**什么是过拟合和欠拟合，怎么解决？**

**过拟合和欠拟合的解释**

* **欠拟合**是指模型在训练集、验证集和测试集上均表现不佳的情况；
* **过拟合**是指模型在训练集上表现很好，到了验证和测试阶段就很差，即模型的泛化能力很差。

**过拟合和欠拟合产生的原因：**

**欠拟合(underfitting)：**

* 模型复杂度过低
* 特征量过少

**过拟合(overfitting)：**

* 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

**解决欠拟合(高偏差)的方法**

**1、模型复杂化**  
对同一个[算法](https://edu.csdn.net/course/detail/40020?utm_source=glcblog&spm=1001.2101.3001.7020)复杂化。例如回归模型添加更多的高次项，增加决策树的深度，增加神经网络的隐藏层数和隐藏单元数等  
弃用原来的算法，使用一个更加复杂的算法或模型。例如用神经网络来替代线性回归，用[随机森林](https://edu.csdn.net/cloud/ml_summit?utm_source=glcblog&spm=1001.2101.3001.7020)来代替决策树等  
**2、增加更多的特征，使输入数据具有更强的表达能力**  
特征挖掘十分重要，尤其是具有强表达能力的特征，往往可以抵过大量的弱表达能力的特征。  
特征的数量往往并非重点，质量才是，总之强特最重要。  
能否挖掘出强特，还在于对数据本身以及具体应用场景的深刻理解，往往依赖于经验。  
**3、调整参数和超参数**  
超参数包括：  
神经网络中：学习率、学习衰减率、隐藏层数、隐藏层的单元数、Adam优化算法中的β1和β2参数、batch\_size数值等。  
其他算法中：随机森林的树数量，k-means中的cluster数，正则化参数λ等。  
**4、增加训练数据往往没有用**  
欠拟合本来就是模型的学习能力不足，增加再多的数据给它训练它也没能力学习好。  
**5、降低正则化约束**  
正则化约束是为了防止模型过拟合，如果模型压根不存在过拟合而是欠拟合了，那么就考虑是否降低正则化参数λ或者直接去除正则化项

**解决过拟合(高方差)的方法：**

**1、增加训练数据数**

* 发生过拟合最常见的现象就是数据量太少而模型太复杂
* 过拟合是由于模型学习到了数据的一些噪声特征导致，增加训练数据的 量能够减少噪声的影响，让模型更多地学习数据的一般特征
* 增加数据量有时可能不是那么容易，需要花费一定的时间和精力去搜集处理数据
* 利用现有数据进行扩充或许也是一个好办法。例如在图像识别中，如果没有足够的图片训练，可以把已有的图片进行旋转，拉伸，镜像，对称等，这样就可以把数据量扩大好几倍而不需要额外补充数据
* 注意保证训练数据的分布和测试数据的分布要保持一致，二者要是分布完全不同，那模型预测真可谓是对牛弹琴了。  
  **2、使用正则化约束**  
  代价函数后面添加正则化项，可以避免训练出来的参数过大从而使模型过拟合。使用正则化缓解过拟合的手段广泛应用，不论是在线性回归还是在神经网络的梯度下降计算过程中，都应用到了正则化的方法。常用的正则化有l1正则和l2正则，具体使用哪个视具体情况而定，一般l2正则应用比较多。  
  **3、减少特征数**  
  欠拟合需要增加特征数，那么过拟合自然就要减少特征数。去除那些非共性特征，可以提高模型的泛化能力.  
  **4、调整参数和超参数**  
  不论什么情况，调参是必须的  
  **5、降低模型的复杂度**  
  欠拟合要增加模型的复杂度，那么过拟合正好反过来。  
  **6、使用Dropout**  
  这一方法只适用于神经网络中，即按一定的比例去除隐藏层的神经单元，使神经网络的结构简单化。  
  **7、提前结束训练**  
  即early stopping，在模型迭代训练时候记录训练精度(或损失)和验证精度(或损失)，如果模型训练的效果不再提高，比如训练误差一直在降低但是验证误差却不再降低甚至上升，这时候便可以结束模型训练了。