



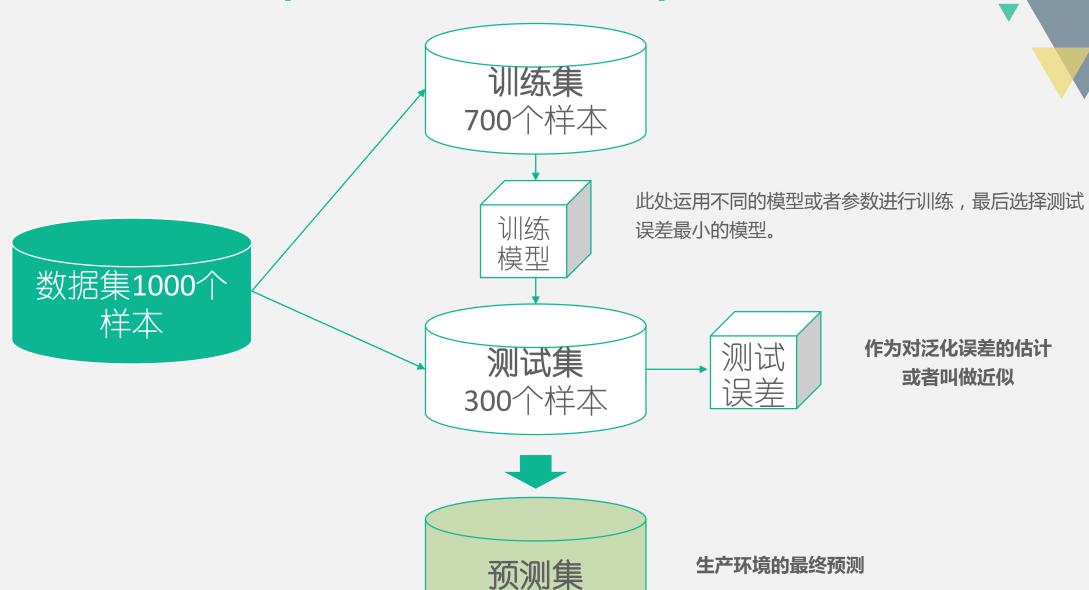
01)留出法回顾

02)交叉验证法回顾

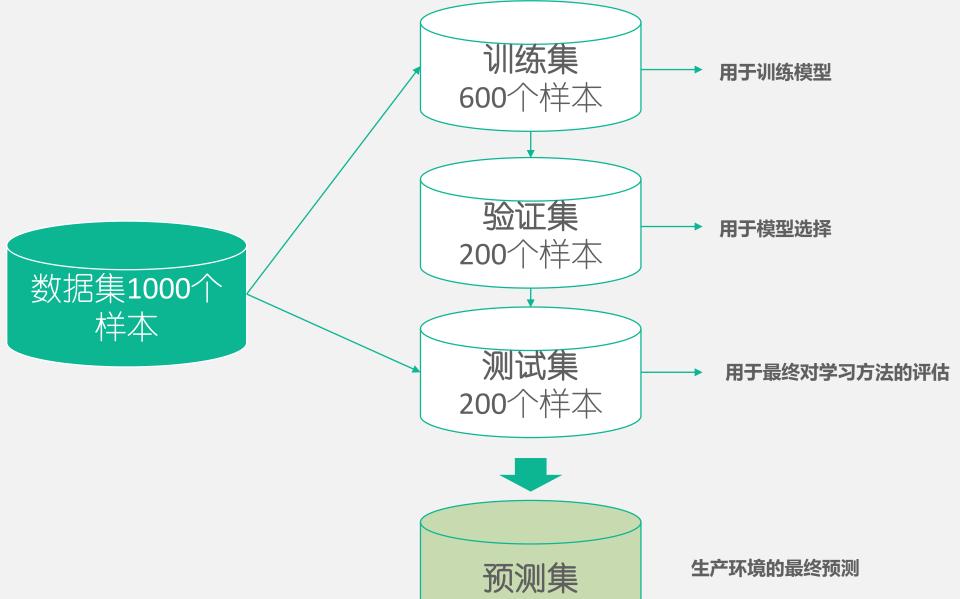
03 Stacking集成学习方法介绍



留出法1回顾(该方法不建议采用)



留出法2回顾(吴恩达推荐的数据划分方法)





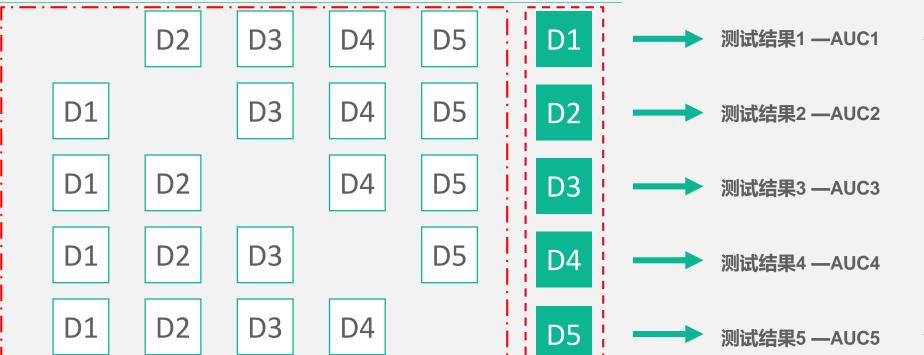


K折交叉验证

- ▲ 这里以5次5折交叉验证为例讲解
- ▲ 5次就是随机的对可能的5种样本划分方法进行重复进行划分



D1 D2 D3 D4 其中一种划分方法举例:将1000个样本通过分层采样的 5个大小一致的互斥子集并划分训练集和测试集。



每一次结果都会有一个评分, 比如是AUC值/均方误差,那 么最终的AUC值就是对所有 5次评估结果取平均值。

训练集

测试集

PART Stacking集成学习方法介绍 3

图例: 指测试数据

Stacking集成方法(一)

▲ 步骤一

假设训练数据包含10000行(有label), 而测试数据包含2500行(无label)。

现在将训练数据划分为训练集8000行,验证集2000行。

1

XGBOOST

验证集2000

测试集2000

测试集2000

测试集2000

测试集2000

对验证集预测生成

预测结果

A1(2000行)

对**测试数据**预测生成预测结果

B1(2500行)

XGBOOST

测试集2000

验证集2000

测试集2000

测试集2000

测试集2000

对**验证集**预测生成 预测结果

A2(2000行)

对**测试数据**预测生成预测结果

B2 (2500行)

XGBOOST

测试集2000

测试集2000

验证集2000

测试集2000

测试集2000

对**验证集**预测生成 预测结果

A3(2000行)

对**测试数据**预测生成预测结果

B3(2500行)

XGBOOST

测试集2000

测试集2000

测试集2000

验证集2000

测试集2000

对**验证集**预测生成 预测结果

A4(2000行)

对**测试数据**预测生成预测结果

B4(2500行)

XGBOOST

测试集2000

测试集2000

测试集2000

测试集2000

验证集2000

对验证集预测生成 预测结果

A5(2000行)

对**测试数据**预测生成预测结果

B5 (2500行)

合成为一个10000列的预测结果, 实际上是原来训练数据的预测结果, 对应真实值label。

2

A1(2000行)

A2(2000行)

A3(2000行)

A4(2000行)

A5(2000行)

1000 0行1 列 ,记为

a1

3

B1 + B2 + B3 + B4 + B5

5

b1 (2500行)

图例: 指测试数据。

Stacking集成方法(二)

▲ 步骤一

假设训练数据包含10000行(有label), 而测试数据包含2500行(无label)。 现在将训练数据划分为训练集8000行, 验证集2000行。

1

LR

验证集2000

测试集2000

测试集2000

测试集2000

测试集2000

对验证集预测生成

预测结果

A1(2000行)

对**测试数据**预测生成预测结果

B1(2500行)

LR

测试集2000

验证集2000

测试集2000

测试集2000

测试集2000

对验证集预测生成

预测结果

A2(2000行)

对**测试数据**预测生成预测结果

B2 (2500行)

LR

测试集2000

测试集2000

验证集2000

测试集2000

测试集2000

对**验证集**预测生成 预测结果

A3(2000行)

对<mark>测试数据</mark>预测生 成预测结果

B3(2500行)

LR

测试集2000

测试集2000

测试集2000

验证集2000

测试集2000

对验证集预测生成预测结果

A4 (2000行)

对**测试数据**预测生成预测结果

B4(2500行)

LR

测试集2000

测试集2000

测试集2000

测试集2000

验证集2000

对**验证集**预测生成 预测结果

A5(2000行)

对**测试数据**预测生成预测结果

B5 (2500行)

合成为一个10000列的预测结果, 实际上是原来训练数据的预测结果, 对应真实值label。

2

A1(2000行)

A2(2000行)

A3(2000行)

A4(2000行)

A5(2000行)

1000 0行1 列 ,记为 a2

3

B1 + B2 + B3 + B4 + B5

E

b2(2500行)

图例: 指测试数据

Stacking集成方法(三)

步骤—

而测试数据包含2500行(无label)。 假设训练数据包含10000行(有label), 现在将训练数据划分为训练集8000行,验证集2000行。

RF

验证集2000

测试集2000

测试集2000

测试集2000

测试集2000

对验证集预测生成

预测结果

A1(2000行)

对测试数据预测生 成预测结果

B1(2500行)

RF

测试集2000

验证集2000

测试集2000

测试集2000

测试集2000

对验证集预测生成

预测结果

A2(2000行)

对测试数据预测生

成预测结果

测试集2000

测试集2000

RF

验证集2000

测试集2000

测试集2000

对验证集预测生成 预测结果

A3(2000行)

对测试数据预测生 成预测结果

B3(2500行)

RF

测试集2000

测试集2000

测试集2000

验证集2000

测试集2000

对验证集预测生成 预测结果

(2000行)

对测试数据预测生 成预测结果

RF

测试集2000

测试集2000

测试集2000

测试集2000

验证集2000

对验证集预测生成 预测结果

A5(2000行)

对测试数据预测生 成预测结果

合成为一个10000列的预测结果, 实际上是原来训练数据的预测结果, 对应真实值label。

A1(2000行)

A2(2000行)

A3(2000行)

A4(2000行)

A5(2000行)

1000 0行1 列 ,记为 a3

对5个测试数据进 行求平均,得到一 个2500行1列数组。

b3(2500行)

B2 (2500行)

B4(2500行)

B5(2500行)

图例: 指测试数据:

Stacking集成方法(四)

▲ 步骤一

假设训练数据包含10000行(有label), 而测试数据包含2500行(无label)。

现在将训练数据划分为训练集8000行,验证集2000行。

1

GBDT

验证集2000

测试集2000

测试集2000

测试集2000

测试集2000

对验证集预测生成

预测结果

A1(2000行)

对**测试数据**预测生成预测结果

B1(2500行)

GBDT

测试集2000

验证集2000

测试集2000

测试集2000

测试集2000

对**验证集**预测生成 预测结果

A2(2000行)

对**测试数据**预测生成预测结果

B2 (2500行)

GBDT

测试集2000

测试集2000

验证集2000

测试集2000

测试集2000

对**验证集**预测生成 预测结果

A3(2000行)

对**测试数据**预测生成预测结果

B3 (2500行)

GBDT

测试集2000

测试集2000

测试集2000

验证集2000

测试集2000

对**验证集**预测生成 预测结果

A4 (2000行)

对**测试数据**预测生成预测结果

B4(2500行)

GBDT

测试集2000

测试集2000

测试集2000

测试集2000

验证集2000

对验证集预测生成 预测结果

A5(2000行)

对**测试数据**预测生成预测结果

B5 (2500行)

合成为一个10000列的预测结果, 实际上是原来训练数据的预测结果, 对应真实值label。

2

A1(2000行)

A2(2000行)

A3(2000行)

A4(2000行)

A5(2000行)

1000 0行1 列 ,记为 a4

3

B1 + B2 + B3 + B4 + B5

5

b4 (2500行)

图例: 指测试数据。

Stacking集成方法(五)

▲ 步骤一

假设训练数据包含10000行(有1abe1), 而测试数据包含2500行(无1abe1)。 现在将训练数据划分为训练集8000行, 验证集2000行。

1

SVM

验证集2000

测试集2000

测试集2000

测试集2000

测试集2000

对验证集预测生成

<u>预测结果</u>

A1(2000行)

对**测试数据**预测生成预测结果

B1(2500行)

SVM

测试集2000

验证集2000

测试集2000

测试集2000

测试集2000

对**验证集**预测生成 预测结果

A2(2000行)

对**测试数据**预测生成预测结果

B2 (2500行)

SVM

测试集2000

测试集2000

验证集2000

测试集2000

测试集2000

对**验证集**预测生成 预测结果

A3(2000行)

对**测试数据**预测生成预测结果

B3(2500行)

SVM

测试集2000

测试集2000

测试集2000

验证集2000

测试集2000

对**验证集**预测生成 预测结果

A4 (2000行)

对**测试数据**预测生成预测结果

B4(2500行)

SVM

测试集2000

测试集2000

测试集2000

测试集2000

验证集2000

对**验证集**预测生成 预测结果

A5(2000行)

对**测试数据**预测生成预测结果

B5 (2500行)

合成为一个10000列的预测结果, 实际上是原来训练数据的预测结果, 对应真实值label。

2

A1(2000行)

A2(2000行)

A3(2000行)

A4(2000行)

A5(2000行)

1000 0行1 列 ,记为

a5

3

B1 + B2 + B3 + B4 + B5

5

b5(2500行)

Stacking集成方法(六)

- ▲ 把a1, a2, a3, a4, a5并列合并得到一个10000行五列的矩阵作为训练集;
- ▲ 把b1, b2, b3, b4, b5并列合并得到一个2500行五列的矩阵作为测试集。
- ▲ 让下一层的模型, 基于他们进一步训练。

LR模型对该层进行训练生成最终结果



V

label

训练数据的 原label

无label待预测

引测试数据的 最终预测结果