特征工程与数据预处理

张翼鹏

October 16, 2019

目录

1 特征工程是什么

“数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已。”

特征工程指的是把原始数据转变为模型的训练数据的过程，目的是最大限度地从原始数据中提取特征以供算法和模型使用。

2 数据预处理

未经处理的特征可能有以下问题：

- **不属于同一量纲**：即特征的规格不一样，不能够放在一起比较。无量纲化可以解决这一问题。
- **信息冗余**：对于某些定量特征，其包含的有效信息为区间划分，例如学习成绩，假若只关心“及格”或“不及格”，那么需要将定量的分数，转换成“1”和“0”表示及格和不及格。二值化可以解决这一问题。
- **定性特征不能直接使用**：某些机器学习算法和模型只能接受定量特征的输入，那么需要将定性特征转换为定量特征。最简单的方式是为每一种定性值指定一个定量值，但是这种方式过于灵活，增加了调参的工作。通常使用哑编码的方式将定性特征转换为定量特征。
- **存在缺失值**：缺失值需要补充。

2.1 无量纲化

2.1.1 标准化

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

令每一**列**数据符合标准正态分布。

2.1.2 归一化

$$x' = \frac{x - \min}{\max - \min}$$

将每一**列**数据缩放到$[0, 1]$区间内。

2.1.3 正则化

$$x' = \frac{x}{\|x\|_p}$$

将每一行数据缩放到单位范数。（该步骤只在少数问题中使用，比如后面需要使用点积或其它核方法计算两个样本之间的相似性时。主要应用于文本分类和聚类中。）

2.2 定量特征二值化（离散化）

设定一个阈值，大于阈值的赋值为1，小于等于阈值的赋值为0：

$$x' = \begin{cases} 1 & x > \text{threshold} \\ 0 & x \leq \text{threshold} \end{cases}$$

进一步，我们用此法同样可以将数据离散化，比如讨论年龄段问题时将$[0, 100]$区间离散化，均匀分成10个互不相交的小区间，分别赋值$0 - 9$。

2.3 定性特征哑编码

假设有N种定性值，则将这一个特征扩展为N种特征，即升维。当原始特征值为第i种定性值时，第i个扩展特征赋值为1，其他扩展特征赋值为0。

哑编码的方式相比直接指定的方式，不同的特征值之间没有序关系且互不影响，不用增加调参的工作。对于线性模型来说，使用哑编码后的特征可以达到非线性的效果。

3 特征选择

特征选择的原因：

- 降低复杂度、降维，减少数据冗余，使模型泛化能力更强，避免过拟合
- 增强对特征和特征值之间的理解。

特征选择的依据：

- **特征是否发散**：如果一个特征不发散，例如方差接近于0，也即样本在这个特征上基本上没有差异，这个特征对于样本的区分并无作用。
- **特征与目标的相关性**：与目标相关性高的特征，应当优选选择。

- **特征与特征的相关性**：若两个不同的特征相关性很强，说明二者携带的信息有很大的重复，易造成信息冗余，此时需要舍弃一个特征，或利用二者构造一个新的特征。

根据特征选择的形式可以将特征选择方法分为3种：

- **Filter（过滤法）**：按照发散性或者相关性对各个特征进行评分，根据评分选择特征。

- **Wrapper（包装法）**：根据目标函数（通常是预测效果评分），每次选择若干特征，或者排除若干特征。

- **Embedded（嵌入法）**：先使用某些机器学习的算法和模型进行训练，得到各个特征的权值系数，根据系数从大到小选择特征。类似于Filter方法，但是是通过训练来确定特征的优劣。

3.1 过滤法

3.1.1 移除低方差的特征

设定阈值，将方差小于该阈值的特征剔除。

3.1.2 单变量特征选择

单独地计算每个特征的某个统计指标，根据该指标来判断哪些特征重要，剔除那些不重要的特征。

- 分类问题（y离散）可采用：卡方检验，互信息。

- 回归问题（y连续）可采用：皮尔森相关系数，最大信息系数。

1、卡方检验

$$\chi^2 = \sum \frac{(A-E)^2}{E}$$

作用：检验定性自变量对定性因变量的相关性。A为观察值，E为期望值。检验统计量越大，说明两变量相关关系越强。

2、皮尔森相关系数

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

作用：衡量变量之间的线性相关性，结果的取值区间为$[-1,1]$。

优势：计算速度快；取值区间是$[-1,1]$，使其能够表示更丰富的关系。

缺陷：只对线性关系敏感，如果关系是非线性的，即便两个变量具有一一对应的关系，皮尔森相关性也可能会接近0。

3、互信息

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

作用：评价定性自变量与定性因变量的相关性。

缺陷：没有办法归一化，在不同数据集上的结果无法做比较；对于连续变量的计算不是很方便，通常变量需要先离散化，但互信息的结果对离散化的方式很敏感。

4、最大互信息系数（MIC）

$$MIC=\max_{\mathrm{m}*n<B}\frac{\max\limits_{\text{不同划分}}\left(\sum\limits_{1\leq x\leq m}\sum\limits_{1\leq y\leq n}p(x,y)\log\frac{p(x,y)}{\sum_{1\leq k\leq m}p(k,y)\sum_{1\leq k\leq n}p(x,k)}\right)}{\log\min\{m,n\}}$$

只讨论数据的两个属性（随机变量），数据点分布在二维空间中，用$m*n$的网格划分数据空间，将落在第(x,y)格子中的数据点的频率作为$P(x,y)$ 的估计，然后计算离散化后的随机变量的互信息。因为$m*n$的网格划分方式不止一种，所以我们要获得使互信息最大的网格划分，然后使用归一化因子，将互信息的值转化为$(0,1)$区间之内。最后，找到能使归一化互信息最大的网格分辨率（即m和n的值，$B=n^{0.6}$），作为MIC的度量值。

优势：若具有足够的样本，可以捕获广泛的关系，而不限于特定的函数类型；对类型不同但噪声程度同等的关系给予相近的分数。

缺陷：当零假设不成立时，MIC的统计可能会受到影响。

3.2 包装法

3.2.1 递归特征消除

对模型进行多轮训练，每轮训练后，移除若干权重（即系数）低的特征，再基于新的特征集进行下一轮训练，直到特征数量达到要求。

3.3 嵌入法

有些机器学习方法本身就具有对特征进行打分的机制，或者很容易将其运用到特征选择任务中。嵌入法即用基于机器学习模型的方法来选择特征。

3.3.1 基于$L1$或$L2$范数的特征选择

对所有原始特征进行训练，给代价函数加入$L1$或$L2$惩罚项：

$$\lambda||w||_1 \qquad \lambda||w||_2$$

使用$L1$范数可以快速让某些特征的权重收缩到0；使用$L2$范数可以集体减小多个特征的权重。

4 降维

特征选择完成后，就可以直接训练模型了。但是由于特征矩阵有时过大，导致计算量大，训练时间长，因此降低特征矩阵维度也是必不可少的。常见的降维方法除了以上提到的基于$L1$惩罚项的模型以外，另外还有主成分分析法（PCA）和线性判别分析法（LDA），其本质都是将原始的样本映射到维度更低的样本空间中。

4.1 PCA

如果要将原始D维数据投影到M维子空间当中，PCA的做法是计算原始数据的协方差矩阵S，求其前M大特征值对应的单位特征向量，然后将其作用到原始样本矩阵上，得到降维后的样本矩阵。

PCA的目标是去掉原始数据冗余的维度，让映射后的样本具有最大的发散性，是一种无监督的降维方法。

4.2 LDA

首先计算类间散度矩阵S_b：

$$S_b = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T$$

其中μ_0是第0类样本的均值，μ_1是第1类样本的均值。

然后计算类内散列矩阵S_w：

$$S_w = \sum_{x \in X_0} (x - \mu_0)(x - \mu_1)^T + \sum_{x \in X_1} (x - \mu_0)(x - \mu_1)^T$$

其中X_0是第0类样本的集合，X_1是第1类样本的集合。

最后${S_w}^{-1}S_b$的最大特征值所对应的特征向量w即为最佳投影方向。

LDA的目标是使相同类别的数据分布更紧凑，不同类别的数据尽量相互远离，让映射后的样本有最好的分类性能，是一种有监督的降维方法。