1.欠拟合

训练误差和验证误差都很大，这种情况称为欠拟合。出现欠拟合的原因是模型尚未学习到数据的真实结构。因此，模拟在训练集和验证集上的性能都很差。

解决办法

1 做特征工程，添加跟多的特征项。如果欠拟合是由于特征项不够，没有足够的信息支持模型做判断。

2 增加模型复杂度。如果模型太简单，不能够应对复杂的任务。可以使用更复杂的模型，减小正则化系数。具体来说可以使用核函数，集成学习方法。

3 集成学习方法boosting（如GBDT）能有效解决high bias

2.过拟合（overfiting / high variance）

模型在训练集上表现很好，但是在验证集上却不能保持准确，也就是模型泛化能力很差。这种情况很可能是模型过拟合。

造成原因主要有以下几种：

1 训练数据集样本单一，样本不足。如果训练样本只有负样本，然后那生成的模型去预测正样本，这肯定预测不准。所以训练样本要尽可能的全面，覆盖所有的数据类型。

2 训练数据中噪声干扰过大。噪声指训练数据中的干扰数据。过多的干扰会导致记录了很多噪声特征，忽略了真实输入和输出之间的关系。
3 模型过于复杂。模型太复杂，已经能够死记硬背记录下了训练数据的信息，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模型都有稳定的输出。模型太复杂是过拟合的重要因素。

针对过拟合的上述原因，对应的**预防和解决办法**如下：
1 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，模型参数跳的非常复杂。
2 增加样本，要覆盖全部的数据类型。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。
3 正则化。在模型算法中添加惩罚函数来防止过拟合。常见的有L1，L2正则化。
4 集成学习方法bagging(如随机森林）能有效防止过拟合
5 减少特征个数(不是太推荐)
注意：降维不能解决过拟合。降维只是减小了特征的维度，并没有减小特征所有的信息。