# 第一章 特征工程

## 1.特征归一化

目的：消除数据特征之间的量纲影响，将所有的特征都统一到一个大致相同的数值区间内，使模型更快地通过梯度下降找到最优解

方法：

（1）线性函数归一化 (Min-Max Scaling)

将原始数据映射到 [0,1] 区间内

X_norm = (X - X_min) / (X_max - X_min)

（2）零均值归一化（Z-Score Normalization）

将原始数据映射到均值为0，标准差为1的分布上

假设原始特征的均值为μ，标准差为σ

z = (x - μ) / σ

通过梯度下降求解的模型通常需要归一化（线性回归，逻辑回归，SVM，神经网络等），但决策树模型并不需要，因为决策树模型进行节点分裂是通过信息增益来计算的。

## 2.类别型特征

类别型特征原始输入一般是字符串形式，除了决策树模型能直接处理字符串形式输入，对于逻辑回归、SVM等模型来说，类别型特征必须转换成数值型特征

**序号编码（Ordinal Encoding）**

用于处理类别间具有大小关系的特征，比如成绩按高中低的排序关系，编码成3,2,1的数值ID，仍然保留大小关系

**独热编码（One-hot Encoding）**

用于处理类别间不具有大小关系的特征，例如血型（A, B, AB, O），将每个血型变成一个4维的系数向量，A血型为（1,0,0,0），B血型为（0,1,0,0）如此类推

对于类别取值较多的情况使用独热编码注意以下问题：

（1）使用稀疏变量来节省空间

（2）配合特征选择来降低维度。因为使用独热编码后，变相的增加了特征维度，可以考虑配合特征选择来降低维度

**二进制编码（Binary Encoding）**

二进制编码分为两步：

（1）给每个类别赋予一个ID

（2）将ID转换为二进制表示

如血型A,B,AB,O分别赋予1,2,3,4，将其转变为对应的二进制编码：

1 -> 001

2 -> 010

3 -> 011

4 -> 100

## 3.高维组合特征的处理

把一阶离散特征两两组合，构成高阶组合特征

| 是否点击 | 语言 | 类型 |
| --- | --- | --- |
| 0 | 中文 | 电影 |
| 1 | 英文 | 电影 |
| 1 | 中文 | 电视剧 |
| 0 | 英文 | 电视剧 |

对“语言”和“类型”进行组合

| 是否点击 | 语言=中文<br>类型=电影 | 语言=英文<br>类型=电影 | 语言=中文<br>类型=电视剧 | 语言=英文<br>类型=电视剧 |
| --- | --- | --- | --- | --- |
| 0 | 1 | 0 | 0 | 0 |

进行特征组合会出现维度爆炸的问题。比如在推荐系统中，用户ID和物品ID进行组合（参考书中P7），假设用户有M个，物品有N个，进行组合后维度有M*N，在互联网环境下，维度会非常高，模型几乎无法学习。解决的办法是将用户和物品分别用k维的低维向量（k<<m, k<<n），则学习的维度变为m*k(物品的k维向量)+n*k(用户的k维向量)

## 4.组合特征

如何有效地找到组合特征?

基于决策树的特征组合寻找方法：每一条路径可以看作是一种特征组合的方式