Notes 14: Weak and strong learning

1. Weak learning

Recall PAC learning definition (henceforth **strong** PAC learning):
Algorithm A PAC learns $\mathcal{C}$ if
  for any concept $c \in \mathcal{C}$ and any distribution $\mathcal{D}$ over X
  for **any** confidence parameter $\delta > 0$ and **any** accuracy parameter $\varepsilon > 0$
  when A takes m samples from $\mathrm{EX}(c, \mathcal{D})$
  with prob. $\geqslant 1 - \delta$, A outputs hypothesis with error $\leqslant \varepsilon$
A needs to work for **arbitrarily small** $\delta > 0$ and $\varepsilon > 0$: stringent requirement!
What if A only is guaranteed to work for **some** $\delta > 0$ and $\varepsilon > 0$? (much weaker guarantee)
Turns out A can be boosted to a strong learning algorithm

---

2. Boosting confidence

Suppose algorithm A, with probability $\geqslant 2/3$, outputs hypothesis with error $\leqslant \varepsilon$ (for any $\varepsilon > 0$)
  A's confidence δ bounded away from 0
Can be converted to strong PAC algorithm (with arbitrarily small δ and ε):

Strong PAC algorithm B
> Repeat $t = 1, \ldots, R$ times:
>   Run A on independent samples, with accuracy being $\varepsilon/2$, to get hypothesis h_t
> Draw m' more samples S to evaluate hypothese $h_1, \ldots, h_R$
> Output the hypothesis with least empirical error on S

$R \stackrel{\text{def}}{=} \frac{3}{2} \ln \frac{2}{\delta} = O\left(\ln \frac{1}{\delta}\right)$ so that

$$\mathbb{P}\left[\text{none of } h_1, \ldots, h_R \text{ has error} \leqslant \frac{\varepsilon}{2}\right] \leqslant \left(1 - \frac{2}{3}\right)^{3/2 \ln(2/\delta)} \leqslant e^{-\ln(2/\delta)} = \frac{\delta}{2}$$

$m' \stackrel{\text{def}}{=} O\left(\frac{1}{\varepsilon} \ln \frac{1}{\delta}\right)$ so that
Chernoff + Union Bound: with prob. $\geqslant 1 - \delta/2$,
  all bad hypotheses among $h_1, \ldots, h_R$ have empirical error $\geqslant \frac{5}{6}\varepsilon$; and
  some $\frac{\varepsilon}{2}$-accurate hypothesis among $h_1, \ldots, h_R$ has empirical error $\leqslant \frac{4}{6}\varepsilon$
Hence any hypothesis with least empirical error must have (true) error $\leqslant \varepsilon$
  Algorithm B succeeds with prob $\geqslant 1 - \delta$
A uses $m = \mathrm{poly}\left(\frac{1}{\varepsilon}\right)$ samples $\implies$ B uses $Rm + m' = \mathrm{poly}\left(\frac{1}{\varepsilon}, \ln \frac{1}{\delta}\right)$ samples
A runs in $T = \mathrm{poly}\left(\frac{1}{\varepsilon}\right)$ time $\implies$ B runs in $RT + m' \,\mathrm{poly}\left(\frac{1}{\varepsilon}\right) = \mathrm{poly}\left(\frac{1}{\varepsilon}, \ln \frac{1}{\delta}\right)$ time
Summary: $O\left(\ln \frac{1}{\delta}\right)$ calls to A; $O\left(\frac{1}{\varepsilon} \ln \frac{1}{\delta}\right)$ further samples to test the hypotheses

---

3. Boosting accuracy

Call algorithm A weak PAC learning algorithm with **advantage** γ if
  for any $c \in \mathcal{C}$, for any distribution $\mathcal{D}$, for any $\delta > 0$
  with probability $\geqslant 1 - \delta$, output hypothesis h with $\mathrm{err}_{\mathcal{D}}(h, c) \leqslant \frac{1}{2} - \gamma$
Getting advantage $\gamma = 0$ (i.e. $\mathrm{err}_{\mathcal{D}}(h, c) = \frac{1}{2}$) is trivial: just output uniformly random guess
Goal: Turn any weak PAC algorithm A with advantage γ into strong PAC algorithm
  with $\mathrm{poly}\left(\frac{1}{\gamma}, \frac{1}{\varepsilon}, \frac{1}{\delta}\right)$ overhead in #samples and running time
Will show efficient boosting algorithm B with following structure

Boosting algorithm B

Draw independent training samples $S = \{(x^1, c(x^1)), \ldots, (x^m, c(x^m))\}$ from $\mathrm{EX}(c, \mathcal{D})$
Initially set $\mathcal{D}_1 =$ uniform distribution over S
Repeat $t = 1, \ldots, R$ times:
    Run A on independent samples from $\mathrm{EX}(c, \mathcal{D}_t)$ to get hypothesis h_t
    Adjust $\mathcal{D}_t$ according to h_t to get updated distribution $\mathcal{D}_{t+1}$ over S
Combine hypotheses $h_1, \ldots, h_R$ to get hypothesis h

Missing details:
    What are $\mathcal{D}_2, \mathcal{D}_3, \ldots$?
    How to combine $h_1, \ldots, h_R$ into h?
    Why $\mathrm{err}_{\mathcal{D}}(h, c) \leqslant \varepsilon$?

History: Theory influenced practical algorithms!
    Kearns and Valiant (1989): introduced weak learning, showing weaking learning may still be hard
    Freund and Schapire (1990): weak and strong learning are equivalent in distribution-free setting
    Freund and Schapire (1995): AdaBoost, now part of many machine learning libraries