Boosting是一族可将弱分类器提升为强分类器的算法。这族算法的工作机制如下：

先从初始训练集训练出一个(弱)分类器，再根据这个分类器的表现对训练样本分布进行调整，使得先前分类器错分的训练样本在后续得到更多的关注，然后基于调整后的样本分布来训练下一个分类器。如此重复进行，直至获得的分类器的数目达到事先给定的值T, 最终将这T个分类器进行加权整合，得到一个强分类器。

AdaBoost(Adaptive Boosting的缩写)是Boosting族算法中最著名的代表。

从偏差-方差权衡的角度看，AdaBoost主要关注降低偏差，因此AdaBoost能基于泛化性能相当弱的分类器构建出很强的集成分类器。

算法 1 (AdaBoost)

输入：训练集$D=\{(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_m,y_m)\}$; 分类器算法$\mathfrak{L}$; 训练轮数$T$;

过程:

(a) $\mathcal{D}_1(\boldsymbol{x})=1/m$.

(b) 对$t=1,\cdots,T$, 执行:

(c) $h_t=\mathfrak{L}(D,\mathcal{D}_t)$;

(d) $\epsilon_t=P_{\boldsymbol{x}\sim\mathcal{D}_t}(h_t(\boldsymbol{x})\neq f(\boldsymbol{x}))$;

(e) 如果$\epsilon_t>0.5$, 则停止; 否则, 继续执行;

(f) $\alpha_t=\frac{1}{2}\ln(\frac{1-\epsilon_t}{\epsilon_t})$;

(g) 令

$$\begin{aligned}\mathcal{D}_{t+1}&=\frac{\mathcal{D}_t(\boldsymbol{x})\exp(-\alpha_t f(\boldsymbol{x})h_t(\boldsymbol{x}))}{Z_t}\\&=\frac{\mathcal{D}_t(\boldsymbol{x})}{Z_t}\times\begin{cases}\exp(-\alpha_t), & \text{如果 } h_t(\boldsymbol{x})=f(\boldsymbol{x})\\ \exp(\alpha_t), & \text{如果 } h_t(\boldsymbol{x})\neq f(\boldsymbol{x})\end{cases},\end{aligned}$$

其中Z_t是某一常数;

(h) 循环结束.

输出: $H(\boldsymbol{x})=\operatorname{sign}(\sum_{t=1}^{T}\alpha_t h_t(\boldsymbol{x}))$.

弱可学习性是否等价于强可学习性?

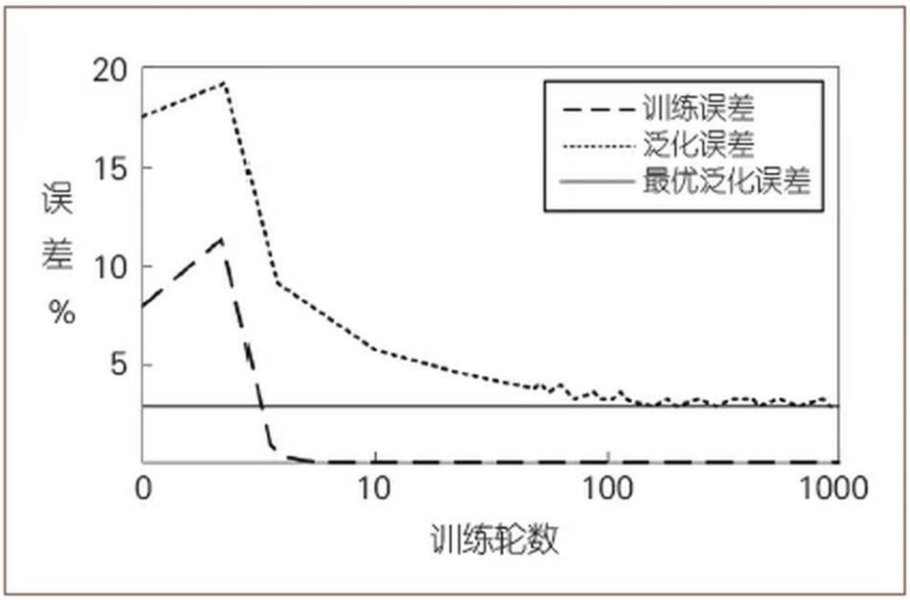
在人脸识别领域被誉为“第一个实时人脸检测器”就是基于 AdaBoost 研制的。

AdaBoost 的训练误差随训练轮数增加而指数级下降, 意味着算法收敛很快。对于泛化性能, 算法在处理新的、没见过的数据时的性能，AdaBoost 的泛化误差≤训练误差

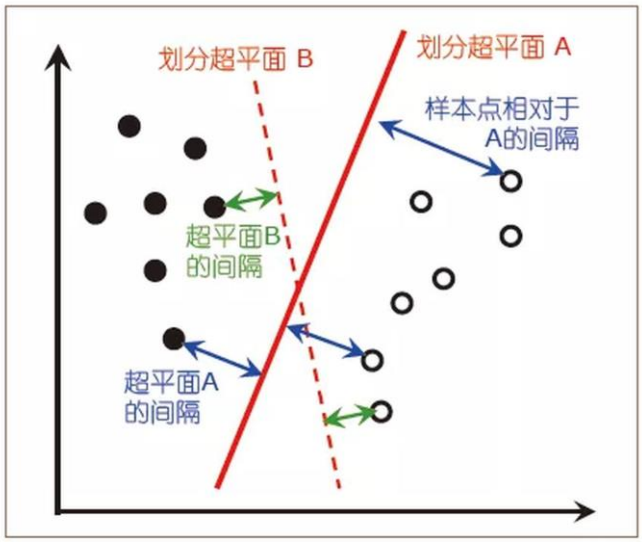
$$+\ \tilde{O}\left(\sqrt{\frac{\ln|\mathcal{H}|\,T}{m}}\right)$$

如果模型过于复杂，导致学得“过度”了，学到了本不该学的训练样本的“特性”而非样本总体的“共性”。训练样本多些好，模型复杂度小些好。

AdaBoost 呈现出一个奇异的现象：没有发生过拟合



在训练误差到达 0 之后继续训练，虽然模型复杂度在增大，但泛化误差却仍会继续下降。不符合奥卡姆剃刀原理

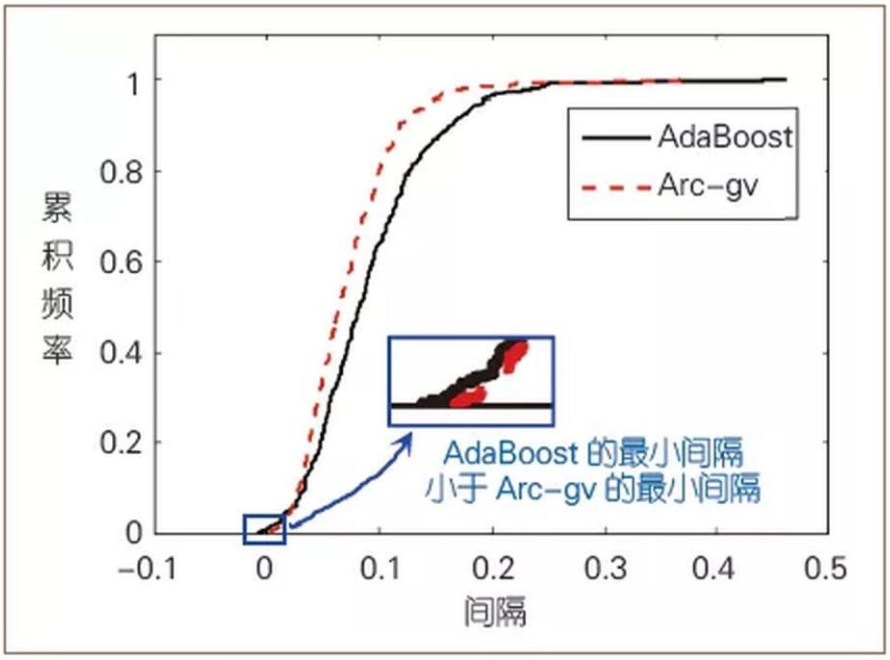


AdaBoost 为什么没有发生过拟合：这是因为即便训练误差达到 0，间隔仍有可能增大。如图超平面 B 已经把两类训练样本点完全分开，其训练误差为 0；继续训练可能找到超平面 A，训练误差仍为 0，但是 A 的间隔比 B 更大，所以泛化误差可以进一步减小。

最小间隔就是最优解，算法 Arc-gv。可以从理论上证明，这个算法能够找到最小间隔达即最优解。，实验却显示出 Arc-gv 的泛化误差大于 AdaBoost。

Boosting 间隔理论体系除了间隔，必然还会涉及到训练样本数和模型复杂度。要讨论间隔对泛化误差的影响，就必须把训练样本数和模型复杂度“固定”住。前者容易：指定训练样本的个数即可；后者则必须专门处理。

决策树模型复杂度由叶结点的数目决定，发现 AdaBoost 决策树虽然与 Arc-gv 决策树的叶结点数目相同，但树的层数却更多。



虽然 Arc-gv 的最小间隔始终大于 AdaBoost，但是若考虑样本总体，则 AdaBoost 的间隔比 Arc-gv 更大一些。从图中可以看到 AdaBoost 的曲线更“靠右”，这意味着有更多的样本点取得较大的间隔值。 “最小间隔”并非 Boosting 间隔理论体系的关键，重要的是间隔的总体分布。或许“间隔均值”或“间隔中位数”是关键物理量。

2008 年提出了“均衡间隔”的概念。2013 年，AdaBoost 在训练过程中随着轮数的增加，不仅使平均间隔增大，还同时使间隔方差变小。同时也意味着 AdaBoost 最终仍有可能发生过拟合，只不过很迟——当平均间隔已无法再增大、间隔方差也无法进一步减小时。若能最大化“平均间隔”同时最小化“间隔方差”，得到的分类器会更好

“AdaBoost 为何未发生过拟合”的答案就是“最大化平均间隔同时最小化间隔方差”