weak classifier)進行合成變成一個強分類器(Strong classifier)，和Bagging不同的是分類器之間是有關聯性的，是透過將舊分類器的錯誤資料權重提高，然後再訓練新的分類器，這樣新的分類器就會學習到錯誤分類資料(misclassified data)的特性，進而提升分類結果。



AdaBoost

- AdaBoost算法，是一種改進的Boosting分類算法。方式是提高被前幾個分類器線性組合的分類錯誤樣本的權重，這樣做可以讓每次訓練新的分類器的時後都聚焦在容易分類錯誤的訓練樣本上。
- 每個弱分類器使用加權投票機制取代平均投票機制，只的準確率較大的弱分類器有較大的權重，反之，準確率低的弱分類器權重較低。
- **AdaBoost的手法**: 讓判斷錯誤的train data提高權重，讓產生新的權重的training set，但在新的分類器上就去加強學這些權重較大的training set。

Algorithm Adaboost - Example



courtesy to Alexander Ihler http://sli.ics.uci.edu/Classes/2012F-273a?action=download&upname=10-ensembles.pdf

AdaBoost algorithm flow

給定一組訓練資料 $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

假設第 k 次的權重為w_k^i，第一次分類器每個樣本的權重設為一樣$w_1^i = 1/n$。

假設我們要訓練 L 個分類器。

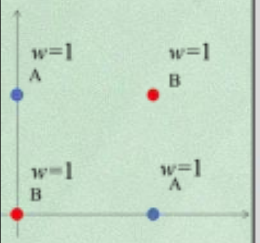
For k=1: L

1. 用權重w_k^i訓練 weak 分類器$f_k(x)$
2. ε_k為第 k 次分類器的訓練誤差
3. $\alpha_k = 0.5 * ln\left((1 - \varepsilon_k)/\varepsilon_k\right)$
4. $$w_{k+1}^i = \begin{cases} w_k^i * e^{\alpha_k} & if\ f_k(x_i) \neq y_i \\ w_k^i * e^{-\alpha_k} & if\ f_k(x_i) = y_i \end{cases}$$

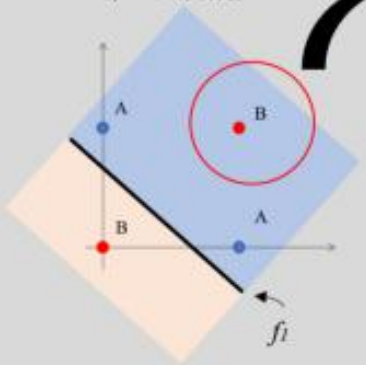
把L個分類器的結果作加權合成

$$H(x) = sign\left(\sum_{k=1}^{L} \alpha_k f_k(x)\right)$$

- 錯誤率越低的分類器在最後結果的合成上要佔較大的權重，下面假設錯誤率分別為0.1, 0.2, 0.4下的權重變化：
- ε_k=0.1, α_k=0.5*ln((1–0.1))/0.1)=1.0986
  ε_k=0.2, α_k=0.5*ln((1–0.2))/0.2)=0.6931
  ε_k=0.4, α_k=0.5*ln((1–0.4))/0.4)=0.2027



第一個分類器



四筆資料一筆紅色的B判錯

錯誤率: $\varepsilon 1 = 0.25$

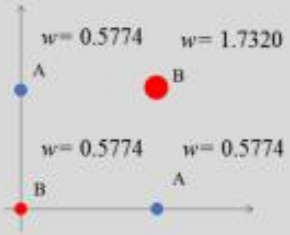
$d1 = 1.7321$

$\alpha 1 = 0.5493$

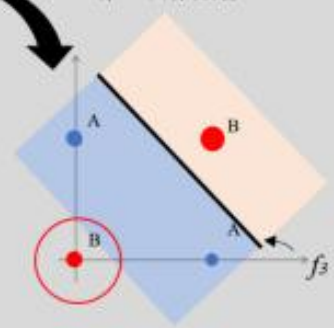
判對的權重更新: $1*e^{-\alpha_1}=0.5774$

判錯的權重更新: $1*e^{\alpha_1} = 1.7320$

權重更新



第二個分類器



四筆資料1筆判錯(1個紅色)

錯誤率: $\varepsilon 2 = \frac{\sum_i w_2^i \delta(f_2(x_i) \neq y_i)}{\sum_i w_2^i}$

=(0.5774)/(0.5774+0.5774+0.5774+1.7320)

= 0.1667

$d2 = 0.3727$

$\alpha 2 = 0.2603$

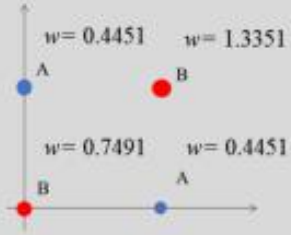
藍色判對的A權重更新: $0.5774*e^{-\alpha_2} = 0.4451$

藍色判對的A權重更新: $0.5774*e^{\alpha_2} = 0.4451$

紅色判對的B權重更新: $1.7320*e^{-\alpha_2} = 1.3351$

紅色判錯的B權重更新: $0.5774*e^{\alpha_2} = 0.7491$

權重更新

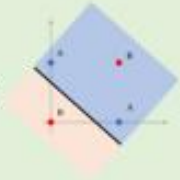
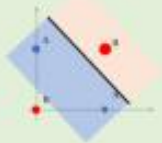


分類器合成

$\varepsilon_1 = 0.25, \alpha_1 = 0.5 * ln((1-0.25)/0.25) = 0.5493$

$\varepsilon_2 = 0.1667, \alpha_2 = 0.5 * ln((1-0.1667)/0.1667) = 0.8046$

$$H(x) = sign\left(\sum_{k=1}^{L} \alpha_k f_k(x)\right)$$

$H(x) = sign(0.5493$  $+ 0.8046$  $)$