机器学习模型的笔记

1数据处理的管道

建立处理节点，处理流程，方便替换节点和组织流程。

1.1管道的四大组织构件：

1.1.1.Pipeline串联管道

from sklearn.pipeline import Pipeline

管道类生成对象。pipe=Pipeline(steps=[('pca',PCA()),('svc',SVC())])。setps参数是列表，嵌套元组，元组内第一个是自命名，第二个是生成数据处理类的实例，生成的新管道函数的对象，同时具有fit 和 transform的方法。前n-1个step中的类都必须有transform函数，最后一步可有可无，一般最后一步为模型。

1.1.2.ColumnTransformer

from sklearn.compose import ColumnTransformer按不同列应用不同的估计转化器，最后列合并。不同于管道函数按照先后顺序。对于dataframe不同的列是包含列名的列表，对于array可以是列序号的列表，均输出合并后特征向量的二位数组。

num\_attribs = list(housing\_num)

cat\_attribs = ["ocean\_proximity"]

full\_pipeline = ColumnTransformer([

("num", num\_pipeline, num\_attribs),

("cat", OneHotEncoder(), cat\_attribs),

])

housing\_prepared = full\_pipeline.fit\_transform(housing)

1.1.3.FeatureUnion 并联管道

把若干节点并联成一个新的结点。这个并联的各个结点同时处理同一个输入数组，生成各自的数组结果后，列拼接成为一个更大的特征向量数组。

from sklearn.pipeline import FeatureUnion

from sklearn.decomposition import PCA

from sklearn.decomposition import KernelPCA

estimators = [('linear\_pca', PCA()), ('kernel\_pca', KernelPCA())]

combined = FeatureUnion(estimators)

# **三者结合很有用,但是ColumnTransformer 可以实现 FeatureUnion 的功能**

1.1.4.FunctionTransformer

可以将任何构造函数包装成一个类,统一 fit 、transform方法，放进pipeline使用非常方便。但是记住构造函数设计用来接受数组，输出数组。构造函数的参数，通过FunctionTransformer创造对象的时候除了传入函数，另外指定一个参数kw\_args={"add\_bedrooms\_per\_room": False}来定义

from sklearn.preprocessing import FunctionTransformer

def add\_extra\_features(X, add\_bedrooms\_per\_room=True):

rooms\_per\_household = X[:, rooms\_ix] / X[:, household\_ix]

population\_per\_household = X[:, population\_ix] / X[:, household\_ix]

if add\_bedrooms\_per\_room:

bedrooms\_per\_room = X[:, bedrooms\_ix] / X[:, rooms\_ix]

return np.c\_[X, rooms\_per\_household, population\_per\_household,

bedrooms\_per\_room]

else:

return np.c\_[X, rooms\_per\_household, population\_per\_household]

attr\_adder = FunctionTransformer(add\_extra\_features, validate=False,

kw\_args={"add\_bedrooms\_per\_room": False})

attr\_adder.fit\_transform(housing.values)

1.2 Pipline总结：

Pipleline中最后一个之外的所有estimators（估计器）外都必须同时是转换器（transformers），最后一个estimator可以是任意类型（classifier，regresser），没有模型也没关系当成转换器用,如果最后一个estimator是个分类器，则整个pipeline就可以作为分类器使用，如果最后一个estimator是个聚类器，则整个pipeline就可以作为聚类器使用。

组成管道对象后，管道对象的方法以最后一个节点的方法为准，运用fit方法，除了最后一步，前面步骤都会用fit\_transform方法，如果管道对象最后一步是模型，用predict方法，前面处理步骤自动会运用transform方法。

注意：如果最后一步是模型的管道对象fit()会传入两个参数，只对前面那个X用前面数据处理，y不会被处理。但有的估计器需要y参数。

pipline.get\_params()获得所有节点超参数。

pipline.set\_params(lr\_\_fit\_intercept = False)修改某一结点的超参数

Pipline[‘节点名’]直接得到这个结点对象。

pipline.score(X\_test, Y\_test) 缺省方法评分。可以用sklearn中都在sklearn.metrics类下更多的评分办法。

from sklearn.model\_selection import cross\_val\_score

scores = cross\_val\_score(dtc,X,Y,cv=5,scoring='f1') 用训练集进行交叉验证。模型评估效果更客观。

管道对象除第一个节点的输入其他都采用数组。

1.3管道构造过程

1.3.1在进入管道前先用from sklearn.model\_selection import train\_test\_split 随机划分或者按重要属性分层划分出测试训练集集后 ，单对训练集进行处理 ，因为划分函数不是估计转换器不能进入管道操作。时间序列如何来划分？时间序列如何交叉验证。

训练集首先进行数据探索，寻找相关关系。确定模型 确定是否需要增加计算指标

1.3.2正式开始处理数据

拆分出输入变量和目标变量，，只看输入变量（输出变量也需要处理？）

检查缺失值，变量多可以选择扔掉缺失值多的变量，样本多可以选择仍掉有缺失值的样本，

训练集中的输入变量中的数值变量和类别变量分开，采用不同的处理方法，用ColumnTransformer对象函数来处理

一般采用from sklearn.impute import SimpleImputer 打头 因为输入df，输出数组。优先进行缺失值填补。数值型采用平均数，类别值采用众数

类别变量进行一个编码转化from sklearn.preprocessing import OrdinalEncoder 或者OneHotEncoder(sparse=False),接受的Series、df 、数组。输出数组

数值型变量的计算转化，用 FunctionTransformer包装转化函数，转化函数考虑输入的是数组，因为在管道内。输出全部数值型数据。

管道估计器填补缺失值，计算指标，标准化。标准化类、编码转化。管道内的节点全部拥有除了最后一个其他必须拥有fit transform 方法。

2.管道模型的评分测试。

2.1人为验证数据集。指定的训练集fit，再用指定的验证集preidct，后用评分函数。训练集中的80%用来fit，剩下的20%作为验证集来评估，此方式不推荐。

评分函数在sklearn中都在sklearn.metrics类下，务必记住哪些评分适合分类，那些适合回归，不能混着用。

分类模型：

accuracy\_score(y\_test,y\_pre)准确率得分，是模型分类正确的数据除以样本总数 ，model.score(x\_test,y\_test)，效果一样。

classification\_report(y\_test,y\_log\_pre)。其中的各项得分的avg/total 是每一分类占总数的比例加权算出来的

print(classification\_report(y\_test,y\_log\_pre))

precision recall f1-score support

0 0.87 0.94 0.90 105

1 0.91 0.79 0.85 73

avg / total 0.88 0.88 0.88 178

confusion\_matrix(y\_true, y\_pred)，混淆矩阵，用来评估分类的准确性。有的分类问题，实际样本中1000个A，10个B，如果最后分类大多数B都被预测错误了，但依据其他评估方法，得分反而很高(因为A的数目相对太多导致的)，召回率和精确率的区别。

回归模型

mean\_absolute\_error（MAE、平均绝对误差）

mean\_squared\_error（MSE、均方误差）

r2\_score（R^2、可决系数）

2.2训练集自动划分验证集，交叉验证 传入全管道全量训练集的x数组 和 y目标值。交叉验证函数自动输出每折的评分。样本条件所限的话，取消测试集评估，直接上线预测。

from sklearn.model\_selection import cross\_val\_score

scores = cross\_val\_score(tree\_reg, housing\_prepared, housing\_labels,

scoring="neg\_mean\_squared\_error", cv=10)

因为评分函数结果都一致为越大越小，对于越小越好的mse，评分函数给出负值。

lin\_rmse\_scores = np.sqrt(-lin\_scores).mean()

交叉验证cross\_val\_score的scoring参数：

分类：accuracy(准确率)、f1、f1\_micro、f1\_macro（这两个用于多分类的f1\_score）、precision(精确度)、recall(召回率)、roc\_auc

回归：neg\_mean\_squared\_error（MSE、均方误差）、r2

聚类：adjusted\_rand\_score、completeness\_score等 【这一块我没怎么用过】

3.在交叉验证后确定了管道中的所有估计器和预测器后，超参数的调优.

3.1网格搜索

再确定了管道流程特备是模型后，需要对模型调优，进行超参数搜索。已选择最佳超参数。

param\_grid = dict(features\_\_pca\_\_n\_components=[1, 2, 3],

#注意估计器名称后两个下划线

features\_\_univ\_select\_\_k=[1, 2],

svm\_\_C=[0.1, 1, 10])

grid\_search = GridSearchCV(pipeline, param\_grid=param\_grid, verbose=10)

grid\_search.fit(X, y)

print(grid\_search.best\_estimator\_)

1. **常用模型**

**空值、离群值的判断**

**2.1回归任务**

* + 1. **时间序列**

**2.1.1.1 时间序列独特的离群值判断：**

am = sup\_water['最低温度'].rolling(window=3,min\_periods=0,center=True).mean()

diff = sup\_water['最低温度'].values-am

diff.abs()>diff.std(ddof=0)\*4

sup\_water.loc[diff.abs()>diff.std(ddof=0)\*4,]

sup\_water['上限']=am+diff.std(ddof=0)\*3

sup\_water['下限']=am-diff.std(ddof=0)\*3

**2.1.1.2时间序列的交叉验证原理：**

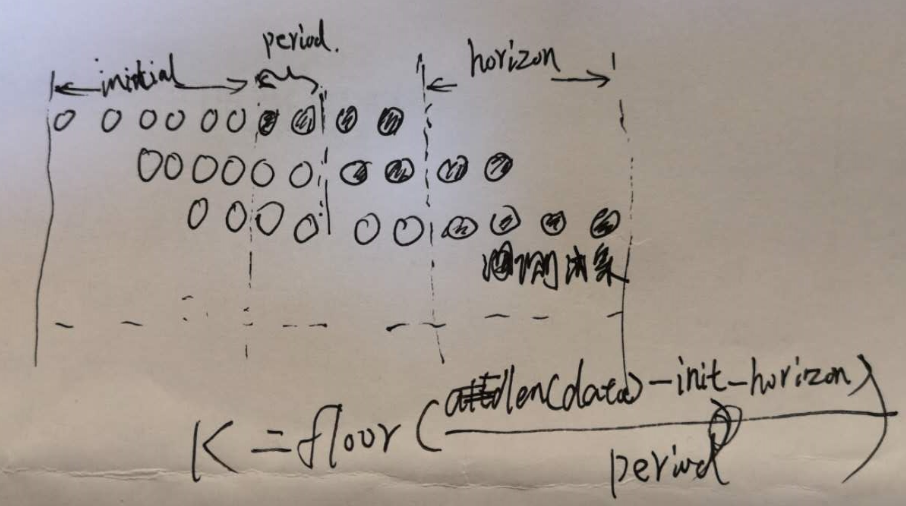
因为时间序列的样本之间是无法交换的，所以没办法像KFold交叉验证一样把数据集切分成若干份训练集和测试集（每份训练集的模型超参数是一样的）。一个比较好的思路是按照时间顺序设置。这里有三个参数：

horizon: 模型预测的范围，如从cutoff点开始数未来30天

period: 每两个 cutoff 点之间的间隔，也就是间隔多少期开始新的一份训练集。cutoff为horizon的开始点。

initial: 用于训练的日期范围，如730天

K计算方法如下图



计算举例：

#initial = 730,horizon = 365,K=10 ,period=7

Len(data)=dt.date(2020,3,12)-dt.date(2017,1,1)=1167

K=(1167-730-365)/7=10

可以自己写一个函数将数据集分多对训练集和测试集

horizon,initial,period=30,1500,30

df\_train\_list,df\_test\_list=ts\_model\_selection.train\_test\_split(df,

horizon='{}days'.format(horizon),

initial='{}days'.format(initial),

period='{}days'.format(period))

for k in range(len(df\_train\_list)):

#print(k)

df\_train,df\_test=df\_train\_list[k],df\_test\_list[k]

……对每一个df\_train训练并做预测。然后和df\_test的y进行对比计算mape或者rmse，收集在一个list里面。可以计算平均值。

**2.1.1.3** Fbprophet

prophet有自动交叉验证函数

#划分数据

from fbprophet.diagnostics import cross\_validation

df\_cv = cross\_validation(m, initial='730 days', period='7 days', horizon = '365 days')

Arima

前提时间序列是稳定的

1.常量的均值 弹簧的轴不歪，原因趋势（trend）-数据随着时间变化。比如说升高或者降低。解决办法：差分

2.常量的方差 弹簧的横向均匀，不扩大。原因数据在特定的时间段内变动。比如说节假日，或者活动。解决办法：对数变换，或者平滑法具体分为移动平均法和指数平均法。

3.与时间独立的自协方差 弹簧没有被暴力拉扯。原因周期变了。

可以通过观察，计算累计方差，累计平均数画线观察。也可以通过teststationarity(ts)检验。

更成熟，灵活的方法。将时间序列按照不同因素分解不同序列相乘或者相加。一般分成三种序列，趋势序列，周期序列，残差序列。运用sm.tsa.seasonal\_decompose（）函数。可以分别对趋势、周期、残差进行预测。再相乘或者相加。

from statsmodels.tsa.arima\_model import ARIMA 使用前 先用移动平均剥离周期性，输出的预测值是差分值+残差，。真实数据需要去反差分，再反移动平均，甚至np.exp（）反log去还原。

缺点，特殊假期无法模拟，外生变量影响没有体现，不能捕捉非线性，，用于规则变化周期性，线性增长残差平稳的。

Sarimax 模型不用反差分，不用反移动平均。更方便。缺点参数太多不好确定，使用网格搜索。

Ar，ma ，都要求平稳序列。一切源于稳定的时间序列的均值开始，如果差分均值就是差值，ar 是均值加上前p期数值乘以各自系数的回归方程。ma是前q期的均值，加上每期实际数与预测数的误差项目乘以系数的回归方程，得到本期的预测值。

Holt

Expotion

是否就是其他模型的变种。

2.1.2线性回归

sklearn.linear\_model.LinearRegression()

.intercept\_[0]看截距，.coef\_[0][0]看系数。

2.1.3分类回归树

随机森林

神经网络

支持向量机

分类任务

聚类任务