# 机器学习（西瓜书）

# 注　解

（第 12 章　计算学习理论）

https://blog.csdn.net/jbb0523

# 前言

经常听人说南大周老师所著的《机器学习》（以下统称为西瓜书）是一本入门教材，是一本科普性质的教科书。在该书第十次印刷之际，周老师在“如何使用本书”中也提到“这是一本入门级教科书”。然而，本人读起来却感觉该书远不止“科普”“入门”那么简单，书中的很多公式需要思考良久方能推导，很多概念需要反复咀嚼才能消化。边读边想着是不是应该将自己学习时遇到的一些知识难点的解析分享出来，以帮助更多的人入门。自己的确也随手做过一些笔记，但由于怀疑这仅是自己的个别现象，毕竟读书期间，思考更多的是如何使用单片机、DSP、ARM、FGPA 等，而这些基本是不需要推导任何公式的，因此作罢。偶然间在周老师的新浪微博看到如下对话：


此时方知，可能“读不懂”并不是个别现象。因此决定写一本“西瓜书注解”或者称为“西瓜书读书笔记”，对自己研读西瓜书时遇到的“台阶”进行解释和推导，以帮助更多的人能够更快地进入到这个领域。另外，近期越来越强地意识到，扎扎实实地推导一些基础算法的公式，无论是对于理解算法本身机理还是进行学术研究，都是非常有必要的。

自己会根据个人学习进度和研究需要按章发布，不知道能不能坚持写完，加油！

毕竟自己也是一名初学者，所以可能一些概念解释并不完整、一些公式推导并不优美，甚至会存在错误，这是不可避免的，不接受漫骂，但欢迎将问题反馈给我，共同学习进步！

（网盘链接：https://pan.baidu.com/s/1QtEiNnk8jMzmbs0KPBN-_w）

# 第 12 章目录

# 第 12 章 计算学习理论

正如本章开篇所述，计算学习理论研究目的是分析学习任务的困难本质，为学习算法提供理论保证，并根据分析结果指导算法设计。例如，本章的定理 12.1、定理 12.3、定理 12.6 所表达意思的共同点是，泛化误差与经验误差之差的绝对值以很大概率$(1-\delta)$很小，且这个差的绝对值随着训练样本个数$(m)$的增加而减小，随着模型复杂度（定理 12.1 为假设空间包含的假设个数$|\mathcal{H}|$，定理 12.3 中为假设空间的 VC 维，定理 12.6 中为(经验)Rademacher 复杂度）的减小而减小。因此，若想要得到一个泛化误差很小的模型，足够的训练样本是前提，最小化经验误差是实现途径，另外还要选择性能相同的模型中模型复杂度最低的那一个；“最小化经验误差”即常说的经验风险最小化，“选择模型复杂度最低的那一个”即结构风险最小化，可以参见 6.4 节最后一段（第 133 页）的描述，尤其是式(6.42)所表达的含义。

## 12.1 基础知识

统计学中有**总体集合**和**样本集合**之分，比如要统计国内本科生对机器学习的掌握情况，此时**全国所有的本科生**就是总体集合，但总体集合往往太大而不具有实际可操作性，一般都是取总体集合的一部分，比如从双一流 A 类、双一流 B 类、一流学科建设高校、普通高校中各找一部分学生（即样本集合）进行调研，以此来了解国内本科生对机器学习的掌握情况。

在机器学习中，样本空间（参见 1.2 节）对应**总体集合**，而我们手头上的样例集$D$对应**样本集合**，样例集$D$是从样本空间中采样而得，分布$\mathcal{D}$可理解为当从样本空间采样获得样例集$D$时每个样本被采到的概率，我们用$\mathcal{D}(t)$表示样本空间第$t$个样本被采到的概率。

### 1、式(12.1)的解释

该式为泛化误差的定义，$P_{\boldsymbol{x}\sim\mathcal{D}}(h(\boldsymbol{x})\neq y)$表示从样本空间中按分布$\mathcal{D}$取一个样本$\boldsymbol{x}$，预测的类别标记$h(\boldsymbol{x})$不等于其真实类别标记$y$的概率，即泛化误差。在本章接下来的内容中，有时也会省略“$\boldsymbol{x}\sim\mathcal{D}$”而直接将此概率记为$P(h(\boldsymbol{x})\neq y)$，例如式(12.10)。

### 2、式(12.2)的解释

该式为经验误差的定义，符号$\mathbb{I}(\cdot)$的含义在书的目录之前的主要符号表中有解释。

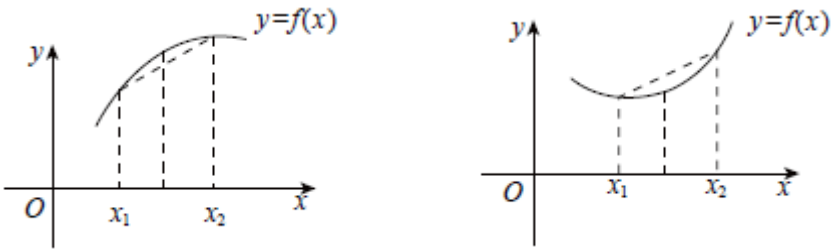
### 3、式(12.3)的解释

概率$P_{\boldsymbol{x}\sim\mathcal{D}}(h_1(\boldsymbol{x})\neq h_2(\boldsymbol{x}))$表示从样本空间中按分布$\mathcal{D}$取一个样本$\boldsymbol{x}$，映射$h_1,h_2$各自预测的类别标记$h_1(\boldsymbol{x}),h_2(\boldsymbol{x})$不相同的概率。

### 4、Jensen 不等式的解释

Jensen 不等式是一个很常用很重要的不等式，也很好理解，这里仅解释一下凹凸函数。

按式(12.4)的表达形式，凹函数和凸函数分别如下图的左图和右图所示：




凹函数满足$f\left(\frac{x_1+x_2}{2}\right) > \frac{f(x_1)+f(x_2)}{2}$，凸函数满足$f\left(\frac{x_1+x_2}{2}\right) < \frac{f(x_1)+f(x_2)}{2}$；但不同文献中有关凹凸函数定义有时正好相反，所以实际使用时要具体问题具体分析。

## 5、Hoeffding 不等式的解释

随机变量的观测值是随机的，进一步地，随机过程的每个时刻都是一个随机变量。

式中，$\frac{1}{m}\sum_{i=1}^{m} x_i$表示$m$个独立随机变量各自的某次观测值的平均，$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}(x_i)$表示$m$个独立随机变量各自的期望的平均。

式(12.5)表示事件$\frac{1}{m}\sum_{i=1}^{m} x_i - \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}(x_i) \geqslant \epsilon$出现的概率不大于$(\text{i.e.,} \leqslant) e^{-2m\epsilon^2}$；

式(12.6)的事件$\left|\frac{1}{m}\sum_{i=1}^{m} x_i - \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}(x_i)\right| \geqslant \epsilon$等价于以下事件：

$$\frac{1}{m}\sum_{i=1}^{m} x_i - \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}(x_i) \geqslant \epsilon \quad \vee \quad \frac{1}{m}\sum_{i=1}^{m} x_i - \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}(x_i) \leqslant -\epsilon$$

其中，$\vee$表示逻辑或（以上其实就是将绝对值表达式拆成两部分而已）。这两个子事件并无交集，因此总概率等于两个子事件概率之和；而$\frac{1}{m}\sum_{i=1}^{m} x_i - \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}(x_i) \leqslant -\epsilon$与式(12.5)表达的事情对称，因此概率相同。

Hoeffding 不等式表达的意思是$\frac{1}{m}\sum_{i=1}^{m} x_i$和$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}(x_i)$两个值应该比较接近，二者之差大于$\epsilon$的概率很小（不大于$2e^{-2m\epsilon^2}$）。

Hoeffding 不等式知道结论即可，如果对其证明感兴趣，可以参见 Hoeffding 在 1963 年发表的论文[Wassily Hoeffding (1963) Probability Inequalities for Sums of Bounded Random Variables, Journal of the American Statistical Association, 58:301, 13-30]的第 4 部分“4. PROOFS OF THE THEOREMS OF SECTION 2”，式(12.5)为第 2 部分“2. SUMS OF INDEPENDENT RANDOM VARIABLES”的 Theorem 1:

*Theorem 1.* If $X_1, X_2, \cdots, X_n$ are independent and $0 \leq X_i \leq 1$ for $i = 1, \cdots, n$, then for $0 < t < 1 - \mu$

$$\Pr\{\overline{X} - \mu \geq t\} \leq \left\{\left(\frac{\mu}{\mu+t}\right)^{\mu+t}\left(\frac{1-\mu}{1-\mu-t}\right)^{1-\mu-t}\right\}^n \tag{2.1}$$

$$\leq e^{-nt^2 g(\mu)} \tag{2.2}$$

$$\leq e^{-2nt^2}, \tag{2.3}$$

值得一提的是，原论文中并未找到式(12.6)。

除 1963 年原论文外，有关 Hoeffding 不等式还可以参见 CSDN 博客“机器学习数学原理（8）——霍夫丁不等式”（https://blog.csdn.net/z_x_1996/article/details/73564926）。

### 6、McDiarmid 不等式的解释

首先，函数$f$有$m$个输入变量，相比于$f(x_1,\dots,x_m)$，$f(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_m)$将第$i$个输入由$x_i$改为$x_i'$；

其次，$|f(x_1,\dots,x_m)-f(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_m)|$表示改变第$i$个输入后函数值发生的变化量；

最后，$\sup$表示上确界(supremum)；若不理解其含义，简单理解为求最大值就好了，但上确界和最大值的确是不一样的；例如，$f(x)=1-e^{-x}$的上确界为 1，但却没有最大值，我们只能说当$x$趋近于正无穷时，$f(x)$的极限为 1，但$f(x)$却取不到这个极限值。

由于$x_1,\dots,x_m$为随机变量，$f(x_1,\dots,x_m)$表示某次观测时对应的函数值；每次观测中，随机变量的取值不同则对应的函数值$f(x_1,\dots,x_m)$也不同，$\mathbb{E}(f(x_1,\dots,x_m))$表示函数值的期望。对比 Hoeffding 不等式，式(12.7)类似式(12.5)，式(12.8)类似式(12.6)。

## 12.2 PAC 学习

本节内容几乎都是概念，多读几遍，好好琢磨一下就好。

概率近似正确(Probably Approximately Correct, PAC)学习，可以读为[pæk]学习。

本节第 2 段讨论的目标概念，可简单理解为真实的映射函数；

本节第 3 段讨论的假设空间，可简单理解为学习算法不同参数时的存在，例如线性分类超平面$f(\boldsymbol{x})=\boldsymbol{w}^\top\boldsymbol{x}+b$，每一组$(\boldsymbol{w},b)$取值就是一个假设；

本节第 4 段讨论的可分的(separable)和不可分的(non-separable)，例如西瓜书第 100 页的图 5.4，若假设空间是线性分类器，则(a)(b)(c)是可分的，而(d)是不可分的；当然，若假设空间为椭圆分类器（分类边界为椭圆），则(d)也是可分的；

本节第 5 段提到的“等效的假设”指的是第 7 页图 1.3 中的 A 和 B 两条曲线都可以完美拟合有限的样本点，故称之为“等效”的假设；另外本段最后还给出了概率近似正确的含义，即“以较大概率学得误差满足预设上限的模型”。

定义 12.1 PAC 辨识的式(12.9)表示输出假设$h$的泛化误差$E(h)\leqslant\epsilon$的概率不小于$1-\delta$；即“学习算法$\mathfrak{L}$能以较大概率（至少$1-\delta$）学得目标概念$c$的近似（误差最多为$\epsilon$）”。

定义 12.2 PAC 可学习的核心在于，需要的样本数目$m$是$1/\epsilon,1/\delta,\mathrm{size}(\boldsymbol{x}),\mathrm{size}(c)$的多项式函数。

定义 12.3 PAC 学习算法的核心在于，完成 PAC 学习所需要的时间是$1/\epsilon,1/\delta,\mathrm{size}(\boldsymbol{x}),\mathrm{size}(c)$的多项式函数。

定义 12.4 样本复杂度指完成 PAC 学习过程需要的最少的样本数量，而在实际中当然也希望用最少的样本数量完成学习过程。

在定义 12.4 之后，抛出来三个问题：

(a)研究某任务在什么样的条件下可学得较好的模型？（定义 12.2）

(b)某算法在什么样的条件下可进行有效的学习？（定义 12.3）

(c)需多少训练样例才能获得较好的模型？（定义 12.4）

有限假设空间指$\mathcal{H}$中包含的假设个数是有限的，反之则为无限假设空间；无限假设空间更为常见，例如能够将图 5.4(a)(b)(c)中的正例和反例样本分开的线性超平面个数是无限多的。

# 12.3 有限假设空间

本节内容分两部分，第 1 部分“可分情形”时，可以达到经验误差$\widehat{E}(h)=0$，做的事情是以$1-\delta$概率学得目标概念的$\epsilon$近似，即式(12.12)；第 2 部分“不可分情形”时，无法达到经验误差$\widehat{E}(h)=0$，做的事情是以$1-\delta$概率学得$\min_{h\in\mathcal{H}} E(h)$的$\epsilon$近似，即式(12.20)。无论哪种情形，对于$h\in\mathcal{H}$，可以得到该假设的泛化误差$E(h)$与经验误差$\widehat{E}(h)$的关系，即“当样例数目$m$较大时，$h$的经验误差是泛化误差很好的近似”，即式(12.18)；实际研究中经常需要推导类似的泛化误差上下界。

**1、式(12.10)的解释**

第 1 个等号是由于$P(h(\boldsymbol{x})=y)+P(h(\boldsymbol{x})\neq y)=1$恒成立；
第 2 个等号是由于$P(h(\boldsymbol{x})\neq y)$就是泛化误差$E(h)$，即式(12.1)；
第 3 个等号是由于式(12.10)之前已假设泛化误差$E(h)>\epsilon$，因此$1-E(h)<1-\epsilon$。

**2、式(12.11)的解释**

由于$D$包含的$m$个样例是从$\mathcal{D}$独立同分布采样而得，因此

$$P\left((h(\boldsymbol{x}_1)=y_1)\wedge\ldots\wedge(h(\boldsymbol{x}_m)=y_m)\right)=\prod_{i=1}^{m}P\left(h(\boldsymbol{x}_i)=y_i\right)$$

其中，∧表示逻辑与，$\prod$表示连乘；上述推导使用了概率论中的结论：若事件 A 与事件 B 独立，则 P(AB)=P(A)P(B)，其中 P(AB)还表示为 P(A∧B)。

结合式(12.10)的结论（对分布$\mathcal{D}$上随机采样而得的任何样例均成立），则有

$$\prod_{i=1}^{m}P\left(h(\boldsymbol{x}_i)=y_i\right)=\prod_{i=1}^{m}\left(1-P(h(\boldsymbol{x}_i)\neq y_i)\right)<\prod_{i=1}^{m}(1-\epsilon)=(1-\epsilon)^m$$

**3、式(12.12)的推导**

①概率$P(h\in\mathcal{H}:E(h)>\epsilon\wedge\widehat{E}(h)=0)$表示：对于假设空间$\mathcal{H}$的假设$h$，事件$E(h)>\epsilon$和事件$\widehat{E}(h)=0$同时成立的概率；注意，此概率与$P(\exists h\in\mathcal{H}:E(h)>\epsilon\wedge\widehat{E}(h)=0)$不同，后者等价于如下概率（假设空间$\mathcal{H}$中共包含$|\mathcal{H}|$个假设）：

$$P\left(\left(E(h_1)>\epsilon\wedge\widehat{E}(h_1)=0\right)\vee\ldots\vee\left(E(h_{|\mathcal{H}|})>\epsilon\wedge\widehat{E}(h_{|\mathcal{H}|})=0\right)\right)$$

其中∃表示存在，$\mathcal{H}=\{h_1,h_2,\ldots,h_{|\mathcal{H}|}\}$，∨表示逻辑或。前者表示假设空间$\mathcal{H}$中的某个特定假设$h$满足$E(h)>\epsilon$和$\widehat{E}(h)=0$成立；后者表示假设空间$\mathcal{H}$中至少存在一个假设$h$满足$E(h)>\epsilon$和$\widehat{E}(h)=0$成立，如果每个假设$h$满足$E(h)>\epsilon$和$\widehat{E}(h)=0$成立的事件之间互斥，那么该概率就是每个假设$h$满足$E(h)>\epsilon$和$\widehat{E}(h)=0$成立的概率之和，即若事件 A 与事件 B 互斥，则 P(A+B)=P(A)+P(B)，其中 P(A+B)还表示为 P(A∨B)，更一般的结论则是 P(A+B)=P(A) + P(B)−P(AB)，其中 P(AB)还表示为 P(A∧B)，∧为逻辑与。

②根据式(12.11)可知，对于假设空间$\mathcal{H}$的某个假设$h$，事件$E(h)>\epsilon$和事件$\widehat{E}(h)=0$同时成立的概率$P(E(h)>\epsilon\wedge\widehat{E}(h)=0)<(1-\epsilon)^m$，其中事件$\widehat{E}(h)=0$表示经验误差为 0，即为式(12.11)中的概率$P$括号中的事件$(h(\boldsymbol{x}_1)=y_1)\wedge\ldots\wedge(h(\boldsymbol{x}_m)=y_m)$。

③对于式(12.12)前的一句话：“我们事先并不知道学习算法$\mathfrak{L}$会输出$\mathcal{H}$中的哪个假设，但仅需保证泛化误差大于$\epsilon$，且在训练集上表现完美的所有假设出现**概率之和**不大于$\delta$即可”，假设空间$\mathcal{H}$中共包含$|\mathcal{H}|$个假设（本节标题即为有限假设空间），**概率之和**如下：

$$P(h\in\mathcal{H}:E(h)>\epsilon\wedge\widehat{E}(h)=0)\leqslant\sum_{i=1}^{|\mathcal{H}|}P\left(E(h_i)>\epsilon\wedge\widehat{E}(h_i)=0\right)$$

④这$|\mathcal{H}|$个假设各自满足事件$E(h)>\epsilon$和事件$\widehat{E}(h)=0$成立的概率均小于$(1-\epsilon)^m$，即对于$\forall h_i\in\mathcal{H},P\left(E(h_i)>\epsilon\wedge\widehat{E}(h_i)=0\right)<(1-\epsilon)^m$，因此式(12.12)第 1 个小于号成立。

⑤因为$|\mathcal{H}|$和$m$均为正数，因此第 2 个小于号实际需要证明$(1-\epsilon)^m<e^{-m\epsilon}$，而这又等价于证明$(1-\epsilon)<e^{-\epsilon}$；

⑥令函数$f(\epsilon)=1-\epsilon-e^{-\epsilon}$，其中定义域为泛化误差$\epsilon\in[0,1]$；对函数求导得$f'(\epsilon)=-1+e^{-\epsilon}$，在定义域$\epsilon\in[0,1]$，$f'(\epsilon)\leqslant 0$成立，也就是说函数将$f(\epsilon)$是单调递减函数；而$f(0)=0$（最大值），因此$f(\epsilon)\leqslant 0$，即$(1-\epsilon)\leqslant e^{-\epsilon}$，等号当且仅当$\epsilon=0$时成立。

## 4、式(12.13)的解释

该式是“令式(12.12)不大于$\delta$”，就是一个定义而已。

## 5、式(12.14)的推导

①式(12.13)两边同时乘以$e^{m\epsilon}$，得$|\mathcal{H}|\leqslant\delta e^{m\epsilon}$，即$\delta e^{m\epsilon}\geqslant|\mathcal{H}|$

②$\delta e^{m\epsilon}\geqslant|\mathcal{H}|$两边同时取自然对数，得$\ln\delta+\ln e^{m\epsilon}\geqslant\ln|\mathcal{H}|$，即$m\epsilon\geqslant\ln|\mathcal{H}|-\ln\delta$，亦即$m\geqslant\frac{1}{\epsilon}(\ln|\mathcal{H}|+\ln\frac{1}{\delta})$；除移项等规则外，使用的对数性质包括$\ln(ab)=\ln a+\ln b$（第 1 个式子），$\ln e^x=x$（第 2 个式子）以及$-\ln x=\ln\frac{1}{x}$（第 3 个式子）。

## 6、引理 12.1 的解释

根据式(12.2)，$\widehat{E}(h)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(h(\boldsymbol{x}_i)\neq y_i)$，而指示函数$\mathbb{I}(\cdot)$取值非 0 即 1，也就是说$0\leq\mathbb{I}(h(\boldsymbol{x}_i)\neq y_i)\leq 1$；对于式(12.1)的$E(h)$实际上表示$\mathbb{I}(h(\boldsymbol{x}_i)\neq y_i)$为 1 的期望$\mathbb{E}(\mathbb{I}(h(\boldsymbol{x}_i)\neq y_i))$（泛化误差表示样本空间中任取一个样本，其预测类别不等于真实类别的概率），当假设$h$确定时，泛化误差固定不变，因此可记为$E(h)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}(\mathbb{I}(h(\boldsymbol{x}_i)\neq y_i))$。

此时，将$\widehat{E}(h)$和$E(h)$代入式(12.15)到式(12.17)，对比式(12.5)和式(12.6)的 Hoeffding 不等式可知，式(12.15)对应式(12.5)，式(12.16)与式(12.15)对称，式(12.17)对应式(12.6)。

## 7、式(12.18)的推导

令$\delta = 2e^{-2m\epsilon^2}$，则$\epsilon = \sqrt{\frac{\ln(2/\delta)}{2m}}$；式(12.17)可表示为$P\left(|E(h) - \widehat{E}(h)| \geqslant \epsilon\right) \leqslant \delta$，而这等价于$P\left(|E(h) - \widehat{E}(h)| \leqslant \epsilon\right) \geqslant 1 - \delta$，即$|E(h) - \widehat{E}(h)| \leqslant \epsilon$成立的概率不小于$1 - \delta$；注意，严格来说事件$|E(h) - \widehat{E}(h)| \geqslant \epsilon$的对立事件应该为事件$|E(h) - \widehat{E}(h)| < \epsilon$，这里暂且不讨论这种情况。

对于不等式$|E(h) - \widehat{E}(h)| \leqslant \epsilon$，其等价于

$$-\epsilon \leqslant E(h) - \widehat{E}(h) \leqslant \epsilon \Leftrightarrow \widehat{E}(h) - \epsilon \leqslant E(h) \leqslant \widehat{E}(h) + \epsilon$$

将$\epsilon = \sqrt{\frac{\ln(2/\delta)}{2m}}$代入即得式(12.18)。

## 8、定理 12.1 的推导

直接依次解释该定理证明中的每一步：

①$P(\exists h \in \mathcal{H} : |E(h) - \widehat{E}(h)| > \epsilon)$表示假设空间$\mathcal{H}$中（至少）存在（一个）假设$h$，使得事件$|E(h) - \widehat{E}(h)| > \epsilon$成立的概率；具体来说，在假设空间$\mathcal{H}$中，只存在一个假设$h$使得事件$|E(h) - \widehat{E}(h)| > \epsilon$成立也行，存在多个假设$h$使得事件$|E(h) - \widehat{E}(h)| > \epsilon$成立亦可。

②综上所述，事件$\exists h \in \mathcal{H} : |E(h) - \widehat{E}(h)| > \epsilon$等价于$|\mathcal{H}|$个事件的并集（逻辑或），即$|E(h_1) - \widehat{E}(h_1)| > \epsilon$，……，$|E(h_{|\mathcal{H}|}) - \widehat{E}(h_{|\mathcal{H}|})| > \epsilon$的并集。（回忆：对于事件 A 和事件 B 来说，P(A+B) = P(A) + P(B) - P(AB)，其中 P(A+B)可写为 P(A∨B)，P(AB)可写为 P(A∧B)；又因为概率肯定为非负值，即 P(AB)≥0，因此 P(A+B)≤P(A)+P(B)成立）

$$\begin{aligned}
&P\left(\exists h \in \mathcal{H} : \left|E(h) - \widehat{E}(h)\right| > \epsilon\right) \\
=&P\left(\left(\left|E_{h_1} - \widehat{E}_{h_1}\right| > \epsilon\right) \vee \ldots \vee \left(\left|E_{h_{|\mathcal{H}|}} - \widehat{E}_{h_{|\mathcal{H}|}}\right| > \epsilon\right)\right) \\
\leqslant &\sum_{h \in \mathcal{H}} P\left(\left|E(h) - \widehat{E}(h)\right| > \epsilon\right),
\end{aligned}$$

因此，上式中 “=” 成立是由于两个事件等价，“≤” 成立则因 P(A+B)≤P(A)+P(B)。

由于$\sum_{h \in \mathcal{H}} P(|E(h) - \widehat{E}(h)| > \epsilon) = \sum_{i=1}^{|\mathcal{H}|} P(|E(h_i) - \widehat{E}(h_i)| > \epsilon)$，而式(12.17)对任意$h \in \mathcal{H}$均成立，因此$\sum_{i=1}^{|\mathcal{H}|} P(|E(h_i) - \widehat{E}(h_i)| > \epsilon) \leqslant \sum_{i=1}^{|\mathcal{H}|} 2e^{-2m\epsilon^2} = 2|\mathcal{H}|e^{-2m\epsilon^2}$，即下式成立：

$$\sum_{h \in \mathcal{H}} P\left(\left|E(h) - \widehat{E}(h)\right| > \epsilon\right) \leqslant 2|\mathcal{H}| \exp(-2m\epsilon^2)$$

即$P(\exists h \in \mathcal{H} : |E(h) - \widehat{E}(h)| > \epsilon) \leqslant 2|\mathcal{H}|e^{-2m\epsilon^2}$成立。

事件$\exists h \in \mathcal{H}: |E(h)-\widehat{E}(h)| > \epsilon$的对立事件为假设空间$\mathcal{H}$中不存在假设$h$使$|E(h)-\widehat{E}(h)| > \epsilon$成立，即事件$\forall h \in \mathcal{H}: |E(h)-\widehat{E}(h)| \leqslant \epsilon$，$\forall$表示“任意”。结合刚刚证明得到的结论，可得$P(\forall h \in \mathcal{H}: |E(h)-\widehat{E}(h)| \leqslant \epsilon) \geqslant 1-2|\mathcal{H}|e^{-2m\epsilon^2}$。

令$\delta = 2|\mathcal{H}|e^{-2m\epsilon^2}$，推导可得$\epsilon$的表达式：

$$\Rightarrow \frac{\delta}{2} = |\mathcal{H}|e^{-2m\epsilon^2} \quad \Rightarrow \ln(\delta/2) = \ln|\mathcal{H}| + \ln e^{-2m\epsilon^2}$$

$$\Rightarrow -\ln(2/\delta) - \ln|\mathcal{H}| = -2m\epsilon^2 \quad \Rightarrow \ln(2/\delta) + \ln|\mathcal{H}| = 2m\epsilon^2$$

$$\Rightarrow \frac{\ln(2/\delta)+\ln|\mathcal{H}|}{2m} = \epsilon^2 \quad \Rightarrow \epsilon = \sqrt{\frac{\ln|\mathcal{H}|+\ln(2/\delta)}{2m}}$$

即$P\left(\forall h \in \mathcal{H}: |E(h)-\widehat{E}(h)| \leqslant \sqrt{\frac{\ln|\mathcal{H}|+\ln(2/\delta)}{2m}}\right) \geqslant 1-\delta$，也就是定理 12.1 的结论。

### 9、定义 12.5 的解释

注意，定义 12.5 属于 12.3.2 节“不可分情形”。这里的不可知 PAC 可学习与 12.2 节定义 12.2 中的 PAC 可学习相对应，因为对于“可分情形”，式(12.20)中的$\min_{h' \in \mathcal{H}} E(h') = 0$，即等价于定义 12.1 的式(12.9)。

“可分情形”中，学习算法$\mathfrak{L}$学得目标概念$c$的$\epsilon$近似；“不可分情形”中，学习算法$\mathfrak{L}$学得假设空间$\mathcal{H}$中泛化误差最小的假设（即$\arg\min_{h \in \mathcal{H}} E(h)$，表示在$\mathcal{H}$的所有假设中最好的那一个）的$\epsilon$近似。

## 12.4 VC 维

不同于 12.3 节的有限假设空间，从本节开始，本章剩余内容均针对无限假设空间。

### 1、式(12.21)的解释

本式给出增长函数(growth function)的定义。增长函数的概念很容易理解，表示假设空间$\mathcal{H}$对$m$个示例所能赋予标记的最大可能结果数。这里解释一下式(12.21)符号的含义。

对于包含$m$个示例的数据集$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$，还可以精练地记为$D = \{(\boldsymbol{x}_i, y_i) \mid 1 \leqslant i \leqslant m\}$；要注意地是，集合元素具有互异性，即“一个集合中，任何两个元素都认为是不相同的，即每个元素只能出现一次”（参见百度百科）。对于假设空间$\mathcal{H}$来说，每个假设$h$都能对$D$中的示例赋予一种标记结果$h|_D = \{(h(\boldsymbol{x}_1), h(\boldsymbol{x}_2), \ldots, h(\boldsymbol{x}_m))\}$，但其中或许有些标记结果相同，因此当写为$\{(h(\boldsymbol{x}_1), h(\boldsymbol{x}_2), \ldots, h(\boldsymbol{x}_m)) \mid h \in \mathcal{H}\}$（或$\{(h_i(\boldsymbol{x}_1), h_i(\boldsymbol{x}_2), \ldots, h_i(\boldsymbol{x}_m)) \mid 1 \leqslant i \leqslant |\mathcal{H}|\}$）时（即构成集合时），相同的标记结果会被去除掉，所得集合只含不同的标记结果。

符号$|\cdot|$表示集合的势(cardinality)，严格定义比较复杂，对于有限集合，简单理解为集合包含元素的个数即可，比如$|\mathcal{H}|$假设空间中的假设个数；作者在西瓜书开篇的主要符号表也提到$|\{\cdots\}|$表示集合$\{\cdots\}$中元素个数。

### 2、定理 12.2 的解释

西瓜书中只给出了定理内容，截止到第 30 印刷，定理内容如下：

**定理 12.2** 对假设空间 $\mathcal{H}$, $m \in \mathbb{N}$, $0 < \epsilon < 1$ 和任意 $h \in \mathcal{H}$ 有

$$P(|E(h) - \hat{E}(h)| > \epsilon) \leqslant 4\Pi_{\mathcal{H}}(2m)\exp\left(-\frac{m\epsilon^2}{8}\right). \tag{12.22}$$

值得一提的是，**该定理内容应该有误；具体来说应该定理中的“任意”改为“存在”**。

该定理在[Vapnik and Chervonenkis, 1971]（英译版，原文为俄语）中的形式如下：

**Theorem 2** *The probability that the relative frequency of at least one event in class S differs from its probability in an experiment of size l by more then $\varepsilon$, for $l \geqq 2/\varepsilon^2$, satisfies the inequality*

$$\mathbf{P}(\pi^{(l)} > \varepsilon) \leqq 4m^S(2l)e^{-\varepsilon^2 l/8}.$$

注意定理描述中使用的是“at least one event in class S”，因此应该是 class S 中“存在”one event 而不是 class S 中的“任意”event。

另外，该定理为基于增长函数对无限假设空间的泛化误差分析，与上一节有限假设空间的定理 12.1。在证明定理 12.1 的式(12.19)过程中，实际证明的结论是

$$P(\exists h \in \mathcal{H} : |E(h) - \hat{E}(h)| > \epsilon) \leqslant 2|\mathcal{H}|e^{-2m\epsilon^2}$$

根据该结论可得式(12.19)的原型（式(12.19)就是将$\epsilon$用$\delta$表示）：

$$P(\forall h \in \mathcal{H} : |E(h) - \hat{E}(h)| \leqslant \epsilon) \leqslant 1 - 2|\mathcal{H}|e^{-2m\epsilon^2}$$

这是因为事件$\exists h \in \mathcal{H} : |E(h) - \hat{E}(h)| > \epsilon$与事件$\forall h \in \mathcal{H} : |E(h) - \hat{E}(h)| \leqslant \epsilon$为对立事件。

注意到当使用$|E(h) - \hat{E}(h)| > \epsilon$表达时对应于“存在”，当使用$|E(h) - \hat{E}(h)| \leqslant \epsilon$表达时则对应于“任意”。

综上所述，式(12.22)使用$|E(h) - \hat{E}(h)| > \epsilon$，所以这里应该对应于“存在”。

## 3、式(12.23)的解释

在[Vapnik and Chervonenkis, 1971]中并未找到 VC 维的明确定义，不过话说一般也都是后人用作者的名字命名某定义或定理，很少有作者自己这么做的。

式(12.23)中的$\{m : \Pi_{\mathcal{H}}(m) = 2^m\}$表示一个集合，集合的元素是能使$\Pi_{\mathcal{H}}(m) = 2^m$成立的所有$m$；最外层的$\max$表示取集合的最大值。注意，这里仅讨论二分类问题。

VC 维的概念还是很容易理解的，有个常见的思维误区西瓜书也指出来了，即“这并不意味着所有大小为$d$的示例集都能被假设空间$\mathcal{H}$打散”，也就是说只要“存在大小为$d$的示例集能被假设空间$\mathcal{H}$打散”即可，这里的区别与前面“定理 12.2 的解释”中提到的“任意”与“存在”的关系一样。

例如对于图 12.1(a)，二维平面上的直线可以打散如此分布的大小为 3 的示例集，但并不是所有的大小为 3 的示例集都可以被打散，当三个示例分布在一条直线上时：

(a) (b) (c) (d)

则二维平面上的直线不能打散图(c)的情况（其实这与例 12.1 类似）。

对于大小为 4 的示例集，你能找出一种分布被直线打散么？不要想着把两个点重叠在一起，因为当这两个点标记不同时，任何假设空间都不能将其打散；也不要想着将三个点放在

一条直线上，前面已经说过这种情况无法打散。

## 4、引理 12.2 证明的解释

本引理的证明建议参考[Mohri et al., 2012]中第 45 页到第 47 页 Theorem 3.5 的证明（或第 2 版[Mohri et al., 2018]中第 41 页到第 42 页 Theorem 3.17 的证明），该书第 2 版下载链接 https://cs.nyu.edu/~mohri/mlbook/，该书第 1 版 PDF 自行网上搜索。

本引理证明应该是西瓜书作者参考整理[Mohri et al., 2012]的思路之后，换了一种更通俗易懂的方式呈现给读者的，从对$\mathcal{H}_{D'|D}$的解释可见一斑；书中证明最后一句话“由集合$D$的任意性，引理 12.2 得证”更是可以看出这一点；但对式(12.26)和式(12.27)的解释，尤其是式(12.27)的解释，又显得很难理解。接下来开始解释引理的证明。

首先，解释一下这里的数学归纳法的归纳假设的含义，即“假设定理对$(m-1,d-1)$和$(m-1,d)$成立”，这里的两个假设分别表示：①若假设空间$\mathcal{H}_1$的 VC 维为$d-1$，则对包含$m-1$个示例的示例集有$\Pi_{\mathcal{H}_1}(m-1)\leqslant\sum_{i=0}^{d-1}\binom{m-1}{i}$成立；②若假设空间$\mathcal{H}_2$的 VC 维为$d$，则对包含$m-1$个示例的示例集有$\Pi_{\mathcal{H}_2}(m-1)\leqslant\sum_{i=0}^{d}\binom{m-1}{i}$成立。基于这两个归纳假设，数学归纳法需要进一步证明的是：若假设空间$\mathcal{H}_3$的 VC 维为$d$，则对包含$m$个示例的示例集有$\Pi_{\mathcal{H}_3}(m)\leqslant\sum_{i=0}^{d}\binom{m}{i}$成立。这里用$\mathcal{H}_1$，$\mathcal{H}_2$，$\mathcal{H}_3$区别每个结论中的假设空间。

对于证明开始的$\mathcal{H}_{|D}$，$\mathcal{H}_{|D'}$，$\mathcal{H}_{D'|D}$，前两个容易理解，第三个可举例如下（$m=3$）：

若$\mathcal{H}_{|D}=\{(--+),(---),(-++),(-+-),(+-+),(++-)\}$

$\mathcal{H}_{|D'}=\{(--),(-+),(+-),(++)\}$

则$\mathcal{H}_{D'|D}=\{(--),(-+)\}$

其中，$\mathcal{H}_{D'|D}$定义式表示，在$\mathcal{H}_{|D'}$的元素当中（$(y_1,y_2,\ldots,y_{m-1})\in\mathcal{H}_{|D'}$），假设空间$\mathcal{H}$中存在假设$h,h'$，它们在$D'=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_{m-1}\}$上的预测相同（$h(\boldsymbol{x}_i)=h'(\boldsymbol{x}_i)=y_i$），但在$D$相比于$D'$新增的样本$\boldsymbol{x}_m$上的预测不同（$h(\boldsymbol{x}_m)\neq h'(\boldsymbol{x}_m)$）。在上例中，对于$\mathcal{H}_{|D'}$中的$(--)$，在$\mathcal{H}_{|D}$中有$(--+)$和$(---)$，前两个样本预测相同，但第 3 个样本预测不同，因此$\mathcal{H}_{D'|D}$包含元素$(--)$；同理，$\mathcal{H}_{D'|D}$还包含元素$(-+)$。

理解了$\mathcal{H}_{|D}$，$\mathcal{H}_{|D'}$，$\mathcal{H}_{D'|D}$的含义之后，易知式(12.25)显然是成立的；从以上例子中，也可以清晰地看到式(12.25)关系成立，即$6=4+2$。

在解释式(12.26)和式(12.27)之前，再一次讨论$\mathcal{H}_{|D}$，$\mathcal{H}_{|D'}$，$\mathcal{H}_{D'|D}$的含义。①$\mathcal{H}$表示假设空间，其包含的每个假设$h$可以对示例集赋予一种标记，$\mathcal{H}$中所有假设对示例集所能赋予的标记种类越多，则$\mathcal{H}$的表示能力越强；②$\mathcal{H}_{|D}$表示$\mathcal{H}$对当前示例集$D$所能赋予的所有可能的标记，如西瓜书第 275 页的定义；③进一步地，可以认为$\mathcal{H}_{|D}$为限制在$D$上由$\mathcal{H}$诱导的假设空间，英文表达为“the set of concepts $\mathcal{H}$ induces by restriction to $D$”；④由于限制在$D$上，则$\mathcal{H}_{|D}$的表示能力小于$\mathcal{H}$的表示能力，即$\mathrm{VC}(\mathcal{H}_{|D})\leqslant\mathrm{VC}(\mathcal{H})$；⑤对$\mathcal{H}_{|D'}$有类似于$\mathcal{H}_{|D}$相同的讨论，但由于$D'$相比于$D$缺第$m$个示例$\boldsymbol{x}_m$，因此$\mathcal{H}_{|D'}$表示能力小于$\mathcal{H}_{|D}$的表示能力，即$\mathrm{VC}(\mathcal{H}_{|D'})\leqslant\mathrm{VC}(\mathcal{H}_{|D})$；⑥对于$\mathcal{H}_{D'|D}$，它实际是$\mathcal{H}_{|D'}$的子集，若包含$m-1$个示例的示例集$Q$能被$\mathcal{H}_{D'|D}$打散，则$\mathcal{H}_{|D}$肯定可以打散示例集$Q\cup\{\boldsymbol{x}_m\}$，这是因为根据$\mathcal{H}_{D'|D}$的定义，样本$\boldsymbol{x}_m$可以被$\mathcal{H}_{|D}$赋予任何标记（本节只讨论二分类）；⑦即$\mathcal{H}_{|D}$能打散的示例个数比$\mathcal{H}_{D'|D}$能打散的示例个数多一个，根据 VC 维的定义有$\mathrm{VC}(\mathcal{H}_{D'|D})\leqslant\mathrm{VC}(\mathcal{H}_{|D})-1$。

重新给出两个归纳假设：“假设定理对$(m-1,d-1)$和$(m-1,d)$成立”，又已知条件$\mathrm{VC}(\mathcal{H})=d$，可得$\mathrm{VC}(\mathcal{H}_{|D'})\leqslant d$和$\mathrm{VC}(\mathcal{H}_{D'|D})\leqslant d-1$：

由增长函数的概念可知$|\mathcal{H}_{|D'}|\leqslant\Pi_{\mathcal{H}_{|D'}}(m-1)$，再结合第 2 个归纳假设“$(m-1,d)$

成立”及$\mathrm{VC}(\mathcal{H}_{|D'}) \leqslant d$，可得$\Pi_{\mathcal{H}_{|D'}}(m-1) \leqslant \sum_{i=0}^{d}\binom{m-1}{i}$，即式(12.26)。

同理，由增长函数的概念可知$|\mathcal{H}_{D'|D}| \leqslant \Pi_{\mathcal{H}_{D'|D}}(m-1)$，再结合第 1 个归纳假设“$(m-1,d-1)$成立”及$\mathrm{VC}(\mathcal{H}_{D'|D}) \leqslant d-1$，可得$\Pi_{\mathcal{H}_{D'|D}}(m-1) \leqslant \sum_{i=0}^{d-1}\binom{m-1}{i}$，即式(12.27)。

书中对式(12.26)和式(12.27)的推导略难理解。其实，书中式(12.26)的推导比较容易理解，但类似地理解式(12.27)就有些困难了，尤其是其中的$|\mathcal{H}_{D'|D}| \leqslant \Pi_{\mathcal{H}}(m-1)$，这里的增长函数中再使用$\mathcal{H}$就显得不是很恰当了；而$\mathcal{H}$的 VC 维为$d$，为什么又进一步进行缩放得到最终的结果就很让人费解了。而若将此处调整为$|\mathcal{H}_{D'|D}| \leqslant \Pi_{\mathcal{H}_{D'|D}}(m-1)$，再结合$\mathcal{H}_{D'|D}$的 VC 为不大于$d-1$，根据第 1 个归纳假设显然可进一步缩放后得式(12.27)。

最后的推导，将式(12.26)和式(12.27)的结论代入式(12.25)即可得第 1 行的“$\leqslant$”关系；第 2 行的“$=$”关系是将$\sum_{i=0}^{d-1}\binom{m-1}{i}$改写为$\sum_{i=0}^{d}\binom{m-1}{i-1}$，将求和符号拆开即可直观地看出后者多出来的一项$\binom{m-1}{-1}$，而此项在边注中已经注释到其值等于 0；最后的结论推导如下：

$$
\begin{aligned}
\binom{m-1}{i}+\binom{m-1}{i-1} &= \frac{(m-1)!}{(m-1-i)!i!}+\frac{(m-1)!}{(m-1-i+1)!(i-1)!} \\
&= \frac{(m-1)!(m-i)}{(m-i)(m-1-i)!i!}+\frac{(m-1)!i}{(m-1-i+1)!(i-1)!i} \\
&= \frac{(m-1)!(m-i)+(m-1)!i}{(m-i)!i!} \\
&= \frac{(m-1)!(m-i+i)}{(m-i)!i!}=\frac{(m-1)!m}{(m-i)!i!} \\
&= \frac{m!}{(m-i)!i!}=\binom{m}{i}
\end{aligned}
$$

## 5、推论 12.2 证明的解释

证明第 1 行的“$\leqslant$”：此即引理 12.2 的式(12.24)

证明第 2 行的“$\leqslant$”：由于$m \geqslant d \Rightarrow \frac{m}{d} \geqslant 1$

证明第 3 行的“$=$”：将与求和变量$i$无关的$\left(\frac{m}{d}\right)^d$提到求和号外面

证明第 4 行的“$\leqslant$”：由于$m \geqslant d$，所以将求和上限由$d$改为$m$时，其值会变大

证明第 5 行的“$=$”：由二项式定理$(x+y)^m=\sum_{i=0}^{m}\binom{m}{i}x^{m-i}y^i$，令$x=1,y=\frac{d}{m}$即得

证明第 6 行的“$\leqslant$”：由于$(1+\frac{d}{m})^m=(1+\frac{d}{m})^{\frac{m}{d}d}$，且$f(x)=(1+\frac{1}{x})^x$在定义域$(0,+\infty)$为单调递增函数，$\lim_{x\to+\infty}f(x)=e$，$e$表示自然常数，因此$f(\frac{m}{d})<e$，进而$(1+\frac{d}{m})^m<e^d$，代入即得结果；另外，这里的“等于”应该是取不到的~

有关$f(x)=(1+\frac{1}{x})^x$的单调性证明，可将函数写为$f(x)=e^{x\ln(1+\frac{1}{x})}$，证明导数大于 0.

在[Mohri et al., 2012]中讲解该推论之前，给出了两个结论，如下图所示：

The significance of Sauer's lemma can be seen by corollary 3.3, which remarkably shows that growth function only exhibits two types of behavior: either $\mathrm{VCdim}(H) = d < +\infty$, in which case $\Pi_H(m) = O(m^d)$, or $\mathrm{VCdim}(H) = +\infty$, in which case $\Pi_H(m) = 2^m$.

***Corollary 3.3***
*Let $H$ be a hypothesis set with $VCdim(H) = d$. Then for all $m \geq d$,*

$$\Pi_H(m) \leq \left(\frac{em}{d}\right)^d = O(m^d). \tag{3.30}$$

即对任意$m \in \mathbb{N}$，若有$\Pi_{\mathcal{H}}(m) = \mathcal{O}(m^d)$，则$\mathrm{VC}(\mathcal{H}) = d$（即上式推论 12.2 的结论）；若有$\Pi_{\mathcal{H}}(m) = 2^m$，则$\mathrm{VC}(\mathcal{H}) = +\infty$（根据定义 12.7 中 VC 维的定义易知）。

## 6、定理 12.3 证明的解释

将推论 12.2 代入定理 12.2，得$P(|E(h) - \widehat{E}(h)| > \epsilon) \leqslant 4(\frac{2em}{d})^d \exp(-\frac{m\epsilon^2}{8})$

令$4(\frac{2em}{d})^d \exp(-\frac{m\epsilon^2}{8}) = \delta$：

两边同乘$\frac{1}{\delta}\exp(\frac{m\epsilon^2}{8})$，得$\frac{4}{\delta}(\frac{2em}{d})^d = e^{\frac{m\epsilon^2}{8}}$

两边同时取自然对数$\ln(\cdot)$，得$\ln\frac{4}{\delta} + d\ln\frac{2em}{d} = \frac{m\epsilon^2}{8}$

两边同乘$\frac{8}{m}$然后再开方，得$\epsilon = \sqrt{\frac{8d\ln\frac{2em}{d} + 8\ln\frac{4}{\delta}}{m}}$

注意：式(12.22)中为$\Pi_{\mathcal{H}}(2m)$，结合式(12.28)则为$\Pi_{\mathcal{H}}(2m) \leqslant (\frac{2em}{d})^d$。

在证明结束后接下来一段话中提到，收敛速率为$O(\frac{1}{\sqrt{m}})$，其实分子也包含$m$，只是取了对数，变化不如分母快而已；最后提到的“分布无关、数据独立”意思是这个泛化误差界跟具体的数据集类型无关，只与数据集包含的样本个数$m$有关，例如正反样本 1:1 数据集和正反样本 1:10 数据集，只要包含的样本个数相同，这里的泛化误差界就是相同的。

## 7、定理 12.4 证明的解释

定理 12.4 的结论既适用于 PAC 可学习（12.2 节定义 12.2），亦适用于不可知 PAC 可学习（12.3 节定义 12.5），二者区别在于目标概念$c$是否属于假设空间$\mathcal{H}$。对于 PAC 学习，目标概念$c$属于假设空间$\mathcal{H}$，式(12.31)的$E(g) = 0$；对于不可知 PAC 学习，目标概念$c$不属于假设空间$\mathcal{H}$，式(12.31)的$E(g) \neq 0$。

注意，式(12.32)对应的两个式子对应前面的关键字“令”，即这里是符号的定义。

①根据推论 12.1，结合式(12.32)可得$\widehat{E}(g) - \frac{\epsilon}{2} \leqslant E(g) \leqslant \widehat{E}(g) + \frac{\epsilon}{2}$，再接下来提到的“至少以$1 - \delta/2$的概率成立”，实际也就是“至少以$1 - \delta'$的概率成立”；

②上式所述结论写为概率形式为$P\left(\left|E(g) - \widehat{E}(g)\right| \leqslant \frac{\epsilon}{2}\right) \geqslant 1 - \delta/2$，将绝对值拆开：

$$P\left(\left(E(g)-\widehat{E}(g)\leqslant\frac{\epsilon}{2}\right)\wedge\left(E(g)-\widehat{E}(g)\geqslant-\frac{\epsilon}{2}\right)\right)\geqslant 1-\delta/2$$

对于事件 A 和 B 来说，一定有 P(A)≥P(AB), P(B)≥P(AB)，P(AB)即 P(A∧B)，因此

$$P\left(E(g)-\widehat{E}(g)\leqslant\frac{\epsilon}{2}\right)\geqslant 1-\delta/2\text{和}P\left(E(g)-\widehat{E}(g)\geqslant-\frac{\epsilon}{2}\right)\geqslant 1-\delta/2$$

其中$E(g)-\widehat{E}(g)\geqslant-\frac{\epsilon}{2}$可等价写为$-E(g)\leqslant-\left(\widehat{E}(g)-\frac{\epsilon}{2}\right)$，在证明的最后一部分会用到！

③对于式(12.34)下方的$P\left(E(h)-\widehat{E}(h)\leqslant\frac{\epsilon}{2}\right)\geqslant 1-\frac{\delta}{2}$，之所以这里是$\frac{\epsilon}{2}$是因为式(12.34)的定义，之所以不等式右边是$1-\delta/2$是因为前面已令$\delta'=\delta/2$，**之所以相比式(12.29)少了绝对值，我认为是作者的笔误**，因为$P\left(E(h)-\widehat{E}(h)\leqslant\frac{\epsilon}{2}\right)$与$P\left(|E(h)-\widehat{E}(h)|\leqslant\frac{\epsilon}{2}\right)$肯定是不同的，事件$|E(h)-\widehat{E}(h)|\leqslant\frac{\epsilon}{2}$只是事件$E(h)-\widehat{E}(h)\leqslant\frac{\epsilon}{2}$的一部分而已，即

$$P\left(|E(h)-\widehat{E}(h)|\leqslant\frac{\epsilon}{2}\right)=P\left(\left(E(h)-\widehat{E}(h)\geqslant-\frac{\epsilon}{2}\right)\wedge\left(E(h)-\widehat{E}(h)\leqslant\frac{\epsilon}{2}\right)\right)$$

对于事件 A 和 B 来说，一定有 P(A)≥P(AB), P(B)≥P(AB)，P(AB)即 P(A∧B)，即事件$|E(h)-\widehat{E}(h)|\leqslant\frac{\epsilon}{2}$只是事件$E(h)-\widehat{E}(h)\leqslant\frac{\epsilon}{2}$与$E(h)-\widehat{E}(h)\geqslant-\frac{\epsilon}{2}$的交集部分。因此

$$P\left(E(h)-\widehat{E}(h)\leqslant\frac{\epsilon}{2}\right)\geqslant 1-\delta/2\text{和}P\left(E(h)-\widehat{E}(h)\geqslant-\frac{\epsilon}{2}\right)\geqslant 1-\delta/2$$

其中，$E(h)-\widehat{E}(h)\leqslant\frac{\epsilon}{2}$可等价写为$E(h)\leqslant\widehat{E}(h)+\frac{\epsilon}{2}$，在证明的最后一部分会用到！

④由$E(h)\leqslant\widehat{E}(h)+\frac{\epsilon}{2}$和$-E(g)\leqslant-\left(\widehat{E}(g)-\frac{\epsilon}{2}\right)$，两个不等式两侧相加，得

$$\begin{aligned}E(h)-E(g)&\leqslant\widehat{E}(h)+\frac{\epsilon}{2}-\left(\widehat{E}(g)-\frac{\epsilon}{2}\right)\\&=\widehat{E}(h)-\widehat{E}(g)+\epsilon\end{aligned}$$

根据证明开头所述的“假设$\mathfrak{L}$为满足经验风险最小化原则的算法，$h$为学习算法$\mathfrak{L}$输出的假设”，再结合由式(12.30)定义的满足 ERM 原则的算法，可以知道$h$是假设空间中经验风险最小的一个假设；虽然$g$表示$\mathcal{H}$中具有最小泛化误差的假设，但仍应有$\widehat{E}(h)\leqslant\widehat{E}(g)$，等价于$\widehat{E}(h)-\widehat{E}(g)\leqslant 0$，因此$\widehat{E}(h)-\widehat{E}(g)+\epsilon\leqslant\epsilon$ 。

⑤$E(h)\leqslant\widehat{E}(h)+\frac{\epsilon}{2}$和$-E(g)\leqslant-\left(\widehat{E}(g)-\frac{\epsilon}{2}\right)$均“以至少$1-\delta/2$的概率成立”，两个不等式两侧相加所得不等式将“以至少$1-\delta$的概率成立”，原因参见如下定理：

【定理】已知事件$a\leqslant b$成立的概率不小于$1-\frac{\delta}{2}$，事件$c\leqslant d$成立的概率不小于$1-\frac{\delta}{2}$，则事件$a+c\leqslant b+d$成立的概率不小于$1-\delta$。

【分析】已知条件可概括为$P(a\leqslant b)\geqslant 1-\frac{\delta}{2},\ P(c\leqslant d)\geqslant 1-\frac{\delta}{2}$，

求证问题可概括为$P(a+c\leqslant b+d)\geqslant 1-\delta$

【证明】将事件$a\leqslant b$和$c\leqslant d$同时成立记为$a\leqslant b\cap c\leqslant d$（交集）

将事件$a \leqslant b$和$c \leqslant d$至少有一个成立记为$a \leqslant b \cup c \leqslant d$（并集）

则

$$\begin{aligned} P(a \leqslant b \cap c \leqslant d) &= P(a \leqslant b) + P(c \leqslant d) - P(a \leqslant b \cup c \leqslant d) \\ &\geqslant (1 - \frac{\delta}{2}) + (1 - \frac{\delta}{2}) - P(a \leqslant b \cup c \leqslant d) \\ &\geqslant 2 - \delta - 1 = 1 - \delta \end{aligned}$$

其中，第 1 行就是事件交集与并集的关系；

第 2 行将已知条件代入；

第 3 行是由于概率$P(a \leqslant b \cup c \leqslant d)$不大于 1；

又因为$P(a + c \leqslant b + d) \geqslant P(a \leqslant b \cap c \leqslant d)$，这是由于若$a \leqslant b$和$c \leqslant d$同时成立，则$a + c \leqslant b + d$肯定成立，但$a + c \leqslant b + d$成立并不能得出$a \leqslant b$和$c \leqslant d$同时成立，即$a \leqslant b$和$c \leqslant d$同时成立是$a + c \leqslant b + d$成立的充分非必要条件。

综上可得，$P(a + c \leqslant b + d) \geqslant 1 - \delta$，证毕！

根据(不可知)PAC 可学习（定义 12.2 和定义 12.5）的定义，要求所需样本数量$m$是$1/\epsilon, 1/\delta, \text{size}(\boldsymbol{x}), \text{size(c)}$的多项式函数，而这可以从式(12.32)和式(12.34)中得出。

# 12.5 Rademacher 复杂度

上一节中介绍的基于 VC 维的泛化误差界是分布无关、数据独立的，本节将要介绍的 Rademacher 复杂度则在一定程度上考虑了数据分布。

## 1、式(12.36)的解释

第 1 个等号：即式(12.2)经验误差的定义；

第 2 个等号：本章开篇第 2 段已提到“本章主要讨论二分类问题，若无特别说明，$y_i \in \mathcal{Y} = \{-1, +1\}$”，即本式中的$y_i$和$h(\boldsymbol{x}_i)$的取值均为$-1, +1$两个值；因此，当$y_i$和$h(\boldsymbol{x}_i)$取值相同（即预测正确）时，$y_i h(\boldsymbol{x}_i) = +1$，反之（即预测错误），$y_i h(\boldsymbol{x}_i) = -1$。综合以上讨论，$\mathbb{I}(h(\boldsymbol{x}_i) \neq y_i) = \frac{1 - y_i h(\boldsymbol{x}_i)}{2}$，即第 2 个等号成立；

第 3 个等号：$\sum_{i=1}^{m} \frac{1 - y_i h(\boldsymbol{x}_i)}{2} = \sum_{i=1}^{m} \frac{1}{2} - \sum_{i=1}^{m} \frac{y_i h(\boldsymbol{x}_i)}{2} = \frac{m}{2} - \frac{1}{2} \sum_{i=1}^{m} y_i h(\boldsymbol{x}_i)$。

## 2、式(12.37)的解释

根据式(12.36)，显然有

$$\begin{aligned} \arg\min_{h \in \mathcal{H}} \widehat{E}(h) &= \arg\min_{h \in \mathcal{H}} \frac{1}{2} - \frac{1}{2m} \sum_{i=1}^{m} y_i h(\boldsymbol{x}_i) \\ &= \arg\max_{h \in \mathcal{H}} \frac{1}{2m} \sum_{i=1}^{m} y_i h(\boldsymbol{x}_i) \end{aligned}$$

其中，以上推导过程类似于$f(x) = 1 - 2x^2$的极大值点位置与$g(x) = x^2$极小值点位置相同，均为$x = 0$；符号 arg max 的含义参见：https://en.wikipedia.org/wiki/Arg_max，表示求解函数最大值点对应的自变量取值，arg min 是其反义符号。

## 3、式(12.38)的解释

相比于式(12.37)，样例真实标记$y_i$换为了 Rademacher 随机变量$\sigma_i$，$\arg\max_{h\in\mathcal{H}}$换为了上确界$\sup_{h\in\mathcal{H}}$。该式表示，对于样例集$D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$，假设空间$\mathcal{H}$中的假设对其预测结果$\{h(\boldsymbol{x}_1),h(\boldsymbol{x}_2),\ldots,h(\boldsymbol{x}_m)\}$与随机变量集合$\boldsymbol{\sigma}=\{\sigma_1,\sigma_2,\ldots,\sigma_m\}$的契合程度。接下来解释一下该式的含义。

$\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(\boldsymbol{x}_i)$中的$\boldsymbol{\sigma}=\{\sigma_1,\sigma_2,\ldots,\sigma_m\}$表示单次随机生成的结果（生成后就固定不动），而$\{h(\boldsymbol{x}_1),h(\boldsymbol{x}_2),\ldots,h(\boldsymbol{x}_m)\}$表示某个假设$h\in\mathcal{H}$的预测结果，至于$\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(\boldsymbol{x}_i)$的取值则取决于本次随机生成的$\boldsymbol{\sigma}$和假设$h$的预测结果的契合程度。

进一步地，$\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(\boldsymbol{x}_i)$中的$\boldsymbol{\sigma}=\{\sigma_1,\sigma_2,\ldots,\sigma_m\}$仍表示单次随机生成的结果（生成后就固定不动），但此时需求解的是假设空间$\mathcal{H}$中所有假设与$\boldsymbol{\sigma}$最契合的那个$h$。

例如，$\boldsymbol{\sigma}=\{-1,+1,-1,+1\}$（即$m=4$，这里$\boldsymbol{\sigma}$仅为本次随机生成结果而已，下次生成结果可能是另一组结果），假设空间$\mathcal{H}=\{h_1,h_2,h_3\}$，其中

$$\begin{aligned}\{h_1(\boldsymbol{x}_1),h_1(\boldsymbol{x}_2),h_1(\boldsymbol{x}_3),h_1(\boldsymbol{x}_4)\}&=\{-1,-1,-1,-1\}\\ \{h_2(\boldsymbol{x}_1),h_2(\boldsymbol{x}_2),h_2(\boldsymbol{x}_3),h_2(\boldsymbol{x}_4)\}&=\{-1,+1,-1,-1\}\\ \{h_3(\boldsymbol{x}_1),h_3(\boldsymbol{x}_2),h_3(\boldsymbol{x}_3),h_3(\boldsymbol{x}_4)\}&=\{+1,+1,+1,+1\}\end{aligned}$$

易知$\frac{1}{m}\sum_{i=1}^{m}\sigma_i h_1(\boldsymbol{x}_i)=0$，$\frac{1}{m}\sum_{i=1}^{m}\sigma_i h_2(\boldsymbol{x}_i)=\frac{2}{4}$，$\frac{1}{m}\sum_{i=1}^{m}\sigma_i h_3(\boldsymbol{x}_i)=0$，因此

$$\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(\boldsymbol{x}_i)=\frac{2}{4}$$

## 4、式(12.39)的解释

相比于式(12.38)，本式增加了对变量$\boldsymbol{\sigma}$求期望$\mathbb{E}_{\boldsymbol{\sigma}}[\cdot]$，此时变量$\boldsymbol{\sigma}$将包含所有可能随机生成结果，即共有$2^m$个不同的$\boldsymbol{\sigma}$。实际上，这$2^m$个不同的$\boldsymbol{\sigma}$对于$D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$来说相当于所有可能的类别指派，即 12.4 节中提到的所有“对分”。因此本式的求解步骤为：针对每个$\boldsymbol{\sigma}$（共$2^m$个不同的$\boldsymbol{\sigma}$），遍历所有$h\in\mathcal{H}$计算式(12.38)；然后将$2^m$个结果求平均即可。

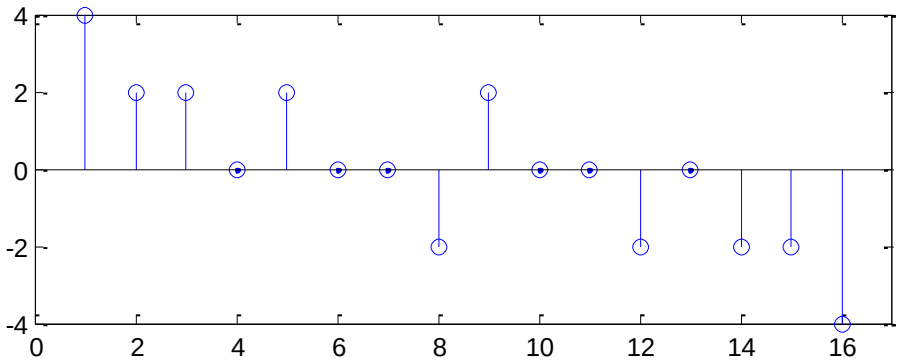
接下来解释本式接下来一段话中的两个问题：

(1)“当$|\mathcal{H}|=1$时，$\mathcal{H}$中仅有一个假设，这时可计算出式(12.39)的值为 0”

设$\mathcal{H}$中仅有的一个假设$\{h(\boldsymbol{x}_1),h(\boldsymbol{x}_2),h(\boldsymbol{x}_3),h(\boldsymbol{x}_4)\}=\{-1,-1,-1,-1\}$，即$m=4$，此时共有$2^4=16$个不同的$\boldsymbol{\sigma}$，运行以下 MATLAB 程序：

```
clc;clear all;close all;
a1 = [-1,+1,-1,+1,-1,+1,-1,+1,-1,+1,-1,+1,-1,+1,-1,+1];
a2 = [-1,-1,+1,+1,-1,-1,+1,+1,-1,-1,+1,+1,-1,-1,+1,+1];
a3 = [-1,-1,-1,-1,+1,+1,+1,+1,-1,-1,-1,-1,+1,+1,+1,+1];
a4 = [-1,-1,-1,-1,-1,-1,-1,-1,+1,+1,+1,+1,+1,+1,+1,+1];
Sigma = [a4',a3',a2',a1'];
h = [-1,-1,-1,-1];
S = zeros(16,1);
for i=1:16
    S(i) = Sigma(i,:)*h';
end
```

stem(S);xlim([0,17]);




可以发现，16 个不同的$\boldsymbol{\sigma}$与$\{-1,-1,-1,-1\}$的内积（即$\sum_{i=1}^{m}\sigma_i h(\boldsymbol{x}_i)$）关于点(8.5, 0)旋转对称，也就是说以坐标系中的点(8.5, 0)为中心，旋转其中的一半之后正好与另一半重合。因此这 16 个值的平均值等于 0。可以尝试将$\{h(\boldsymbol{x}_1),h(\boldsymbol{x}_2),h(\boldsymbol{x}_3),h(\boldsymbol{x}_4)\}$改为其它值（例如式(12.38)注解中最后的例子中的$h_2$或$h_3$的输出），其平均值也等于 0；这是因为$2^m$个不同的$\boldsymbol{\sigma}$中总是两两互为相反数序列的，例如$\{-1,-1,-1,-1\}$和$\{+1,+1,+1,+1\}$，再例如$\{-1,-1,+1,+1\}$和$\{+1,+1,-1,-1\}$，每对相反数序列与$\{h(\boldsymbol{x}_1),h(\boldsymbol{x}_2),h(\boldsymbol{x}_3),h(\boldsymbol{x}_4)\}$求内积后的值一定是相反数，因此最终总体上求期望时结果一定等于 0。

(2)“当$|\mathcal{H}|=2^m$且$\mathcal{H}$能打散$D$时，对任意$\boldsymbol{\sigma}$总有一个假设使得$h(\boldsymbol{x}_i)=\sigma_i\ (i=1,\ldots,m)$，这时可计算出式(12.39)的值为 1”

此时，针对$2^m$个不同的$\boldsymbol{\sigma}$中的任意一个，$\mathcal{H}$中一定存在$\{h(\boldsymbol{x}_1),h(\boldsymbol{x}_2),h(\boldsymbol{x}_3),h(\boldsymbol{x}_4)\}$与$\boldsymbol{\sigma}$完全相同，即$\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(\boldsymbol{x}_i)=1$，进而针对所有$\boldsymbol{\sigma}$求期望时也等于 1。

## 5、定义 12.8 的解释

式(12.40)相比于式(12.39)就是将$\mathcal{X}$和$\mathcal{H}$替换为$\mathcal{Z}$和$\mathcal{F}$；但是这里的$\mathcal{F}$不同于$\mathcal{H}$的是，在式(12.39)中假设空间$\mathcal{H}$中的假设$h$输出均为$-1,+1$的二值数据，而这里的函数空间$\mathcal{F}$中的函数$f$输出均为实值。这也是接下来定理 12.5 和定理 12.6 的不同之处。

## 6、定义 12.9 的解释

定义 12.8 中的$\widehat{R}_Z(\mathcal{F})$针对某个集合$Z$，而定义 12.9 则可以简单理解为在空间$\mathcal{Z}$中根据分布$\mathcal{D}$采样获得多个不同的集合$Z$，求平均值。当然，期望和平均值是不一样的，当集合$Z$的个数增大时，求得的平均值将越来越接近真实的期望，即大数定理。

## 7、定理 12.5 的解释

个人感觉，这个定理就是为下一个定理 12.6 做准备的，并没有特别的其它意义。其中：

$\mathbb{E}[f(\boldsymbol{z})]$表示对于$f\in\mathcal{F}$和$\boldsymbol{z}\in\mathcal{Z}$的期望，简写为$\mathbb{E}[f]$；如果令$f(\boldsymbol{z})=\mathbb{I}(h(\boldsymbol{z})\neq y)$，其中$y$表示$\boldsymbol{z}$的真实类别标记，则$\mathbb{E}[f(\boldsymbol{z})]$就是式(12.1)的泛化误差。

$\frac{1}{m}\sum_{i=1}^{m}f(\boldsymbol{z}_i)$表示在示例集$Z=\{\boldsymbol{z}_1,\boldsymbol{z}_2,\ldots,\boldsymbol{z}_m\}$上的函数平均值；类似于$\mathbb{E}[f(\boldsymbol{z})]$，如果令$f(\boldsymbol{z})=\mathbb{I}(h(\boldsymbol{z})\neq y)$，则$\frac{1}{m}\sum_{i=1}^{m}f(\boldsymbol{z}_i)$就是式(12.2)的经验误差。

因此，式(12.42)和式(12.43)实际上表示“泛化”和“经验”的关系，但因为这里的函数

空间$\mathcal{F}$是区间$[0,1]$上的实值函数，所以需要进一步将结论推广至二分类问题，即定理 12.6。

在本定理证明结束之后，提到“定理 12.5 中的函数空间$\mathcal{F}$是区间$[0,1]$上的实值函数，因此定理 12.5 只适用于回归问题”，但这里有个问题是，若$f \in \mathcal{F}$就是回归预测函数，那么$\mathbb{E}[f(\boldsymbol{z})]$和$\frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{z}_i)$并没有实际意义吧（回归函数值求平均？），但若$f \in \mathcal{F}$表示回归预测值与真实值的误差函数，那又如何保证函数空间$\mathcal{F}$是区间$[0,1]$上的实值函数呢？

## 8、定理 12.5 的证明

本定理的证明可以参考定理 12.5 上面一段话最后给出的参考文献[Mohri et al., 2012]（该书 2018 年已出第 2 版，官网 https://cs.nyu.edu/~mohri/mlbook/即可下载），定理 12.5 在第 1 版中是第 35 页的 Theorem 3.1，在第 2 版中是第 31 页的 Theorem 3.3。

①$\widehat{E}_Z(f)=\frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{z}_i)$和$\Phi(Z)=\sup_{f\in\mathcal{F}}\mathbb{E}[f]-\widehat{E}_Z(f)$是符号$\widehat{E}_Z(f)$和$\Phi(Z)$的定义，但$\Phi(Z)=\sup_{f\in\mathcal{F}}\mathbb{E}[f]-\widehat{E}_Z(f)$应该写为$\Phi(Z)=\sup_{f\in\mathcal{F}}\left(\mathbb{E}[f]-\widehat{E}_Z(f)\right)$更清晰。

②关于$\Phi(Z')-\Phi(Z)\leqslant\frac{1}{m}$的证明，第 1 个等号就是将$\Phi(Z)$的定义式代入即可；第 2 行的小于等于号在原文献证明中专门提到“since the difference of suprema does not exceed the supremum of the difference”，即“上确界的差”不超过“差的上确界”；第 3 个等号就是将$\widehat{E}_Z(f)$的定义式代入，并且已经定义示例集$Z$与$Z'$只有第$m$个样本不同，因此相减之后只剩$f(\boldsymbol{z}_m)-f(\boldsymbol{z}'_m)$，分母的$m$来自$\widehat{E}_Z(f)$的定义；最后一行的小于等于号是因为$f(\boldsymbol{z}_m)$和$f(\boldsymbol{z}'_m)$位于区间$[0,1]$，因此二者之差最大值在$f(\boldsymbol{z}_m)=1$和$f(\boldsymbol{z}'_m)=0$时取得，最大值为 1。

③同理，可得$\Phi(Z)-\Phi(Z')\leqslant\frac{1}{m}$，而当$\Phi(Z')-\Phi(Z)\leqslant\frac{1}{m}$和$\Phi(Z)-\Phi(Z')\leqslant\frac{1}{m}$同时成立时，$-\frac{1}{m}\leqslant\Phi(Z)-\Phi(Z')\leqslant\frac{1}{m}$成立，即$|\Phi(Z)-\Phi(Z')|\leqslant\frac{1}{m}$成立。

④有关式(12.44)结果的证明：结论$|\Phi(Z)-\Phi(Z')|\leqslant\frac{1}{m}$表明$\Phi(Z)$满足 McDiarmid 不等式的条件（第 268 页），即

$$\sup_{x_1,\ldots,x_m,x_i'}|f(x_1,\ldots,x_m)-f(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_m)|\leqslant c_i$$

其中这里的$c_i=\frac{1}{m},\ i=1,\ldots,m$；根据式(12.7)，即

$$P\left(\Phi(Z)-\mathbb{E}_Z[\Phi(Z)]\geqslant\epsilon\right)\leqslant\exp\left(\frac{-2\epsilon^2}{\sum_i c_i^2}\right)$$

令$\exp\left(\frac{-2\epsilon^2}{\sum_i c_i^2}\right)=\delta$，由$\sum_i c_i^2=\sum_{i=1}^{m}\frac{1}{m^2}=\frac{1}{m}\Rightarrow\exp\left(-2m\epsilon^2\right)=\delta$，得$\epsilon=\sqrt{\frac{\ln(1/\delta)}{2m}}$，即

$$P\left(\Phi(Z)-\mathbb{E}_Z[\Phi(Z)]\geqslant\sqrt{\frac{\ln(1/\delta)}{2m}}\right)\leqslant\delta$$

等价于

$$P\left(\Phi(Z)-\mathbb{E}_Z[\Phi(Z)]\leqslant\sqrt{\frac{\ln(1/\delta)}{2m}}\right)\geqslant 1-\delta$$

即$\Phi(Z)-\mathbb{E}_Z[\Phi(Z)]\leqslant\sqrt{\frac{\ln(1/\delta)}{2m}}$以至少$1-\delta$的概率成立，至此式(12.44)有关结果证毕。

⑤接下来，证明$\mathbb{E}_Z[\Phi(Z)]$的上界，即$\mathbb{E}_Z[\Phi(Z)]\leqslant 2R_m(\mathcal{F})$：

第 1 行的等号即代入$\Phi(Z)$的定义；

第 2 行的等号中的$Z'$应该不再是前面刚开始证明时提到的那个$Z'$了，而是从$\mathcal{Z}$中以 i.i.d. 采样得到的示例集，因此$\mathbb{E}_{Z'}[\widehat{E}_{Z'}(f)]=E(f)$，又由于$Z$与$Z'$独立，因此$\mathbb{E}_{Z'}[\widehat{E}_Z(f)]=\widehat{E}_Z(f)$，故$\mathbb{E}_{Z'}[\widehat{E}_{Z'}(f)-\widehat{E}_Z(f)]=\mathbb{E}_{Z'}[\widehat{E}_{Z'}(f)]-\mathbb{E}_{Z'}[\widehat{E}_Z(f)]=E(f)-\widehat{E}_Z(f)$；

第 3 行的小于等于号如左侧边注所说：“利用 Jensen 不等式(12.4)和上确界函数的凸性”，有$\sup_{f\in\mathcal{F}}\mathbb{E}_{Z'}[\widehat{E}_{Z'}(f)-\widehat{E}_Z(f)]=\mathbb{E}_{Z'}[\sup_{f\in\mathcal{F}}(\widehat{E}_{Z'}(f)-\widehat{E}_Z(f))]$，其中，相比于式(12.4)，$\sup_{f\in\mathcal{F}}$对应于$f(\cdot)$，$\widehat{E}_{Z'}(f)-\widehat{E}_Z(f)$对应于$x$，而$\mathbb{E}_{Z,Z'}[\cdot]$即$\mathbb{E}_Z[\mathbb{E}_{Z'}[\cdot]]$的简写；

第 4 行的等号就是将$\widehat{E}_Z(f)$的定义式代入；

第 5 行引入了 Rademacher 随机变量，而这并不影响期望值；这是因为当$\sigma_i=+1$时，$\sigma_i(f(\boldsymbol{z}_i')-f(\boldsymbol{z}_i))=f(\boldsymbol{z}_i')-f(\boldsymbol{z}_i)$，当$\sigma_i=-1$时，$\sigma_i(f(\boldsymbol{z}_i')-f(\boldsymbol{z}_i))=f(\boldsymbol{z}_i)-f(\boldsymbol{z}_i')$，相当于交换了$Z$和$Z'$的第$i$个示例，又由于求期望时是针对所有可能的$Z$和$Z'$，因此引入 Rademacher 随机变量并不影响期望值；

第 6 行的小于等于号使用了性质$\sup(U+V)\leqslant\sup(U)+\sup(V)$；而且第 1 项只含$\boldsymbol{z}_i'$，第 2 项只含$\boldsymbol{z}_i$，因此第 1 项下标只有$\boldsymbol{\sigma},Z'$，第 2 项下标只有$\boldsymbol{\sigma},Z$；

第 7 行的等号正如边注所说$\sigma_i$与$-\sigma_i$分布相同（求期望时要针对所有可能的$\boldsymbol{\sigma}$），$Z$和$Z'$也都是从$\mathcal{Z}$中以 i.i.d.采样得到的示例集；

第 8 行就是式(12.41)关于$R_m(\mathcal{F})$的定义，其中$\mathbb{E}_{\boldsymbol{\sigma},Z}[\cdot]=\mathbb{E}_Z[\mathbb{E}_{\boldsymbol{\sigma}}[\cdot]]$。

有关这个证明的每一行推导原理在参考文献[Mohri et al., 2012]中都有详细解释。

⑥根据式(12.44)可知

$$\Phi(Z)\leqslant\mathbb{E}_Z[\Phi(Z)]+\sqrt{\frac{\ln(1/\delta)}{2m}}\leqslant 2R_m(\mathcal{F})+\sqrt{\frac{\ln(1/\delta)}{2m}}$$

而根据前面定义$\Phi(Z)=\sup_{f\in\mathcal{F}}\mathbb{E}[f]-\widehat{E}_Z(f)$，因此

$$\mathbb{E}[f]-\widehat{E}_Z(f)\leq\sup_{f\in\mathcal{F}}\mathbb{E}[f]-\widehat{E}_Z(f)\leqslant 2R_m(\mathcal{F})+\sqrt{\frac{\ln(1/\delta)}{2m}}$$

即$\mathbb{E}[f]\leqslant\widehat{E}_Z(f)+2R_m(\mathcal{F})+\sqrt{\frac{\ln(1/\delta)}{2m}}$，其中$\mathbb{E}[f]$为$\mathbb{E}[f(\boldsymbol{z})]$的简写，**即式(12.42)得证。**

⑦关于式(12.45)，上方一段应该是“**由定义 12.8 可知**”，在式(12.40)中，由于$f(\boldsymbol{z})$取值在区间$[0,1]$内，因此改变$Z$中的一个示例造成的改变最多为 1（当$f(\boldsymbol{z}_i)$由 0 变至 1 时），因此对$\widehat{R}_Z(\mathcal{F})$的改变最多为$1/m$。注意到$R_m(\mathcal{F})=\mathbb{E}_Z\left[\widehat{R}_Z(\mathcal{F})\right]$，根据式(12.7)可知

$$P\left(\widehat{R}_Z(\mathcal{F})-\mathbb{E}_Z\left[\widehat{R}_Z(\mathcal{F})\right]\geqslant\epsilon\right)=P\left(\widehat{R}_Z(\mathcal{F})-R_m(\mathcal{F})\geqslant\epsilon\right)\leqslant\exp\left(\frac{-2\epsilon^2}{\sum_i c_i^2}\right)$$

其中$c_i=\frac{1}{m}$，$i=1,\ldots,m$；令$\exp\left(\frac{-2\epsilon^2}{\sum_i c_i^2}\right)=\frac{\delta}{2}$，解得$\epsilon=\sqrt{\frac{\ln(2/\delta)}{2m}}$，即

$$P\left(\widehat{R}_Z(\mathcal{F})-R_m(\mathcal{F})\geqslant\sqrt{\frac{\ln(2/\delta)}{2m}}\right)\leqslant\frac{\delta}{2}$$

注意到式(12.7)中$f(x_1,\ldots,x_m)$与$\mathbb{E}(f(x_1,\ldots,x_m))$大小关系应该是对称的，类似于由式(12.5)对应着第 272 页引理 12.1 中的式(12.15)和式(12.16)一样，因此下式应该也成立：

$$P\left(R_m(\mathcal{F})-\widehat{R}_Z(\mathcal{F})\geqslant\sqrt{\frac{\ln(2/\delta)}{2m}}\right)\leqslant\frac{\delta}{2}\Leftrightarrow P\left(R_m(\mathcal{F})-\widehat{R}_Z(\mathcal{F})\leqslant\sqrt{\frac{\ln(2/\delta)}{2m}}\right)\geqslant 1-\frac{\delta}{2}$$

即$R_m(\mathcal{F})-\widehat{R}_Z(\mathcal{F})\leqslant\sqrt{\frac{\ln(2/\delta)}{2m}}$以至少$1-\frac{\delta}{2}$的概率成立。

⑧对于接下来的式子，只需将式(12.44)原来的$\delta$换为$\frac{\delta}{2}$即可，因为这里要得到“以至少$1-\frac{\delta}{2}$的概率成立”。

⑨对于最后的式(12.46)，综合以上推导结果即可（⑧⑤⑦）：

$$\begin{aligned}\Phi(Z)&\leqslant\mathbb{E}_Z[\Phi(Z)]+\sqrt{\frac{\ln(2/\delta)}{2m}}\leqslant 2R_m(\mathcal{F})+\sqrt{\frac{\ln(2/\delta)}{2m}}\\&\leqslant 2\left(\widehat{R}_Z(\mathcal{F})+\sqrt{\frac{\ln(2/\delta)}{2m}}\right)+\sqrt{\frac{\ln(2/\delta)}{2m}}=2\widehat{R}_Z(\mathcal{F})+3\sqrt{\frac{\ln(2/\delta)}{2m}}\end{aligned}$$

上式中，$\Phi(Z)\leqslant\mathbb{E}_Z[\Phi(Z)]+\sqrt{\frac{\ln(2/\delta)}{2m}}$以至少$1-\frac{\delta}{2}$的概率成立，而$\mathbb{E}_Z[\Phi(Z)]\leqslant 2R_m(\mathcal{F})$，因此$\Phi(Z)\leqslant 2R_m(\mathcal{F})+\sqrt{\frac{\ln(2/\delta)}{2m}}$以至少$1-\frac{\delta}{2}$的概率成立；而式(12.45)的$R_m(\mathcal{F})\leqslant\widehat{R}_Z(\mathcal{F})+\sqrt{\frac{\ln(2/\delta)}{2m}}$以至少$1-\frac{\delta}{2}$的概率成立，因此式(12.46)以至少$1-\delta$的概率成立。该结论可以参见如下定理：

【定理】已知事件$a\leqslant b$成立的概率不小于$1-\frac{\delta}{2}$，事件$b\leqslant c$成立的概率不小于$1-\frac{\delta}{2}$，则事件$a\leqslant c$成立的概率不小于$1-\delta$。

【分析】此问题与定理 12.4 的证明思路基本一致，这里仍将：

已知条件概括为$P(a\leqslant b)\geqslant 1-\frac{\delta}{2},\ P(b\leqslant c)\geqslant 1-\frac{\delta}{2}$

求证问题概括为$P(a\leqslant c)\geqslant 1-\delta$

【证明】

$$\begin{aligned}P(a\leqslant b\cap b\leqslant c)&=P(a\leqslant b)+P(b\leqslant c)-P(a\leqslant b\cup b\leqslant c)\\&\geqslant(1-\frac{\delta}{2})+(1-\frac{\delta}{2})-P(a\leqslant b\cup b\leqslant c)\\&\geqslant 2-\delta-1=1-\delta\end{aligned}$$

又因为$P(a\leqslant c)\geqslant P(a\leqslant b\cap b\leqslant c)$，这是由于若$a\leqslant b$和$b\leqslant c$同时成立，则$a\leqslant c$肯定成立，但$a\leqslant c$成立并不能得出$a\leqslant b$和$b\leqslant c$同时成立，即$a\leqslant b$和$b\leqslant c$同时成立是$a<c$成立的充分非必要条件（例如$a=4,b=3,c=5$时，$a\leqslant c$成立，但只有$b\leqslant c$成立而$a\leqslant b$不成立）。

综上可得，$P(a\leqslant c)\geqslant 1-\delta$，证毕！

⑩将前面定义的$\Phi(Z)=\sup_{f\in\mathcal{F}}\mathbb{E}[f]-\widehat{E}_Z(f)$代入式(12.46)，得

$$\sup_{f\in\mathcal{F}}\mathbb{E}[f]-\widehat{E}_Z(f)\leqslant 2\widehat{R}_Z(\mathcal{F})+3\sqrt{\frac{\ln(2/\delta)}{2m}}$$

又因为$\mathbb{E}[f]-\widehat{E}_Z(f)\leq\sup_{f\in\mathcal{F}}\mathbb{E}[f]-\widehat{E}_Z(f)$（上确界的定义），因此

$$\mathbb{E}[f]-\widehat{E}_Z(f)\leqslant 2\widehat{R}_Z(\mathcal{F})+3\sqrt{\frac{\ln(2/\delta)}{2m}}$$

即$\mathbb{E}[f]\leqslant\widehat{E}_Z(f)+2\widehat{R}_Z(\mathcal{F})+3\sqrt{\frac{\ln(2/\delta)}{2m}}$，其中$\mathbb{E}[f]$为$\mathbb{E}[f(\boldsymbol{z})]$的简写，**即式(12.43)得证。**

## 9、定理 12.6 的解释

针对二分类问题，定理 12.5 给出了“泛化误差”和“经验误差”的关系，即：
式(12.47)基于 Rademacher 复杂度$R_m(\mathcal{H})$给出了泛化误差$E(h)$的上界；

式(12.48)基于经验 Rademacher 复杂度$\widehat{R}_D(\mathcal{H})$给出了泛化误差$E(h)$的上界。

读到这里可能大家都会有疑问：定理 12.6 的设定其实也适用于定理 12.5，即值域为二值的$\{-1,+1\}$也属于值域为连续值的$[0,1]$的一种特殊情况，这一点从接下来的式(12.49)的转换可以看出。那么，为什么还要针对二分类问题专门给出定理 12.6 呢？

根据(经验)Rademacher 复杂度的定义可以知道，$R_m(\mathcal{H})$和$\widehat{R}_D(\mathcal{H})$均大于零（参见前面有关式(12.39)的解释，书中式(12.39)下面的一行也提到该式取值范围是$[0,1]$）；因此，相比于定理 12.5 来说，定理 12.6 的上界更紧，因为二者的界只有中间一项关于(经验)Rademacher 复杂度的部分不同，在定理 12.5 中是两倍的(经验)Rademacher 复杂度，而在定理 12.6 中是一倍的(经验)Rademacher 复杂度，而(经验)Rademacher 复杂度大于零。

因此，为二分类问题量身定制的定理 12.6 相比于通用的定理 12.5 来说，二者的区别在于定理 12.6 考虑了二分类的特殊情况，得到了比定理 12.5 更紧的泛化误差界，仅此而已。

## 10、定理 12.6 的证明

①首先通过式(12.49)将值域为$\{-1,+1\}$的假设空间$\mathcal{H}$转化为值域为$[0,1]$的函数空间$\mathcal{F}_{\mathcal{H}}$；

②接下来是该证明最核心部分，即证明式(12.50)的结论$\widehat{R}_Z(\mathcal{F}_{\mathcal{H}})=\frac{1}{2}\widehat{R}_D(\mathcal{H})$：

第 1 行等号就是定义 12.8；第 2 行等号就是根据式(12.49)将$f_h(\boldsymbol{x}_i,y_i)$换为$\mathbb{I}(h(\boldsymbol{x}_i)\neq y_i)$；第 3 行等号类似于式(12.36)的第 2 个等号；第 4 行等号说明如下：

$$\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\frac{1-y_ih(\boldsymbol{x}_i)}{2}=\sup_{h\in\mathcal{H}}\frac{1}{2m}\sum_{i=1}^{m}\sigma_i+\sup_{h\in\mathcal{H}}\frac{1}{2m}\sum_{i=1}^{m}\frac{-y_i\sigma_ih(\boldsymbol{x}_i)}{2}$$

其中$\sup_{h\in\mathcal{H}}\frac{1}{2m}\sum_{i=1}^{m}\sigma_i$与$h$无关，所以$\sup_{h\in\mathcal{H}}\frac{1}{2m}\sum_{i=1}^{m}\sigma_i=\frac{1}{2m}\sum_{i=1}^{m}\sigma_i$，即第 4 行等号；

第 5 行等号是由于$\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\sum_{i=1}^{m}\sigma_i\right]=0$，例如当$m=2$时，所有可能得$\boldsymbol{\sigma}$包括$(-1,-1)$，$(-1,+1),(+1,-1)$和$(+1,+1)$，求期望后显然结果等于 0；第 6 行等号正如边注所说，“$-y_i\sigma_i$与$\sigma_i$分布相同”（原因跟定理 12.5 中证明$\mathbb{E}_Z[\Phi(Z)]\leqslant 2R_m(\mathcal{F})$相同，即求期望时要针对所有可能的$\boldsymbol{\sigma}$参见，第 282 页第 8 行）；第 7 行等号再次使用了定义 12.8。

③关于式(12.51)，根据式(12.50)的结论，可证明如下：

$$R_m(\mathcal{F}_{\mathcal{H}}) = \mathbb{E}_Z\left[\widehat{R}_Z(\mathcal{F}_{\mathcal{H}})\right] = \mathbb{E}_D\left[\tfrac{1}{2}\widehat{R}_D(\mathcal{H})\right] = \tfrac{1}{2}\mathbb{E}_D\left[\widehat{R}_D(\mathcal{H})\right] = \tfrac{1}{2}R_m(\mathcal{H})$$

其中第 2 个等号由$Z$变为$D$只是符号根据具体情况的适时变化而已。

④最后，将式(12.49)定义的$f_h$替换定理 12.5 中的函数$f$，则

$$\mathbb{E}\left[f(\boldsymbol{z})\right] = \mathbb{E}\left[\mathbb{I}(h(\boldsymbol{x}) \neq y)\right] = E(h)$$

$$\tfrac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{z}_i) = \tfrac{1}{m}\sum_{i=1}^{m} \mathbb{I}(h(\boldsymbol{x}_i) \neq y_i) = \widehat{E}(h)$$

将式(12.51)代入式(12.42)，即用$\frac{1}{2}R_m(\mathcal{H})$替换式(12.42)的$R_m(\mathcal{F})$，式(12.47)得证；

将式(12.50)代入式(12.43)，即用$\frac{1}{2}\widehat{R}_D(\mathcal{H})$替换式(12.43)的$\widehat{R}_Z(\mathcal{F})$，式(12.48)得证。

这里有个疑问在于，定理 12.5 的前提是“实值函数空间$\mathcal{F}: \mathcal{Z} \to [0,1]$”，而式(12.49)得到的函数$f_h(\boldsymbol{z})$的值域实际为$\{0,1\}$，仍是离散的而非实值的；当然，定理 12.5 的证明也只需要其函数值在$[0,1]$范围内即可，并不需要其连续。

## 11、定理 12.7 的证明

在[Mohri et al., 2012]第 2 版（https://cs.nyu.edu/~mohri/mlbook/）中，该定理描述如下：

**Corollary 3.8** *Let $\mathcal{G}$ be a family of functions taking values in $\{-1,+1\}$. Then the following holds:*

$$\mathfrak{R}_m(\mathcal{G}) \leq \sqrt{\frac{2\log \Pi_{\mathcal{G}}(m)}{m}}. \tag{3.21}$$

**Proof:** For a fixed sample $S = (x_1, \ldots, x_m)$, we denote by $\mathcal{G}_{|S}$ the set of vectors of function values $(g(x_1), \ldots, g(x_m))^\top$ where $g$ is in $\mathcal{G}$. Since $g \in \mathcal{G}$ takes values in $\{-1,+1\}$, the norm of these vectors is bounded by $\sqrt{m}$. We can then apply Massart's lemma as follows:

$$\mathfrak{R}_m(\mathcal{G}) = \mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{u \in \mathcal{G}_{|S}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i u_i\right]\right] \leq \mathop{\mathbb{E}}_{S}\left[\frac{\sqrt{m}\sqrt{2\log|\mathcal{G}_{|S}|}}{m}\right].$$

By definition, $|\mathcal{G}_{|S}|$ is bounded by the growth function, thus,

$$\mathfrak{R}_m(\mathcal{G}) \leq \mathop{\mathbb{E}}_{S}\left[\frac{\sqrt{m}\sqrt{2\log \Pi_{\mathcal{G}}(m)}}{m}\right] = \sqrt{\frac{2\log \Pi_{\mathcal{G}}(m)}{m}},$$

which concludes the proof. □

其中，证明中提到的 Massart's lemma 如下：

**Theorem 3.7 (Massart's lemma)** *Let $\mathcal{A} \subseteq \mathbb{R}^m$ be a finite set, with $r = \max_{\mathbf{x} \in \mathcal{A}} \|\mathbf{x}\|_2$, then the following holds:*

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\sup_{\mathbf{x} \in \mathcal{A}}\sum_{i=1}^{m} \sigma_i x_i\right] \leq \frac{r\sqrt{2\log|\mathcal{A}|}}{m}, \tag{3.20}$$

*where $\sigma_i$s are independent uniform random variables taking values in $\{-1,+1\}$ and $x_1, \ldots, x_m$ are the components of vector $\mathbf{x}$.*

**Proof:** The result follows immediately from the bound on the expectation of a maximum given by Corollary D.11 since the random variables $\sigma_i x_i$ are independent and each $\sigma_i x_i$ takes values in $[-|x_i|, |x_i|]$ with $\sqrt{\sum_{i=1}^{m} x_i^2} \leq r^2$. □

其中，Corollary D.11 如下：

**Corollary D.11 (Maximal inequality)** *Let $X_1 \dots X_n$ be $n \geq 1$ real-valued random variables such that for all $j \in [n]$, $X_j = \sum_{i=1}^{m} Y_{ij}$ where, for each fixed $j \in [n]$, $Y_{ij}$ are independent zero mean random variables taking values in $[-r_i, +r_i]$, for some $r_i > 0$. Then, the following inequality holds:*

$$\mathbb{E}\left[\max_{j\in[n]} X_j\right] \leq r\sqrt{2\log n},$$

*with $r = \sqrt{\sum_{i=1}^{m} r_i^2}$.*

**Proof:** By the independence of the $Y_{ij}$s for fixed $j$ and Hoeffding's lemma (Lemma D.1), the following inequality holds for all $j \in [n]$:

$$\mathbb{E}[e^{tX_j}] = \mathbb{E}\left[\prod_{i=1}^{m} e^{tY_{ij}}\right] = \prod_{i=1}^{m} \mathbb{E}[e^{tY_{ij}}] \leq \prod_{i=1}^{m} e^{\frac{t^2 r_j^2}{2}} = e^{\frac{t^2 r^2}{2}}, \tag{D.26}$$

with $r^2 = \sum_{i=1}^{m} r_i^2$. The result then follows immediately by Theorem D.10. □

其中，Hoeffding's lemma (Lemma D.1)如下：

**Lemma D.1 (Hoeffding's lemma)** *Let $X$ be a random variable with $E[X] = 0$ and $a \leq X \leq b$ with $b > a$. Then, for any $t > 0$, the following inequality holds:*

$$\mathbb{E}[e^{tX}] \leq e^{\frac{t^2(b-a)^2}{8}}. \tag{D.2}$$

该引理证明太长了，就不给出来了，还是自己看书吧……

### 12、式(12.53)的推导

由推论12.2式(12.28)有$\Pi_{\mathcal{H}}(m) \leqslant \left(\frac{e\cdot m}{d}\right)^d$，即$\ln \Pi_{\mathcal{H}}(m) \leqslant d \ln \frac{e\cdot m}{d}$；由定理12.7式(12.52)有$R_m(\mathcal{H}) \leqslant \sqrt{\frac{2\ln \Pi_{\mathcal{H}}(m)}{m}}$，结合$\ln \Pi_{\mathcal{H}}(m) \leqslant d \ln \frac{e\cdot m}{d}$，即$R_m(\mathcal{H}) \leqslant \sqrt{\frac{2d \ln \frac{e\cdot m}{d}}{m}}$，将此结论代入式(12.47)即得式(12.53)。

## 12.6 稳定性

上上节中介绍的基于 VC 维的泛化误差界是分布无关、数据独立的，上一节介绍的 Rademacher 复杂度则在一定程度上考虑了数据分布，但二者得到的结果均与具体学习算法无关；本节将要介绍的稳定性分析可以获得与算法有关的分析结果。算法的“稳定性”考察的是算法在输入发生变化时，输出是否会随之发生较大的变化。

### 1、泛化/经验/留一损失的解释

根据式(12.54)上方关于损失函数的描述：“刻画了假设$\mathfrak{L}_D$的预测标记$\mathfrak{L}_D(\boldsymbol{x})$与真实标记$y$之间的差别”，这里针对的是二分类，预测标记和真实标记均只能取$-1$和$+1$两个值，它们之间的“差别”又能是什么呢？

因此，当“差别”取为$\mathbb{I}(\mathfrak{L}_D(\boldsymbol{x}), y)$时，式(12.54)的泛化损失就是式(12.1)的泛化误差，式(12.55)的经验损失就是式(12.2)的经验误差，如果类似于式(12.1)和式(12.2)继续定义留一误差，那么式(12.56)就对应于留一误差。

### 2、定义 12.10 的解释

式(12.57)就是$\mathfrak{L}$关于损失函数$\ell$满足$\beta$-均匀稳定性的定义而已。它表示基于训练集$D$学得的假设$\mathfrak{L}_D$与基于移除第$i$个样本后训练集$D^{\setminus i}$学得的假设$\mathfrak{L}_{D^{\setminus i}}$之间的泛化损失（定义参见式(12.54)）的差异。

为了接下来讨论定理 12.8，这里继续对$\beta$的取值进行讨论。当训练集$D$包含的样本数$m$增多时，根据式(12.57)的含义，$\beta$一定会变小；例如，替换包含 10 个样本的训练集$D$中的某个样本和替换包含 10000000 个样本的训练集$D$中的某个样本，肯定替换操作对后者带来的影响更小（当然，若你非要用蝴蝶效应来说事儿，我也没办法）。

定义 12.10 下方的不等式就是绝对值不等式(参见百度百科)的具体应用，推导如下：

根据绝对值不等式$|a \pm b| \leqslant |a| + |b|$，可得

$$|a-b| = |(a-c)-(b-c)| \leqslant |a-c| + |b-c|$$

将$a$替换为$\ell(\mathfrak{L}_D, \boldsymbol{z})$，$b$替换为$\ell(\mathfrak{L}_{D^i}, \boldsymbol{z})$，$c$替换为$\ell(\mathfrak{L}_{D^{\setminus i}}, \boldsymbol{z})$，即得第 1 个不等式关系；

为避免符号混乱，令不等式中的$D_1 = D, D_2 = D^i$，分别将不等式中的$D_1$和$D_2$作为式(12.57)中的$D$，而由$D^i$和$D^{\setminus i}$的定义可知，无论$D$取$D_1$还是$D_2$时，$D^{\setminus i}$都是相同的（因为$D_1$和$D_2$除第$i$个样本外，其余样本都相同），因此若算法$\mathfrak{L}$关于损失函数$\ell$满足$\beta$-均匀稳定性时，根据式(12.57)的定义，$|\ell(\mathfrak{L}_D, \boldsymbol{z}) - \ell(\mathfrak{L}_{D^{\setminus i}}, \boldsymbol{z})|$和$|\ell(\mathfrak{L}_{D^i}, \boldsymbol{z}) - \ell(\mathfrak{L}_{D^{\setminus i}}, \boldsymbol{z})|$均满足小于等于$\beta$，因此二者之和小于等于$2\beta$。

$|\ell(\mathfrak{L}_D, \boldsymbol{z}) - \ell(\mathfrak{L}_{D^i}, \boldsymbol{z})|$表示基于训练集$D$学得的假设$\mathfrak{L}_D$与基于替换第$i$个样本后训练集$D^i$学得的假设$\mathfrak{L}_{D^i}$之间的泛化损失的差异。

## 3、定理 12.8 的解释

西瓜书在该定理下方已明确给出该定理的意义，即“定理 12.8 给出了基于稳定性分析推导出的学习算法$\mathfrak{L}$学得假设的泛化误差界”，式(12.58)和式(12.59)分别基于经验损失和留一损失给出了泛化损失的上界。接下来讨论两个相关问题：

(1)定理 12.8 的条件包括损失函数有界，即$0 \leqslant \ell(\mathfrak{L}_D, \boldsymbol{z}) \leqslant M$；如本节第 1 条注解“泛化/经验/留一损失的解释”中所述，若“差别”取为$\mathbb{I}(\mathfrak{L}_D(\boldsymbol{x}), y)$，则泛化损失对应于泛化误差，此时上限$M = 1$。

(2)在前面泛化误差上界的推导中（例如定理 12.1、定理 12.3、定理 12.6、定理 12.7），上界中与样本数$m$有关的项收敛率均为$O(1/\sqrt{m})$，但在该定理中却是$O(\beta\sqrt{m})$；一般来讲，随着样本数$m$的增加，经验误差/损失应该收敛于泛化误差/损失，因此这里假设$\beta = 1/m$（书中式(12.59)下方第 3 行写为$\beta = O(1/m)$），而在第 2 条注解“定义 12.10 的解释”中已经提到$\beta$的取值的确会随着样本数$m$的增多会变小，虽然书中并没有严格去讨论$\beta$随$m$增多的变化规律，但至少直觉上是对的。

## 4、经验损失最小化

顾名思义，“经验损失最小化”指通过最小化经验损失来求得假设函数。

这里，“对于损失函数$\ell$，若学习算法$\mathfrak{L}$所输出的假设满足经验损失最小化，则称算法$\mathfrak{L}$满足经验风险最小化原则，简称算法是 ERM 的”。

在第 278 页，若学习算法$\mathfrak{L}$输出的假设$h$满足式(12.30)，则也称$\mathfrak{L}$为满足经验风险最小化原则的算法。而很明显，式(12.30)是在最小化经验误差。

那么最小化经验误差和最小化经验损失有什么区别么？

在第 286 页左下角边注中提到，“最小化经验误差和最小化经验损失有时并不相同，这是由于存在某些病态的损失函数$\ell$使得最小化经验损失并不是最小化经验误差”。

对于“误差”、“损失”、“风险”等概念的辨析，参见第 2 章 2.1 节的注解。

## 5、定理 12.9 的证明的解释

本定理的证明过程与第 278 页定理 12.4 的证明过程有很多类似之处。

①将$\epsilon'=\frac{\epsilon}{2}$代入$\frac{\delta}{2}=2\exp\left(-2m(\epsilon')^2\right)$，得$\frac{\delta}{2}=2\exp\left(-\frac{m\epsilon^2}{2}\right)$，解得$m=\frac{2}{\epsilon^2}\ln\frac{4}{\delta}$。

②对应 Hoeffding 不等式(12.6)，$\ell(g,\mathcal{D})$对应$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}(x_i)$，$\widehat{\ell}(g,D)$对应$\frac{1}{m}\sum_{i=1}^{m}x_i$ 。

③对于方程$\frac{2}{m}+(4+M)\sqrt{\frac{\ln(2/\delta)}{2m}}=\frac{\epsilon}{2}$，两边同乘$m$，得$2+(4+M)\sqrt{\frac{m\ln(2/\delta)}{2}}=\frac{\epsilon}{2}m$，整理移项得

$$\frac{\epsilon}{2}(\sqrt{m})^2-(4+M)\sqrt{\frac{\ln(2/\delta)}{2}}\sqrt{m}-2=0$$

这是一个一元二次方程，根据求根公式得

$$\sqrt{m}=\frac{(4+M)\sqrt{\frac{\ln(2/\delta)}{2}}\pm\sqrt{(4+M)^2\frac{\ln(2/\delta)}{2}-4\times\frac{\epsilon}{2}\times(-2)}}{2\times\frac{\epsilon}{2}}$$

由于$\sqrt{m}\geqslant 0$，因此上式中只能取分子为加号的那一项，即

$$\sqrt{m}=\frac{(4+M)\sqrt{\frac{\ln(2/\delta)}{2}}+\sqrt{(4+M)^2\frac{\ln(2/\delta)}{2}-4\times\frac{\epsilon}{2}\times(-2)}}{2\times\frac{\epsilon}{2}}$$

两边平方，可得$m$与变量$1/\delta, 1/\epsilon$之间的关系（参见 P269 定义 12.2），即$O\left(\frac{1}{\epsilon^2}\ln\frac{1}{\delta}\right)$。

④由$|\ell(g,\mathcal{D})-\widehat{\ell}(g,D)|\leqslant\frac{\epsilon}{2}$，可知$-\frac{\epsilon}{2}\leqslant\ell(g,\mathcal{D})-\widehat{\ell}(g,D)\leqslant\frac{\epsilon}{2}$；根据左侧不等号关系（即$-\frac{\epsilon}{2}\leqslant\ell(g,\mathcal{D})-\widehat{\ell}(g,D)$），可得$-\ell(g,\mathcal{D})\leqslant-(\widehat{\ell}(g,D)-\frac{\epsilon}{2})$；

结合$\ell(\mathfrak{L},\mathcal{D})\leqslant\widehat{\ell}(\mathfrak{L},D)+\frac{\epsilon}{2}$，两个不等式两侧相加，得最后一个式子：

$$\begin{aligned}\ell(\mathfrak{L},\mathcal{D})-\ell(g,\mathcal{D})&\leqslant\widehat{\ell}(\mathfrak{L},D)+\frac{\epsilon}{2}-\left(\widehat{\ell}(g,D)-\frac{\epsilon}{2}\right)\\&\leqslant\widehat{\ell}(\mathfrak{L},D)-\widehat{\ell}(g,D)+\epsilon\\&\leqslant\epsilon\end{aligned}$$

注意：**这里共有 3 个“$\leqslant$”，但第 2 个应该是“$-$”，这应该是作者的笔误。**至于为何该式“以至少$1-\delta$的概率成立”，可以参见定理 12.4 的证明，思路相同（【定理】已知事件$a\leqslant b$成立的概率不小于$1-\frac{\delta}{2}$，事件$c\leqslant d$成立的概率不小于$1-\frac{\delta}{2}$，则事件$a+c\leqslant b+d$成立的概率不小于$1-\delta$）。

## 12.7 本章小节

相信本章是很多读者望而却步的一章，相信本章是很多以本书为教材的老师在讲课时会直接跳过的一章，相信本章是很多资深机器学习研究者都不予理睬的一章……

但既然已经入了机器学习这个坑，总感觉如果不去了解一点儿有关本章的内容，少点什么似的……

本章前两节介绍概念，12.3 节讨论有限假设空间的泛化误差界；但现实学习任务所面临的通常是无限假设空间，因此 12.4 节基于 VC 维讨论无限假设空间的泛化误差界；但基于 VC 维的泛化误差界是分布无关、数据独立的，因此 12.5 节继续基于 Rademacher 复杂度讨论泛化误差界；但无论是基于 VC 维还是 Rademacher 复杂度的泛化误差界，所得到的结果

均与具体学习算法无关，因此 12.6 节继续基于稳定性(stability)讨论泛化误差界。

本章很多内容参考了[Mohri et al., 2012]，可以使用 bing 或 google 搜索该书的 PDF 版；该书 2018 年已出第 2 版，官网 https://cs.nyu.edu/~mohri/mlbook/即可下载；该书第 1 版已有中译版，参见[张文生 等译. 机器学习基础. 机械工业出版社，2019.]。