第 15 章 验证

15.1 模型选择问题

命题 15.1.1 (模型选择的两大陷阱)

设共有 M 个候选模型 $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_M$，对应算法 $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_M$。目标是选出 $\mathcal{H}_{m^*}$，使得 $g_{m^*} = \mathcal{A}_{m^*}(\mathcal{D})$ 的 E_{out} 最小。

(a) 按最小 E_{in} 选择

$$m^* = \arg\min_{1\leq m\leq M} E_{\text{in}}\big(\mathcal{A}_m(\mathcal{D})\big)$$

这种方法存在严重问题：

- 更复杂的模型（如 Φ_{1126} 比 Φ_1）总能获得更低的 E_{in}
- 无正则化的模型（$\lambda = 0$ 比 $\lambda = 0.1$）总能获得更低的 E_{in}
- 等价于在整个并集 $\bigcup_m \mathcal{H}_m$ 上最小化 E_{in}
- 导致过拟合和泛化性能下降

(b) 按最小 E_{test} 选择

$$m^* = \arg\min_{1\leq m\leq M} E_{\text{test}}\big(\mathcal{A}_m(\mathcal{D})\big)$$

其中 $\mathcal{D}_{\text{test}}$ 为独立测试集。根据有限模型族的 Hoeffding 不等式：

$$E_{\text{out}}(g_{m^*}) \leq E_{\text{test}}(g_{m^*}) + O\left(\sqrt{\frac{\log M}{N_{\text{test}}}}\right)$$

但在实际应用中：

- 测试集 $\mathcal{D}_{\text{test}}$ 往往无法获取
- 这种方法虽然理论上完美，但在实践中不可行
- 甚至可能被视为” 作弊” 行为

结论：

$$\text{仅凭}E_{\text{in}}\text{ 选择模型} \rightarrow \text{灾难；仅凭}E_{\text{test}}\text{ 选择模型} \rightarrow \text{不可能。}$$

命题 15.1.2 (误差对比：E_{in}、E_{test} 与 E_{val})

	E_{in}	E_{test}	E_{val}
数据来源	训练集 $\mathcal{D}$	独立测试集 $\mathcal{D}_{\text{test}}$	验证集 $\mathcal{D}_{\text{val}} \subset \mathcal{D}$
可行性	随时可用	通常不可获取	随时可用
数据状态	被算法用于训练	从未被使用	未被算法使用
用途	训练/选择模型	最终泛化估计	模型选择
风险	过拟合（已污染）	无偏估计	合法“作弊”

结论：

利用E_{val}进行模型选择，既可行又“合法作弊”。

例题 15.1 选择题：假设集的最小误差假设判定

对于 $\mathcal{X}=\mathbb{R}^d$，考虑两个假设集 $\mathcal{H}_+$ 和 $\mathcal{H}_-$。$\mathcal{H}_+$ 包含所有 $w_1 \geq 0$ 的感知机，$\mathcal{H}_-$ 包含所有 $w_1 \leq 0$ 的感知机。记 g_+ 和 g_- 分别为各自假设集中最小 E_{in} 的假设。下列哪项陈述正确？

1) 若 $E_{\text{in}}(g_+) < E_{\text{in}}(g_-)$，则 g_+ 是所有感知机中的最小 E_{in} 假设。
2) 若 $E_{\text{test}}(g_+) < E_{\text{test}}(g_-)$，则 g_+ 是所有感知机中的最小 E_{test} 假设。
3) 两个假设集是不相交的。
4) 以上均不正确。

解答　正确选项为 $\boxed{1}$。两假设集的交集为 $w_1=0$ 的感知机，并集是所有感知机。

- 选项 1：因两假设集的并集是所有感知机，若 $E_{\text{in}}(g_+) < E_{\text{in}}(g_-)$，则 g_+ 是所有感知机中最小 E_{in} 的假设。
- 选项 2：测试误差与训练误差不同，无法据此判定 g_+ 是所有感知机中最小 E_{test} 的假设。
- 选项 3：两假设集有交集（$w_1=0$ 的感知机），并非不相交。 ■

15.2 验证

命题 15.2.1 (验证集模型选择法则)

1. 验证集划分　将数据集 $\mathcal{D}$（大小为 N）划分为两部分：

$$\mathcal{D}_{\text{train}} \quad (\text{大小}N-K) \qquad \mathcal{D}_{\text{val}} \quad (\text{大小}K)$$

其中 $\mathcal{D}_{\text{val}} \subset \mathcal{D}$ 称为验证集，是手头用于模拟测试集的样本，满足：

$$\mathcal{D}_{\text{val}} \overset{\text{iid}}{\sim} P(\mathbf{x}, y)$$

为了保证验证集“干净”，只允许将 $\mathcal{D}_{\text{train}}$ 输入给学习算法 $\mathcal{A}_m$ 进行模型拟合。

2. 模型选择流程　对每个模型空间 $\mathcal{H}_m$ 及其算法 $\mathcal{A}_m$（$m=1,\ldots,M$）执行：

1) 使用训练集学习：$g_m^- = \mathcal{A}_m(\mathcal{D}_{\text{train}})$；
2) 计算其在验证集上的误差：$E_{\text{val}}(g_m^-)$；
3) 选择验证误差最小的模型：

$$m^* = \arg\min_{1\leq m\leq M} \left(E_m = E_{\text{val}}(g_m^-)\right)$$

3. 泛化误差界　对所有 m，有如下 Hoeffding 泛化保证：

$$E_{\text{out}}(g_m^-) \leq E_{\text{val}}(g_m^-) + O\left(\sqrt{\frac{\log M}{K}}\right)$$

特别地，对模型选择结果 m^* 有：

$$E_{\text{out}}(g_{m^*}^-) \leq E_{\text{val}}(g_{m^*}^-) + O\left(\sqrt{\frac{\log M}{K}}\right)$$

4. 最终模型训练　选定 m^* 后，可在整个数据集 $\mathcal{D}$ 上重新训练得到最终模型：

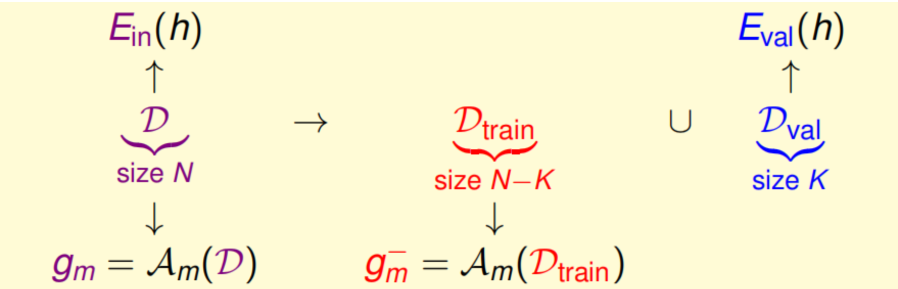
$$g_{m^*} = \mathcal{A}_{m^*}(\mathcal{D})$$

基于学习曲线的启发式收益，有：

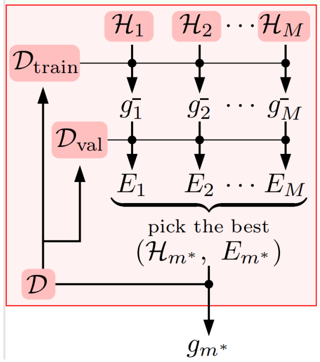
$$E_{\text{out}}(g_{m^*}) \leq E_{\text{out}}(g_{m^*}^-)$$

从而整体泛化误差满足：

$$E_{\text{out}}(g_{m^*}) \leq E_{\text{out}}(g_{m^*}^-) \leq E_{\text{val}}(g_{m^*}^-) + O\left(\sqrt{\frac{\log M}{K}}\right)$$



(a) 训练集与验证集划分示意图



(b) 基于验证集的模型选择

图 15.2.1: 模型选择流程示意图例

命题 15.2.2 (实践中的验证：子模型 vs 全模型，以及 K 的两难)

固定总数据量 N，当验证集大小 K 从 5 增至 25 时，四种策略的 E_{out} 变化规律如下：

- **in-sample**（直接用 $\mathcal{D}$ 训练并选模型）
  - 对应模型为 $g_{\hat{m}}$，其泛化误差恒定为 $E_{\text{out}}(g_{\hat{m}}) \approx 0.56$
  - 误差偏高的原因是模型发生了过拟合
- **sub-g**（用 $\mathcal{D}_{\text{train}}$ 和 $\mathcal{D}_{\text{val}}$ 选模型，最终仍用 $\mathcal{D}_{\text{train}}$ 训练）
  - 对应模型为 $g_{m^*}^-$，其 E_{out} 曲线先降低后升高
  - 在某些 K 取值下表现甚至劣于 in-sample 策略
  - 核心原因：训练数据减少 $\Rightarrow$ 所有 g_m 变差
- **full-g**（用 $\mathcal{D}_{\text{train}}$ 和 $\mathcal{D}_{\text{val}}$ 选模型，最后用完整 $\mathcal{D}$ 重新训练）
  - 对应模型为 g_{m^*}，其 E_{out} 曲线呈单调下降趋势，在 K 较大时会略微上升
  - 满足泛化误差关系：$E_{\text{out}}(g_{m^*}) \leq E_{\text{out}}(g_{m^*}^-)$
  - 经验取值：$K \approx \frac{N}{5}$ 时可接近最优泛化误差
- **optimal**（作弊选择，使用测试集误差 E_{test} 选择）
  - 理论上可达到最优性能，但实际不可行
  - 因测试集在模型训练阶段应保持不可见性

关于 K 选择的两难困境：

- K 偏大：评估准确性提高，但训练数据不足导致模型性能下降
- K 偏小：训练效果较好，但验证集对泛化误差的评估准确性降低

实战中建议取 $K \approx \frac{N}{5}$，以平衡评估精度与训练效果。

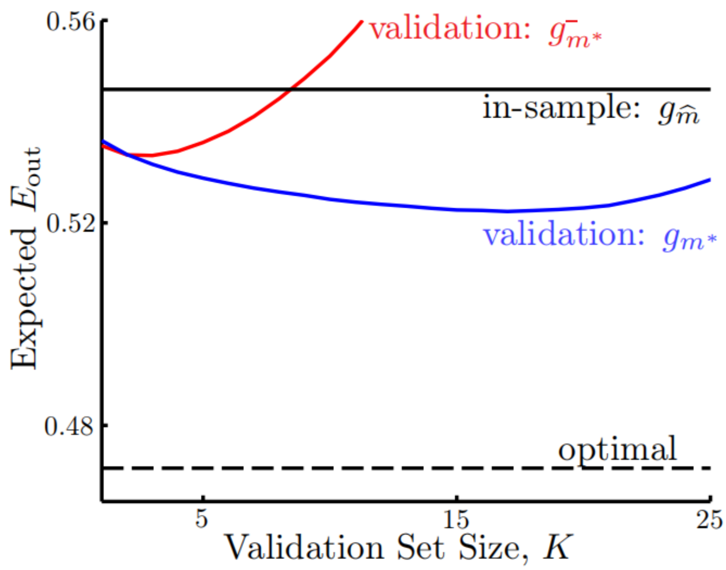


图 15.2.2: 不同模型期望样本外误差与验证集大小关系图

例题 15.2 选择题：验证过程的总计算代价

对于一个学习模型，使用 N 个样本的训练时间为 N^2 秒。当使用 $K = \frac{N}{5}$ 进行验证，并训练 25 个不同参数的模型以得到最终的 g_{m^*} 时，总时间为：

1) $6N^2$
2) $17N^2$
3) $25N^2$
4) $26N^2$

解答 正确选项为 $\boxed{2}$。验证过程分为两步：

1. 训练 25 个模型得到 g^-_m：每个模型的有效训练样本为 $N - K = \frac{4N}{5}$，时间为 $\left(\frac{4N}{5}\right)^2 = \frac{16}{25}N^2$，总时间为 $25 \times \frac{16}{25}N^2 = 16N^2$。
2. 训练最终模型 g_{m^*}：时间为 N^2。

总时间为 $16N^2 + N^2 = 17N^2$。 ■

15.3 留一交叉验证

命题 15.3.1 (极端情形：$K = 1$ 的留一交叉验证)

设总样本数为 N。当验证集大小取 $K = 1$ 时，交叉验证过程如下：

1. 单次验证过程：每一轮从样本集中留出一个样本 $(\boldsymbol{x}_n, y_n)$ 作为验证点，利用剩余的 $N - 1$ 个样本训练模型 g^-_n。

2. 单点验证误差定义：

$$E^{(n)}_{\text{val}}\left(g^-_n\right) = \text{err}\left(g^-_n(\boldsymbol{x}_n), y_n\right) = e_n$$

其中，$\text{err}(\cdot,\cdot)$ 表示模型预测值与真实值之间的误差度量函数。

3. LOOCV 估计量：留一交叉验证（Leave-One-Out Cross Validation, LOOCV）的误差估计量定义

为所有单点验证误差的平均值：

$$E_{\text{loocv}}(\mathcal{H},\mathcal{A}) \;=\; \frac{1}{N}\sum_{n=1}^{N} e_n \;=\; \frac{1}{N}\sum_{n=1}^{N} \operatorname{err}\big(g_n^-(\boldsymbol{x}_n), y_n\big)$$

4. 估计目标：我们期望该估计量能够近似反映最终模型的泛化误差：

$$E_{\text{loocv}}(\mathcal{H},\mathcal{A}) \;\approx\; E_{\text{out}}(g)$$

其中 g 表示使用全部 N 个样本训练得到的最终模型，$E_{\text{out}}(g)$ 为其泛化误差。
核心思想：每次仅留一个样本用于验证，循环执行 N 次后取平均。

命题 15.3.2 (留一交叉验证（LOOCV）的示意与理论保证)

1. 示意图　假设存在两个候选模型，分别为线性模型 $\mathcal{H}_{\text{linear}}$ 和常数模型 $\mathcal{H}_{\text{constant}}$。当对 3 个样本分别执行 LOOCV 操作时，单点误差依次记为 e_1、e_2、e_3，则有：

$$E_{\text{loocv}}(\text{linear}) = \frac{e_1+e_2+e_3}{3},$$
$$E_{\text{loocv}}(\text{constant}) = \frac{e_1+e_2+e_3}{3} \text{ (对应常数模型)}.$$

通过对这两个模型的 LOOCV 误差进行比较，我们可以选出：

$$m^* = \arg\min_{1\le m\le M} E_{\text{loocv}}(\mathcal{H}_m,\mathcal{A}_m)$$

2. 理论保证　设 g_n^- 是在数据集 $\mathcal{D}\setminus\{(\boldsymbol{x}_n,y_n)\}$（其样本数量为 $N-1$）上训练得到的模型，那么：

$$E_{\text{loocv}}(\mathcal{H},\mathcal{A}) = \frac{1}{N}\sum_{n=1}^{N} \operatorname{err}\big(g_n^-(\boldsymbol{x}_n), y_n\big) = \frac{1}{N}\sum_{n=1}^{N} E_{\text{out}}(g_n^- \mid \mathcal{D}_n)$$

其中 $\mathcal{D}_n = \mathcal{D}\setminus\{(\boldsymbol{x}_n,y_n)\}$。对上述等式两边取期望可得：

$$\mathbb{E}_{\mathcal{D}}\big[E_{\text{loocv}}(\mathcal{H},\mathcal{A})\big] = \mathbb{E}_{\mathcal{D}}\big[E_{\text{out}}(g^-)\big]$$

这里 g^- 表示在样本数量为 $N-1$ 的随机数据集上训练得到的模型。由此可见，LOOCV 被认为是 $E_{\text{out}}(g)$ 的“几乎无偏估计”。

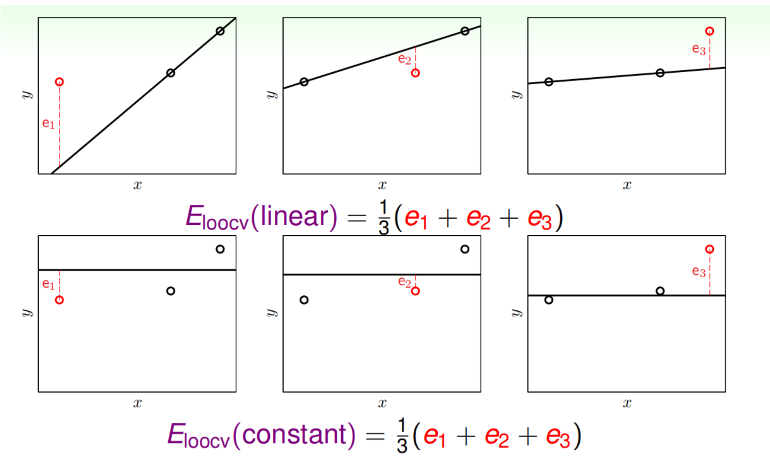


图 15.3.1: 留一交叉验证示意图

定义 15.3.1 (留一交叉验证（Leave-One-Out Cross-Validation, LOOCV））

留一交叉验证是一种穷尽式（exhaustive）交叉验证方法，适用于样本量较小的数据集。其流程如下：

1. 设数据集含 N 个样本。
2. 对每一个样本 $i = 1, 2, \dots, N$：
   - 以样本 (x_i, y_i) 为验证集；
   - 其余 $N-1$ 个样本构成训练集，训练模型 g_i^-；
   - 计算验证误差 $e_i = \mathrm{err}\big(g_i^-(x_i), y_i\big)$。
3. 最终性能估计取所有轮次的平均值

$$\text{LOOCV 误差} \triangleq \frac{1}{N}\sum_{i=1}^{N} e_i.$$

特性

- 数据利用率高：几乎利用全部数据进行训练。
- 无随机性：每次划分唯一确定。
- 计算代价大：需训练 N 次模型，N 较大时开销显著。
- 期望无偏：$\mathbb{E}[\text{LOOCV 误差}] = \mathbb{E}[E_{\text{out}}(g^-)]$，其中 g^- 为在 $N-1$ 个样本上训练的模型。♣

例题 15.3 选择题：LOOCV 平方误差计算

考虑三个样本 $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), (\boldsymbol{x}_3, y_3)$，其中 $y_1 = 1, y_2 = 5, y_3 = 7$。使用 LOOCV 估计最优常数预测（平方误差最小）的性能，其 E_{loocv} 为：

1) 0
2) $\frac{56}{9}$
3) $\frac{60}{9}$
4) 14

解答 正确选项为 $\boxed{4}$。LOOCV 每次留下一个样本，其余样本的平均为预测值：

- 留下 y_1：训练集平均为 6，误差 $(6-1)^2 = 25$；
- 留下 y_2：训练集平均为 4，误差 $(4-5)^2 = 1$；
- 留下 y_3：训练集平均为 3，误差 $(3-7)^2 = 16$。

平均误差为 $\frac{25+1+16}{3} = 14$。■

15.4 V 折交叉验证

命题 15.4.1 (留一交叉验证（LOOCV）的缺点)

1. 计算代价高　LOOCV 要求将每个样本依次留出作为验证集，总共需要训练 N 次模型，其计算公式为：

$$E_{\text{loocv}}(\mathcal{H}, \mathcal{A}) = \frac{1}{N}\sum_{n=1}^{N} \mathrm{err}\big(g_n^-(\boldsymbol{x}_n), y_n\big)$$

其中，g_n^- 表示第 n 次在 $N-1$ 个样本上训练得到的模型。

在实际应用中，只有当存在像线性回归的解析解这类特殊情况时，LOOCV 才有可能在较大的 N

下可行，否则其计算量会大到几乎无法处理。

2. 方差大、稳定性差　由于每次验证仅使用单个样本，这使得 LOOCV 的估计量对异常值极为敏感，从而导致：

$$\mathrm{Var}\big(E_{\text{loocv}}\big) \gg \mathrm{Var}\big(E_{\text{k-fold}}\big)$$

命题 15.4.2 (V 折交叉验证（V-Fold Cross Validation）)

1. 计算量缩减思路

- 留一交叉验证：将 $\mathcal{D}$ 均分为 N 份，依次取 1 份验证、$N-1$ 份训练，共训练 N 次
- V 折交叉验证改进：将 $\mathcal{D}$ 随机均分为 V 份

$$\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \cdots \cup \mathcal{D}_V, \quad |\mathcal{D}_i| \approx \frac{N}{V}$$

- 循环验证过程：第 v 折用 $\mathcal{D}_v$ 验证，其余 $V-1$ 份训练得 g_v^-
- V 折误差计算公式：

$$E_{\text{CV}}(\mathcal{H}, \mathcal{A}) = \frac{1}{V}\sum_{v=1}^{V} \frac{1}{|\mathcal{D}_v|} \sum_{(x,y)\in\mathcal{D}_v} \mathrm{err}\big(g_v^-(x), y\big)$$

2. 模型选择

- 通过交叉验证误差选择最优模型：

$$m^* = \arg\min_{1\le m\le M} E_{\text{CV}}(\mathcal{H}_m, \mathcal{A}_m)$$

3. 实用经验

- 实际应用中，推荐选择以下参数：
- $V=10$（十折交叉验证）或 $V=5$（五折交叉验证），通常能取得良好效果

4. 验证的本质与报告规范

- 各阶段核心任务：
  - 训练阶段：在假设空间内挑选假设
  - 验证阶段：在候选模型中挑选最终模型
  - 测试阶段：仅作评估，不参与任何选择
- 结果报告原则：
  - 验证误差通常比测试误差乐观，因此应报告最终测试集结果，而非最佳验证结果

定义 15.4.1 (V 折交叉验证（V-Fold Cross-Validation）)

V 折交叉验证（又称 k 折交叉验证）是一种重抽样（resampling）模型验证技术，用于评估统计模型对未知数据的泛化能力。

操作步骤

1. 将含有 N 个样本的数据集随机划分为 V 个互不相交、大小大致相等的子集（称为“折”或“folds”）。
2. 依次取第 v 折（$v=1,\ldots,V$）作为验证集，其余 $V-1$ 折作为训练集。
3. 在训练集上训练模型，并在验证集上计算预测误差 e_v。

4. 重复上述过程 V 次，最终交叉验证误差为

$$E_{\text{CV}} = \frac{1}{V}\sum_{v=1}^{V} e_v.$$

常用设置实践中常取 $V=5$ 或 $V=10$，分别称为 5 折或 10 折交叉验证。

特性

- 每个样本恰好被用作验证一次，有效利用数据；
- 通过取平均降低因单次划分带来的方差；
- 便于与留一法（$V=N$）或随机子抽样法比较，计算代价介于二者之间。

例题 15.4 选择题：10 折交叉验证的总训练时间

对于一个学习模型，使用 N 个样本的训练时间为 N^2 秒。当使用 10 折交叉验证训练 25 个不同参数的模型以得到最终的 g_{m^*} 时，总训练时间为：

1) $\frac{47}{2}N^2$
2) $47N^2$
3) $\frac{407}{2}N^2$
4) $407N^2$

解答 正确选项为 $\boxed{3}$。10 折交叉验证每次训练使用 $\frac{9N}{10}$ 个样本，时间为 $\left(\frac{9N}{10}\right)^2$。25 个模型的总训练时间为：

$$25 \times 10 \times \left(\frac{9N}{10}\right)^2 = 25 \times \frac{81}{10}N^2 = 202.5N^2$$

最终模型训练时间为 N^2，总时间为：

$$202.5N^2 + N^2 = \frac{407}{2}N^2$$

■

15.5 总结

笔记 [验证]

- 模型选择问题：用 E_{in} 危险，用 E_{test} 又不诚实。
- 验证：在训练集 $\mathcal{D}_{\text{train}}$ 上训练模型 $\mathcal{A}_m$，用验证误差 E_{val} 选模型，最后把最佳模型 $\mathcal{A}_{m^*}$ 放回完整数据 $\mathcal{D}$ 再训练。
- 留一交叉验证：计算量巨大，但估计几乎无偏。
- V 折交叉验证：计算与性能之间的折中方案。