特征空间介绍

九章-仲青

模型是什么?

- y = f(X)
- F: 逻辑回归/决策树/随机森林/GBDT/神经网络等等
- Y 有哪些?
  - 分类问题
  - 回归问题
- X：命名为 特征空间

特征空间的含义

- 多大?
  - 特征类别
  - 特征个数
- 特征类型
  - Categorical
  - Continuous
  - Mix
  - 可扩展的其他类型

性别
城市
年龄

电商的 item_id
商品价格

搜索词

如何用代码表达特征空间?

- Categorical
  - LabelEncoder/OneHotEncoder
- Continuous
  - 保持不变
- Mix??
  - 自定义代码，满足灵活性和扩展性

自定义特征空间的逻辑

- 字符串形式的样本
- 特征类定义
  - 填充特征空间
    - 特征的实质含义
    - 特征的数字ID
    - 特征覆盖了多少样本, 覆盖数特别小的特征需要过滤
  - 利用填充好的特征集合 转化字符串形式为可训练的格式 (这里是 libsvm 格式)

- Libsvm format: y qid:xx fid:fvalue .. Fid:fvalue

代码实例

- 演示代码实例