四种变量

根据变量的度量性质,统计学将变量分为四种: 定类型变量(nominal scale),定序型变量(ordinal scale),定距型变量(interval scale)和定比型变量(ratio scale).更详细解释参看维基百科(https://en.wikipedia.org/wiki/Level_of_measurement).

定类型变量的编码

在机器学习中,经常存在定类型变量(nominal scale或qualitative variable). 比如,一个人可能有 ["男", "女"], ["同性恋", "异性恋", "双性恋"]等定类型的特征.这些特征能够被有效地编码成整数,比如["男", "异性恋"]可以被表示为[0, 1],["女", "双性恋"]表示为[1, 2].

这个整数特征表示并不能直接作为特征在机器学习中使用,因为这样的连续输入,模型会认为类别之间是定距或者定比的(有截距时通常定距,无截距时通常定比).

假设一个定类型数据有K种类别,将其转换为能够被机器学习模型使用的编码有

- 独热编码(one-hot encoding), 在OneHotEncoder中实现.每个类别被转化为一个K维向量,每维对应一种类别.每种类别转化为该类对应的维度为1,其他维度为0的K维向量.因为有且只有一个维度为1,其他为0,所以叫做独热.
- 虚拟编码(dummy coding/reference cell coding), 选出一种类别作为参照取值(叫做参照类).每个类别被转化为一个K-1维向量,每个维度对应除了参照类之外的一种类别.每种非参照类类别转化为该类对应的维度为1,其他维度为0的K-1维向量.参照类转化为K-1维零向量.因为选出一个类别作为参照类,而每个类被分为一个组(cell),所以叫做reference cell coding.
- 效应编码(effect coding),和虚拟编码一样选出一个参照类,除了参照类的取值编码方式一样,只是参照类转化为K-1维全是-1的向量.因为在传统的方差分析中,类别均值相对总体均值的偏差叫做治疗效果(treatment effect).医学统计中,每个类别代表服用一种药物,每种类别的效果就是服用对应药物相比平均情况的治疗效果.而效应编码中,类别系数正好是treatment effect,所以叫做effect coding.

线性回归时比较

使用不同定类型变量编码时,类别作为特征时的没有正则化的线性回归

编码方式	截距	i类系数	i类编码,共K类	编码维度
dummy coding	参照类均值	i类的均值 - 参照类均值	若不是参照类,i-dim为1,其他为0;否则全为 0	K-1
effect coding	总体均值	i类均值 - 总体均值	若不是参照类,i-dim为1,其他为0;否则全部为-1	K-1
one-hot encoding	c	i类均值-c	i-dim为1,其他为0	K

其中,总体均值(grand mean)指的是所有类均值的均值, c为任意常数

P.S. 在一些教材中,类别的均值叫做level,标准差叫做scale.

N.B.: 在线性回归中,如果不设定截距为空,则one-hot编码会引起某些参数的singularity,因为one-hot编码下,K维向量的和恒为1,引发perfect multicollinearity.

以上编码方式的python实现

one-hot encoding, multi-hot的实现可以参看 sklearn.preprocessing中OneHotEncoder,LabelBinarizer, MultiLabelBinarizer和LabelEncoder (https://scikit-learn.org/stable/modules/classes.html#module-sklearn.preprocessing)

dummy coding的实现可以参看pandas.get_dummies (https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get_dummies.html).

effect coding的实现接口没找到,可以通过dummy coding来转化(全零向量转化为全-1向量).

In [ ]: