模型评估与调参

文章很多知识点和代码选自：万字长文总结机器学习的模型评估与调参
部分公式截图来自网上

pipeline

Transformer：转换器，本质上将一个df转化为另一个df，格式数值可能有所变化

Estimator：评估器，本质上由数据生成了转化器，df训练生成模型

Pipline：多个转换器和评估器组合到一起生成工作流，共享一个API

```
from sklearn.preprocessing import StandardScaler # 用于进行数据标准化
from sklearn.decomposition import PCA # 用于进行特征降维
from sklearn.linear_model import LogisticRegression # 用于模型预测
from sklearn.pipeline import Pipeline
pipe_lr = Pipeline([('scl', StandardScaler()),
                    ('pca', PCA(n_components=2)),
                    ('clf', LogisticRegression(random_state=1))])
pipe_lr.fit(X_train, y_train)
print('Test Accuracy: %.3f' % pipe_lr.score(X_test, y_test))
y_pred = pipe_lr.predict(X_test)
```

当管道pipe_lr执行fit方法时：
1）StandardScaler执行fit和transform方法；
2）将转换后的数据输入给PCA；
3）PCA同样执行fit和transform方法；
4）最后数据输入给LogisticRegression，训练一个LR模型

Pipline执行时会依次执行里面所有的分类器和转化器，管道里有多少转化器都可以

K折交叉验证

交叉验证本质是选出用于模型评估的测试数据

步骤：

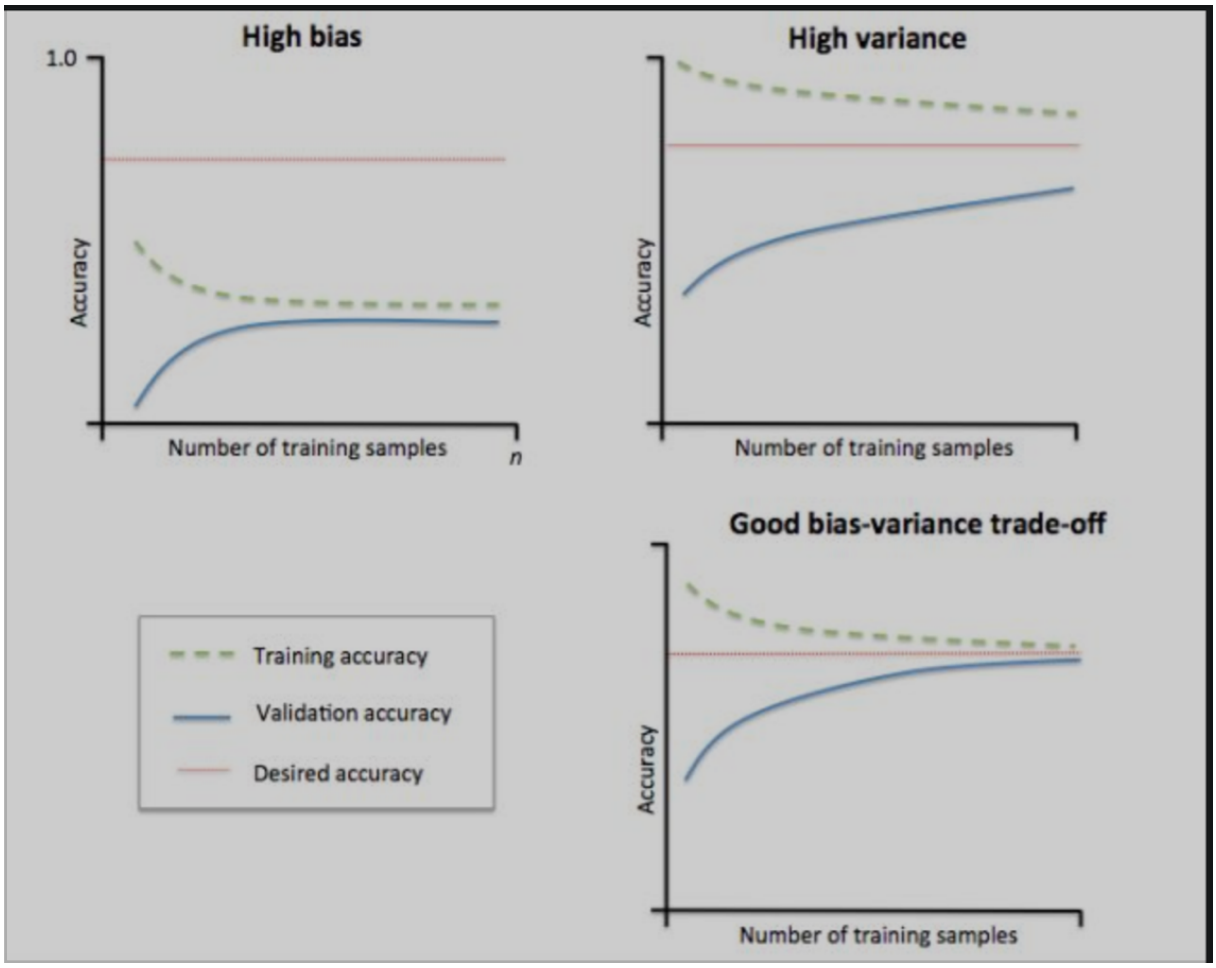
- 不重复抽样将数据分成K份
- K-1份训练，一份测试
- 重复K次

- 将平均值作为结果

曲线对比

绘制学习曲线和验证曲线，确定问题是高偏差还是高方差，针对不同情况采取不同策略

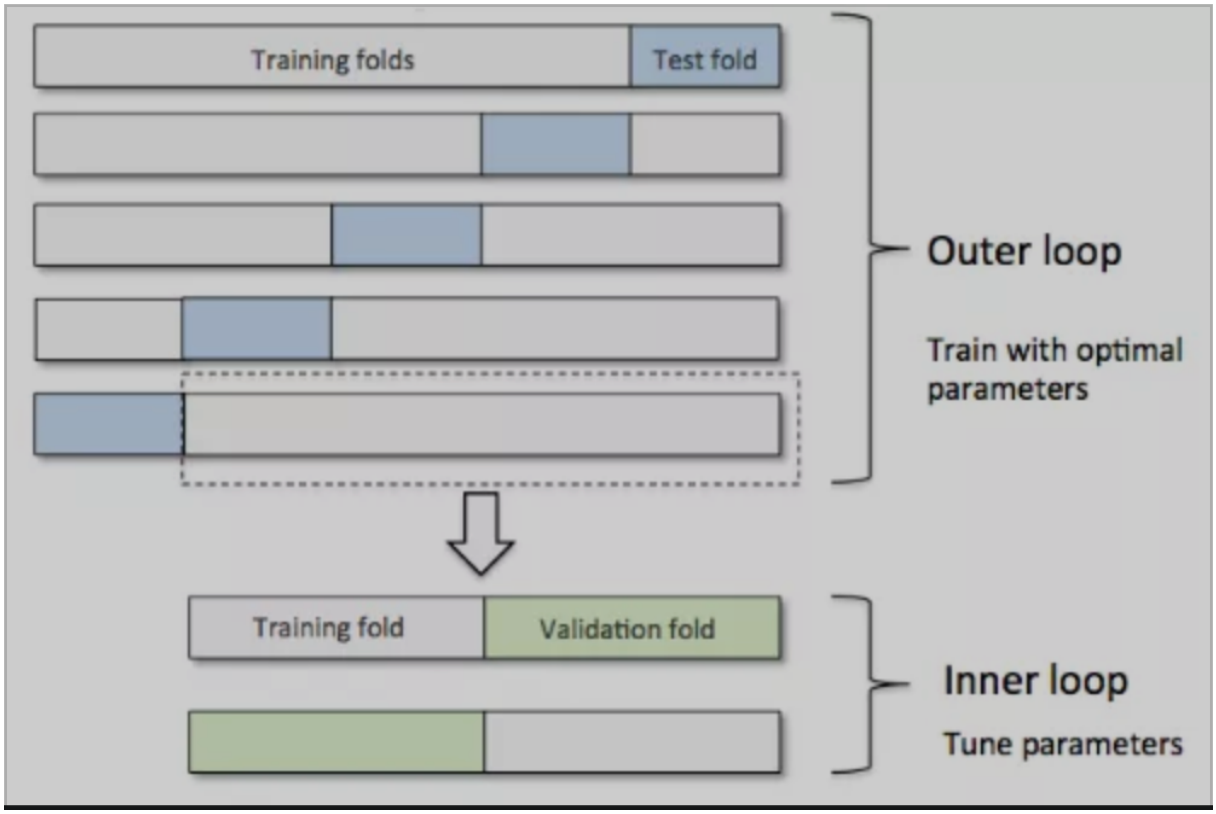
- 高偏差（模型过于简单）
  - 更多数据
  - 更多特征
  - 特征处理方式
  - 数据清洗方式（距离算法异常值影响）
  - 算法参数调整
  - 算法选择问题
- 高方差（模型太复杂）
  - 减少特征纬度
  - 根据不同算法不同策略
    - 树形
    - 线性
    - 聚类
    - 神经网络



网格搜索

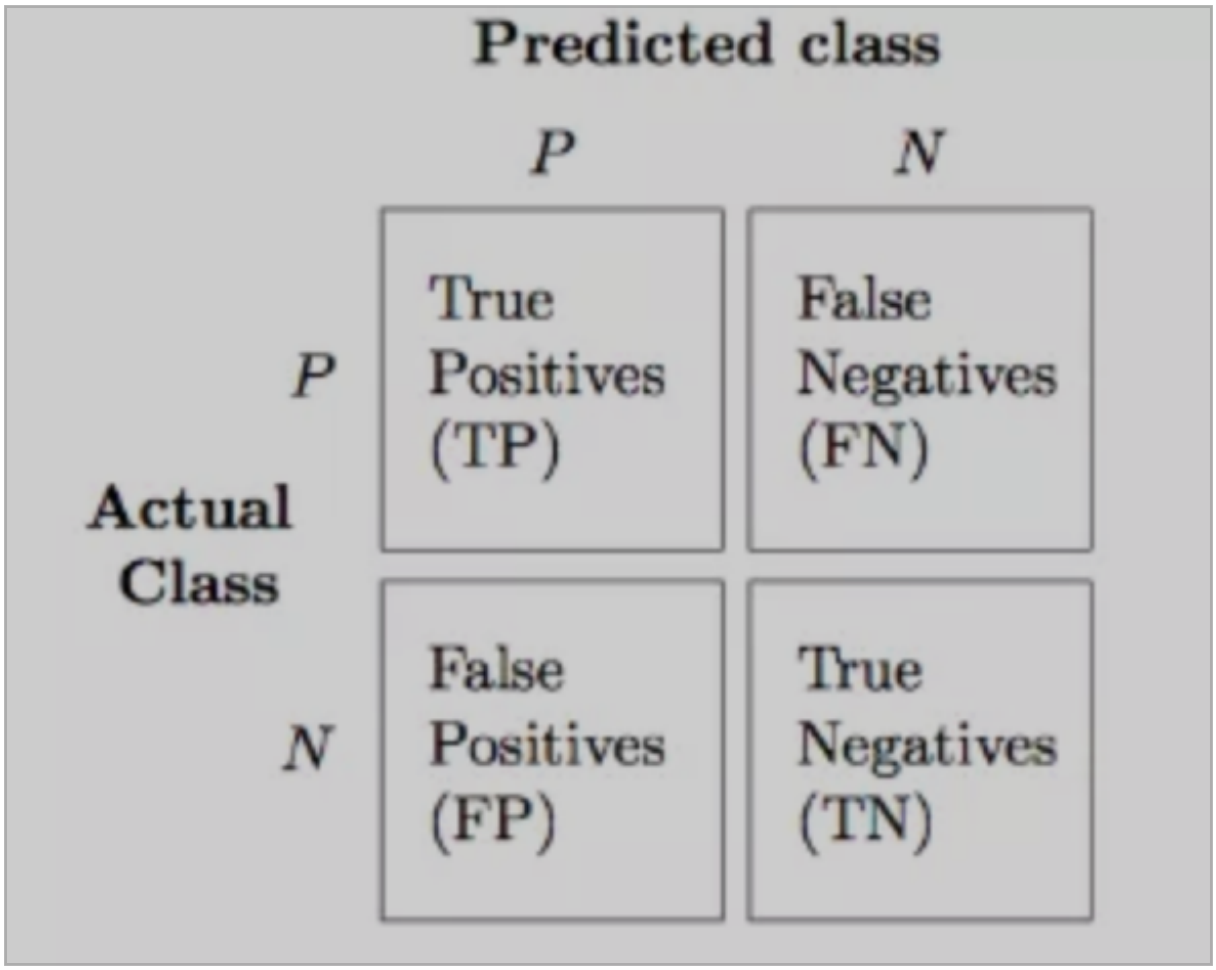
- 暴力寻找超参数（学习率、正则系数、决策树深度）

嵌套交叉验证


Training folds
Test fold
Outer loop
Train with optimal parameters
Training fold
Validation fold
Inner loop
Tune parameters

评价指标

- 混淆矩阵



TP(True Positive): 真实为0，预测也为0

FN(False Negative): 真实为0，预测为1

FP(False Positive): 真实为1，预测为0

TN(True Negative): 真实为1，预测也为1

- AUC

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- 准确率（预测为真实际为真占总数的比例）

$$Precision = \frac{TP}{TP + FP}$$

- 召回率：（多少真值被预测出来了）

$$Sensitivity = Recall = \frac{TP}{TP + FN}$$

- ROC曲线（横坐标假正率，纵坐标真正率）
- ROC曲线面积为AUC，面积越大分类模型性能越好，一次函数约等于瞎猜

真正率(true positive rate,TPR)，指的是被模型正确预测的正样本的比例：

$$TPR = \frac{TP}{TP + FN}$$

假正率(false positive rate,FPR)，指的是被模型错误预测的正样本的比例：

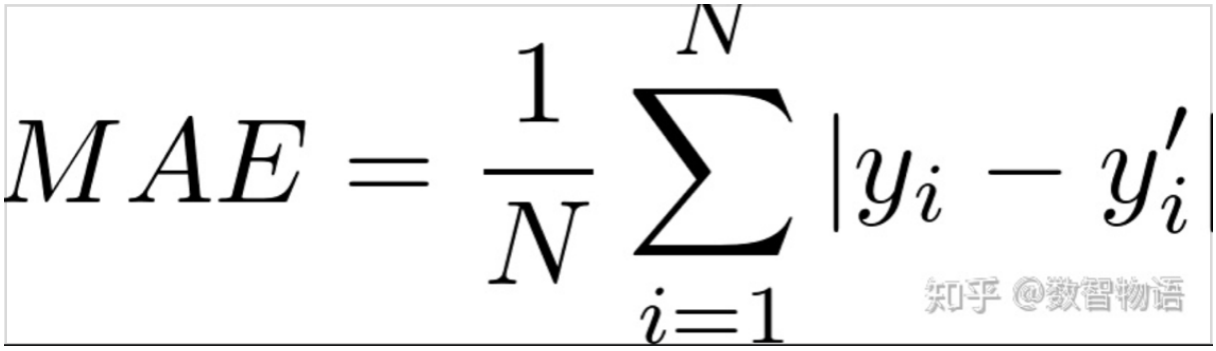
$$FPR = \frac{FP}{TN + FP}$$

不同模型评估方式

- 分类
  - 混淆矩阵那一套
  - AUC
  - ROC曲线
  - 召回率（疾病预测类）

- 准确率

- 回归
  - MAE 平均绝对误差



  - MSE 均方误差