人工智能之机器学习

Stacking

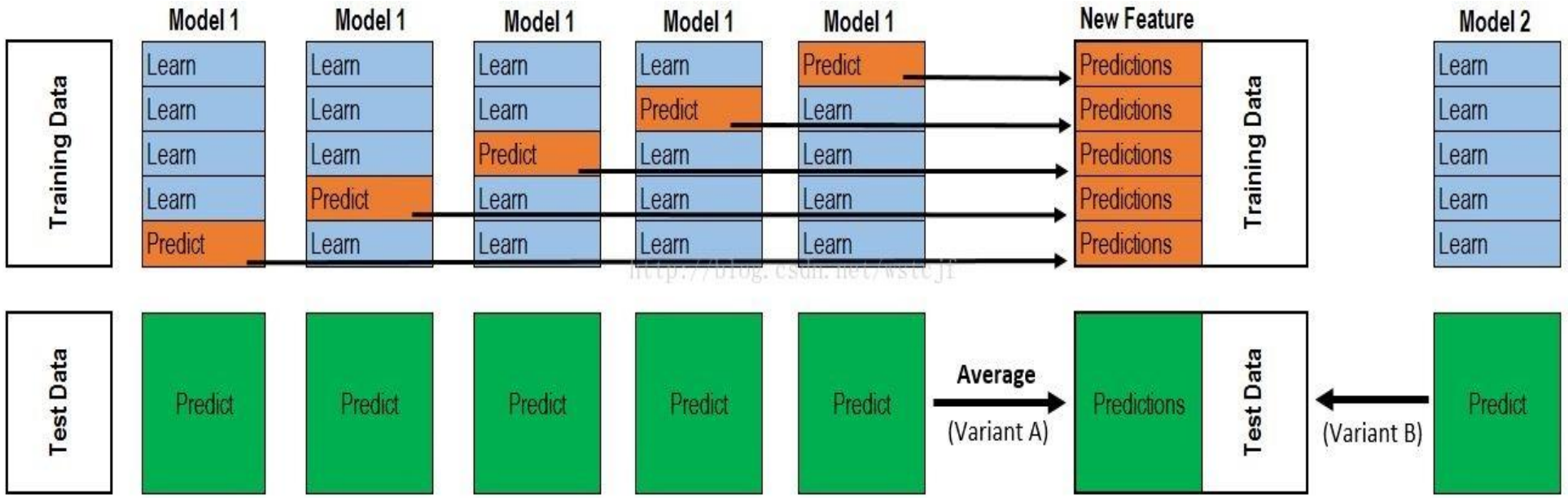
主讲人：李老师

Stacking概述

- Stacking(有时候也称之为stacked generalization)是指训练一个模型用于组合(combine)其他各个模型。即首先我们先训练多个不同的模型，然后再以之前训练的各个模型的输出为输入来训练一个模型，以得到一个最终的输出。
- 如果可以选用任意一个组合算法，那么理论上，Stacking可以表示前面提到的各种Ensemble方法。然而，实际中，我们通常使用单层logistic回归作为组合模型。
- 注意：Stacking有两层，一层是不同的基学习器（classifiers/regressors），第二个是用于组合基学习器的元学习器（meta_classifier/meta_regressor）

Stacking原理讲解

- 直观理解



Stacking原理讲解

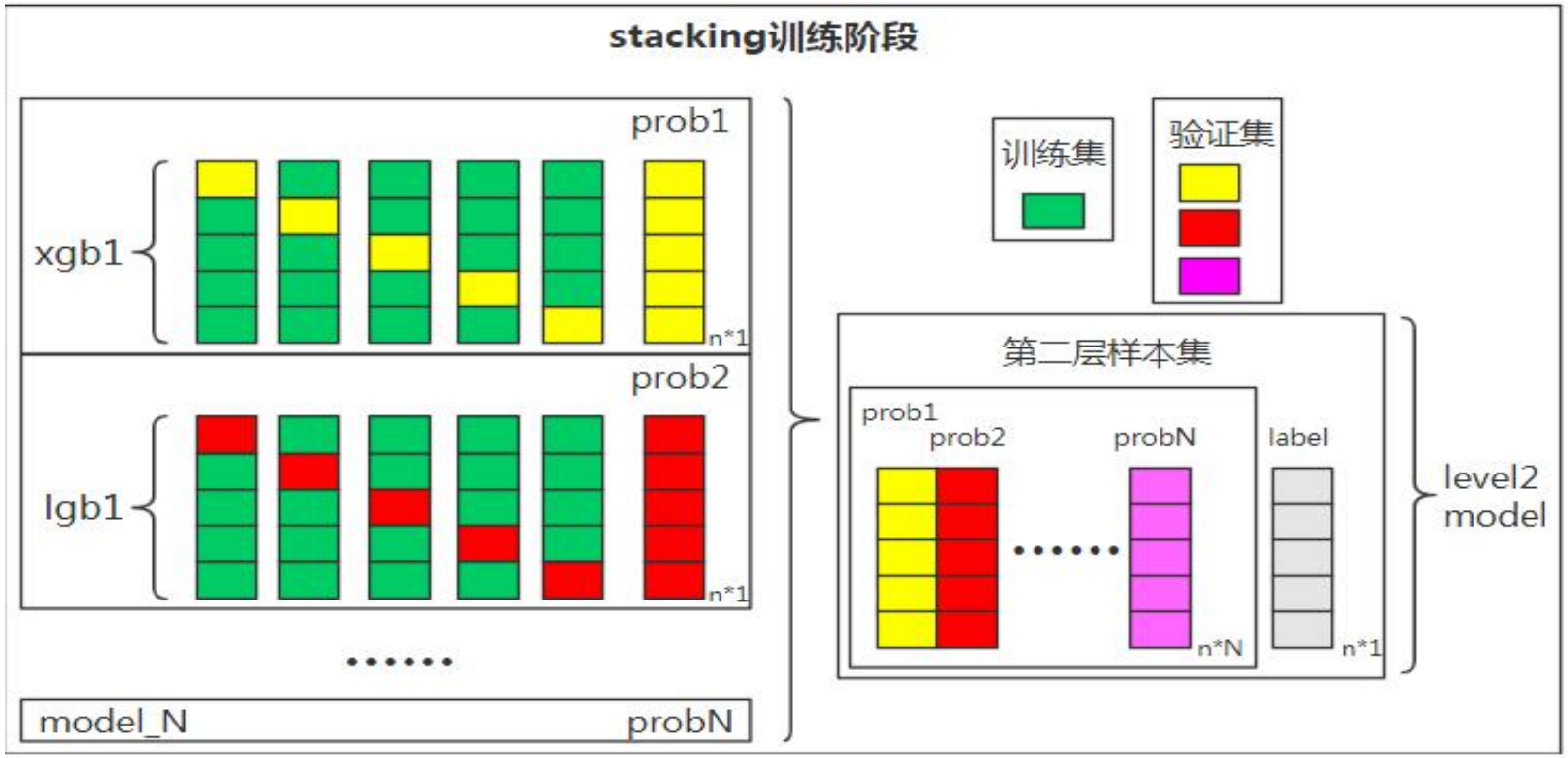
- 图中上半部分是用一个基础模型进行5折交叉验证，如：用XGBoost作为基础模型Model1，5折交叉验证就是先拿出四折作为training learn（蓝色部分），另外一折作为testing predict（橙色部分）。注意：在stacking中此部分数据会用到整个traing data。如：假设我们整个training data包含10000行数据，testing data包含2500行数据，那么每一次交叉验证其实就是对training data进行划分，在每一次的交叉验证中training learn将会是8000行，testing predict是2000行。
- 每一次的交叉验证包含两个过程，1. 基于training learn训练模型；2. 基于training learn训练生成的模型对testing predict进行预测。在整个第一次的交叉验证完成之后我们将会得到关于当前testing predict的预测值，这将会是一个一维2000行的数据，记为a1。注意！在这部分操作完成后，我们还要对数据集原来的整个testing data进行预测，这个过程会生成2500个预测值，这部分预测值将会作为下一层模型testing data的一部分，记为b1（绿色部分）。因为我们进行的是5折交叉验证，所以以上提及的过程将会进行五次，最终会生成针对testing data数据预测的5列2000行的数据a1,a2,a3,a4,a5，对testing set的预测会是5列2500行数据b1,b2,b3,b4,b5。

Stacking原理讲解

- 在完成对Model1的整个步骤之后，我们可以发现a1,a2,a3,a4,a5其实就是对原来整个training data的预测值，将他们拼凑起来，会形成一个10000行一列的矩阵，记为A1。而对于b1,b2,b3,b4,b5这部分数据，我们将各部分相加取平均值，得到一个2500行一列的矩阵，记为B1。
- 以上就是stacking中一个模型的完整流程，stacking中同一层通常包含多个模型，假设还有Model2: LR，Model3：RF，Model4: GBDT，Model5：SVM，对于这四个模型，我们可以重复以上的步骤，在整个流程结束之后，我们可以得到新的A2,A3,A4,A5,B2,B3,B4,B5矩阵。
- 在此之后，我们把A1,A2,A3,A4,A5并列合并得到一个10000行五列的矩阵作为新的training data，B1,B2,B3,B4,B5并列合并得到一个2500行五列的矩阵作为新的testing data。让下一层的模型（元学习器），基于他们进一步训练。

Stacking原理讲解

- 训练阶段



Stacking原理讲解

- 预测阶段