1. **过拟合的原因:**
2. 数据集太小
3. 选择的模型太复杂
4. 数据采集有偏差

* 建模样本选取有误，如样本数量太少，选样方法错误，样本标签错误等，导致选取的样本数据不足以代表预定的分类规则
* 样本噪音干扰过大，使得机器将部分噪音认为是特征从而扰乱了预设的分类规则
* 假设的模型无法合理存在，或者说是假设成立的条件实际并不成立
* 参数太多，模型复杂度过高
* 对于决策树模型，如果我们对于其生长没有合理的限制，其自由生长有可能使节点只包含单纯的事件数据(event)或非事件数据(no event)，使其虽然可以完美匹配（拟合）训练数据，但是无法适应其他数据集
* 对于神经网络模型：a)对样本数据可能存在分类决策面不唯一，随着学习的进行,，BP算法使权值可能收敛过于复杂的决策面；b)权值学习迭代次数足够多(Overtraining)，拟合了训练数据中的噪声和训练样例中没有代表性的特征

**一般来说防止过拟合的方法有：**

1. 获取更多数据
2. 减少特征变量
3. 限制权值（正则化）
4. 贝叶斯方法
5. 结合多种模型
6. 正则化（Regularization）（L1和L2）
7. 数据扩增，即增加训练数据样本
8. Dropout
9. Early stopping
10. **欠拟合的原因:**
11. 选择的模型过于简单
12. 特征量太少

**如何防止欠拟合:**

1. 引入新的特征
2. 添加多项式特征
3. 减少正则化参数
4. L0,L1,L2正则化，也叫L0,L1,L2范数

* 增加新特征，可以考虑加入进特征组合、高次特征，来增大假设空间
* 添加多项式特征，这个在机器学习算法里面用的很普遍，例如将线性模型通过添加二次项或者三次项使模型泛化能力更强
* 减少正则化参数，正则化的目的是用来防止过拟合的，但是模型出现了欠拟合，则需要减少正则化参数
* 使用非线性模型，比如核SVM 、决策树、深度学习等模型
* 调整模型的容量(capacity)，通俗地，模型的容量是指其拟合各种函数的能力
* 容量低的模型可能很难拟合训练集；使用集成学习方法，如Bagging ,将多个弱学习器Bagging

**偏差（Bias）**

首先error=bias+variance

Bias反映的是模型在样本上的输出与真实值之间的误差，即模型本身的精确度。

**方差（Variance）**

Variance反映的是模型每一次输出结果与模型输出期望之间的误差，即模型的稳定性。