



## 两层stacking结构 理解



CONTENTS

01 ▶ 留出法回顾

02 ▶ 交叉验证法回顾

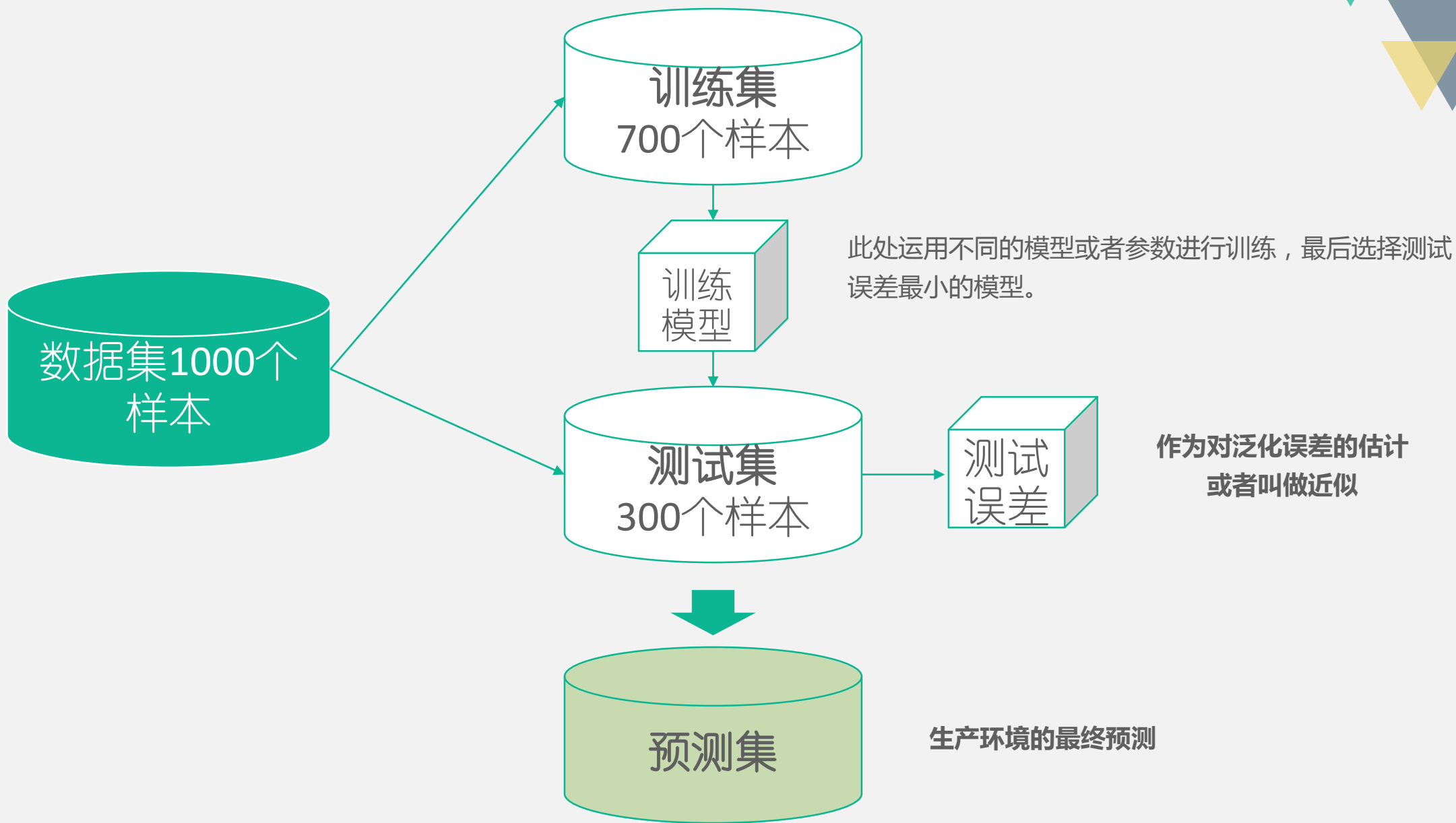
03 ▶ Stacking集成学习方法介绍

PART

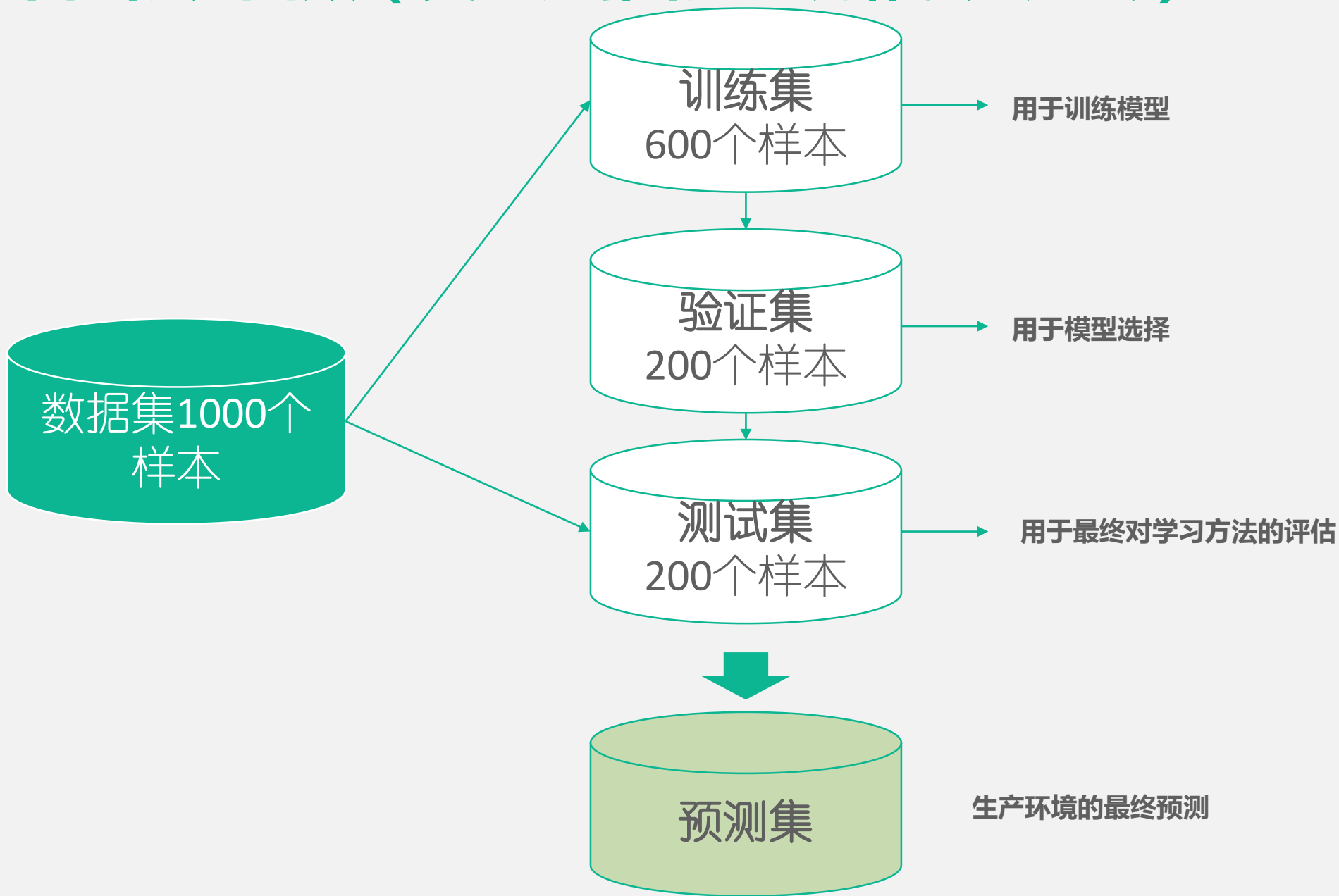
1

留出法回顾

# 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）



PART

2

交叉验证法回顾

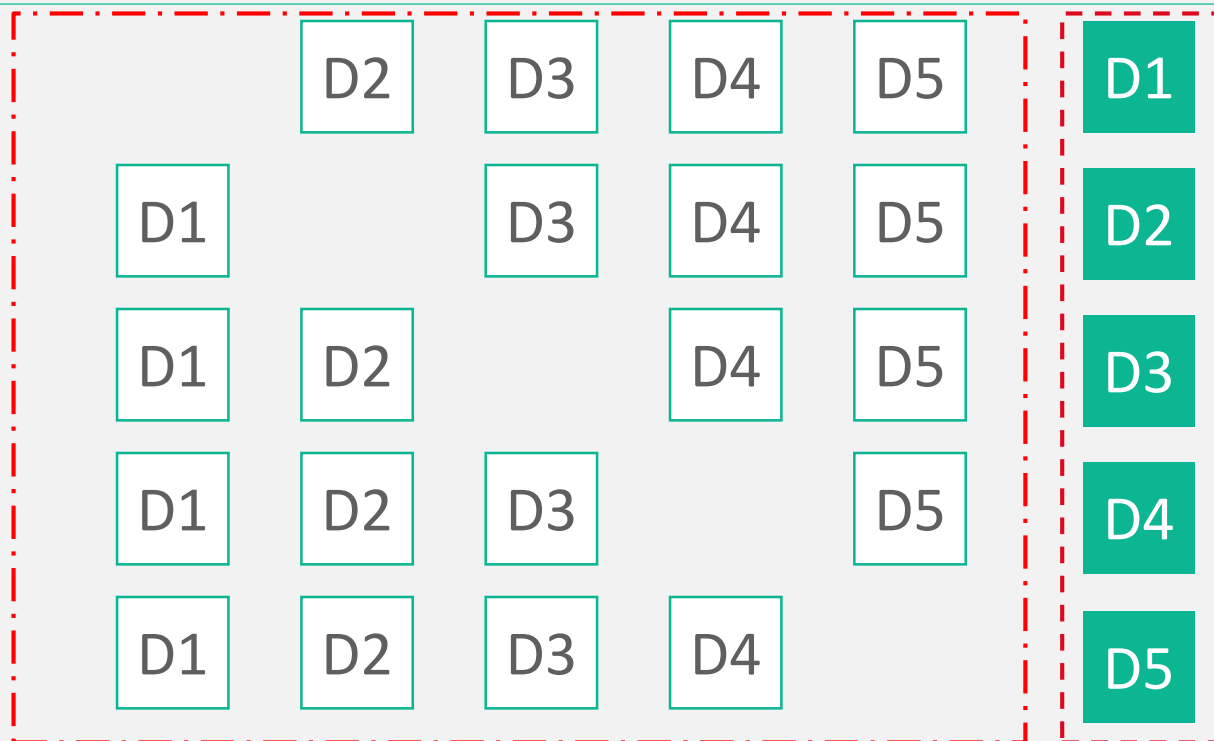
# K折交叉验证

- ▲ 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

数据集D共1000个样本



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。



训练集



测试集

- 测试结果1 —AUC1
- 测试结果2 —AUC2
- 测试结果3 —AUC3
- 测试结果4 —AUC4
- 测试结果5 —AUC5

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。



PART



3

Stacking集成学习方法介绍



图例：  
指测试数据

# Stacking集成方法（一）

## ▲ 步骤一

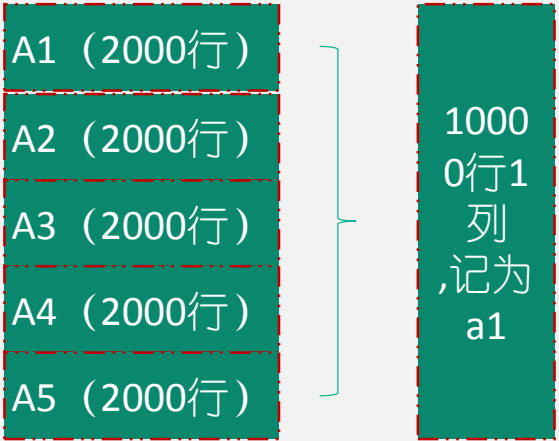
假设训练数据包含10000行(有label)，而测试数据包含2500行(无label)。现在将训练数据划分为训练集8000行，验证集2000行。

1

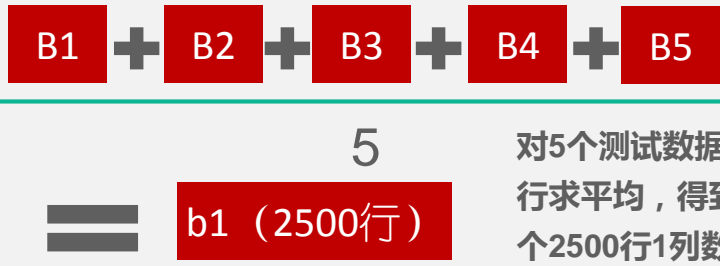


合成为一个10000列的预测结果，实际上是原来训练数据的预测结果，对应真实值label。

2



3



对5个测试数据进行求平均，得到一个2500行1列数组。

# Stacking集成方法（二）

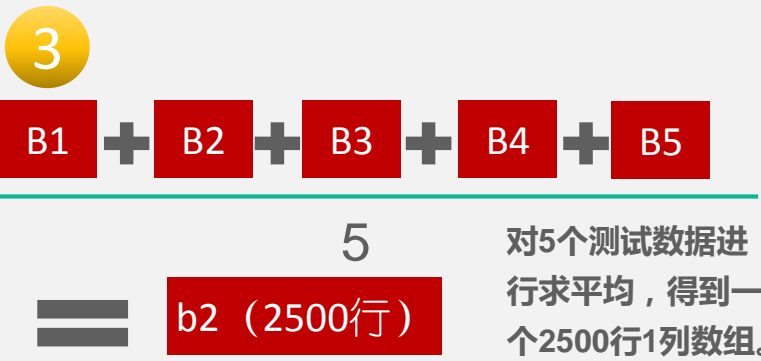
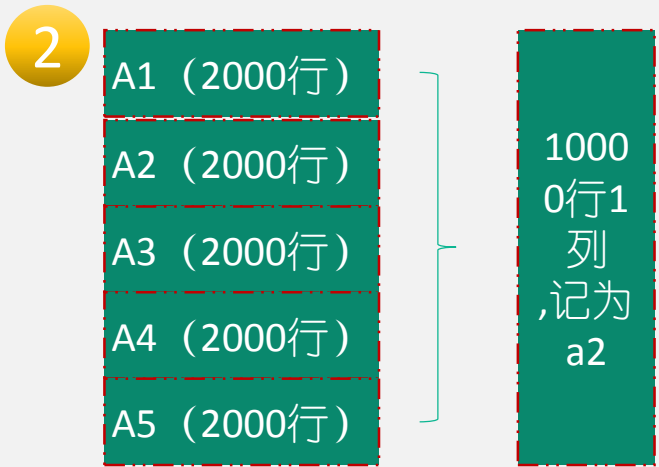
## ▲ 步骤一

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



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。



# Stacking集成方法（三）

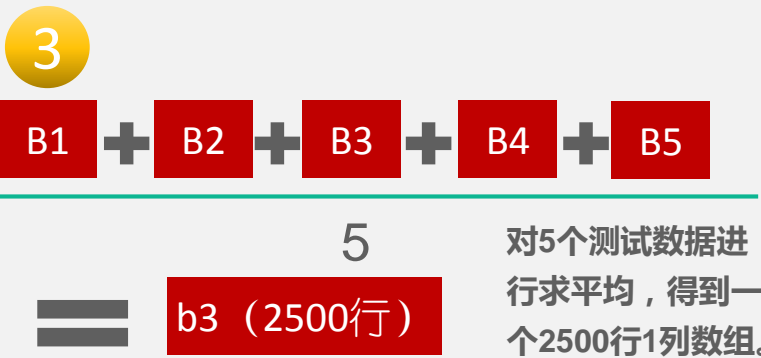
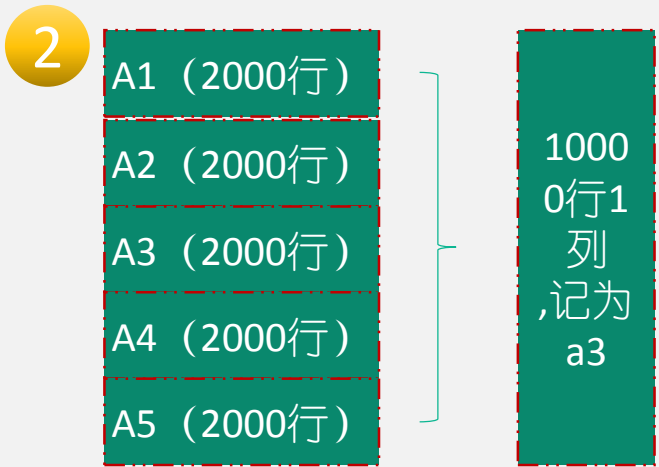
## ▲ 步骤一

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



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。



图例：  
指测试数据

# Stacking集成方法（四）

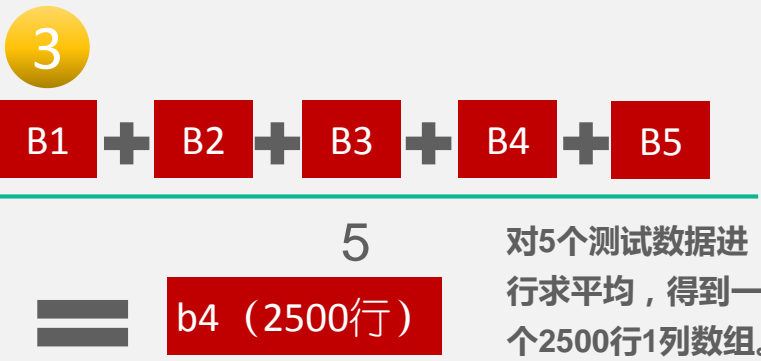
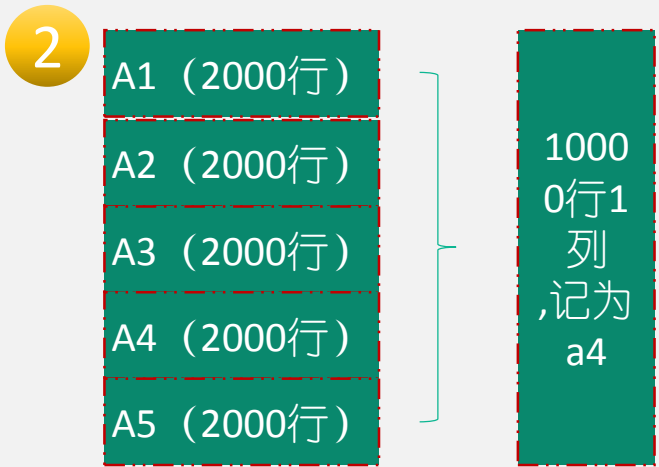
## ▲ 步骤一

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



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。



# Stacking集成方法（五）

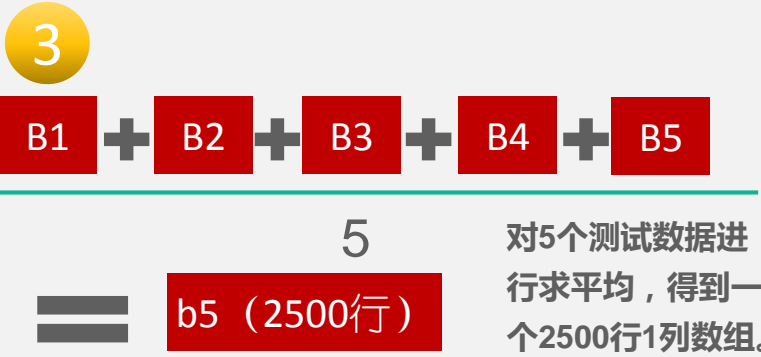
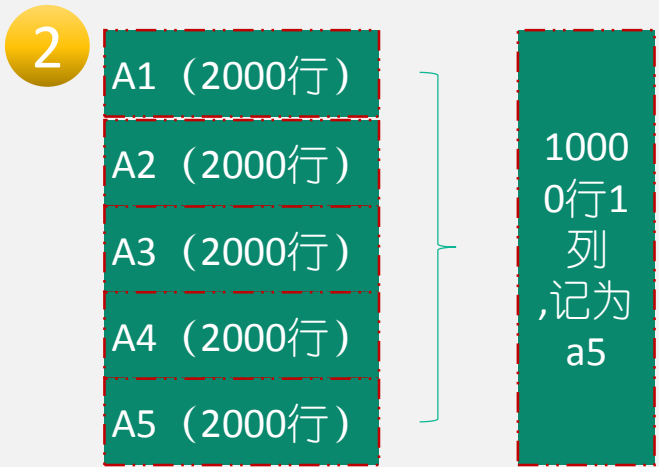
## ▲ 步骤一

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



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。



# 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

