模式识别与机器学习 23-24

第十二章作业

韦诗睿

202328002509044

https://github.com/RuiNov1st/UCAS_PRML_2324

2023 年 12 月 30 日

第十二章 集成学习

1. 模型复杂度过低/过高通常会导致 Bias 和 Variance 怎样的问题？

答：

Bias（偏差）和 Variance（方差）是对于泛化误差的分解，其中 Bias 度量了模型的期望预测和真实结果的偏离程度，反映了模型本身的拟合能力；而 Variance 刻画了数据扰动所造成的影响。

模型复杂度过低，则模型拟合数据的能力较弱，无法很好地学习数据的特征，因此 Bias 较高；同时模型对于数据的变化不敏感，学习结果较稳定，因此 Variance 较低。

反之，当模型复杂度较高时，模型可以很好地拟合数据，Bias 较低；同时模型也很容易对数据集的改变或对噪声敏感，学习结果不稳定，因此 Variance 较高。

2. 怎样判断、怎样缓解过拟合/欠拟合问题？

答：

判断：可以通过检查模型在训练集和验证集上的损失判断是否过拟合。若模型在训练集和验证集上的损失都较高，则说明模型此时还未能很好地学习数据的特征，此时模型欠拟合；若模型在训练集上的损失较低，而验证集上的损失升高，则说明模型此时过度拟合了训练数据导致无法适应新数据的变化，此时模型过拟合。

缓解：对于欠拟合问题，可以增加模型的复杂度和增加训练次数，以此提高模型学习数据的能力来缓解；对于过拟合问题，可以增加训练数据量，选择较简单的模型或在损失函数中增加正则项降低模型复杂度，以此减少模型对于训练数据的过度学习来缓解。

3. 比较 Bagging 和 Boosting 算法的异同。

答：

相同：Bagging 和 Boosting 都是集成学习的方法，通过集成多个弱学习器的结果达到一个强学习器的效果。

不同：(1)Bagging 通过 Boostrap 采样方法对数据集进行独立采样，在每个样本数据集上独立并行训练弱分类器，所有弱分类器的结果求均值/投票集成得到最后的结果。由于独立采样，因此 Bagging 可以降低方差，而偏差不变。适用于对偏差低、方差高的模型进行融合，如决策树和神经网络，代表模型是随机森林。(2) Boosting 则是通过对弱分类器进行顺序训练，希望下一个弱分类器能够弥补上一个分类器的错误，不同分类器之间可以实现互补，所有分类器的结果加权集成得到最后的结果。由于顺序训练且实现互补，因此 Boosting 可以降低偏差，但无法降低方差。代表方法是 AdaBoost 和 Gradient Boosting。

4. 简述 Adaboosting 的流程。

答：

AdaBoost 算法通过增加数据集中上一个弱学习器分错的样本的权重，使得下一个弱学习器着重学习这些错误的样本，以此达到与上一个弱学习器互补的目的。算法流程如下：对于二分类问题：

(1) 为数据样本赋予相同的初始权重 $\frac{1}{N}$

(2) 使用权重在指数损失函数下训练弱分类器，要求该分类器在数据集上的正确率 $> 50\%$；

(3) 在权重下计算分类器的错误率 ε，由此计算数据加权值 $d = \sqrt{\frac{1-\varepsilon}{\varepsilon}}$ 和模型加权值 $\alpha = \log d$；

(4) 更新数据集样本权重：分类器正确分类的样本权重除以 d，错误分类的样本权重乘 d，注意归一化；
(5) 使用新的样本权重重复上述步骤。最后集成的强分类器为多个弱分类器结果在各自 α 下的加权值。
上述算法流程图如下：

Algorithm 1 AdaBoost

Input: 数据集 $(x_1, y_1), ..., (x_N, y_N)$，其中 $y_i \in \{-1, 1\}$
Output: 集成分类器 $f(x)$

训练集样本的初始权重：$w_1 = \frac{1}{N}$
for each $m = 1$ to M **do**
    使用权重 w_m 训练弱分类器 $f_m(x)$
    计算 $f_m(x)$ 在 w_m 上的误差：$\varepsilon_m = \sum_{i, f_m(x_i) \neq y_i} w_{m,i}$
    样本权重值计算：$d_m = \sqrt{\frac{1-\varepsilon_m}{\varepsilon_m}}$；模型权重计算：$\alpha_m = \log d_m = \frac{1}{2} \log \frac{1-\varepsilon_m}{\varepsilon_m}$；归一化因子计算：$Z_m = \sum_{i, f_m(x_i) \neq y_i} w_{m,i} d_m + \sum_{i, f_m(x_i) = y_i} w_{m,i} / d_m$
    // 样本权值更新：
    for each $i = 1$ to N **do**
        if $f_m(x_i) = y_i$ **then**
            $w_{m+1,i} = \frac{w_{m,i}/d_m}{Z_m}$
        else
            $w_{m+1,i} = \frac{w_{m,i} d_m}{Z_m}$
        end if
    end for
end for
集成分类器：$f(x) = sgn(\sum_{m=1}^{M} \alpha_m f_m(x))$