Chapter 8

8.1 $P(H(n) \le k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}$

Hoeffding inequality:

$$P\left(H(n) \le (p-\delta)n\right) \le e^{-2\delta^2 n} \qquad (8.44)$$

(another version:) if $P(x_i \in [a_i, b_i]) = 1$, $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$,

then $P(\bar{x} - E(\bar{x}) \ge t) \le \exp\left\{\frac{-2t^2n^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right\}$

$$P(|\bar{x} - E(\bar{x})| \ge t) \le 2\exp\left\{\frac{-2t^2n^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right\}.$$

① through (8.44),

$$P(H(x) \ne f(x)) = \sum_{k=0}^{\lfloor T/2 \rfloor} \binom{T}{k} (1-\varepsilon)^k \varepsilon^{(T-k)}$$

$$P(h_i(x) \ne f(x)) = \varepsilon$$

$$H(x) = \operatorname{sign}\left(\sum_{i=1}^{T} h_i(x)\right)$$

---

assume that $k^+(x) \triangleq \sum_{i=1}^{T} \max(h_i(x), 0)$, $k^-(x) = T - k^+(x)$

$$P(H(x) \ne f(x)) = P\left(k^+(x) \ge \frac{T}{2}\right)$$

$$= P\left(k^-(x) < \frac{T}{2}\right) \qquad \text{set } \frac{T}{2} = (\varepsilon - \delta)\cdot T, \text{ that is } \delta = \varepsilon - \frac{1}{2}$$

$$\le \exp\left\{-2\left(\frac{1}{2} - \varepsilon\right)^2 T\right\}$$

$$= \exp\left\{-\frac{1}{2}(1-2\varepsilon)^2 T\right\}.$$

② through another expression,

$X_i \triangleq h_i(x)$, $\bar{X} > 0$ denotes that $H(x) = f(x)$

$$P(H(x) \ne f(x)) = P(\bar{X} < 0) \qquad E(\bar{X}) = E(h_i(x)) = 1 - 2\varepsilon$$

$$= P\left(\bar{X} - E(\bar{X}) < 2\varepsilon - 1\right)$$

$$\le \exp\left\{\frac{-2(2\varepsilon-1)^2 \cdot T^2}{4T}\right\}$$

$$= \exp\left\{-\frac{1}{2}(1-2\varepsilon)^2 \cdot T\right\}. \quad \square$$

8.2 没看懂 $e(f(x)\cdot H(x))$ 是什么东西.

8.4 Gradient Boosting:

$\mathcal{L}$ is loss function, for example: $\mathcal{L}_{MSE}(y, F(x)) = \frac{1}{n}\cdot\sum_{i=1}^{n}(y_i - F(x_i))^2$

F^* is the estimation function,

$F^* = \arg\min_F \mathcal{L}(y, F(x))$

use steepest method, we have

$F_m = F_{m-1} - \gamma_m \cdot \left.\frac{\partial \mathcal{L}(y,F)}{\partial F}\right|_{F=F_{m-1}}$,

we need to calculate $h_m = \left.\frac{\partial \mathcal{L}(y_i,F)}{\partial F}\right|_{F=F_{m-1}}$, h_m is the basic learner.

1. initializing: $F_0(x_i) = \arg\min_\gamma \mathcal{L}(y_i, \gamma) \quad \forall i \le n$

2. $r_{im} = \left.\frac{\partial \mathcal{L}(y_i, F(x_i))}{\partial F(x_i)}\right|_{F=F_{m-1}(x_i)}$, $\forall i \le n$,

use $\{(x_i, r_{im})\}_{i=1}^{n}$ as data, to train h_m.

(Remark: if $h_m(x_i) = r_{im}$, $\forall i \le n$. $h_m(x_i)$ is the best learner)

3. $\gamma_m = \arg\min_\gamma \mathcal{L}(y, F_{m-1}(x) + \gamma h_m(x))$

4. $F_m(x) = F_{m-1}(x) + \gamma_m \cdot h_m(x)$.

Diff between Gradient Boosting and AdaBoost:

1. Gradient Boost 生成数据 $\{r_{im}\}$ 通过 steepest method and Loss function.

得到 h_m, 作为下一个 basic learner, 和 γ_m 参数

2. AdaBoost, 通过 Loss function 得到下一个数据集的 weight.

得到 basic function h_m 和参数 α_m.

不同点是, Gradient Boosting 类似于 AdaBoost 的推广版本,

可以选择不同的 梯度算法, 以及可以适用于 Data 不支持 weight 的情况.

(Steepest gradient, Newton method)

8.6 因为 Naive Bayesian Classifier 是通过

$P(x_i|c) \approx \frac{N(x_i,c)}{N(c)}$ 出现频率估计的，

Bagging 方法不能降低 $\frac{N(x_i,c)}{N(c)}$ 的方差，所以效果不好

（方差保持不变）

$$\begin{cases} Var\left(\frac{\sum \Pi(x_i|c)}{n}\right) = \frac{1}{n} Var(\Pi(x_i|c)) \text{ 以 sustained stably.} \\ P(x_i|c) = E(\Pi(x_i|c)). \end{cases}$$

8.8. MultiBoosting 算法 和 Iterative Bagging 算法：

1. MultiBoosting ： AdaBoost 作为基学习器易过拟合，(over fitted)

Bagging 选取的 Data 需要尽可能平均.

2. Iterative Bagging ： Bagging 作为基学习器，因为 Data 为 weighted data，

当 Data 样本量较小时，应当停止 Bagging 迭代，

当样本量较小时，容易欠拟合，(under fitted)