特征工程（基础）

知识点

数据类型

在表格类数据建模过程中，常常需要处理的数据类型：

1. 文本；
    - 需要做特征表示；（词袋模型、tf-idf等）
2. 类别变量；
    - 类别数量较少可直接onehot；
    - 类别数量较多可以尝试均值编码等方案；
    - lightgbm可以支持直接的类别型特征输入，xgb等其他模型需要onehot；
3. 排序变量；

    一般可以按照连续性变量处理；
4. 连续变量；
    - 线性模型、逻辑回归等需要进行标准化，缺失处理；
    - 树模型不需要进行标准化，一般不需要处理缺失；

数据流程

一般在数据处理与特征工程的工作流程为：

1. 异常值处理；
2. 特征构造；
3. 分布调整与标准化
4. 缺失处理；

在具体项目中，2，3，4可以调整顺序；

数据处理

- 为什么要处理缺失
- 缺失值处理
- 数据标准化方法
    - 分布良好的数据可以直接进行中心标准化 (x - mu) / std
    - 有偏分布可以尝试做log等非线性单调变换后再进行中心标准化 (x - mu) / std
    - 对于较为特殊的分布可以尝试RankGauss标准化；查看sklearn QuantileTransformer文档

特征挖掘

- 特征工程是什么
- python 字符串方法

- 正则表达式
- 文本特征稀疏表示：词袋、ngram、tf-idf

  阅读文章到tf-idf，后面部分先不看。
- 类别型特征
- 类别型特征：均值编码
- [阅读]CTR如何构造特征，Louis回答
- 特征选择概览
- 特征选择实战
- [阅读]特征工程概览
- [阅读]高阶特征工程

QA

1. 两个类别型变量构造笛卡尔特征组合为什么能提升模型表现?

练习

利用文本、特征组合、模型融合等技术，将Titanic项目做到 0.80-0.85得分，越高越好。