第 23 章 混合与袋装

23.1 集成的动机

命题 23.1.1 (基于数学符号的预测融合框架)

设你有 T 位“朋友”模型 $g_1,\dots,g_T$，每个模型对输入 x 给出股价涨跌预测 $g_t(x)\in\{-1,+1\}$。

1. 选择最可信的朋友

$$G(x)=g_{t^*}(x),\quad t^*=\arg\min_{t\in\{1,\dots,T\}} \mathrm{E}_{\mathrm{val}}(g_t^-).$$

2. 均匀融合

$$G(x)=\mathrm{sign}\Big(\sum_{t=1}^{T}1\cdot g_t(x)\Big).$$

3. 非均匀加权融合

$$G(x)=\mathrm{sign}\Big(\sum_{t=1}^{T}\alpha_t\, g_t(x)\Big),\quad \alpha_t\ge 0.$$

- 仅选最优：$\alpha_t=\mathbb{I}\big[t=\arg\min_s \mathrm{E}_{\mathrm{val}}(g_s^-)\big]$；
- 均匀权重：$\alpha_t=1$；
- 任意非负权重：例如 α_t 与 $\mathrm{E}_{\mathrm{val}}(g_t^-)$ 成反比。

4. 条件融合

$$G(x)=\mathrm{sign}\Big(\sum_{t=1}^{T}q_t(x)\, g_t(x)\Big),\quad q_t(x)\ge 0,$$

其中 $q_t(x)$ 可按 α_t 或任意非负函数设定。

总结　上述框架构成丰富的融合模型族（**aggregation models**），从简单选择到复杂条件加权，灵活权衡偏差与方差。

命题 23.1.2 (验证选模 vs. 融合：动机与优势)

1. 验证选模（**Selection by Validation**）

$$G(x)=g_{t^*}(x),\quad t^*=\arg\min_{t\in\{1,\dots,T\}} \mathrm{Eval}(g_t^-).$$

- 简单、常用；
- 若以 $E_{\mathrm{in}}(g_t^-)$ 代替 $\mathrm{Eval}(g_t)$，则因模型复杂度惩罚不足而需付出“复杂度代价”；
- 依赖单个强假设保证 $\mathrm{E}_{\mathrm{val}}$（从而 E_{out}）小。

2. 融合（**Aggregation**）的动机

- 不必强求单个强模型，可利用多个（可能较弱）假设；
- 通过均匀或非均匀加权，融合效果可优于任何单一成员；
- 可视作隐式特征变换或正则化；
- 适当融合 $\Rightarrow$ 性能提升。

结论　融合策略将“选最优”升级为“组合众长”，在偏差-方差权衡中获得更优泛化。

例题 23.1 选择题：决策树桩的均匀混合

考虑三个决策树桩假设 $g_1(x) = \text{sign}(1-x)$、$g_2(x) = \text{sign}(1+x)$ 和 $g_3(x) = -1$。均匀混合这三个假设后的结果 $G(x)$ 为：

1) $2[\![|x| \le 1]\!] - 1$
2) $2[\![|x| \ge 1]\!] - 1$
3) $2[\![x \le -1]\!] - 1$
4) $2[\![x \ge +1]\!] - 1$

解答 正确选项为 $\boxed{1}$。分析各区间的预测值，其中 $G(x) = \frac{g_1(x)+g_2(x)+g_3(x)}{3}$：

- $x < -1$ 时，$G(x) = -\frac{1}{3}$；
- $-1 \le x \le 1$ 时，$G(x) = \frac{1}{3}$；
- $x > 1$ 时，$G(x) = -\frac{1}{3}$。

结果对应 $2[\![|x| \le 1]\!] - 1$。 ■

23.2 均匀混合（Uniform Blending）

命题 23.2.1 (均匀融合（Uniform Blending）)

分类情形

给定已知模型 $g_1, \ldots, g_T$，每人一票：

$$G(x) = \text{sign}\Big(\sum_{t=1}^{T} 1 \cdot g_t(x)\Big).$$

- 若所有 g_t 相同（独裁）：效果与单个 g_t 无异；
- 若 g_t 差异显著（多样性 + 民主）：多数票可纠正少数错误。

多类推广

$$G(x) = \arg\max_{k\in\{1,\ldots,K\}} \sum_{t=1}^{T} \mathbb{I}[g_t(x) = k].$$

回归情形

$$G(x) = \frac{1}{T}\sum_{t=1}^{T} g_t(x).$$

- 若所有 g_t 相同：效果等同单个模型；
- 若 g_t 差异显著：某些高估，某些低估 $\to$ 平均反而更准确。

结论 即使最简单的均匀融合，在模型多样性充足时，也能优于任意单一假设。 ♠

定理 23.2.1 (均匀融合的偏差-方差分解定理)

设均匀融合后的平均模型为

$$G(x) = \frac{1}{T}\sum_{t=1}^{T} g_t(x),$$

并记对 $t = 1, \ldots, T$ 的算术平均为 $\text{avg}[\cdot]$。则对任意输入点 x 有如下均方误差分解：

$$\text{avg}\big[(g_t(x) - f(x))^2\big] = \text{avg}\big[(g_t(x) - G(x))^2\big] + (G(x) - f(x))^2$$

对整个输入分布取期望（即取 E_{out}）后，得到

$$\frac{1}{T}\sum_{t=1}^{T} E_{\text{out}}(g_t) = \text{avg}\big[E_{\text{out}}(g_t)\big] = \text{avg}\big[E_{\text{out}}\big((g_t - G)^2\big)\big] + E_{\text{out}}(G) \geq E_{\text{out}}(G).$$

即融合后的误差不超过单个模型的平均误差。

证明 为便于书写，在固定的输入点 x 处，去掉显式的 x 依赖写作 g_t, f, G。从代数展开开始：

$$\begin{aligned}\text{avg}\big[(g_t - f)^2\big] &= \text{avg}\big[g_t^2 - 2g_t f + f^2\big] \\ &= \text{avg}(g_t^2) - 2\ \text{avg}(g_t)\, f + f^2.\end{aligned}$$

由定义 $\text{avg}(g_t) = G$，因此

$$\text{avg}\big[(g_t - f)^2\big] = \text{avg}(g_t^2) - 2Gf + f^2.$$

将 $\text{avg}(g_t^2)$ 写为关于 G 的展开：

$$\text{avg}(g_t^2) = \text{avg}\big[(g_t - G + G)^2\big] = \text{avg}\big[(g_t - G)^2 + 2(g_t - G)G + G^2\big].$$

利用线性算子与 $\text{avg}(g_t - G) = \text{avg}(g_t) - G = 0$，可得

$$\text{avg}(g_t^2) = \text{avg}\big[(g_t - G)^2\big] + G^2.$$

把此式代回前式：

$$\begin{aligned}\text{avg}\big[(g_t - f)^2\big] &= \big(\text{avg}[(g_t - G)^2] + G^2\big) - 2Gf + f^2 \\ &= \text{avg}[(g_t - G)^2] + (G^2 - 2Gf + f^2) \\ &= \text{avg}[(g_t - G)^2] + (G - f)^2,\end{aligned}$$

得到点态（fixed-x）分解：

$$\text{avg}\big[(g_t - f)^2\big] = \text{avg}\big[(g_t - G)^2\big] + (G - f)^2.$$

对输入 x 按数据分布取期望（即对整个样本空间取 $E_x[\cdot]$ 或记为 $E_{\text{out}}(\cdot)$），并用线性性交换 avg 与 E_x 的顺序：

$$\begin{aligned}\text{avg}\big[E_{\text{out}}(g_t)\big] &= \text{avg}\big[E_x\big[(g_t(x) - f(x))^2\big]\big] \\ &= E_x\big[\text{avg}[(g_t(x) - f(x))^2]\big] \\ &= E_x\big[\text{avg}[(g_t(x) - G(x))^2]\big] + E_x\big[(G(x) - f(x))^2\big] \\ &= \text{avg}\big[E_{\text{out}}\big((g_t - G)^2\big)\big] + E_{\text{out}}(G).\end{aligned}$$

由于 $\text{avg}[E_{\text{out}}((g_t - G)^2)] \geq 0$，于是

$$\text{avg}\big[E_{\text{out}}(g_t)\big] \geq E_{\text{out}}(G).$$

■

命题 23.2.2 (特殊 g_t：虚拟迭代过程的偏差-方差分解)

考虑虚拟迭代过程：对 $t = 1, 2, \ldots, T$，

- 从分布 $\mathcal{P}^N$ 独立同分布抽取大小为 N 的数据集 $\mathcal{D}_t$；
- 以算法 $\mathcal{A}$ 得到模型 $g_t = \mathcal{A}(\mathcal{D}_t)$。

令平均模型

$$G=\frac{1}{T}\sum_{t=1}^{T}g_t,\quad \bar{g}=\lim_{T\to\infty}G=\mathbb{E}_{\mathcal{D}\sim\mathcal{P}^N}\big[\mathcal{A}(\mathcal{D})\big].$$

则任意 g_t 的期望风险（泛化误差）可分解为

$$\mathbb{E}_{\mathcal{D}}\big[\mathrm{E}_{\mathrm{out}}(g_t)\big]=\underbrace{\mathrm{Bias}^2(\bar{g})}_{\text{共识模型的偏差}}+\underbrace{\mathbb{E}_{\mathcal{D}}\big[\,\mathrm{Var}(g_t)\big]}_{\text{对共识的期望方差}}.$$

结论　均匀混合（uniform blending）通过降低方差项，提高整体性能的稳定性。

例题 23.2 选择题：线性回归假设的均匀混合

设 $g_1(\mathbf{x}),g_2(\mathbf{x}),\ldots,g_T(\mathbf{x})$ 是 T 个线性回归假设，其中 $g_t(\mathbf{x})=\mathbf{w}_t^T\mathbf{x}$。对这些假设进行均匀混合得到 $G(\mathbf{x})=\frac{1}{T}\sum_{t=1}^{T}g_t(\mathbf{x})$，则 $G(\mathbf{x})$ 是：

1) 常数函数
2) 线性函数
3) 二次函数
4) 无正确选项

解答　正确选项为 $\boxed{2}$。将线性回归假设代入均匀混合公式：$G(\mathbf{x})=\frac{1}{T}\sum\limits_{t=1}^{T}\mathbf{w}_t^T\mathbf{x}=\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t\right)^T\mathbf{x}$，令 $\mathbf{w}=\frac{1}{T}\sum\limits_{t=1}^{T}\mathbf{w}_t$，则 $G(\mathbf{x})=\mathbf{w}^T\mathbf{x}$，仍为线性函数。 ■

23.3 线性/任意混合

算法 23.3.1: 线性混合（Linear Blending）

设定　已知 T 个基预测器 $g_t(x)$，给每个 g_t 赋予权重 $\alpha_t\geq 0$（投票/票数），构造最终分类器

$$G(x)=\mathrm{sign}\Big(\sum_{t=1}^{T}\alpha_t\,g_t(x)\Big).$$

回归情形　将线性混合视为线性回归 + 特征变换：

$$\min_{\alpha_t\geq 0}\sum_{n=1}^{N}\Big(y_n-\sum_{t=1}^{T}\alpha_t\,g_t(x_n)\Big)^2,$$

等价于

$$\min_{\mathbf{w}\geq 0}\sum_{n=1}^{N}\Big(y_n-\mathbf{w}^\top\Phi(x_n)\Big)^2,\quad \Phi_t(x)=g_t(x)\ (\text{把}g_t\ \text{当作新特征}).$$

符号约束的讨论

- 若允许 $\alpha_t<0$，则 $\alpha_t g_t(x)=|\alpha_t|\,(-g_t(x))$，即反向投票——对错误率极高的基模型（如 99% 错误的涨跌分类器）直接取反即可。
- 实践中常放宽到 $\alpha_t\in\mathbb{R}$，再经正则化控制。

总结　线性混合 = 线性模型 + 以 g_t 为特征变换 + 权重约束（非负或正则化）。

命题 23.3.1 (线性混合 vs. 选择（Selection）及任意混合（Any Blending）)

1. 线性混合 **vs.** 选择

- 选择（**Selection**）：从 T 个假设类 $\mathcal{H}_t$ 中各自训练出 g_t，然后再选择 $g_{\hat{t}}$：
$$\hat{t} = \arg\min_t \mathrm{E}_{\text{in}}(g_t).$$
复杂度代价为
$$d_{\text{VC}}\left(\bigcup_{t=1}^{T} \mathcal{H}_t\right) \text{（表示在所有候选模型的并集中支付的 VC 维成本）}$$
- 线性混合（**Linear Blending**）：将选择视为特例如下：
$$\text{令}\alpha_t = \begin{cases} 1, & \text{if } t = \hat{t}, \\ 0, & \text{otherwise.} \end{cases}$$
因此线性混合的复杂度代价
$$\geq d_{\text{VC}}\left(\bigcup_{t=1}^{T} \mathcal{H}_t\right).$$

2. 任意混合（**Any Blending / Stacking**）

给定 $g_1^-, g_2^-, \ldots, g_T^-$（在训练集 $\mathcal{D}_{\text{train}}$ 上训练），构造新特征
$$\Phi(x) = \left(g_1^-(x), g_2^-(x), \ldots, g_T^-(x)\right)^\top.$$
将验证集 $\mathcal{D}_{\text{val}}$ 映射为
$$\left\{(z_n = \Phi(x_n),\, y_n)\right\}_{n=1}^{N_{\text{val}}}.$$

- 线性混合：在新特征上求解
$$\alpha = \arg\min_\alpha \sum_{(z,y)\in\mathcal{D}_{\text{val}}} \ell(y,\, \alpha^\top z) \quad\Longrightarrow\quad G_{\text{LinB}}(x) = \alpha^\top \Phi(x).$$
- 任意混合（**Stacking**）：允许在特征 $\Phi(x)$ 上使用任意元学习器 $\mathcal{A}$：
$$\tilde{g} = \mathcal{A}(\{(z_n, y_n)\}), \quad G_{\text{AnyB}}(x) = \tilde{g}(\Phi(x)).$$

注意事项：

- 任意混合灵活且强大，可实现“条件混合”（conditional blending）；
- 但如同所有模型组合一样，存在过拟合风险，需结合交叉验证与正则化方法使用。

23.4 袋装（Bagging）

命题 23.4.1 (为均匀聚合而构造多样性)

要获得有效的均匀聚合，需要保证 g_t 之间的差异。常见策略：

- 不同模型：$g_1 \in \mathcal{H}_1$，$g_2 \in \mathcal{H}_2, \ldots, g_T \in \mathcal{H}_T$；
- 不同超参：如 GD 的学习率 $\eta \in \{0.001, 0.01, \ldots, 10\}$；
- 算法随机性：随机 PLA 配合不同随机种子；

- 数据随机性：交叉验证内部产生的子模型 g_v^-。

命题 23.4.2 (重访偏差-方差与自助法)

设共识模型

$$\bar{g} = \mathbb{E}_{\mathcal{D}\sim P^N}\big[\mathcal{A}(\mathcal{D})\big].$$

其期望性能满足

$$\mathbb{E}_{\mathcal{D}}\big[\mathrm{E}_{\text{out}}(g)\big] = \underbrace{\mathrm{Bias}^2(\bar{g})}_{\text{共识偏差}} + \underbrace{\mathbb{E}_{\mathcal{D}}\big[(g-\bar{g})^2\big]}_{\text{方差}}.$$

但 $\bar{g}$ 需无穷多 $\mathcal{D}_t \sim P^N$ 才能精确获得。

自助法（**Bootstrapping**）：从手头的 D 中重采样生成 $D_t \overset{\text{i.i.d.}}{\approx} P^N$，从而

$$\bar{g} \approx \frac{1}{T}\sum_{t=1}^{T}\mathcal{A}(D_t), \quad D_t \leftarrow \mathrm{Bootstrap}(D).$$

算法 23.4.2: 自助聚合 Bootstrap Aggregation（装袋法 bagging）

虚拟迭代过程（理想不可行）

1. 对 $t = 1, 2, \ldots, T$：从总体分布 P^N 独立同分布抽取大小为 N 的数据集 $\mathcal{D}_t$；
2. 由基算法 $\mathcal{A}$ 得到模型 $g_t = \mathcal{A}(\mathcal{D}_t)$；
3. 均匀聚合：$G = \dfrac{1}{T}\displaystyle\sum_{t=1}^{T} g_t$。

实际迭代过程（Bootstrap 近似）

1. 对 $t = 1, 2, \ldots, T$：从手头的数据集 D 中有放回地重采样大小为 N'（通常 $N' = N$）的数据集 $\tilde{\mathcal{D}}_t$；
2. 由基算法 $\mathcal{A}$ 得到模型 $g_t = \mathcal{A}(\tilde{\mathcal{D}}_t)$；
3. 均匀聚合：$G = \dfrac{1}{T}\displaystyle\sum_{t=1}^{T} g_t$。

总结　Bagging 是一种简单的元算法（meta-algorithm），在任意基算法 $\mathcal{A}$ 之上，通过 Bootstrap 重采样构造多样性，从而降低方差、提升稳定性。

例题 23.3 选择题：自助法重采样的概率

使用自助法从包含 N 个样本的数据集 $\mathcal{D}$ 中有放回地抽取 N 个样本组成新数据集 $\tilde{\mathcal{D}}_t$，则 $\tilde{\mathcal{D}}_t$ 与 $\mathcal{D}$ 完全相同的概率为：

1) 0
2) $\frac{1}{N^N}$
3) $\frac{N!}{N^N}$
4) 1

解答　正确选项为 $\boxed{3}$。自助法的重采样规则是有放回地抽取 N 个样本。要使 $\tilde{\mathcal{D}}_t$ 与 $\mathcal{D}$ 完全相同，需满足两个条件：

1. 每个样本被选中一次：每次抽样的概率为 $\frac{1}{N}$，独立抽样 N 次的概率为 $\left(\frac{1}{N}\right)^N$。
2. 排列组合：N 个样本的全排列数为 $N!$，即所有可能的顺序都需被考虑。

因此，概率为：$\left(\frac{1}{N}\right)^N \times N! = \frac{N!}{N^N}$ ■

23.5 总结

笔记 [混合与袋装]

- 集成的动机：聚合得到的 G 可以兼具强性能与稳健性。
- 均匀混合（Uniform Blending）：多样假设各投一票，输出平均值。
- 线性/任意混合：将假设视为特征变换进行两层学习。
- 袋装（Bagging）：通过自助采样构造多样假设，实现 Bootstrap Aggregation。

笔记 [总体结论] 混合与袋装通过融合多个模型的预测结果，借助多样性弥补单一模型的不足，在降低方差、提升泛化性能上效果显著。这一策略保留了基模型的特性，却也可能增加计算成本与过拟合风险。因此，需通过合理设计融合规则（如均匀、加权）和构造多样模型（如自助采样）平衡性能与复杂度，是增强模型稳健性的重要集成学习手段。