4 建模与调参

代码在自己的文件夹中
参考链接：https://blog.csdn.net/weixin_43532000/article/details/105228284
参考链接：https://github.com/datawhalechina/team-learning/blob/master/%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98%E5%AE%9E%E8%B7%B5%EF%BC%88%E4%BA%8C%E6%89%8B%E8%BD%A6%E4%BB%B7%E6%A0%BC%E9%A2%84%E6%B5%8B%EF%BC%89/Task4%20%E5%BB%BA%E6%A8%A1%E8%B0%83%E5%8F%82%20.md

1. 线性回归模型：
    - 线性回归对于特征的要求；
    - 处理长尾分布；
    - 理解线性回归模型；
2. 模型性能验证：
    - 评价函数与目标函数；
    - 交叉验证方法；
    - 留一验证方法；
    - 针对时间序列问题的验证；
    - 绘制学习率曲线；
    - 绘制验证曲线；
3. 嵌入式特征选择：
    - Lasso 回归；
    - Ridge 回归；
    - 决策树；
4. 模型对比：
    - 常用线性模型；
    - 常用非线性模型；
5. 模型调参：
    - 贪心调参方法；
    - 网格调参方法；
    - 贝叶斯调参方法；

4.1 模型的介绍，

（1） 线性回归模型
https://zhuanlan.zhihu.com/p/49480391
（2） 决策树模型
https://zhuanlan.zhihu.com/p/65304798
（3） GBDT 模型
https://zhuanlan.zhihu.com/p/45145899
（4） XGBoost 模型
https://zhuanlan.zhihu.com/p/86816771
（5） LightGBM 模型
https://zhuanlan.zhihu.com/p/89360721

4.2 模型在训练的过程中往往会表现出两种特征：

过拟合与欠拟合
过拟合的原因：

1.模型没有很好或足够数量的训练训练集

2.训练数据和测试数据有偏差

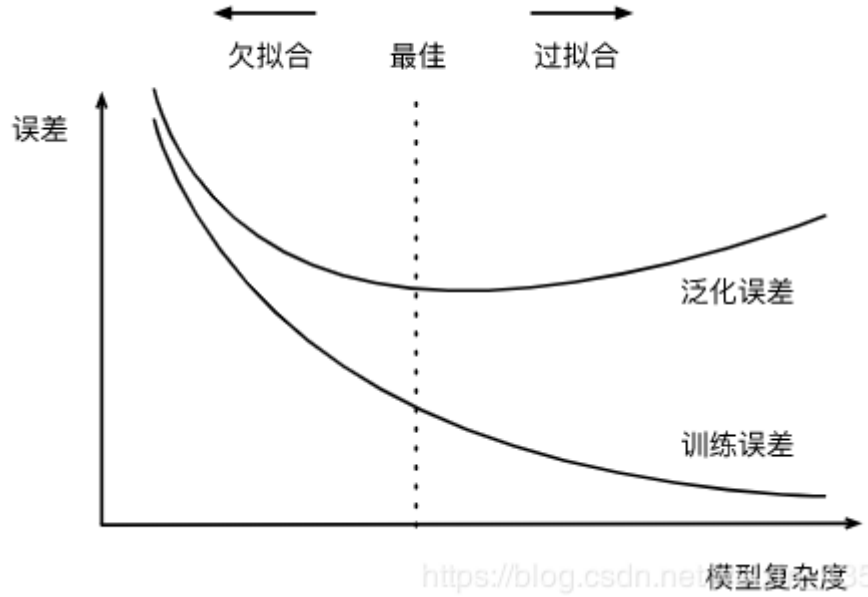
3.模型的训练过度，过于复杂，没有学到主要的特征

欠拟合的原因：

1.模型没有很好或足够数量的训练训练集

2.模型的训练特征过于简单

模型的复杂度：



4.3 数据内存的处理

在开展建模过程中的一项重要任务就是调整数据的类型，能够帮助数据在运算过程中所占用的内存空间；调节成字节小的数据类型，降低在计算机的内存占用，提高训练速度；

4.4 正式建模前的预分析

（1）简单的建模分析，先引入了线性回归+五折交叉验证+模拟真实业务的情况开展分析。
（2）查看训练的线性回归模型的截距（intercept）与权重(coef)

（3）如果发现预测的数据呈现长尾分布的话，用运行对数函数进行分布状态的转换。

（4）五折交叉验证：
在使用训练集对参数进行训练的时候，经常会发现人们通常会将一整个训练集分为三个部分（比如 mnist 手写训练集）。一般分为：训练集（train_set），评估集（valid_set），测试集（test_set）这三个部分。这其实是为了保证训练效果而特意设置的。其中测试集很好理解，其实就是完全不参与训练的数据，仅仅用来观测测试效果的数据。而训练集和评估集则牵涉到下面的知识了。
因为在实际的训练中，训练的结果对于训练集的拟合程度通常还是挺好的（初始条件敏感），但是对于训练集之外的数据的拟合程度通常就不那么令人满意了。因此我们通常并不会把所有的数据集都拿来训练，而是分出一部分来（这一部分不参加训练）对训练集生成的参数进行测试，相对客观的判断这些参数对训练集之外的数据的符合程度。这种思想就称为交叉验证（Cross Validation）

（5）模拟真实业务情况：

但在事实上，由于我们并不具有预知未来的能力，五折交叉验证在某些与时间相关的数据集上反而反映了不真实的情况。通过 2018 年的二手车价格预测 2017 年的二手车价格，这显然是不合理的，因此我们还可以采用时间顺序对数据集进行分隔。在本例中，我们选用靠前时间的 4/5 样本当作训练集，靠后时间的 1/5 当作验证集，最终结果与五折交叉验证差距不大

（6）绘制学习曲率与验证曲线

4.5 多种模型的的对比

多种模型对比

（1）线性模型 & 嵌入式特征选择

在过滤式和包裹式特征选择方法中，特征选择过程与学习器训练过程有明显的分别。而嵌入式特征选择在学习器训练过程中自动地进行特征选择。嵌入式选择最常用的是 L1 正则化与 L2 正则化。在对线性回归模型加入两种正则化方法后，他们分别变成了岭回归与 Lasso 回归。

（2）非线性模型

除了线性模型以外，还有许多我们常用的非线性模型如下。我们选择了部分常用模型与线性模型进行效果比对。

（3）模型调参

在此我们介绍了三种常用的调参方法如下：

贪心算法 https://www.jianshu.com/p/ab89df9759c8

网格调参 https://blog.csdn.net/weixin_43172660/article/details/83032029

贝叶斯调参 https://blog.csdn.net/linxid/article/details/81189154