Notes for Fundations of Data Science 3

//** 笔记基本根据上课的顺序完成，但是对一些问题做了注释和补充 **//

1 机器学习与感知器算法

在 d 维空间中，最简单的封隔器是线性分隔器或半空间，详细的概念之后介绍。他们分类的标准是特征权值之和与阈值的大小关系（二分类）(阈值门/感知器)。拟合半空间或线性分隔器的问题包括 n 个样本 $x_1, x_2, \cdots, x_n$，在 d 维空间中，每个样本都有标签 1 和-1，任务是找到一个 d 维向量和阈值 t，使得标记为 +1 的 x_i:

$$\vec{w} \cdot \vec{x_i} > t \tag{1}$$

对于每个标记为-1 的 x_i,

$$\vec{w} \cdot \vec{x_i} < t \tag{2}$$

满足上述不等式的向量-阈值对 (w,t) 被称为**线性分隔器**。上述公式是一个关于未知数 w 和 t 的线性规划，可以通过通用的线性规划算法来解决。然而，当存在一个有很大“余地”或边际的可行解 w 时，感知器算法可以更快地得到结果。

我们首先进行技术上的修改，为每个 x_i 和 w 添加一个额外的坐标，写作 $\hat{x_i} = (x_i, 1)$ 和 $\hat{w} = (w, -t)$。假设 l_i 是 x_i 上 ±1 标签，不等式 (1)(2) 可以重写为:

$$(\hat{w} \cdot \hat{x_i}) l_i > 0 \qquad (1 \le i \le n)$$

这种变化使得原点被包含入。

感知器算法:

$w \longleftarrow 0$

`while there exists` x_i `with` $x_i l_i \cdot w \le 0$`, update` $w \longleftarrow w + x_i l_i$

显然对于每个 x_i 每次修改会使得 $(w \cdot x_i) l_i$ 增加值 $x_i \cdot x_i l_i^2 = |x_i|^2$，直观上对于某个参与变化的样本是有益的，但不一定对 x_j 有益。以下我们说明可以找到这种操作的步数的上界。如果权重 $w*$ 满足 $(w* \cdot x_i) l_i > 0$ 对于所有的 i，则任何样本 x_i 到样本分隔器 $w* \cdot x = 0$ 的最小距离称为样本分隔器的边距。将 w* 进行缩放，使得 $(w* \cdot x_i) l_i \ge 1$ 对于所有的 i 成立。那么分隔器的边距至少为 $\frac{1}{|w*|}$。如果所有的点都位于半径为 r 的球内，则 $r|w*|$ 是球的半径与边距的比例。下面的定理就给出了更新步骤次数的上界:

定理：
如果存在一个 w^* 满足 $(w^* \cdot x_i)l_i \geq 1$ 对于所有的 i，则感知器算法在最多 $r^2|w^*|^2$ 次更新中找到一个 w，使得 $(w \cdot x_i)l_i > 0$ 对于所有的 i，其中 $r = \max_i\{|x_i|\}$

定理证明：
假设 w^* 就是满足上述的，我们考察 $w^T w$ 和 $|w|^2$ 的变化，每次至少将 $w^T w^*$ 增加 1：

$$(w + l_i x_i)^T w^* = w^T w^* + x_i^T l_i w^* \geq w^T w^* + 1$$

中间的不等式是因为只有 x_i 满足 $l_i w \cdot x_i \geq 0$ 才会进行更新。
如果感知器算法进行了 m 次更新，则 $w^T w^* \geq m$，并且 $|w|^2 \leq mr^2$。于是，$|w||w^*| \geq m$ 和 $|w| \leq r\sqrt{m}$。
因此：

$$m \leq |w||w^*| \qquad \frac{m}{|w^*|} \leq |w| \qquad \frac{m}{|w^*|} \leq r\sqrt{m}$$

于是：

$$m \leq r^2|w^*|^2$$

即为所证。

2 一般的学习模型

/* 本部分相当于从机器学习的基础理论对教材和课程的一些补充 */

一般的统计学习模型
这里我们先定义一个规则，$h : \mathcal{X} \to \mathcal{Y}$，该函数也被称作预测器 (Predictor)、假设 (Hypothesis) 或分类器 (Classifier)。这个预测器可以用来预测一个新的领域的元素的标签。
分类误差：未能成功预测随机数据点正确标签的概率（随机数据点是从之前提到的潜在分布中生成的），$ERROR_h = Pr(h(x) \neq f(x))$。假设正确的标记函数是 $f(x)$，给定领域子集（Domain Subset）$A \subset \mathcal{X}$，假设概率分布为 $\mathcal{D}$，$\mathcal{D}(A)$ 决定了能够观测到 $x \in A$ 的概率。称 A 为一个事件，将其表达为函数 $\pi : \mathcal{X} \to \{0, 1\}$，也就是说，$A = \{x \in \mathcal{X} : \pi(x) = 1\}$。在这种情况下，也用 $\mathbb{P}_{x\sim\mathcal{D}}[\pi(x)]$ 表示 $\mathcal{D}(A)$，预测准则 (h:$\mathcal{X} \to \mathcal{Y}$) 的错误率被定义为：

$$L_{\mathcal{D},f}(h) \overset{def}{=} \mathbb{P}_{x\sim\mathcal{D}}[h(x) \neq f(x)] \overset{def}{=} \mathcal{D}(\{x : h(x) \neq f(x)\})$$

上述的误差也叫泛化误差、真实误差或损失。

经验风险最小化
首先定义训练误差：

$$L_s(h) \overset{def}{=} \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}$$

该误差也可被称为经验误差或经验风险。我们希望至少从预测器 h 出发能够最小化 $L_s(h)$，这个过程就交经验风险最小化（ERM）。但可能产生过拟合，为了修正 ERM 准则，应该对假设空间加以限制，这个限制就叫归纳偏置。**有限假设类**对于一个类，我们可以限制它的势的上界，即预测器的个数，这样修改后的假设类被称为有限假设类。我们希望说明的是拥有足够多的训练样本的时候，只要 $\mathcal{H}$ 是有限类，$ERM_{\mathcal{H}}$ 将不会过拟合。

可实现性假设： 存在 $h* \in \mathcal{H}$，使得 $L_{\mathcal{D},f}(h*) = 0$，这个假设意味着对于任意随机样本集 S（其中 S 中的实例是根据分布 $\mathcal{D}$ 随机采集，标签由 f 决定）以概率 1 使得 $L_S(h*) = 0$。

独立同分布假设（i.i.d.）： 训练集中的样本根据 $\mathcal{D}$ 独立同分布，然后根据标记函数 f 确定其标签，记为 $S \sim \mathcal{D}^m$，其中，m 是 S 的势，$\mathcal{D}^m$ 表示 m-组的概率，对于 m-组中的每一个元素，都是独立于组中的其他元素而从 $\mathcal{D}$ 中独立抽取的。

置信参数：将采样到非代表性样本的概率表示为 δ，同时 $1-\delta$ 在该预测中被称为置信参数（confidence parameter）。

精度参数（accuracy parameter），用于评估预测质量的参数，记作 ε。如果 $L_{\mathcal{D},f}(h_S) > \varepsilon$，则预测器是失败的。否则认为预测器是近似正确的。于是我们设定 m-组实例预测失败的概率上界，形式上，设 $S|_x = (x_1, x_2, \cdots, x_m)$ 为训练实例集，上界是:

$$\mathcal{D}^m(\{S|_x : L_{\mathcal{D},f}(h_s) > \varepsilon\})$$

设 $\mathcal{H}_B$ 为差的假设集:

$$\mathcal{H}_B = \{h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \varepsilon\}$$

设样本的误导集:

$$M = \{S|_x : \exists h \in \mathcal{H}_B, L_S(h) = 0\}$$

它对于所有误导集中的样本，存在一个差的预测器 h，使得其看上去是一个好的假设。由于假设的可实现性意味着 $L_S(h_S) = 0$，所以，预测器“差”的充要条件是样本全部位于误导样本集 M 中，形式上可以表示为:

$$\{S|_x : L_{\mathcal{D},f}(h_S) > \varepsilon\} \subseteq M$$

并且 M 可以写作:

$$M = \bigcup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\}$$

因此，

$$\mathcal{D}^m(\{S|_x : L_{\mathcal{D},f}(h_S) > \varepsilon\}) \leq \mathcal{D}^m(M) = \mathcal{D}^m(\bigcup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\}) \qquad \cdots (1)$$

（联合界 lemma）： 对于任意集合 A、B 以及分布 $\mathcal{D}$，有:

$$\mathcal{D}(A \bigcup B) \leq \mathcal{D}(A) + \mathcal{D}(B)$$

注意，上面表述中的 $\mathcal{D}$ 表示的是 Pr，即概率，不是先前说的分布。

这个引理的证明也显然的，除非每个事件两两相互独立，否则不能取等。利用联合界引理，我们修改公式 (1) 可得:

$$\mathcal{D}^m(\{S|_x : L_{\mathcal{D},f}(h_S) > \varepsilon\}) \leq \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S|_x : L_S(h) = 0\})$$

对于上述不等式，我们限制不等式右边的被加数。固定某“差”假设 $h \in \mathcal{H}_B$。$L_S(h) = 0$ 等同于 $\forall i$, $h(x_i) = f(x_i)$。由于样本的 i.i.d.，我们有:

$$\begin{aligned}\mathcal{D}^m(\{S|_x : L_S(h) = 0\}) &= \mathcal{D}^m(\{S|_x : \forall i, h(x_i) = f(x_i)\}) \\ &= \Pi_{i=1}^m \mathcal{D}(\{x_i : h(x_i) = f(x_i)\})\end{aligned}$$

对于训练集中的每个独立样本，有：

$$\mathcal{D}(\{x_i : h(x_i) = y_i\}) = 1 - L_{\mathcal{D},f}(h) \leq 1 - \varepsilon$$

利用不等式 $1 - \varepsilon \leq e^{-\varepsilon}$，对任意 $h \in \mathcal{H}_B$ 有：

$$\mathcal{D}^m(\{S|_x : L_S(h) = 0\}) \leq (1 - \varepsilon)^m \leq e^{-\varepsilon m}$$

结合前述不等式有：

$$\mathcal{D}^m(\{S|_x : L_{\mathcal{D},f}(h_S) > \varepsilon\}) \leq |\mathcal{H}_B|e^{-\varepsilon m} \leq |\mathcal{H}|e^{-\varepsilon m}$$

推论：设 $\mathcal{H}$ 是一个有限假设类，$\delta \in (0,1)$，$\varepsilon > 0$，m 为一个整数，以下不等式成立：

$$m \geq \frac{log(|\mathcal{H}|/\delta)}{\varepsilon}$$

从而对于任何标记函数 f，任何分布 $\mathcal{D}$，可实现性假设保证在独立同分布的样本集 S 上（S 的势为 m）最少以 $1 - \delta$ 的概率，对于每个 ERM 假设 h_S，成立：

$$L_{\mathcal{D},f}(h_S) \leq \varepsilon$$

PAC 学习理论

概率近似正确学习 (PAC 可学习)：若存在一个函数 $m_{\mathcal{H}} : (0,1)^2 \longrightarrow \mathbb{N}$ 和一个学习算法，使得对任意 $\varepsilon, \delta \in (0,1)$ 和 $\mathcal{X}$ 上的一个分布 $\mathcal{D}$，任意标号函数 $f : \mathcal{X} \longrightarrow \{0,1\}$，如果在 $\mathcal{H}, \mathcal{D}, f$ 下满足可实现性假设，那么当样本数量 $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ 时，其中样本由分布 $\mathcal{D}$ 独立同分布采样得到并且由函数 f 标记，算法将以不小于 $1 - \delta$ 的概率返回一个假设类 h，使该假设类 h 满足 $L_{\mathcal{D},f}(h) \leq \varepsilon$。这里对 PAC 的定义其实就是上一部分推论中指出的结果。

采样复杂度 由于 PAC 可学习中已经确定了能够概率近似正确的样本复杂度的下界，因此这是一个充分条件，只要达到这个下界就一定可以保证概率近似正确。因此合适的采样复杂度一定先于此达到；对于任何 PAC 可学习假设类，其采样复杂度满足：

$$m_{\mathcal{H}}(\varepsilon, \delta) \leq \left\lceil \frac{log(|\mathcal{H}|)/\delta}{\varepsilon} \right\rceil$$

贝叶斯最优预测器：给定 $\mathcal{X} \times \{0,1\}$ 上的任一概率分布 $\mathcal{D}$，将 $\mathcal{X}$ 映射到 $\{0,1\}$ 的最好的预测器是：

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & (\mathbb{P}[y = 1|x] \geq 1/2) \\ 0 & others \end{cases}$$

不可知 PAC 可学习

若存在一个函数 $m_{\mathcal{H}} : (0,1)^2 \longrightarrow \mathbb{N}$ 和一个学习算法 A，使得对于任意 $\varepsilon, \delta \in (0,1)$ 和 $\mathcal{X} \longrightarrow \mathcal{Y}$ 上的一个分布 $\mathcal{D}$，当样本的数量满足 $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ 时，其中样本由分布 $\mathcal{D}$ 独立同分布采样得到，算法将以不少于 $1 - \delta$ 的概率返回一个假设类 h，使该假设类满足：

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon$$

相当于保证了某个预测器在假设类中是比较好的。

广义损失函数

给定任意集合 $\mathcal{H}$ 和定义域 $\mathbb{Z}$，令 l为$\mathcal{H} \times \mathbb{Z} \longrightarrow \mathbb{R}^+$ 的一个映射函数，称这个函数为损失函数。损失函数被定义为分类器的期望损失，$h \in \mathcal{H}$，$\mathbb{Z}$ 上的概率分布 $\mathcal{D}$，即：

$$L_{\mathcal{D}(h)} \overset{def}{=} \mathop{\mathbb{E}}_{z\sim\mathcal{D}}[l(h,z)]$$

也就是说，目标 z 是从分布 $\mathcal{D}$ 上随机采样获得，为此考察假设类 h 在目标 z 下的期望损失，可以定义经验风险为给定数据集 $S = (z_1, \cdots, z_m) \in \mathbb{Z}^m$ 上的期望损失，即，

$$L_S(h) \overset{def}{=} \frac{1}{m}\sum_{i=1}^{m} l(h, z_i)$$

下面是两种常见的损失函数：

1. **0-1 损失：**随机变量 z 取值序列对集合 $\S \times \mathcal{Y}$，损失函数为：

$$l_{0-1}(h,(x,y)) \overset{def}{=} \begin{cases} 0 & h(x) = y \\ 1 & h(x) \neq y \end{cases}$$

2. **平方损失函数：**

$$l_{sq}(h,(x,y)) \overset{def}{=} (h(x) - y)^2$$

广义损失函数下的不可知 PAC 可学习

对于集合 $\mathbb{Z}$ 和损失函数 $l : \mathcal{H} \times \mathbb{Z} \longrightarrow \mathbb{R}^+$，若存在一个函数 $m_{\mathcal{H}} : (0,1)^2 \longrightarrow \mathbb{N}$ 和一个学习算法，使得对于任意 $\varepsilon, \delta \in (0,1)$，以及 $\mathbb{Z}$ 上的任一分布 $\mathcal{D}$，当样本数量满足 $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ 时，其中样本分布由分布 $\mathcal{D}$，独立同分布采样得到，算法将以不小于 $1-\delta$ 的概率返回一个函数假设类 h，使得该假设类 h 满足：

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon$$

其中 $L_{\mathcal{D}}(h) = \mathbb{E}_{z\sim\mathcal{D}}[l(h,z)]$。

> **可测量性** 在前面的定义中，对于任意 $h \in \mathcal{H}$，我们将 $l(h,\cdot) : \mathcal{H} \times \mathbb{Z} \longrightarrow \mathbb{R}^+$ 视为随机变量，定义 $L_{\mathcal{D}}(h)$ 为该随机变量的期望值；因此，我们需要要求 $l(h,\cdot)$ 是可测的。形式上，我们假定存在一个 $\mathbb{Z}$ 的 $\sigma-$ 代数子集，以及其上的概率分布 $\mathcal{D}$，$\mathbb{R}^+$ 的每个分割的原像在这个 $\sigma-$ 代数里，在 0-1 损失的二分类下，$\sigma-$ 代数在 $\mathcal{X} \times \{0,1\}$ 上，在 l 上的假设相当于假设对于任意的 h，集合 $\{(x, h(x)) : x \in \mathcal{X}\}$ 是 $\sigma-$ 代数。

> **完全与自主表示学习** 在前面的定义中，我们要求算法返回一个假设。在某些情况下，$\mathcal{H}$ 是 $\mathcal{H}'$ 的子集，损失函数可以拓展成一个从 $\mathcal{H}' \times \mathbb{Z}$ 到实数的函数；在这种情况下，我们允许算法返回一个假设 $h' \in \mathcal{H}'$，只要他满足 $L_{\mathcal{D}}(h') \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon$。允许算法从 $\mathcal{H}$ 返回一个假设，称为自主表示学习，完全学习要求算法必须从 $\mathcal{H}$ 中返回一个假设；自主表示学习有时也被称为“不完全学习”，尽管在自主表示学习中并不存在不恰当的情况。

3 学习的一致收敛性和奥卡姆剃刀

ε-代表性样本：如果满足下列不等式：

$$\forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon$$

一个训练集 S 就被称为 ε-代表性样本（关于定义域 $\mathcal{Z}$，假设类 $\mathcal{H}$，损失函数 l，和分布 $\mathcal{D}$）

$\varepsilon/2 - lemma$ 假设一个训练集 S 是 $\varepsilon/2$-代表性的，那么，对于任何一个 $ERM_{\mathcal{H}}(S)$ 的输出，即任意 $h_S \in \underset{h\in\mathcal{H}}{min} L_S(h)$ 都满足：

$$L_{\mathcal{D}}(h_S) \leq \underset{h\in\mathcal{H}}{min} L_{\mathcal{D}}(h) + \varepsilon$$

使用三角不等式可以直接证明。这个引理保证了 ERM 规则是一个不可知 PAC 学习器，应该满足至少在概率 $1-\delta$ 下随机选择一个训练集，它将是 ε-代表性训练集。一致收敛形式化了这个要求：

如果一个假设类 $\mathcal{H}$ 满足如下条件，那么它就有一致收敛性质：存在一个函数 $m^{UC}_{\mathcal{H}} : (0,1)^2 \longrightarrow \mathbb{N}$ 使得对于所有 $\varepsilon, \delta \in (0,1)$ 和在 $\mathbb{Z}$ 上的所有概率分布 $\mathcal{D}$，如果 S 是从 $\mathcal{D}$ 得到的一个独立同分布的满足 $m \geq m^{UC}_{\mathcal{H}}(\varepsilon, \delta)$ 的样本，那么，至少在概率 $1-\delta$ 下，S 是 ε-代表性的。

推论：如果类 $\mathcal{H}$ 关于函数 $m^{UC}_{\mathcal{H}}$ 有一致收敛性，那么这个类是样本复杂度为 $m_{\mathcal{H}}(\varepsilon, \delta) \leq m^{UC}_{\mathcal{H}}(\varepsilon/2, \delta)$ 的不可知 PAC 可学习的；而且在这种情况下，$ERM_{\mathcal{H}}$ 范式是关于 $\mathcal{H}$ 的成功的不可知 PAC 可学习的。

上面两处函数 m^{UC} 都表示是一致收敛（Uniform Convergence）的函数结果。同时结果说明了只要我们确定对于一个有限假设类，一致收敛成立，那么每个有限假设类都是不可知 PAC 可学习的。

对一致收敛性，考察两方面的证明：
固定 ε, δ，我们需要找到一个样本大小 m 以保证下面的条件成立：

- 对于任何 $\mathcal{D}$，至少在概率 $1-\delta$ 下，从 $\mathcal{D}$ 中采样得到的独立同分布的样本的选择 $S = (z_1, z_2, \cdots, z_m)$，对于所有的 $h \in \mathcal{H}$，$|L_S(h) - L_{\mathcal{D}}(H)| \leq \varepsilon$ 成立：

$$\mathcal{D}^m(\{S : \forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon\}) \geq 1 - \delta$$

- $\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_s(h) - L_{\mathcal{D}}(h)| > \varepsilon\}) < \delta$

注意到：

$$\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\} = \bigcup_{h\in\mathcal{H}} \{S : |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}$$

使用联合界引理：

$$D^m(\{S : \exists h \in \mathcal{H}, |L_s(h) - L_{\mathcal{D}}(h)| > \varepsilon\}) \leq \sum_{h\in\mathcal{H}} \mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}) \tag{2}$$

下面考察上述不等式右式子的每个加数，$|L_S(h) - L_{\mathcal{D}}(h)|$ 基本是 0，因为当样本足够大的时候绝对值中的前者实际上集中分布在后者周围。这是大数定理告诉我们的结果。为了提供更加真实的信息，我们引入 Hoeffding 界来衡量这种测度集中度：

Hoeffding 不等式： 令 $\theta_1, \theta_2, \cdots, \theta_m$ 是一个独立同分布的随机变量序列，假设对于所有的 i, $\mathbb{E}[\theta_i] = \mu$ 而且 $\mathbb{P}[a \leq \theta_i \leq b] = 1$，那么，对于所有的 $\varepsilon > 0$，成立：

$$\mathbb{P}\left[|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu| > \varepsilon\right] \leq 2e^{-\frac{2m\varepsilon^2}{(b-a)^2}}$$

证明：

lemma1：Hoeffding 引理 设 X 是一个随机变量，取值于区间 $[a, b]$ 且满足 $\mathbb{E}[X] = 0$，那么，对于任意的 $\lambda > 0$，有：

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2(b-a)^2}{8}}$$

下证该引理： 由于函数 $f(x) = e^{\lambda X}$ 是凸函数，故 $\forall \alpha \in (0,1)$ 和 $x \in [a, b]$，有：

$$f(x) \leq \alpha f(x) + (1-\alpha)f(b)$$

令 $\alpha = \frac{b-x}{b-a} \in [0,1]$，则：

$$e^{\lambda x} \leq \frac{b-x}{b-a}e^{\lambda a} + \frac{x-a}{b-a}e^{\lambda b}$$

取上式的期望，结合 $\mathbb{E}[X] = 0$，可得：

$$\mathbb{E}[e^{\lambda X}] \leq \frac{b-\mathbb{E}[X]}{b-a}e^{\lambda a} + \frac{\mathbb{E}[x]-a}{b-a}e^{\lambda b} = \frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b}$$

记 $h = \lambda(b-a), p = -\frac{a}{b-a}$且$L(h) = -hp + log(1-p+pe^h)$，则上式右边可以写成 $e^{L(h)}$。因此，只需 $L(h) \leq \frac{h^2}{8}$，利用泰勒公式，$L(0) = L'(0) = 0$且$L''(h) \leq 4$ 对所有 h 成立，立证。

回到 Hoeffding 不等式，记 $X_i = \theta_i - \mathbb{E}[\theta_i]$ 且 $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$，由指数函数的单调性和马尔可夫不等式，对 $\forall \lambda, \varepsilon > 0$，有：

$$\mathbb{P}[\bar{X} \geq \varepsilon] = \mathbb{P}[e^{\lambda\bar{X}} \geq e^{\lambda\varepsilon}] \leq e^{-\lambda\varepsilon}\mathbb{E}[e^{\lambda\bar{X}}]$$

由独立性假设有：

$$\mathbb{E}[e^{\lambda\bar{X}}] = \mathbb{E}[\prod_i e^{\frac{\lambda X_i}{m}}] = \prod_i \mathbb{E}[e^{\frac{\lambda X_i}{m}}]$$

利用引理 1，对任意 i 有：

$$\mathbb{E}[e^{\frac{\lambda X_i}{m}}] \leq e^{\frac{\lambda^2(b-a)^2}{8m^2}}$$

从而，

$$\mathbb{P}[\bar{X} \geq \varepsilon] \leq e^{-\lambda\varepsilon}\prod_i e^{\frac{\lambda(a-b)^2}{8m^2}} = e^{-\lambda\varepsilon + \frac{\lambda^2(b-a)^2}{8m}}$$

令 $\lambda = \frac{4m\varepsilon}{(b-a)^2}$，则：

$$\mathbb{P}[\bar{X} \geq \varepsilon] \leq e^{-\frac{2m\varepsilon^2}{(b-a)^2}}$$

类似地，对变量 $-\bar{X}$ 讨论，可得 $\mathbb{P}[\bar{X} \leq -\varepsilon] \leq e^{-\frac{2m\varepsilon^2}{(b-a)^2}}$，结合两方面即证。

回到一致收敛性的问题，注意到涉及的随机变量都是独立同分布的且 $L_S(h) = \frac{1}{m}\sum_{i=1}^{m}\theta_i$，$L_{\mathcal{D}}(h) = \mu$，

假设所有的 θ_i 均采样自 $[0,1]$，可得：

$$\mathcal{D}^m(\{S: L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}) = \mathbb{P}\left[|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu| > \varepsilon\right] \leq 2e^{-2m\varepsilon^2}$$

结合不等式 (2) 有：

$$\begin{aligned}\mathcal{D}^m(\{S: \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}) &\leq \sum_{h\in\mathcal{H}} 2e^{-2m\varepsilon^2} \\ &= 2|\mathcal{H}|e^{-2m\varepsilon^2}\end{aligned}$$

如果选择 $m \geq \frac{log(2|\mathcal{H}|/\delta)}{2\varepsilon^2}$，则有：

$$\mathcal{D}^m(\{S: \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}) \leq \delta$$

推论： 令 $\mathcal{H}$ 是一个有限假设类，$\mathbb{Z}$ 是一个定义域，并且令 $l: \mathcal{H}\times\mathbb{Z} \longrightarrow [0,1]$ 是一个损失函数，那么 $\mathcal{H}$ 具有手链性质，并且样本的复杂度是

$$m_{\mathcal{H}}^{UC}(\varepsilon,\delta) \leq \left\lceil \frac{log(2|\mathcal{H}|/\delta)}{\varepsilon^2} \right\rceil$$

并且利用 ERM 算法，这个类是不可知 PAC 可学习的，样本复杂度是：

$$m_{\mathcal{H}}(\varepsilon,\delta) \leq m_{\mathcal{H}}^{UC}(\varepsilon/2,\delta) \leq \left\lceil \frac{2log(2|\mathcal{H}|/\delta)}{\varepsilon^2} \right\rceil$$

奥卡姆剃刀 (Occam's Razor)：

奥卡姆剃刀原理指的是简单的解释往往优于复杂的解释，因为复杂的解释可能导致过拟合。奥卡姆剃刀原理在机器学习中可以数学化表示如下：

给定任何描述语言，考虑一个分布为 $\mathcal{D}$ 的样本 S，至少有 $1-\delta$ 概率使得任意满足 $L_S(h) = 0$ 的假设 h，如果可以用少于 b 比特描述，则当样本复杂度为：

$$m_{\mathcal{H}}(\varepsilon,\delta) = \frac{1}{\varepsilon}[blog2 + log(1/\delta)]$$

即可满足 $L_{\mathcal{D}}(h) \leq \varepsilon$。

推论：以至少 $1-\delta$ 的概率，所有满足 $L_S(h) = 0$ 且可以用少于 b 比特描述的假设 h，其真实错误率满足：

$$L_{\mathcal{D}}(h) \leq \frac{blog2 + log(1/\delta)}{m_{\mathcal{H}}(\varepsilon,\delta)}$$