机器学习讲稿

问题背景

- 为了环境中微生物 如肠道 土壤中微生物的中功能，需要对DNA进行测序
- 降低分析成本 对于DNA片段不进行组装 直接进行功能预测
- 对于新型冠状肺炎等 病毒的遗传信息 需要对DNA 片段进行分类

Input DNA 片段 Output DNA 类型

数据处理

核心问题: **解决输入序列长度不一致问题**

基于序列特征

- 方法来源 通过文献确定
- 统计序列中 A T C G 以及 从AAA AAT AAC AAG ATA ATT ... GGG的数量
- 进行normalization
- 共计68维数据

通过这种方法将每一个样本的维度统一

one-hot 编码

将 ATCG四个字母都进行one-hot 编码 由于编码后维数较低，所以不进行pre-training 找到每一个碱基对应的隐向量

编码后若一个序列长度为k 则该样本被转化为[1,k,4]维度的数据

此类数据适合RNN LSTM GRU bi-RNN等模型进行训练

模型

决策树类型

- 单一决策树
- gdbt 重点 残差 *可以与resnet做类比* 核心 每一次学习的一定不比上一次学习的差
- Random Forest **bagging**

神经网络类型

- MLP 多层感知机 构建网络 68 256 4 激活函数 Relu 分类效果.....
- RNN ...
  - 模型优化 由于会出现梯度爆炸的问题 尽管可以进行梯度裁剪 但RNN在实际中难以捕捉到时间序列中时间步距离较大的依赖 $\frac{\partial L}{\partial h_t}$ 会出现指数项，这里会出现梯度爆炸
  - LSTM
  - GRU
  - 模型缺点： 模型复杂度明显升高，训练难度增大

训练技巧

- 超参数调优 **最关键参数：学习率** 如果学习率过大，会出现梯度爆炸等问题
- 梯度裁剪 当loss 输出为 nan 时需要进行梯度裁剪
- EarlyStop 当test dataset的 loss 连续20 epoch都没有减小 停止训练
- Dropout 防止过拟合

生成式模型

朴素贝叶斯 朴素贝叶斯是生成方法，也就是直接找出特征输出Y和特征X的联合分布P(X,Y),然后用 $P(Y|X) = \frac{P(X,Y)}{P(X)}$ 得出。

- 分类效果

问题： 假设各个变量相互独立 这个并不科学!

结果分析与收获

模型可改进部分

- 在特征工程中进一步确定各个变量的相互作用关系
- 在RNN LSTM GRU模型中可以进一步优化超参数 如修改网络隐藏层 可以考虑biRNN 双向模型可以捕捉到序列双向特征对于分类的影响
- 在生成方法中可以考虑使用贝叶斯信念网

收获

- 面对真实数据，解决实际问题，如特征工程等
- 对于多个算法进行学习