1 组合模型

在实际使用中, 将多个模型结合起来往往能进一步提高总体性能, 这种方法被称作**组合模型 (Combination Models)** 或**集成学习 (Ensemble Learning)**. 在组合模型中, 我们先训练多个模型, 称为**弱学习器**, 然后再将它们按照一定的方式结合起来.

在大多数情况下, 我们会使用单一的基础学习算法来训练多个弱学习器, 这样组成的模型被称作**同质的**组合模型. 如果使用不同的基础学习算法来训练弱学习器, 则称为**异质的**组合模型.

对弱学习器的选择应当与组合方法相匹配. 对于低偏差高方差的基础模型, 应当使用能减小方差的组合方法; 对于低方差高偏差的基础模型, 应当使用能减少偏差的组合方式.

1.1 委员会方法与自抽样

定义 1.1 委员会方法 训练多个不同模型, 取其均值作为结果的方法称为**委员会方法 (Committees)**.

委员会方法能有效减少模型的方差. 下面从数学上证明这一点.

证明. 假定我们有 M 个模型 $y_1(\boldsymbol{x}), \cdots, y_M(\boldsymbol{x})$, 则委员会的预测结果 y_{com} 为

$$y_{\text{com}}(\boldsymbol{x}) = \frac{1}{M}\sum_{m=1}^{M} y_m(\boldsymbol{x})$$

设第 m 个模型的预测结果 $y_m(\boldsymbol{x})$ 与真值 $h(\boldsymbol{x})$ 的关系为

$$y_m(\boldsymbol{x}) = h(\boldsymbol{x}) + \varepsilon_m(\boldsymbol{x})$$

于是该模型的均方误差为

$$\mathbb{E}_m = \left\langle [y_m(\boldsymbol{x}) - h(\boldsymbol{x})]^2 \right\rangle = \left\langle \varepsilon_m(\boldsymbol{x})^2 \right\rangle$$

于是委员会的均方误差为

$$\mathbb{E}_{\text{com}} = \left\langle \left[\frac{1}{M}\sum_{m=1}^{M} y_m(\boldsymbol{x}) - h(\boldsymbol{x}) \right]^2 \right\rangle = \left\langle \left[\frac{1}{M}\sum_{m=1}^{M} \varepsilon_m(\boldsymbol{x}) \right]^2 \right\rangle$$

我们假定每个模型都是无偏的, 即 $\langle \varepsilon_m(\boldsymbol{x}) \rangle = 0$, 且各模型之间的误差相互独立, 即 $\langle \varepsilon_m(\boldsymbol{x})\varepsilon_n(\boldsymbol{x}) \rangle = 0 (m \neq n)$. 于是

$$\mathbb{E}_{\text{com}} = \frac{1}{M^2}\sum_{m=1}^{M} \left\langle \varepsilon_m(\boldsymbol{x})^2 \right\rangle = \frac{1}{M^2}\sum_{m=1}^{M} \mathbb{E}_m = \frac{1}{M} \left\langle \mathbb{E}_m \right\rangle$$

也即委员会预测的均方误差只有单个模型平均均方误差的 $1/M$. 当然, 实际情况中个模型的误差并非完全独立, 但只要相关性不强, 委员会方法仍然能显著降低方差. □

委员会方法使用的不同模型通常是用单一基础学习算法在不同的数据集进行训练得到的. 如果简单地把原式数据集划分为 M 份, 可能导致每一份的数据量过少而产生偏差. 这就要用到自抽样法.

定义 1.2 自抽样法 **自抽样法 (Bootstrap Sampling)** 是从原始数据集中有放回地随机抽取样本, 组成新的训练集的方法. 具体而言, 对一个样本数为 N 的数据集, 按有放回随机抽样的方法抽取 N 次形成一个新的样本数为 N 的数据集 $\mathcal{D}_m$, 重复 M 次用作训练每个模型的数据集.

自抽样法在统计学上属于非参数的蒙特卡洛方法. 当自抽样法和委员会方法结合使用时, 被称为**袋装法 (Bagging)** 或**自举汇聚法 (Bootstrap Aggregating)**.

1.2 提升法和自适应增强法

前述的委员会方法使用的是并行方式组合模型, 而下面介绍的提升法使用的是串行方式组合模型.

定义 1.3 提升法 **提升法 (Boosting)** 是通过串行地训练多个弱学习器, 并将它们结合成一个强学习器的方法.

提升法的朴素思想类似于错题本, 即每次训练完成之后将预测错误的样本加大权重, 使得下一个弱学习器更关注这些难以预测的样本. 这样经过多次训练后, 最终的强学习器能更好地处理各种样本. 下面以**自适应增强法 (AdaBoost)** 为例介绍提升法的具体过程, 这是一种用于二分类问题的提升算法.

1. 初始化数据集的样本权重:

$$w_n^{(1)} = \frac{1}{N}, \quad n = 1, 2, \cdots, N$$

2. 依次对模型 $m = 1, 2, \cdots, M$ 执行以下步骤:

(a) 采用如下误差函数对模型 m 进行训练:

$$J_m = \sum_{n=1}^{N} w_n^{(m)} I(y_m(\boldsymbol{x}_n) \neq t_n)$$

其中 $I(y_m(\boldsymbol{x}_n) \neq t_n)$ 在预测预测错误时返回 1, 否则返回 0.

(b) 计算模型 m 的加权错误率:

$$\varepsilon_m = \frac{J_m}{\sum_{n=1}^{N} w_n^{(m)}}, \quad \alpha_m = \ln \frac{1 - \varepsilon_m}{\varepsilon_m}$$

这里 ε_m 即为模型 m 预测错误的比例, $\varepsilon_m = 0.5$ 对应于随机猜测. 如果 $\varepsilon_m \geqslant 0.5$, 只需把模型 m 的预测结果取反即可.

(c) 更新样本权重:

$$w_n^{(m+1)} = w_n^{(m)} \exp[\alpha_m I(y_m(\boldsymbol{x}_n) \neq t_n)]$$

即预测正确的样本权重不变, 预测错误的样本权重乘以一个大于 1 的因子 $\exp(\alpha_m)$.

3. 训练完成后, 可以用下式对新数据点进行预测:

$$y_{\text{ada}}(\boldsymbol{x}) = \sum_{m=1}^{M} \alpha_m y_m(\boldsymbol{x})$$

对于二分类问题, 通常取符号函数 $\text{sgn}(y_{\text{ada}}(\boldsymbol{x}))$ 作为最终预测结果.

1.3 决策树和随机森林

1.3.1 决策树

决策树是机器学习中的一种重要模型, 主要用于分类问题, 但也可以用于回归问题. 决策树与人类思考过程有很好的吻合程度, 具有优秀的可解释性.

决策树的主要思想是分治, 即将输入 $\boldsymbol{x}$ 在空间上划分成一些越来越小的区域, 在不同的小区域使用不同的简单模型进行预测.

通常, 决策树的节点通常是基于输入特征的某个阈值进行划分的. 例如, 对于一个二维输入 (x_1, x_2), 可以在 x_1 轴上选择一个阈值 x_1^*, 将数据划分为 $x_1 \leqslant x_1^*$ 和 $x_1 > x_1^*$ 两部分. 然后对每一部分继续选择另一个特征和阈值进行划分, 直到满足某个停止条件 (如达到最大深度或节点中的样本数过少).

决策树的训练过程通常使用贪心算法, 即在每一步选择当前最优的划分方式. 从根节点到叶节点的路径来进行预测.

1.3.2 随机森林

随机森林是基于决策树的一种集成学习方法, 即使用单颗决策树作为弱学习器的 bagging 方法.