防止过拟合方法

1、交叉检验，得到较优的模型参数；
2、特征选择，减少特征数或使用较少的特征组合，对于按区间离散化的特征，增大划分的区间。
3、正则化，常用的有 L_1、L_2 正则。而且 L_1 正则还可以自动进行特征选择。
4、如果有正则项则可以考虑增大正则项参数 lambda.
5、增加训练数据，有限避免过拟合.
6、Bagging ,将多个弱学习器Bagging 一下效果会好很多，比如随机森林等。

解决欠拟合

- SVM：增大惩罚参数C

多特征大数据训练方法

- 随机抽取少量样本
- 试用在线机器学习算法
- PCA降维，减少特征数

倾斜数据集处理（假设正100，负50） （类别不平衡）

- 从正100中抽取50（欠采样）
- 复制负50为两份（过采样）
- 负权重为正的两倍（阈值移动，再缩放？）
- 性能度量使用准确率和召回率的F1度量，而非使用准确度

$$\frac{y'}{1-y'} = \frac{y}{1-y} \times \frac{m^-}{m^+}$$

降维可以用的方法

- 前向特征选择
- 后向特征排除
- 训练后去掉一个特征，交叉验证看测试集表现，有提升就可以去掉
- 去除相关性高的特征

偏差&方差窘境

偏差大是欠拟合（可以增加特征），方差大是过拟合

k折交叉验证评价

- k不是越大越好，开销
- 但大k会有更小的偏差
- 选择k时要最小化数据集之间的方差

5. 根据上述计算结果, 我们可以发现,

$$\mathbb{E}[\bar{x}_m^*] = \mathbb{E}[\bar{x}_m] = \mu \tag{50}$$

$$var[\bar{x}_m^*] = \frac{\sigma^2}{m}[2 - \frac{1}{m}] = (2 - \frac{1}{m})var[\bar{x}_m] \approx 2var[\bar{x}_m] \tag{51}$$

另外我们如果采用 k 折交叉验证法的方式采样, 类似地我们有,

$$\mathbb{E}[\bar{x}_m^{'}] = \mathbb{E}[\bar{x}_m] = \mu \tag{52}$$

$$var[\bar{x}_m^{'}] \approx \frac{k}{k-1}var[\bar{x}_m] \tag{53}$$

综上所述, 虽然通过自助采样法得到的样本均值仍然是总体均值的无偏估计, 但是其方差变为原来的接近两倍, 而这相当于使用 2 折交叉验证采样的效果, 所以一般来说, 自助法采样对数据分布的改变大于交叉验证法.

自助法缺点

- 改变数据集分布，引起了估计偏差（方差变为原来的近两倍）
- 相当于2折交叉验证

可以用神经网络构造的算法

- 线性回归
- 对数几率回归
- ~~KNN~~

杂

- 减少决策树处理大数据时间：减少树深度
- 用非线性可分的SVM目标函数构建线性可分模型：C无穷
- 增加模型复杂度（增加神经网络层数），训练错误率一定降低，但测试集可能增加
- 训练完SVM，非支持向量的样本可以去掉，也可以继续分类
- 使用PCA之前必须规范化数据