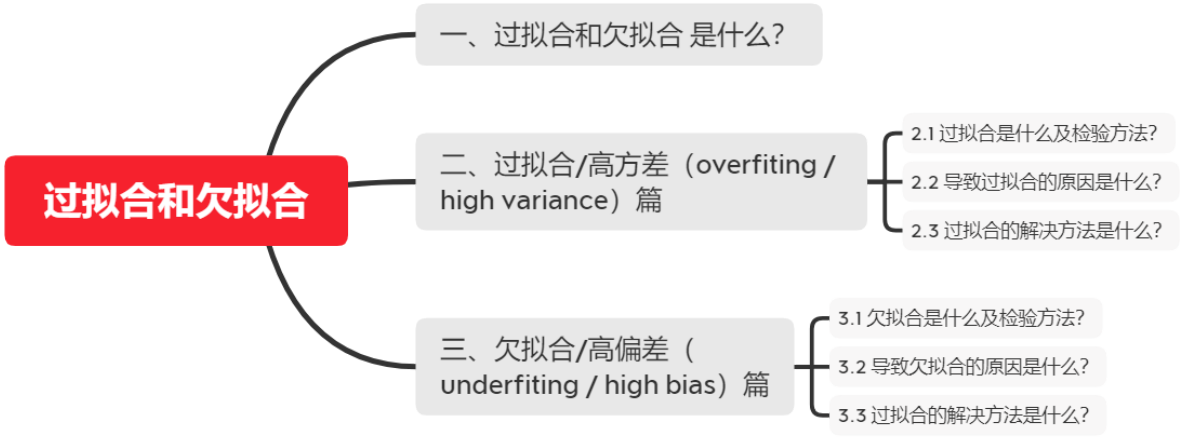
【关于 过拟合和欠拟合】那些你不知道的事



一、过拟合和欠拟合 是什么？

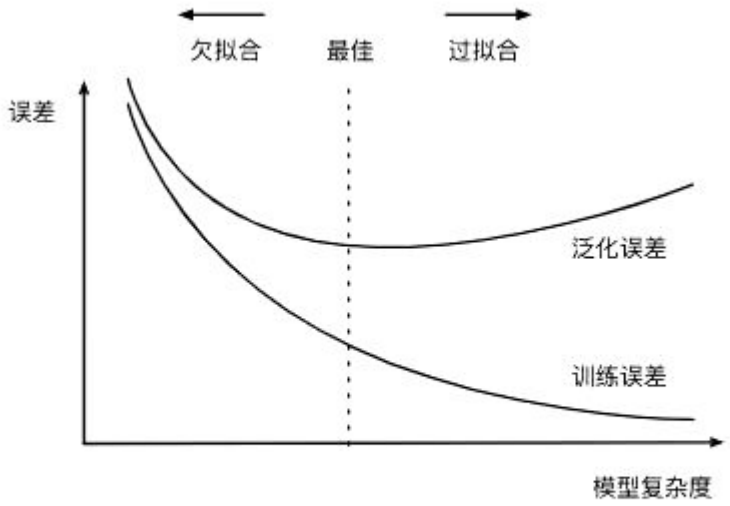


图 3.4 模型复杂度对欠拟合和过拟合的影响

欠拟合和过拟合属于对立情况，都是导致模型泛化能力不高的两种常见原因，均是模型学习能力和数据复杂性失调的表现

二、过拟合/高方差（overfiting / high variance）篇

2.1 过拟合是什么及检验方法？

- 问题表现方式：高方差
  - 如果 训练集 和 测试集 的 误差间 呈现较大的差异时，即为高方差;
  - 在 高方差 时，训练集 训练效果很好， 但是 验证集 的验证效果很差的时候， 即 训练集 和 验证集 呈现出 较大的差异，即模型的泛化能力差。这种现象 称为 过拟合;
- 检验方法：此时，观察模型在训练集和测试集上的损失函数值随着epoch的变化情况，当模型 在 测试集 上的 损失函数值 出现 先下降后上升，那么此时可能出现过拟合。

2.2 导致过拟合的原因是什么？

1. 训练集数量不足，样本类型单一。例如：如果 我们 利用 只包含 负样本的训练集 训练 模型，然后利用训练好的模型 预测 验证集中 的 正样本时，此时就会出现，模型 在 训练的时候，效果特别好，但是在验证的时候效果下降问题。因此，在选取训练集时，应当覆盖所有的数据类型；
2. 训练集中存在噪声。噪声指的是 训练数据中 的 干扰数据，噪声数据 会 误导模型 记录较多 的 错误特征，而 忽略了 真实样本 中 的正确特征信息；
3. 模型复杂度过高。当模型过于复杂时，会导致 模型 **过于充分** 的 学习到 训练数据集中特征信息，但是遇到没有见过的数据的时候不能够变通，泛化能力太差。我们希望模型对不同的数据都有稳定的输出。模型太复杂是过拟合的重要因素。

2.3 过拟合的解决方法是什么？

1. 标注不同类型的样本，是 样本尽可能的均衡。数据经过清洗之后再进行模型训练，防止噪声数据干扰模型；
2. 降低训练模型复杂度。在训练和建立模型的时候，从相对简单的模型开始，不要一开始就把特征做的非常多，模型参数挑的非常复杂；
3. 正则化。在模型算法中添加惩罚函数来防止模型出现过拟合问题。常见的有L1，L2，dropout 正则化等。而且 L1正则还可以自动进行特征选择；
4. 采用 bagging(如随机森林等）集成学习方法 来 防止过拟合；
5. 减少特征个数(不是太推荐，但也是一种方法)。可以使用特征选择，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；
6. 交叉检验。利用 交叉检验的方法，来让模型得到充分的训练，以得到较优的模型参数；
7. 早停策略。本质上是交叉验证策略，选择合适的训练次数，避免训练的网络过度拟合训练数据；
8. DropOut策略。核心思想就是bagging，可以看作是低成本的集成学习。所谓的Dropout指的是在用前向传播算法和反向传播算法训练DNN模型时，一批数据迭代时，随机的从全连接DNN网络中去掉一部分隐藏层的神经元。 在对训练集中的一批数据进行训练时，我们随机去掉一部分隐藏层的神经元，并用去掉隐藏层的神经元的网络来拟合我们的一批训练数据。使用基于dropout的正则化比基于bagging的正则化简单，这显而易见，当然天下没有免费的午餐，由于dropout会将原始数据分批迭代，因此原始数据集最好较大，否则模型可能会欠拟合。

三、欠拟合/高偏差（underfiting / high bias）篇

3.1 欠拟合是什么及检验方法？

- 问题表现：高偏差
  - 如果 训练集 和 测试集 的 误差 收敛 但是收敛值 很高时，即为高偏差;
  - 虽然 训练集 和 测试集 都可以收敛，但是偏差很高，训练集和验证集的准确率都很低，这种现象 称为 欠拟合;
- 检验方法：模型 无法很好的拟合数据，导致 训练集和测试集效果都不佳。

3.2 导致欠拟合的原因是什么？

- 原因：模型没有 充分 学习到 数据中的特征信息，使得 模型 无法很好地拟合数据

3.3 欠拟合的解决方法是什么？

1. 特征工程。添加更多的特征项，eg：特征组合、高次特征 等，来增大假设空间；
2. 集成学习方法。 boosting（如GBDT）能有效解决 high bias；
3. 提高 模型复杂度。当 所采用的模型比较简单，不能够应对复杂的任务。可以考虑 提升模型复杂度，选用复杂度更好、学习能力更强的模型。比如说可以使用 SVM 的核函数，增加了模型复杂度，把低维不可分的数据映射到高维空间，就可以线性可分，减小欠拟合；
4. 减小正则化系数。

参考资料

1. 为什么PCA不被推荐用来避免过拟合？