特征选择和构造

特征选择

为什么要进行特征选择

- 由于数据的冗余性、复杂性、缺失性等原因，数特征不能直接使用。
- 或是数据量太大的情况，影响模型的效率，并且难以解释。

特征选择

如何进行特征选择

- 特征选择就是从原始特征中选取一些最有效的特征来降低维度,，提高模型泛化能力减低过拟合的过程，主要目的是剔除掉无关特征和冗余特征，选出最优特征子集；
- 常见的特征选择方法可以分为3类：过滤式（filter）、包裹式（wrapper）、嵌入式（embedding）。

特征选择

- 1 过滤式
- 通过方差选择法、相关系数法、卡方检验法、互信息法来对特征进行评分，设定阈值或者待选择的阈值的个数来选择；
- 1.1 方差选择法
- 计算各个特征的方差，剔除小于设定的阈值的特征，剔除特征值 波动较小的特征，例如一个特征的所有值都为1，那这个特征对于预测目标变量就没什么作用；方法很简单，但实际应用中只有少数变量才会存在只取某个值的情况，对特征选择作用比较小，可以当做数据预处理部分，之后再用其他方法进行特征选择。

```
from sklearn.feature_selection import VarianceThreshold
var = VarianceThreshold(threshold=0)
var.fit_transform(df)
df = df.iloc[var.get_support(True),:]
#VarianceThreshold返回已经提出方差为0的列，通过get_support[True]定位剩余变量所在的列
```

特征选择

- 1 过滤式
- 1.2 相关系数法

- 皮尔森相关系数衡量的是变量之间的线性相关性，取值范围在-1-+1之间，-1表示完全负相关，+1表示完全正相关，0表示线性无关；
- 可以使用scipy的pearsonr 计算皮尔森相关系数，且它还可以同时计算出p值。
- 也可使用pandas的corr（）函数计算。

```
import numpy as np
from scipy.stats import pearsonr
x = np.random.normal(0,10,300)
y = x + np.random.normal(0,10,300)
pearsonr(x,y)
```

```
irisdata.corr()
```

	sepal length	sepal width	petal length	pretal width
sepal length	1.000000	-0.117570	0.871754	0.817941
sepal width	-0.117570	1.000000	-0.428440	-0.366126
petal length	0.871754	-0.428440	1.000000	0.962865
pretal width	0.817941	-0.366126	0.962865	1.000000

特征选择

- 1 过滤式
- 1.3卡方检验法
- 检验定性自变量对定性因变量的相关性

```
from sklearn.feature_selection import chi2
#chi2要求变量值非负，返回卡方值和P值
from sklearn.feature_selection import SelectKBest
from sklearn.datasets import load_iris

iris = load_iris()

model = SelectKBest(chi2, k=2)
model.fit_transform(iris.data,iris.target)
var = model.get_support(True)
```

特征选择

- 2 包裹式
- 2.1 递归特征消除法

- 根据预测效果(AUC/MSE)或者其他方法对特征组合进行评分，主要方法有递归特征消除法；
- 递归特征消除法的主要思想是反复的构建模型，然后选出最好或最坏的特征，把选出的特征放到一边，然后在剩余的特征上重复这个过程，直到所有特征都遍历了。在这个过程中特征被消除的次序就是特征的排序。

```
from sklearn.feature_selection import RFE
from sklearn.linear_model import LinearRegression
lr = LinearRegression()
rfe = RFE(lr, n_features_to_select=2)
rfe.fit(iris.data, iris.target)
var = rfe.get_support(True)
```

特征选择

- 3 嵌入式
- 3.1 正则化

- 正则化主要包括L1正则化和L2正则化：
- L1正则化将系数W的L1范数作为惩罚项加到损失函数中，L1正则方法具有稀疏解的特性，因此天然具有特征选择的特性，但是不代表没被选到的特征就不重要，有可能是因为两个高度相关的特征最后只保留了一个；另外L1正则化和非正则化模型一样是不稳定的，如果特征集合中具有相关联的特征，当数据发生细微变化时也有可能导致很大的模型差异。

```
from sklearn.linear_model import Lasso
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
x = scaler.fit_transform(iris.data)
y = iris.target
lasso = Lasso(alpha=0.2)
lasso.fit(x,y)
lasso.coef_
```

特征选择

- 3 嵌入式
- 3.1 正则化

- L2正则化将系数向量的L2范数添加到损失函数中，由于L2惩罚项中的系数是二次方的，会让系数的取值变得平均，对于有相关性的变量，会得到相近的系数；L2正则化也较L1稳定；

```
from sklearn.linear_model import Ridge
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
x = scaler.fit_transform(iris.data)
y = iris.target
ridge = Ridge(alpha=10)
ridge.fit(x,y)

ridge.coef_
```

```
array([ 0.01580835, -0.05520001,  0.31571552,  0.41614276])
```

特征选择

- 4 树模型
- 4.1 通过决策树的形式，构成模型，反向推理自变量对因变量的重要性。

```
from sklearn import metrics
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
model = ExtraTreesClassifier()
x = scaler.fit_transform(iris.data)
y = iris.target
model.fit(x, y)
print(model.feature_importances_)
```

```
[0.11000591 0.06020284 0.40528317 0.42450808]
```

特征选择

- 5 RFE搜索算法
- 基于对特征子集的高效搜索，从而找到最好的子集，意味着演化了的模型在这个子集上有最好的质量。递归特征消除算法（RFE）是这些搜索算法的其中之一，Scikit-Learn库同样也有提供。

```
from sklearn.feature_selection import RFE
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
# create the RFE model and select 3 attributes
rfe = RFE(model, 3)
rfe = rfe.fit(x, y)
# summarise the selection of the attributes
print(rfe.support_)
print(rfe.ranking_)
```

```
[False  True  True  True]
[2 1 1 1]
```

特征构造

特征构造

- 特征构造是指从原始数据中构建新的特征，也属于特征选择的一种手段。特征构建工作并不完全依赖于技术，它要求我们具备相关领域丰富的知识或者实践经验，基于业务，花时间去观察和分析原始数据，思考问题的潜在形式和数据结构，从原始数据中找出一些具有物理意义的特征。
- 在实际业务中，通常我们只拥有几个到几十个不等的基础变量，而多数变量没有实际含义，不适合直接建模，如用户地址（多种属性值的分类变量）、用户日消费金额（弱数值变量）。
- 而此类变量在做一定的变换或者组合后，往往具有较强的信息价值，对数据敏感性和机器学习实战经验能起到一定的帮助作用。所以我们需要对基础特征做一些衍生类的工作。

特征构造

- 特征构造
- 1. 特征扩展

- 基于一个特征，使用特征值打平（扩展）的方式衍生多个标注类型的特征，也可以理解为离散化。对于分类变量，直接one-hot编码；对于数值型特征，离散化到几个固定的区间段，然后用one-hot编码。比如信贷场景逾期天数：pandas的get_dummies()可以直接对变量进行one-hot编码，其中prefix是为one-hot编码后的变量进行命名。

PD	stage	M1	M2	M3	M4	M5	M6	M6+
1~30	M1	1	0	0	0	0	0	0
31~60	M2	0	1	0	0	0	0	0
61~90	M3	0	0	1	0	0	0	0
91~120	M4	0	0	0	1	0	0	0
121~150	M5	0	0	0	0	1	0	0
151~190	M6	0	0	0	0	0	1	0
191+	M6+	0	0	0	0	0	0	1

特征构造

- 特征构造
- 2. 特征组合

- 指将两个或多个输入特征通过数学运算进行组合。
- 比如，计算线损率。

Thank you!