特征工程

特正归一化

1. **什么是特征归一化?**
   - 归一化可以使所有特征都统一到一个相同的数值区间内
   - 本质上是一种线性变换（不会改变原始数据的数值排序）
2. **什么是时候需要用归一化?**
   - 涉及距离计算的算法都需要
   - 损失函数中含有正则项是需要：
     - 特征scale越大，系数越小，同样的惩罚下对于该系数的惩罚就会变小
   - 使用梯度下降法求解的模型一般都需要特征归一化（决策树等就不适用）
3. **有哪几种归一化方法？什么情况下使用它们?**
   - 线性函数归一化：min-max scaling，将数据映射到[0,1]的范围内，保证数据有正有负
     - 对输出结果范围有要求
     - 数据比较稳定，不存在极端的最大最小值
     - 如果最大最小值不稳定，可以用经验常量值来代替

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}},$$

   - 零均值归一化：z-score nromalization，将数据映射到N(0,1)的分布上，使各维度数量级相同
     - 数据存在异常值和较多噪音

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

4. **为什么需要对数值类型的特征做归一化?**
   - 特征的scale会影响模型对其的重要性的判断，scale更大的特征容易被误判为高权重
   - 归一化能够加快梯度下降的速度
     - 在梯度下降中，学习速率相同的情况下，数值更大的特征的系数的step会更大，优化曲线是震荡形，需要比别的特征更多迭代次数

类别型特征

1. **在对数据进行预处理时，应该怎样处理类别型特征？**
    - 序号编码：处理类别间具有大小关系的数据
    - 独热编码：使用稀疏向量来表示类别
    - 二进制编码：赋予类别ID并进行哈希映射，得到0/1特征向量，维数少于独热编码
2. **什么是组合特征？**
    - 通过组合2个或多个类别特征，来提高模型对复杂关系的拟合能力
3. **如何处理高维组合特征？**
    - 通过矩阵分解，用k维矩阵代替m维和n维矩阵，把参数规模从m*n降解为m*k+n*k
4. **怎样有效地找到组合特征？**
    - 根据原始数据构造决策树，每一条从根节点到叶节点的路径就是一种特征组合方式

文本表示模型

1. List

Word2Vec

1. List

图像数据

1. 图像数据不足时的处理方法