# 第一次作业

赵丰，2017310711

April 1, 2018

**P1** 对于给定的 $\epsilon, \delta$ 和模型 $c$, 以及分布 $D$, 产生了 $m$ 个平面的样本点，样本点的标签若落在 $c$ 内标 1，$c$ 外标 0。给出算法 $\mathcal{A}$ 求出 $c$ 的一个估计为包含所有标签为 1 的样本点的以原点为圆心的最小的圆。要证明对于 $m > (1/\epsilon)\log(1/\delta)$ 有下式成立

$$\Pr_{S\sim D^m}[R(h_s) > \epsilon] < \delta \tag{1}$$

首先可以知道 $\mathcal{A}$ 产生的估计在样本点有限的情况下总是含在集合 $c$ 内，$x$ 服从分布 $D$, 如果 $P(x \in c) <= \epsilon$, 我们总假设 $\epsilon$ 是小量，这种情况的含义是说我们取的分布 $D$ 产生的点大部分在 $c$ 外，是那些标签为 0 的点，而我们的算法只会把标签为 1 的点错判为标签为 0 的点，因为按照分布 $D$ 产生的标签为 1 的点很少，所以根据代价函数的定义，有 $R(h) <= \epsilon$（很小）恒成立，所以式(1)左边恒为 0，自然成立。

下面假设 $P(x \in c) > \epsilon$, 在这种情况下我们尝试估计 $\Pr_{S\sim D^m}[R(h_s) > \epsilon]$, 因为 $P(x \in c) > \epsilon$，所以 $\exists 0 < r_0 < r$, 使得 $P(x|r_0 \le ||x|| \le r) = \epsilon$。下面假设我们的 $\mathcal{A}$ 产生的圆的半径 $r_1$ 介于 $r_0$ 和 $r$ 之间，那么所有判错的点出现的概率小于等于 $P(x|r_1 \le ||x|| \le r)$ 小于等于 $P(x|r_0 \le ||x|| \le r) = \epsilon$, 这种情况同样得到(1)左边为 0。所以我们讨论最坏的情形，即 $r_1$ 比 $r_0$ 还小，在这种情况下事件 $R(h_s) > \epsilon$ 可以推出任给 m 个样本点产生的 $c$ 的估计的半径 $r_1 < r_0$, 我们计算一下后者的概率，也就是这 $m$ 个样本点没有一个落在 $r_0$ 和 $r$ 之间就会产生这个错判，对于单个 $x$，$P(x|\,||x|| < r_0 或 ||x|| > r) = 1 - \epsilon$, 又因为 $m$ 个样本是独立同分布的，所以估计的半径 $r_1 < r_0$ 的概率为 $(1-\epsilon)^m$, 根据两个随机事件的包含关系不难得到:

$$\Pr_{S\sim D^m}[R(h_s) > \epsilon] \le (1-\epsilon)^m \tag{2}$$

如果我们令 $(1-\epsilon)^m < \delta$，即满足了 $PAC$ 学习的定义，从而推出 $m > (1/\epsilon)\log(1/\delta)$

**P2** Gertrude 选圆的方法对 $c$ 的估计不一定能保证圆心在 $x_0$。Gertude 持有的命题的逆否命题为如果由样本点张成的目标圆与 $r_1, r_2, r_3$ 都有交集，那么 $P(x \in c) > \epsilon$ 小于 $\epsilon$。举一个反例：如图1所示，代价函数的取值为大圆面积减去小圆面积，差集不全含在 $r_1 \cup r_2 \cup r_3$ 内，而根据图中的面积比例差集的面积是有可能大于 $r_1 \cup r_2 \cup r_3$ 面积之和即 $\epsilon$ 的。

因此 Gertrude 的方法有误。

**P3** 本题证明 $\forall \delta > 0$, 存在不快于多项式增长的函数 $p(\frac{1}{\delta})$，使得当 $m > p(\frac{1}{\delta})$ 时下式成立：

$$\Pr_{S\sim D^m}(R(h_S) = 0) > 1 - \delta \tag{3}$$

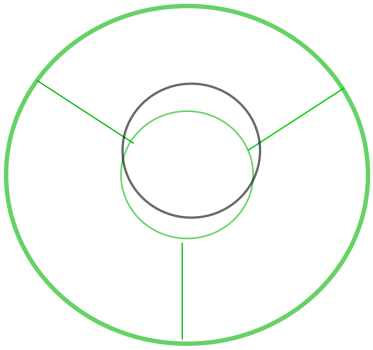

Figure 1: 目标圆在 $c$ 中的补不全落在 $r_1 \cup r_2 \cup r_3$ 示意图

设 $Z = \{(x_1, y_1), \ldots, (x_u, y_u)\}$，其中 $u = |Z|$, 设我们的分布 $D$ 是一般的 Multinomial 分布即满足 $P(X = x_i) = p_i$, 但我们这里假设 $0 < p_i < 1$。因为若某一个 $p_i$ 为零，即 sample 中不可能取到这个点，可以把它从 $Z$ 中剔除，如果某一个 $p_i = 1$，则由概率归一化条件，其他的 $p_j$ 全为零，问题是平凡的。所以我们可以取到 $0 < p_m = \min_{1 \le i \le u}\{p_i\} < 1$, 考虑到

$$R(h) = \Pr_{x \sim D}[h(x) \neq c(x)] = \sum_{i=1}^{u} 1_{h(x_i) \neq y_i} \tag{4}$$

所以 $R(h)$ 可能取到的最小值为 $p_m$，且只能从有限个离散的值中取。我们知道 PAC 学习算法找到的 $h_S$ 可以使得泛化误差尽可能小，即有 $\Pr_{S \sim D^m}(R(h_S) \le \epsilon) \ge 1 - \delta$, 对于 $m > p_1(\frac{1}{\epsilon}, \frac{1}{\delta})$ 成立。这里我们取小量 $\epsilon = \frac{p_m}{2}$，则使得 $\Pr_{S \sim D^m}(R(h_S) \le \epsilon)$ 相当于 $\Pr_{S \sim D^m}(R(h_S) = 0)$。因此我们得到多项式 $p = p_1(\frac{2}{p_m}, \frac{1}{\delta})$，当 $m > p$ 时我们有

$$\Pr_{S \sim D^m}(R(h_S) = 0) > 1 - \delta$$

**P4** 当 $t = 0$ 时结论显然成立; 若 $t > 0$, 记 $Y = X - \mu + \lambda$, 则 $Y$ 均值为 $\lambda$, 方差为 $\sigma^2$,

考虑 $P(Y \ge t + \lambda)$, 由 Markov 不等式我们有:

$$P(Y \ge t + \lambda) \le P(Y^2 \ge (t + \lambda)^2) \tag{5}$$

$$\le \frac{E(Y^2)}{(t + \lambda)^2} \tag{6}$$

$$= \frac{\sigma^2 + \lambda^2}{(t + \lambda)^2} \tag{7}$$

取 $\lambda = \frac{\sigma^2}{t}$，代入上式化简得到

$$P(X - \mu \geq t) = P(Y \geq t + \lambda) \leq \frac{\sigma^2}{t^2 + \sigma^2} \tag{8}$$