特征工程的常用套路

1. 综述

在机器学习中有一句出名的话

> *garbage in, garbage out*

意思是，如果你输入的数据是垃圾，那无论你的模型多么精巧，你最终得到的结果也是垃圾，所以如果将数据清洗得更有意义，是重要的一点。

一般特征工程，主要分为三大步：

- 特征的清洗
- 特征的转换
- 特征的抽取

这里我们主要讲特征的抽取（feature ectract），即从原始的特征中抽取出更有意义的特征。

---

2. 特征抽取套路

- 类别特征
- 连续特征
- 时间序列
- 多值特征
- 时间特征

- 特征组合

2.1 类别特征

类别特征，是指特征值是离散的变量的特征。

1. 1 维编码（有1维的内在顺序）
   e.g 衣服：大，中，小。

2. 高维编码（有高维的内在顺序）
   e.g 国家首都：CH ,US。
   可以通过算法向量化，也可以通过人工向量化。

3. one-hot编码（无内在顺序，且类别数较少）

4. count 编码 (统计类别出现频率)
   PS : 未引入标签信息，

5. target 编码

   PS : 引入了标签信息，本质上 stacking 的弱化版本。需要引入交叉验证防止 data leak。

6. Stacking 编码

   PS : 引入标签信息，通过模型算法进行编码。需要引入交叉验证防止 data leak。

2.2 连续特征

1. 分箱
   PS：连续特征离散化，起到防止过拟合的引入非线性作用。

2. 归一化和标准化
   PS：深度学习需要，树模型不需要。

3. 连续值当作类别值

2.3 时间序列

一般范式：

主体+窗口+函数

函数一般是一些统计相关函数，如 min,max,mean,medium,std,peak 数量,指数平均,lag,加权平均,增长(下降)率,时序 L_2 范数 …

自动时序功能包

2.4 多值特征

一个特征对应多个值

e.g 汽车的温度指标,统计,10–20度出现次数，20–30度出现次数...

多值特征-->直方图特征。

直方图特征-->分布差距特征。

分布差距-->各种散度特征。

2.5 时间特征

- 周几，月份
- 季节
- 一年的第几天
- 一年的第几周
- 时间差
- 月末

- 周末
- 节假日

2.6 特征组合

低阶特征组合：

特征+特征

高阶特征组合(主要适合 CTR 预估类的比赛)：

网络结构	是否需要人工特征	组成	参数个数	需要预训练	低阶特征表达	高阶特征表达
LR	是	LR	1+n	否	是	否
FM	否	FM	1+n+n*k	否	是	是
DNN	否	MLP	n*H1+H1*H2+H2*1	否	否	是
FNN	否	FM+MLP	1+n+n*k+(1+f+f*k)*H1+H1*H2+H2*1	是	否	是
PNN	否	FM+product+MLP	IPNN: 1+n+n*k+(f*k+f*(f-1)/2)*H1+H1*H2+H2*1 OPNN: 1+n+n*k(f*k+f*(f-1)/2*k*k)*H1+H1*H2+H2*1	否	否	是
wide&deep	是	LR+embedding+MLP	1+n+n*l+f*k*H1+H1*H2+H2	否	是	是
deepFM	否	FM+embedding+MLP	1+n+n*k+f*k*H1+H1*H2+H1	否	是	是
NFM	否	FM+embedding+MLP	1+n+n*k+k*H1+H1*H2+H2*1	否	是	是
AFM	否	FM+embedding+attention+MLP	1+n+n*k+k*H1+H1*2+k*1	否	是	是
DCN	否	embedding+cross+MLP+LR	1+n+2*dLc+d*(m+1)+m*(m+1)*(Ld-1)+1+d+m	否	是	是
DIN	否	embeddin+attention+MLP	n*k+attention+f*k*H1+H1*H2+H2*1	否	是	是