Z1 特征工程

- 特征归一化
  - 目的

    将所有的特征统一到一个大致相同的数值区间内

    消除数据特征之间的量纲影响

    使得不同指标之间具有可比性
  - 方法

    线性函数归一化（Min-Max Scaling）

$$X = \frac{X - X_{min}}{X_{max} - X_{min}}$$

    零均值归一化（Z-Score Normalization）

$$z = \frac{x - \mu}{\sigma}$$

> 数据归一化不是万能的。通过梯度下降法求解的模型通常需要归一化，包括线性回归、逻辑回归、支持向量机、神经网络等，但对于决策树模型并不适用。

- 类别型特征
  - 编码方式
    - 序号编码（Ordinal Encoding）

      处理类别间具有大小关系的数据
    - 独热编码（One-hot Encoding）

      处理类别间不具有大小关系的特征
    - 二进制编码（Binary Encoding）

利用二进制对ID进行哈希映射，最终得到0/1特征向量，维数少于独热编码，节省存储空间。

▼ 高维组合特征的处理

为了提高复杂关系的拟合能力，在特征工程中经常会把一阶离散特征两两组合，构成高阶组合特征。

将用户（m个）和物品（n个）分别用k维的低维向量表示，降低参数数量。

▼ 组合特征

举例：一种基于决策树的特征组合寻找方法，每一条从根节点到叶节点的路径都可以看成一种特征组合方式。

▼ 文本表示模型（略）

▼ Word2Vec（略）

CBOW和Skip-gram

▼ 图像数据不足时的方法

▼ 基于模型的方法：降低过拟合风险

简化模型（如将非线性模型简化为线性模型）

添加约束项以缩小假设空间（如L1/L2正则化）

集成学习

迁移学习：借用一个在大规模数据集上预训练好的通用模型，并在针对目标任务的小数据集上进行微调（fine-tune）

Dropout超参数等

▼ 基于数据的方法：数据扩充

GAN

一定程度内的随机旋转、平移、缩放、裁剪、填充、左右翻转等

对像素添加噪声扰动，比如椒盐噪声、高斯白噪声等

颜色变换

改变图像的亮度、清晰度、对比度、锐度等

> 也可以先对图像进行特征提取，然后在图像的特征空间内进行变换，利用一些通用的数据扩充或上采样技术，如SMOTE。